J STATISTIKA

Program Studi Statistika

EISSN : 2654-7511
PISSN : 2089-0028

J STATISTIKA

JURNAL ILMIAH TEORI DAN APLIKASI STATISTIKA

Volume 16 | Nomor 1 | 2023

EDITORIAL TEAM

INTRODUCTION

We are delighted to announce the upcoming publication of Volume 16, Number 1 of JStatistika, affiliated with the Statistics Department at PGRI Adi Buana University Surabaya, scheduled for release in July 2023. This particular issue of the JStatistika Scientific Journal features a diverse array of articles addressing a wide spectrum of topics. One of the highlighted articles delves into **"**Penerapan Model Spasial Menggunakan Matriks Pembobot Queen Contiguity dan Euclidean Distance Terhadap Kasus Gizi Buruk Balita di Provinsi Nusa Tenggara Timur; Peramalan Curah Hujan Harian Kabupaten Jember Dengan Jaringan Saraf Tiruan Dan General Circulation Model; Analisis Regresi Logistik Biner Multilevel pada Status Kemiskinan di Pulau Jawa menggunakan Algoritma MCMC Metropolis-Hasting; Peramalan Nilai Ekspor Migas di Indonesia dengan Model Long ShortTerm Memory (LSTM) dan Gated Recurrent Unit (GRU); Application of Agglomerative Hierarchical Clustering Method for Grouping Non-Cash Food Assistance Recipients in Ngambon Bojonegoro; Multiperiod Logit on Survival Analysis of Financial Distress in Manufacturing Company; Optimasi Produk Plastik pendekatan Taguchi Mixed Level pada Faktor Interaksi Injeksi Molding; Comparison of Support Vector Machine (SVM) and Autoregressive Integrated Moving Average (ARIMA) Methods for Predicting Air Quality Using Python and KNIME; Klasifikasi Pengeluaran per Kapita di Tiga Provinsi Sulawesi menggunakan K-Nearest Neighbor; Perbandingan Kinerja Hasil Luaran Model Jaringan Syaraf Tiruan dan SARIMA Untuk Prediksi Awal Musim Hujan Kota Pangkalpinang; Analisis Faktor-Faktor Yang Mempengaruhi Indeks Pembangunan Manusia Berdasarkan Kabupaten/Kota Di Jawa Tengah; Analisis Faktor-Faktor Yang Mempengaruhi Anemia Pada Ibu Hamil Menggunakan CART; Pengelompokkan Kecamatan di Kabupaten Bima Berdasarkan Jumlah Produksi dan Luas Panen Bawang Merah Tahun 2021 Menggunakan K-Means Clustering; Estimation of Survival Function in Head and Neck Cancer Patients Using the Kaplan-Meier Method; Analisis Tingkat Kepuasan Masyarakat Terhadap Pelayanan BPJS Kesehatan Cabang Utama Surabaya Dengan Metode Customer Satisfaction Indeks dan Importance Performance Analysis**"**

The JStatistika Scientific Journal enthusiastically welcomes and invites contributions in a diverse range of formats, including but not limited to scholarly scientific articles that encompass various facets of statistical science. We eagerly seek research findings, comprehensive reports, insightful case studies, thorough literature reviews, and updates that pertain to the dynamic landscape of statistical science. Our overarching objective is to cultivate a repository of knowledge that is not only current but also invaluable in tackling the ever-evolving and intricate challenges confronting our field. We actively encourage authors to submit their work if it resonates with the most recent advancements and frontiers in statistical science. Our aspiration is to foster an environment where these contributions can flourish, ultimately serving as a

wellspring of cutting-edge insights and understanding. We believe that these insights are instrumental in addressing the multifaceted issues that confront us in today's complex world.

Our editorial team extends a warm and inclusive invitation to scientists and scholars from diverse backgrounds and affiliations, including institutions of higher learning and esteemed research organizations. We seek your valuable contributions, whether they be grounded in empirical research results or rooted in rigorous scholarly studies within the expansive domain of statistics and its myriad practical applications. We hold a deep appreciation for the feedback and perspectives of our esteemed readership. Your input not only enriches the discourse but also plays a pivotal role in our continuous efforts to elevate the quality and relevance of the journal. We earnestly value your insights and ideas, recognizing that they are integral to our ongoing pursuit of excellence. Our ultimate vision is for the articles featured in the JStatistika Scientific Journal to transcend the confines of academia and serve as a wellspring of knowledge that benefits not only scholars and researchers but also professionals actively engaged in the diverse realms of statistical science and its multifaceted real-world applications. Through collaborative efforts and a shared commitment to advancing our understanding of statistics, we aim to make a meaningful impact in the broader scientific community and beyond.

Jstatistika has been indexed by Sinta 4 Kemendikbud, Garuda, Google Scholar, Crossref, Dimensions, One Search, Scilit, Journal Stories, Neliti, Base.

Surabaya, July 2023

Editor in Chief

LIST OF CONTENTS

Penerapan Model Spasial Menggunakan Matriks Pembobot *Queen Contiguity* dan *Euclidean Distance* Terhadap Kasus Gizi Buruk Balita di Provinsi Nusa Tenggara Timur

Kris Suryowati [(1)], Meitriana Nahak [(2)], Rokhana Dwi Bekti [(3)]

[1,2,3] Program Studi Statistika, Fakultas Sains Terapan, Institut Sains & Teknologi AKPRIND Yogyakarta

Jl. Kalisahak No 27 Kota Yogyakarta
e-mail: suryowati@akprind.ac.id[(1)], mitmitmeitriana29@gmail.com[(2)], rokhana@akprind.ac.id[(3)]

ABSTRAK

Gizi buruk balita masih menjadi masalah yang dihadapi oleh negara Indonesia khususnya pada provinsi Nusa Tenggara Timur (NTT). Kasus gizi buruk balita dapat disebabkan oleh berbagai faktor, seperti persentase penduduk miskin, persentase berat badan balita lahir rendah, dan jumlah fasilitas kesehatan. Kasus gizi buruk antar lokasi saling berhubungan menunjukkan adanya efek spasial. Oleh karena itu penelitian ini menggunakan metode model spasial untuk menguji pengaruh tersebut. Model spasial dianataranya *Spatial Autoregressive Model* (SAR), *Spatial Error Model* (SEM), dan *Spatial Durbin Model* (SDM). Matriks pembobot merupakan komponen penting dalam pembentukan model karena menunjukkan hubungan keterkaitan antar lokasi. Penelitian ini menggunakan pembobot *queen contiguity* dan *euclidean distance* kemudian membandingkannya. Sumber data adalah data sekunder tahun 2021, dimana amatan adalah 22 Kabupatenm/Kota di Provinsi NTT. Hasil penelitian menunjukkan bahwa estimasi parameter dan uji hipotesis memberikan hasil yang berbeda di masing-masing pembobot. Hasil uji efek spasial Moran's I dengan pembobot queen contiguity menunjukkan adanya autokorelasi spasial pada jumlah fasilitas kesehatan, sementara itu dengan pembobot euclidean distance adalah persentase berat badan balita lahir rendah. Berdasarkan perbandingan nilai AIC dan MSE, model terbaik yang analisis gizi buruk balita di NTT adalah SDM dengan pembobot queen contiguity.
Kata kunci: Model spasial, pembobot queen contiguity, pembobot euclideance distance, gizi buruk balita

ABSTRACT

Malnutrition of toddlers is still a problem faced by the Indonesian state, especially in the province of East Nusa Tenggara (NTT). Cases of under-five malnutrition can be caused by various factors, such as the percentage of poor people, the percentage of under-fives with low birth weight, and the number of health facilities. Cases of malnutrition between locations are interconnected showing a spatial effect. Therefore this study uses the spatial model method to examine this effect. Spatial models include Spatial Autoregressive Model (SAR), Spatial Error Model (SEM), and Spatial Durbin Model (SDM). The weighting matrix is an important component in model building because it shows the relationship between locations. This study uses queen contiguity and euclidean distance weights and then compares them. The data source is secondary data for 2021, where the observations are 22 Regencies/Cities in NTT. The results showed that parameter estimation and hypothesis testing gave different results in each weighting. The results of the Moran's I spatial effect test with queen contiguity weighting show that there is a spatial autocorrelation in the number of health facilities, meanwhile with the euclidean distance weighting is the percentage of low birth weight children. Based on a comparison of AIC and MSE values, the best model for analyzing under-five malnutrition in NTT is SDM with a queen contiguity weight.
Keywords: *spatial modeling, queen contiguity weight, euclideance distance weight, Toddler Malnutrition*

PENDAHULUAN

Hubungan sebab akibat antara variabel independen terhadap variabel dependen dapat dilakukan melalui analisis regresi. Metode ini menghasilkan suatu persamaan matematis yang menggambarkan pola hubungan tersebut. Salah satu metode untuk mengestimasi nilai-nilai parameter pada persamaan regresi adalah *Ordinary Least Square* (OLS). Metode OLS ini sangat ketat terhadap asumsi, diantaranya residual harus berdistribusi normal, independen, dan identik. Pada beberapa kasus asumsi ini sulit terpenuhi, khususnya pada data spasial yang mengandung pengaruh lokasi geografis dari setiap amatan. Apabila OLS dipaksakan maka akan menghasilkan estimasi parameter yang tidak tepat. Salah satu alternatifnya adalah pemodelan spasial.

Pemodelan spasial [1] merupakan suatu model yang digunakan untuk mengevaluasi hubungan antara satu variabel dengan beberapa variabel lain dengan memberikan efek spasial pada beberapa lokasi yang menjadi pusat pengataman. Pemodelan spasial terdiri atas beberapa metode *Spatial Autoregressive Models* (SAR), *Spatial Error Models* (SEM), dan *Spatial Durbin Model* (SDM) [2]. Metode SAR disebut juga Spatial Lag Model (SLM) adalah salah satu model spasial dengan pendekatan area dengan memperhitungkan pengaruh spasial lag pada variabel dependen saja. Metode SEM merupakan model digunakan saat nilai error pada suatu lokasi berkorelasi dengan nilai error dengan lokasi sekitarnya atau dengan kata lain terdapat korelasi spasial antar eror. Metode SDM merupakan model regresi spasial yang memiliki bentuk seperti *Spatial Autoregressive Model* (SAR) yang memiliki spasial lag pada variabel respon (y). Hanya saja, SDM memiliki ciri khas adanya spasial lag pada variabel prediktor (X) [3].

Komponen yang mendasar dari model spasial adalah matriks pembobot spasial. Matriks ini mencerminkan adanya hubungan antara satu wilayah dengan wilayah lainnya [4]. Berdasarkan tipe pembobotnya, analisis regresi spasial dapat dibedakan menjadi analisis dengan pendekatan titik dan pendekatan area. Pendekatan titik adalah metode yang menggunakan informasi jarak (*distance*) sebagai pembobotnya. Pembentukan matriks pembobot jarak diperoleh dari perhitungan euclidean distance (*jarak euclidean*) antara lokasi penelitian berdasarkan derajat [5]. Sedangkan pendekatan area merupakan pendekatan yang menggambarkan relatif lokasi suatu unit data spasial dengan lokasi lain di suatu area. Dalam pendekatan area, terdapat beberapa jenis pembobot, salah satunya adalah *queen contiguity* [1]. Pembobot *queen contiguity* ini adalah jenis pembobot ketetanggaan yang memperhatikan sisi maupun sudut wilayah satu dengan wilayah yang lain. Untuk pengembangan teori Spasial dalam hal ini perlu dianalisis perbadingan kedua pembobot dalam pemodelan pemodelan spasial SAR, SEM, dan SDM.

Gizi buruk balita masih menjadi masalah yang dihadapi oleh negara Indonesia khususnya pada provinsi NTT. Gizi buruk balita disebabkan oleh berbagai faktor seperti presentase penduduk miskin, presentase berat badan balita lahir rendah (BBLR), dan jumlah fasilitas kesehatan. Berdasarkan data publikasi Badan Pusat Statistik, Provinsi Nusa Tenggara Timur menempati urutan pertama dengan rasio penderita gizi bruk tertinggi pada tahun 2018 sebesar 35,4, tahun 2019 sebesar 30,3%, tahun 2020 sebesar 28,2%, dan tahun 2021 kembali melonjak hingga mencapai 37,8%. Data tersebut menunjukkan bahwa kasus gizi buruk balita di Provinsi NTT mengalami kenaikan dan penurunan, sehingga perlu diketahui yang menjadi faktor-faktor penyebabnya.

Data gizi buruk balita di setiap Kabupaten/Kota memiliki karakterisitik pola spasial, dimana karakteristik antar lokasi dapat saling berhubungan. Hal ini menunjukkan adanya pengaruh spasial

sehingga diperlukan penggunaan metode regresi spsial untuk analisis faktor yang mempengaruhi gizi buruk balita di NTT. Beberapa penelitian yang menggunakan metode spasial untuk analisis angka gizi buruk diantaranya penelitian [6] yang melakukan analisis memodelkan angka gizi buruk di Provinsi Bali dengan pendekatan regresi spasial. Dalam penelitian ini, matriks pembobot yang digunakan adalah *Queen Contiguity* dan model terbaik adalah model SEM. Penelitian [7] yang melakukan analisis spasial gizi kurang balita di Kota Tangerang. Penelitian [8] yang menyusun pemetaan angka gizi buruk pada balita menurut Menggunakan Analisis Regresi spasial.

Penelitian yang telah membandingkan berbagai jenis pembobot pada model spasial diantaranta penelitian [5] yang melakukan pemodelan regresi logistik spasial dengan matriks pembobot spasial queen contiguity dan matriks jarak euclidean. Penelitian [9] yang melakukan kajian pengaruh matriks pembobot spasial dalam model spasial data panel untuk menentukan faktor-faktor yang mempengaruhi tingkat kemiskinan. Penelitian [10] yang membandingkan pembobot invers jarak dan kontinguity pada analissi data DBD. Penelitian [11] membandingkan beberapa pembobot kontinguity (*Rook Contiguity, Bishop Contiguity,* dan *Queen Contiguity*) pada analsisi data IPM menggunakan SEM. Hasil menunjukkan *Rook Contiguity* adalah yang terbaik.

Berdasarkan latar belakang tersebut maka penelitian ini melakukan pemodelan spasial pada kasus gizi buruk balita di Provinsi NTT. Penelitian ini membandingkan pemodelan spasial SAR, SEM, dan SDM ketika menggunakan pembobot queen continguity dan euclidean distace.

METODE

Data yang digunakan dalam penelitian ini adalah data sekunder tahun 2021 yang diperoleh dari publikasi BPS dan Dinas Kesehatan Provinsi NTT. Variabel dependen yang digunakan adalah Persentase Gizi Buruk Balita, variabel independennya adalah Persentase Penduduk Miskin, Persentase Berat Badan Balita Lahir Rendah/BBLR, dan Jumlah Fasilitas Kesehatan Provinsi NTT.

Tahapan penelitian ini diawali dengan dengan mendeskripsikan variabel penelitian dengan analisis deskriptif dan peta tematik, kemudian melakukan analisis regresi linear berganda dengan metode *OLS*. Selanjutnya menentukan pembobot *Queen Contiguity* dan *Euclidean Distance*, melakukan uji efek spasial, hingga estimasi parameter SAR, SEM, dan SDM. Langkah akhir adalah interpretasi dan membandingkan pemodelan menggunakan nilai AIC dan MSE. Analisis regresi linear berganda dengan metode *Ordinary Least Square* dengan persamaan

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \quad (1)$$

Dengan Y adalah variabel dependen, $\beta_0,\ \beta_1,\ \beta_2, \ldots, \beta_k$ adalah Parameter regresi, $X_1, X_2, \ldots., X_k$ adalah variabel independen, dan ε adalah redidual. Pada pemodelan regresi OLS dengan data spasial, asumsi residual distribusi normal, independen, dan identic sering tidak memenuhi. Hal ini karena adanya efek spasial pada data, yaitu adanya dependensi atau hubungan antar amatan. Alternative model yang dapat digunakan adalah regresi spasial. Analisis regresi spasial adalah analisis yang mengevaluasi hubungan antara satu variabel dengan beberapa variabel lain dengan memberkan efek spasial pada beberapa lokasi yang menjadi pusat pengamatan. Secara umum model regresi spasial sebagai berikut [12,13,14]:

$$Y = \rho WY + X\beta + u \quad (2)$$
$$u = \lambda Wu + \varepsilon,$$
$$\varepsilon \sim N(0, \sigma^2 I)$$

Dimana Y adalah matriks variabel respon yang berukuran (n x 1), X adalah matriks variabel bebas berukuran n x (k + 1)), β adalah vektor koefisien parameter regresi berukuran (k + 1) x 1, ρ adalah koefisien autokorelasi lag spasial, λ adalah koefisien autokorelasi lag dan *error* yang bernilai $|\lambda| < 1$, u adalah vektor *error* yang diasumsikan mengandung autokorelasi berukuran n x 1, ε adalah vektor *error* yang berukuran n x 1, I adalah matriks identitas berordo n x n, W adalah matriks pembobot spasial yang berukuran n x n.

Model SAR atau disebut juga *Spatial Lag Model* (SLM) adalah salah satu model spasial dengan pendekatan area dengan memperhitungkan pengaruh spasial lag pada variabel dependen saja. Bentuk umum model SAR seperti pada persamaan (3):

$$y = \rho W_{1y} + X\beta + \varepsilon \quad (3)$$

Model SEM merupakan salah satu bentuk analisis regresi spasial yang terbentuk ketika terjadi autokorelasi pada error. Bentuk umum model SEM seperti pada persamaan (4):

$$y = x\beta + \lambda W_2 u + \varepsilon \quad (4)$$
$$y = x\beta + (1 - \lambda W_2)^2 + \varepsilon$$

Model SDM memiliki bentuk persamaan seperti SAR dengan hanya ada pengaruh spaial lag pada variabel dependen. Namun, SDM merupakan kasus khusus dari model SAR dengan menambahkan pengaruh lag pada variabel independen. Bentuk umum model SDM seperti pada persamaan (5) :

$$Y = \rho W_1 Y + \beta_0 + X\beta_1 + WX\beta_2 + \varepsilon \text{ , } \varepsilon \sim N(0, \sigma^2 I) \quad (5)$$

Matrik pembobot spasial (W) memiliki peran penting pada pemodelan untuk menunjukkan hubungan antar lokasi. Banyak jenis pembobot spasial yang dapat digunakan yaitu pendekatan area dan titik. Matrik pembobot spasial *Queen Contiguity* merupakan salah satu jenis bobot pendekatan area [15]. Tipe ini menentukan daerah pengatamannya berdasarkan sisi-sisi yang saling bersinggungan dengan memperhitungkan sudut. Lokasi yang bersisian atau titik sudutnya bertemu dengan lokasi yang menjadi perhatian diberi pembobotan $W_{ij} = 1$, sedangkan untuk lokasi lainnya adalah $W_{ij} = 0$. Sementara itu matriks *Euclidean Distance* adalah salah satu jenis bobot pendekatan titik. Pembentukan matriks pembobot ini diperoleh dari perhitungan jarak euclidean antara lokasi berdasarkan koordinat *lattitude* dan *longitude* dengan rumus sebagai berikut [16]:

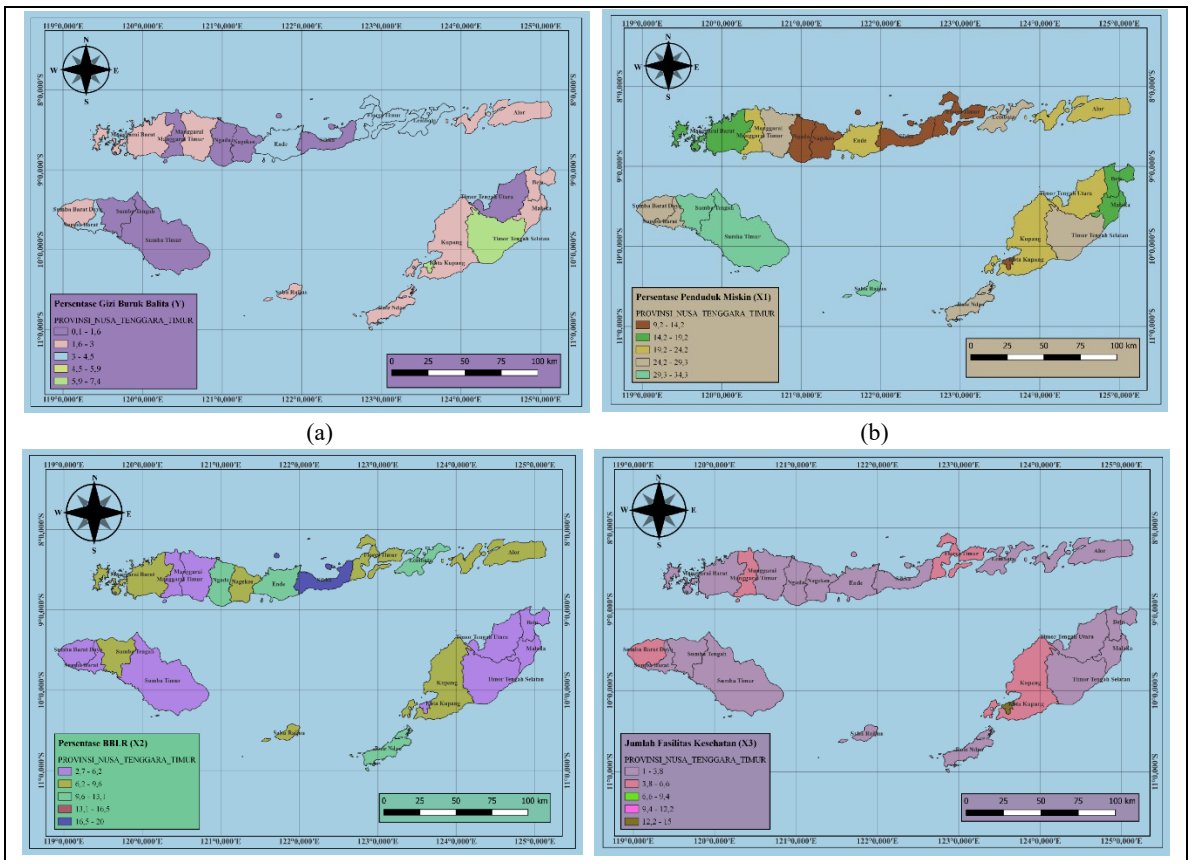
$$d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2} \quad (6)$$

dengan, u_i adalah *longitude* pada lokasi ke-*i*, u_j adalah *longitude* pada lokasi ke-*j*, v_i adalah *latitude* pada lokasi ke-*i*, dan v_j adalah *latitude* pada lokasi ke-*j*. Berdasarkan jarak yang diperoleh, maka akan dibuat matriks pembobot W_{ij} yang terbentuk berdasarkan batas jarak suatu lokasi yang mempengaruhi lokasi lain (r) dengan ketentuan pembentukkan matriks $W_{ij}(d)$ sebagai berikut :

$$W_{ij} = \begin{cases} 1, & jika\ d_{ij} < r \\ 0, & jika\ d_{ij} \geq r \end{cases}$$

HASIL DAN DISKUSI

Rata-rata persentase gizi buruk balita di 22 Kabupaten/Kota Propinsi NTT tahun 2021 adalah 2,447%. Persentase gizi buruk balita tertinggi berada pada kabupaten Timur Tengah Selatan sebesar 7,390% dan persentase gizi buruk terendah berada di kabupaten Ngada sebesar 0,090%. Pola spasial dari masing-masing variabel dapat dilihat melalui peta tematik disajikan pada Gambar 1. Pada data persentase gizi buruk balita, sebagian besar terletak pada kelas interval 1,6% - 3% yaitu kabupaten Sumba Barat Daya, Sabu Raijua, Rote Ndao, Kupang, Malaka, Belu, Alor, Manggarai Timur, dan Manggarai Barat. Hal ini diketahui bahwa wilayah-wilayah yang bertetanggaan sebagian besar memiliki kategori persentase gizi buruk balita yang hamper sama. Begitu juga untuk variable persentase penduduk miskin, persentase berat badan balita lahir rendah, dan jumlah fasilitas kesehatan.



(a) (b)

(c) (d)

Gambar 1. Pola Spasial Data : a) persentase gizi buruk balita, b) persentase penduduk miskin, c) persentase Berat Badan Balita Lahir Rendah, d) Jumlah fasilitas kesehatan di Provinsi NTT

Hasil pemodelan regresi berganda dengan estimasi OLS mempunyai nilai R^2 sebesar 26,36% yang artinya bahwa sebesar 26,36% variabel persentase gizi buruk balita dapat dijelaskan oleh sekelompok variabel independen. Selanjutnya dilakukan uji asumsi klasik. Uji asumsi klasik bertujuan untuk memberikan kepastian bahwa persamaan regresi yang diperoleh memiliki ketepatan dalam estimasi, tidak bias, dan konsisten. Asumsi-asumsi yang harus dipenuhi yaitu normalitas, tidak terdapat gejala heterokedastisitas, autokorelasi, dan multikolinearitas. Berdasarkan pengujian, asumsi yang belum terpenuhi adalah adanya heterogenitas. Hal ini menunjukkan ada indikasi efek spasial.

Langkah selanjutnya adalah melakukan uji Moran's I untuk mengetahui ada tidaknya efek spasial dan pemodelan spasial. Tahap awal yaitu menentukan matriks pembobot spasial meliputi pembobot "*Queen Contiguity* merupakan matriks pembobot yang memperhatikan persinggungan sisi dan sudut" dan "*Euclidean Distance.atau* matriks pembobot jarak". Pada matriks pembobot queen contiguity untuk kabupaten/kota yang saling berdekatan (bersinggungan sisi/sudutnya maka diberi bobot 1) sedangkan kabupaten/kota yang tidak berdekatan (tidak bersinggungan sisi/sudutnya) diberi bobot 0, sehingga matriks beordo 22×22 sebagai berikut :

$$W_{ij} = \begin{bmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 1 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \dots & & & & & \\ 0 & 0 & 1 & 0 & \dots & 0 \end{bmatrix}$$

Selanjutnya matriks pembobot *Euclidean Distance* yang diperoleh dari perhitungan jarak *euclidean* antara lokasi penelitian berdasarkan derajat *latitude* dan *longitude,* sehingga dihasilkan matriks jarak antar kabupaten/kota dengan satuan kilometer sebagai berikut

$$d_{22\times22} = \begin{bmatrix} 0 & 220{,}0968 & 23{,}7114 & \cdots & 436{,}0421 \\ 220{,}0968 & 0 & 205{,}9597 & \cdots & 248{,}4350 \\ 23{,}7114 & 205{,}9597 & 0 & \cdots & 430{,}0099 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 436{,}0421 & 280{,}4751 & 430{,}0099 & \cdots & 0 \end{bmatrix}$$

Kemudian menghitung matriks pembobot W_{jarak} berdasarkan batas jarak suatu lokasi yang mempengaruhi lokasi lain dengan terlebih dahulu menghitungan rata-ratanya dalam hal ini diperoleh $(r) = 438\ km$. Penentuan pembobot W_{jarak} dengan formula sebagai berikut:

$$W_{ij} = \begin{cases} 1, Jika\ d_{ij} < 438 \\ 0, \quad lainnya \end{cases}$$

sehingga diperoleh matriks pembobot jarak sebagai berikut:

$$W_{jarak\ 22\times22} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ \vdots \\ 22 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & \cdots & 22 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & 1 & \cdots & 1 \\ 1 & 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 0 \end{bmatrix} \end{array}$$

Uji indeks moran digunakan untuk menguji dependensi spasial atau korelasi spasial antar lokasi (kabupaten). Pengujian ini dilakukan dengan hipotesis nol (Ho) adalah tidak ada autokorelasi spasial. Pengambilan keputusan adalah Ho ditolak ketika $|Z_{hitung}| > Z_{\frac{\alpha}{2}}$. Dengan menggunakan pembobot *queen contiguity,* variable yang terbutkti mengandung autokorelasi spasial adalah jumlah fasilitas kesehatan. Sementara itu menggunakan pembobot *euclidean distance*, yang mengandung autokorelasi spasial variabel persentase berat badan balita lahir rendah.

Selanjutnya hasil estimasi parameter model SAR menggunakan matriks pembobot queen contiguity dan euclidean distance disajikan pada Tabel 1. Hasil estimasi parameter, AIC dan MSE pada kedua jenis pembobot adalah berbeda. Model SAR dengan pembobot *queen continguity* menghasilkan nilai $\hat{\rho}$ = 0,237 artinya persentase gizi buruk balita suatu kabupaten akan tinggi jika berdekatan dengan kabupaten lain yang memiliki persentase gizi buruk balita yang tinggi juga. Pada model SAR dengan pembobot *Euclidean Distance* menghasilkan nilai $\hat{\rho}$ = -0.0167 artinya persentase gizi buruk balita suatu kabupaten akan tinggi jika berdekatan dengan kabupaten lain yang memiliki persentase gizi buruk balita yang rendah. Nilai AIC dan MSE pada pembobot Queen Continguitu adalah lebih kecil. Variable independen yang berpengaruh pada kedua model SAR adalah sama yaitu persentase penduduk miskin (X_3). Koefisien regresi masing-masing adalah sebesar 0,3596 dan 0,3666. Angka positif menunjukkan bahwa persentase gizi buruk balita tinggi jika persentase penduduk miskin juga tinggi

Tabel 1. Estimasi Parameter Model SAR

Parameter	Estimasi Parameter	*Std.Error*	Z_{value}	P_{value}
Pembobot *Queen Contiguity*				
β_0	0,5228	1,4068	0,3716	0,7101
β_1	0,0399	0,0455	0,8772	0,3804
β_2	-0,0610	0,0771	-0,7913	0,4287
β_3	0,3596	0,1128	3,1886	0,0014 (*)
$\hat{\rho}$ = 0,237, AIC = 87,169, MSE = 1,3192				
Pembobot *Euclidean Distance*				
β_0	1,0503	1,6579	0,6335	0,5264
β_1	0,0539	0,0466	1,1558	0,2477
β_2	-0,0270	0,0858	-0,3145	0,7531

$\boldsymbol{\beta_3}$	0,3666	0,1104	3,3190	0,0009 (*)
$\widehat{\boldsymbol{\rho}}$ = -0,0167 , AIC = 87,995 ,MSE = 1,3584				

Hasil estimasi parameter model SEM menggunakan matriks pembobot queen contiguity dan euclidean distance disajikan pada Tabel 2. Hasil estimasi parameter, AIC dan MSE pada kedua jenis pembobot adalah berbeda. Model SEM dengan pembobot *queen continguity* menghasilkan nilai Nilai λ = 0,45576, artinya persentase gizi buruk balita suatu kabupaten/kota akan tinggi jika berdekatan dengan kabupaten lain yang memiliki residual yang tinggi juga. Sementara itu, model SAR dengan pembobot *Euclidean Distance* menghasilkan nilai λ = -0,14665, artinya persentase gizi buruk balita suatu kabupaten/kota akan tinggi jika berdekatan dengan kabupaten lain yang memiliki nilai residual yang rendah.

Tabel 2. Estimasi Parameter model SEM

Parameter	Estimasi Parameter	*Std.Error*	Z_{value}	P_{value}
Pembobot *Queen Contiguity*				
β_0	0,6644	1,1907	0,5580	0,5768
β_1	0,0296	0,0363	0,8173	0,4138
β_2	-0,0247	0,0659	-0,3754	0,7074
β_3	0,4613	0,1090	4,2324	0,00002313(*)
λ = 0,45576, AIC = 86,338, MSE = 1,2474				
Pembobot *Euclidean Distance*				
β_0	1,6608	1,4508	1,1447	0,2523
β_1	0,0194	0,0457	0,4260	0,6701
β_2	-0,0842	0,0948	-0,8887	0,3741
β_3	0,3313	0,1086	3,0502	0,0022 (*)
λ = -0,14665, AIC = 85,887, MSE = 1,1876				

Keterangan : (*) signifikan pada α = 5%

Nilai AIC dan MSE pada pembobot *Euclideance Distance* adalah lebih kecil. Variable independen yang berpengaruh pada kedua model SEM adalah sama yaitu persentase penduduk miskin (X_3). Koefisien regresi masing-masing adalah sebesar 0,4613 dan 0,3313. Angka positif menunjukkan bahwa prosentase gizi buruk balita tinggi jika persentase penduduk miskin juga tinggi.

Hasil estimasi parameter model SDM menggunakan matriks pembobot queen contiguity dan euclidean distance disajikan pada Tabel 3. Hasil estimasi parameter, AIC dan MSE pada kedua jenis pembobot adalah berbeda. Nilai AIC pada pembobot *Queen Continguity* adalah lebih kecil. Namun demikian, nilai MSE pada pembobot *Euclideance Distance* adalah lebih kecil.

Tabel 3. Hasil Estimasi Parameter model SDM

Parameter	Estimasi Parameter	*Std.Error*	Z_{value}	P_{value}
Pembobot *Queen Contiguity*				

β_0	-0,9053	1,2252	-0,7389	0,4599
β_{11}	0,0849	0,0393	2,1584	0,0308 (*)
β_{12}	-0,0117	0,0617	-0,1908	0,8486
β_{13}	0,4657	0,0983	4,7342	0,0000022 (*)
β_{21}	0,0666	0,0329	2,0240	0,0429 (*)
β_{22}	-0,0591	0,0703	-0,8399	0,4009
β_{23}	-0,6217	0,1986	-3,1294	0,0017 (*)
$\hat{\rho}$ = 0,33361, AIC = 82,346, MSE = 1,0183				
Pembobot *Euclidean Distance*				
β_0	3,3542	1,7073	1,9646	0,0494
β_{11}	-0,0266	0,0457	-0,5827	0,5601
β_{12}	-0,1152	0,1150	-1,0019	0,3163
β_{13}	0,1988	0,1079	1,8415	0,0655 (**)
β_{21}	-0,0575	0,0266	-2,1607	0,0307 (*)
β_{22}	-0,0628	0,0518	-1,2109	0,2259
β_{23}	-0,0411	0,0749	-0,5494	0,5827
$\hat{\rho}$ = -0,18847, AIC = 88,495, MSE = 1,0003				

Variable independen yang berpengaruh pada kedua model SDM adalah berbeda. Pada pembobot *Queen Continguity,* variable independen yang berpengaruh adalah persentase penduduk miskin (X_1), jumlah fasilitas kesehatan (X_3), lag persentase penduduk miskin (WX_1), dan lag jumlah fasilitas kesehatan (WX_3). Suatu kabupaten/kota akan memiliki persentase gizi buruk balita tinggi jika kabupaten/kota tersebut memiliki persentase penduduk miskin tinggi, jumlah fasilitas kesehatan yang tinggi, bertetanggan dengan kabupaten/kota yang memiliki persentase penduduk miskin tinggi, dan bertetanggan dengan kabupaten/kota yang memiliki jumlah fasilitas kesehatan rendah. Pada pembobot *Euclideance Distance,* variable independen yang berpengaruh adalah jumlah fasilitas kesehatan (X_3) dan lag persentase penduduk miskin (WX_1). Suatu kabupaten/kota akan memiliki persentase gizi buruk balita tinggi jika kabupaten/kota tersebut memiliki jumlah fasilitas kesehatan yang tinggi dan bertetanggan dengan kabupaten/kota yang memiliki persentase penduduk miskin rendah.

Pemilihan model terbaik menggunakan nilai AIC dan nilai MSE terkecil seperti pada Tabel 5. Dapat diketahui bahwa model terbaik dari pembobot *queen contiguity* dan *euclidean distance* adalah model *Spatial Durbin Model* (SDM) dengan nilai AIC yang diperoleh pada pembobot *queen contiguity* sebesar 82,346 dan nilai MSE sebesar 1,0187. Dan pada pembobot *euclidean distance* diperoleh nilai MSE sebesar 1,003. Apabila dibandingkan antar pembobot pada SDM, maka pembobot queen continguity adalah yang lebih baik. Persamaan model ini adalah

$$Y_i = 0{,}33361 \sum_{j=1}^{n} W_{ij}Y_j - 0{,}9053 + 0{,}0849X_{1i} - 0{,}0117X_{2i} + 0{,}4657X_{3i} + 0{,}0666 \sum_{j=1}^{n} W_{ij}X_{1j} - 0{,}0591 \sum_{j=1}^{n} W_{ij}X_{2j} - 0{,}6217 \sum_{j=1}^{n} W_{ij}X_{3j}$$

Tabel 5. Pemilihan Model Terbaik

Model	Pembobot *Queen Contiguity*		Pembobot *Euclidean Distance*	
	AIC	MSE	AIC	MSE
SAR	87,169	1,3192	87,995	1,3584
SEM	86,338	1,2474	85,887	1,1876
SDM	82,346	1,0183	88,495	1,0003

KESIMPULAN

Model SAR, SEM, dan SDM dengan pembobot *queen contiguity* dan *euclidean distance* menghasilkan hasil estimasi parameter, signifikansi parameter, nilai AIC dan MSE yang berbeda-beda. Berdasarkan perbandingan nilai AIC dan MSE, model terbaik dari pembobot *queen contiguity* dan *euclidean distance* adalah model *Spatial Durbin Model* (SDM) dengan nilai AIC yang diperoleh pada pembobot *queen contiguity* sebesar 82,346 dan nilai MSE sebesar 1,0187. Dan pada pembobot *euclidean distance* diperoleh nilai MSE sebesar 1,003.

Apabila dibandingkan antar pembobot pada SDM, maka pembobot queen continguity adalah yang lebih baik. Model ini menghasilkan variable independen yang berpengaruh adalah persentase penduduk miskin (X_1), jumlah fasilitas kesehatan (X_3), lag persentase penduduk miskin (WX_1), dan lag jumlah fasilitas kesehatan (WX_3). Suatu kabupaten/kota akan memiliki persentase gizi buruk balita tinggi jika kabupaten/kota tersebut memiliki persentase penduduk miskin tinggi, jumlah fasilitas kesehatan yang tinggi, bertetanggan dengan kabupaten/kota yang memiliki persentase penduduk miskin tinggi, dan bertetanggan dengan kabupaten/kota yang memiliki jumlah fasilitas kesehatan rendah.

DAFTAR PUSTAKA

[1] L. Anselin, *Spatial Econometrics: Methodes and Models*. Kluwer Academic Publisher,1988.
[2] K. Suryowati, R.D. Bekti, R. Fajiriyah, E. Siswoyo, "The effect of regional characteristics and relationship among locations in air pollution using spatial autoregressive (SAR) and spatial durbin models (SDM)," in *Proceeding, Journal of Physics: Conference Series*, 2021.
[3] M. L. Laia, R. Deswanto, E. S. Utami, dan R. D. Bekti, "Metode Spatial Durbin Model Untuk Analisis Demam Berdarah Dengue Di Kabupaten Bantul," *Jurnal Nasional Teknologi Terapan (JNTT)*, vol. 3, no. 2, 2021
[4] Mariana,"Pendekatan Regresi Spasial Dalam Pemodelan Tingkat Pengangguran Terbuka," *Jurnal Matematika* , Volume 1, 2013.

[5] Y. L. Ridwan, "Pemodelan Regresi Logistik Spasial Dengan Matriks Pembobot Queen Contiguity dan Matriks Jarak Euclidean," *Jurnal Mahasiswa Statistik, Universitas Brawijaya*, vol. 2, no. 2, pp. 1175–1182, 2016.

[6] A. A. I. A. Pratami, I. K. G. Sukarsa, N. L. P. Suciptawati, dan S. P. E. N. Kencana, "Memodelkan Angka Gizi Buruk Di Provinsi Bali Dengan Pendekatan Regresi Spasial, " *Jurnal Matematika*, Vol 10, No 2, pp. 103-110, 2021.

[7] N. K. Usada, K.S. Wanodya, dan N. Trisna, "Analisis Spasial Gizi Kurang Balita di Kota Tangerang Tahun 2019,"*Jurnal Biostatistik, Kependudukan, dan Informatika Kesehatan (BIKFOKES)*, vol 2, no 1, pp. 1-15, 2021.

[8] R. Rahmayeti, "Pemetaan Angka Gizi Buruk pada Balita Menurut Kabupaten/Kota di Provinsi Sumatera Barat Menggunakan Analisis Regresi Spasial" (*Doctoral dissertation, Universitas Negeri Padang*), 2019.

[9] F. B. Lega dan R. D. BektiKAJIAN pengaruh matriks pembobot spasial dalam model spasial data panel untuk menentukan faktor-faktor yang mempengaruhi tingkat kemiskinan di Provinsi Nusa Tenggara Timur," *Jurnal Statistika Industri dan Komputasi*, Vol 8, No 1, pp. 1-14, 2023..

[10] K. Suryowati, R. D. Bekti, dan A. Faradila, "A comparison of weights matrices on computation of dengue spatial autocorrelation, " *in IOP Conference Series: Materials Science and Engineering,* 2018, p. 012052).

[11] A. S. Utami, Y. Yundari, dan N. Imro'ah, "perbandingan beberapa matriks pembobot dalam Spatial Error Model Pada IPM Pulau Kalimantan Tahun 2020, *Bimaster: Buletin Ilmiah Matematika, Statistika dan Terapannya*, Vol 11, No 5, 2020.

[12] R. Deswanto, M. L. Laila,. E.S. Utami, dan R.D. Bekti. Pemodelan Spasial Area dengan R. Yogyakarta: Akprind Press.

[13] F. Fauzi, 2016. "Model Regresi Spasial Terbaik Indeks Pembangunan Manusia Provinsi Jawa Tengah. Skripsi. Semarang: FMIPA Universitas Negeri Semarang

[14] R.D. Bekti, G. Nurhadiyanti, dan E. Irwansyah, "Spatial pattern of diarrhea based on regional economic and environment by spatial autoregressive model," in *AIP Conference Proceedings* Vol. 1621, No. 1, pp. 454-461, 2014.

[15] I. R. Akolo, "Perbandingan Matriks Pembobot Rook dan Queen Contiguity dalam Analisis Spatial Autoregressive Model (SAR) dan Spatial Error Model (SEM)," *Jambura Journal of Probability and Statistics*, Vol 3, No 1, pp. 11-18, 2022.

[16] C. Caroline, E. P. Lestari, C. Srimindarti, D. Kusumawati, dan A. N. Safriandono, "Spatial Interaction Pattern of Local Workers in Central Java Province by using the Euclidean Distance Approach," *International Journal of Business & Management Science*, Vol 10, No 2, 2020

Peramalan Curah Hujan Harian Kabupaten Jember Dengan Jaringan Saraf Tiruan Dan *General Circulation Model*

Abduh Riski [(1)], Ahmad Kamsyakawuni [(2)], Cahya Ramadhani Azhar [(3)]

[1,2,3] Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Jember

Jalan Kalimantan No. 37, Jember, Jawa Timur, Indonesia

e-mail: riski.fmipa@unej.ac.id [(1)]

ABSTRAK

Curah hujan memiliki peran penting di beberapa bidang seperti pertanian dan pengairan. Oleh sebab itu diperlukan model peramalan untuk mengetahui curah hujan di masa yang akan datang. Model peramalan dapat dibentuk menggunakan jaringan saraf tiruan (JST) *backpropagation*. Hasil akurasi peramalan JST diukur dengan MAE, korelasi dan RMSE. Data lokal sebagai data target model merupakan data rataan curah hujan harian dari 73 stasiun di wilayah kabupaten Jember mulai dari Oktober 2019 hingga Desember 2020. Data global sebagai data input model menggunakan data *Global Circulation Model* (GCM) model CSIRO-MK3-6-0 dengan eksperimen RCP 2.6. Data GCM direduksi menggunakan *principal component analysis* (PCA) guna menghindari multikolinieritas pada data. Penelitian ini mengkombinasikan jumlah neuron sebesar 10 hingga 100 neuron dan dua fungsi aktivasi pada model JST. Berdasarkan hasil penelitian, model terbaik yang digunakan untuk peramalan adalah model JST dengan 100 neuron dan fungsi aktivasi biner dengan MAE sebesar 6,1205, korelasi sebesar -0,0125, dan RMSE sebesar 9,0251. hasil peramalan curah hujan harian kabupaten Jember untuk bulan Januari 2021 adalah terjadi curah hujan tertinggi pada hari ke-19 sebesar 10,0471 mm/hari dan curah hujan terendah terdapat pada hari ke-2 sebesar 1,3106 mm/hari.
Kata kunci: curah hujan harian; *general circulation model*; jaringan saraf tiruan; peramalan

ABSTRACT

Rainfall is essential in several fields, such as agriculture and irrigation. Therefore, a forecasting model is needed to determine future rainfall. Forecasting models can be formed using artificial neural networks (ANN) backpropagation. MAE, correlation, and RMSE measure the results of ANN forecasting accuracy. Local data as the model target data is the average daily rainfall data from 73 stations in the Jember regency from October 2019 to December 2020. Global data as the input data model uses the Global Circulation Model (GCM) data model CSIRO-MK3-6-0 with the RCP 2.6 experiment. GCM data is reduced using principal component analysis (PCA) to avoid multicollinearity. This study combines the number of neurons of 10 to 100 neurons and two activation functions in the ANN model. Based on the research results, the best model used for forecasting is the ANN model with 100 neurons and a binary activation function with an MAE of 6.1205, a correlation of -0.0125, and an RMSE of 9.0251. the results of forecasting the daily rainfall for the Jember regency for January 2021 are that the highest rainfall occurred on the 19th day at 10.0471 mm/day, and the lowest rainfall occurred on the second day at 1.3106 mm/day.
Keywords: *daily rainfall; general circulation models; artificial neural networks; forecasting*

PENDAHULUAN

Indonesia merupakan negara beriklim tropis dan hanya memiliki dua musim yang didasarkan atas banyaknya curah hujan, kedua musim tersebut adalah musim hujan dan musim kemarau. Salah satu bidang yang bergantung pada kondisi hujan adalah bidang pertanian, karena tinggi rendahnya curah hujan dapat mempengaruhi hasil panen. Kabupaten Jember merupakan daerah yang mayoritas penduduknya bergerak di bidang pertanian. Untuk meminimalisir masalah tersebut, perlu adanya model peramalan yang dapat digunakan untuk meramalkan intensitas curah hujan secara lokal di Kabupaten Jember. Salah satu cara yang dapat dilakukan untuk membuat model peramalan adalah melalui pendekatan data iklim global menggunakan data *General Circulation Model* (GCM).

GCM merupakan alat untuk memperkirakan perubahan iklim global di masa depan yang diukur berdasarkan peningkatan konsentrasi gas rumah kaca di atmosfer. Informasi dari GCM disajikan dalam bentuk *grid*, dimana tiap *grid* memiliki nilai dan persamaan dari parameter yang berbeda tergantung pada resolusi dari penyedia luaran GCM [1]. Model GCM menggunakan luaran *grid-grid* untuk menduga perubahan unsur-unsur cuaca. Data GCM memiliki banyak dimensi yang kemungkinan besar terjadi korelasi antar *grid* dalam domain dan multikolinearitas antar variabel. Semakin besar domain dan semakin banyak variabel yang digunakan, hal ini juga berpengaruh pada kompleksitas model. Sehingga perlu adanya pereduksian variabel tanpa mengurangi variansi data tersebut. *Principal component analysis* (PCA) merupakan metode yang dapat digunakan untuk mereduksi variabel-variabel yang saling berkorelasi. Variabel yang dihasilkan melalui PCA menghasilkan variabel-variabel yang tidak saling berkolerasi. Proporsi keragaman yang dianggap cukup mewakili total keragaman data jika keragaman kumulatif mencapai 70% s/d 80%. Penelitian oleh [2] menghasilkan tingkat akurasi lebih baik menggunakan metode PCA dibanding dengan metode HMM dalam mengenali pola wajah.

Salah satu metode *machine learning* yang dapat digunakan untuk membentuk model peramalan curah hujan harian adalah jaringan saraf tiruan (JST) [3]. Metode pembelajaran JST berbentuk sistem yang memproses informasi dan mempunyai cara kerja seperti jaringan saraf biologi. *Backpropagation* merupakan algoritma yang biasa digunakan dalam pembentukan model JST. Algoritma *backpropagation* ini mengurangi kesalahan atau error dengan menyesuaikan hasil pelatihan berdasarkan target yang ingin dicapai. Algoritma ini memiliki 3 tahap pelatihan, propagasi maju, propagasi mundur, dan modifikasi bobot.

Beberapa penelitian tentang peramalan curah hujan menggunakan JST seperti peramalan curah hujan bulanan kota Palembang menggunakan JST *Backpropagation* menghasilkan akurasi peramalan yang baik dengan tingkat error sebesar 0,2528 [4]. JST juga telah digunakan untuk meramalkan curah hujan bulanan daerah Jember dan menghasilkan RMSE 83,487 [5]. Berdasarkan hal tersebut, pada artikel ini dilakukan peramalan curah hujan harian Kabupaten Jember dengan reduksi variabel GCM menggunakan PCA sebagai metode pra-pemrosesan data dan metode pelatihan menggunakan JST *Backpropagation*. Batasan penelitian ini menggunakan *grid* GCM yang berada pada *latitude*: -19,5852º LS hingga 0,9337º LU dan *longitude*: 103,125º BT hingga 123,750º BB.

METODE PENELITIAN

Terdapat dua jenis data yang digunakan pada artikel ini yaitu data input dan data target. Data input adalah data GCM dari CMIP5 dengan model CSIRO-MK3-6-0, eksperimen RCP 2.6, dan resolusi GCM sebesar 1,875º×1,875º [6]. Data target merupakan data aktual dari 78 stasiun yang tersebar di wilayah kabupaten Jember mulai dari Oktober 2019 hingga Desember 2020. Seluruh data dibagi menjadi dua bagian yaitu data pelatihan sebanyak 366 data (80%) dan data pengujian sebanyak 92 data (20%).

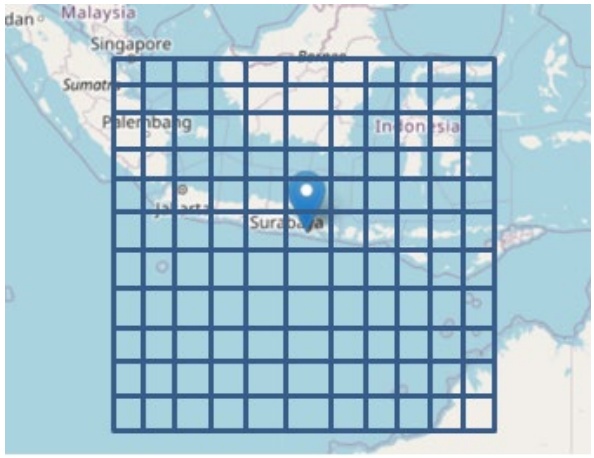
Banyaknya domain GCM dapat memungkinkan terjadinya multikolinieritas antar variabel input. PCA dapat mengatasi masalah multikoliniearitas. Reduksi variabel grid GCM oleh PCA akan menghasilkan variabel baru yang disebut *principal component* (PC). Variabel baru yang dibuat tetap mempertimbangkan varian dari data GCM. PC pertama merupakan kombinasi linier dengan maksimum variansi [7]. Variabel PC ini digunakan sebagai input ke dalam algoritma *backpropagation*.

Selanjutnya dibangun model arsitektur jaringan dengan mengkombinasikan neuron dan fungsi aktivasi ke dalam dua *hidden layer*. Fungsi aktivasi yang digunakan adalah fungsi aktivasi biner dengan rentang nilai [0,1] dan fungsi aktivasi bipolar dengan rentang nilai [-1,1]. Pembelajaran berhenti apabila epoch telah mencapai 1000 atau target minimal error mencapai 10^{-6}. Model terbaik adalah model yang memiliki MAE dan RMSE yang terkecil dan korelasi terbesar. MAE dan RMSE yang kecil menunjukan kemiripan antara data target dan hasil dari pembelajaran. Model JST terbaik kemudian dipilih untuk digunakan sebagai model peramalan.

HASIL DAN PEMBAHASAN

A. Konstruksi Data

Titik wilayah Kabupaten Jember diambil di pusat kota Jember pada *longitude* 113,7047581˚ dan *latitude* -8,17704765˚. Kabupaten Jember terletak di antara grid baris ke-6 kolom ke-6 ($grid_{6,6}$), dengan *longitude* 112,500˚ BT s/d 114,375˚ BB dan *latitude* -8,3937˚ LS s/d -10,2589˚ LU. Penentuan ukuran domain *grid* yang digunakan berukuran $n \times n$ yaitu ukuran domain *grid* 1×1, 3×3, 5×5 hingga 11×11. Ukuran-ukuran domain *grid* tersebut menentukan banyaknya variabel input. Gambaran wilayah dalam bentuk grid dapat dilihat pada Gambar 1.



Gambar 1. Domain *Grid* GCM

B. Pra-Pemrosesan Data

Domain *grid* yang digunakan sebagai variabel input berukuran 1×1, 3×3, 5×5 hingga 11×11. Banyak variabel input beragam sesuai dengan grid GCM yang digunakan. Berdasarkan Tabel 1 persentase varian kumulatif ukuran domain *grid* 3×3 adalah sebesar $\geq 98\%$, dengan PC yang tereduksi menjadi 6 variabel. Peneliti menetapkan PC minimum yang memenuhi persentase varian kumulatif $\geq 98\%$ sebagai PC terpilih. Persentase varian kumulatif setiap ukuran domain *grid* dapat dilihat pada Tabel 1. Jumlah PC masing-masing domain *grid* pada Tabel 1 merupakan jumlah variabel input yang akan digunakan pada proses pelatihan. Pada ukuran domain *grid* 3×3 dengan variabel input sebanyak 6 PC mampu menjelaskan sebesar 98,55% varian dari data GCM.

Tabel 1. Persentase Varian Kumulatif PC Setiap Ukuran Domain *Grid*

Ukuran Domain *Grid*	Jumlah PC	Persentase Varian Kumulatif (%)
1×1	1	100
3×3	6	98,55
5×5	15	98,31
7×7	27	98,00
9×9	44	98,05
11×11	65	98,03

C. Pelatihan Model JST

Hasil model pelatihan dengan fungsi aktivasi biner dengan jumlah iterasi 1000 dan jumlah neuron pada lapisan tersembunyi: 10, 20, 30, sampai dengan 100 memberikan nilai yang beragam. Nilai RMSE, MAE dan koefisien korelasi (COR) pelatihan dan pengujian menggunakan fungsi aktivasi biner dan bipolar dapat dilihat pada Tabel 2.

Tabel 2. Hasil Pelatihan dan Pengujian dengan Fungsi Aktivasi Biner dan Bipolar

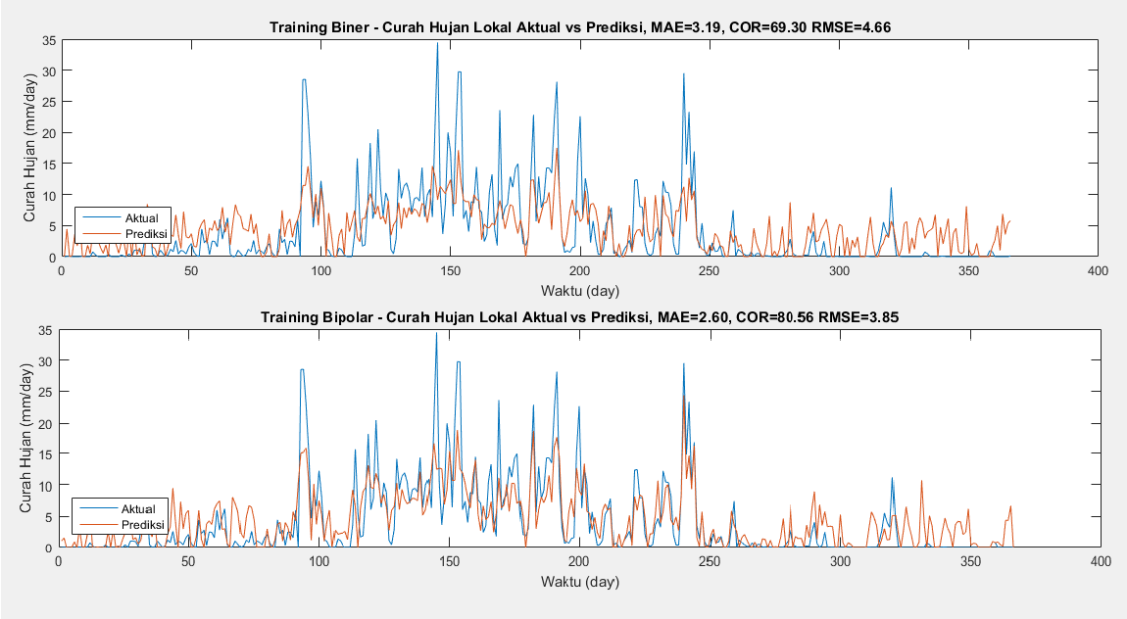
Ukuran Domain	Neuron	Fungsi Aktivasi	Pelatihan			Pengujian		
			MAE	*COR*	*RMSE*	*MAE*	*COR*	*RMSE*
1×1	90	Biner	4,0639	0,3775	5,8693	6,1376	0,0509	8,8596
3×3	90	Biner	3,9840	0,3844	5,8524	6,1711	-0,0778	8,9706
5×5	40	Biner	3,7824	0,4929	5,5446	5,9585	0,0335	9,0930
7×7	70	Biner	3,6154	0,5423	5,3605	5,9891	0,0623	8,9220
9×9	90	Biner	3,5122	0,5930	5,1593	6,0246	0,1123	8,8052
11×11	100	Biner	3,1900	0,6930	4,6633	6,1205	-0,0125	9,0251
1×1	90	Bipolar	3,9233	0,4006	5,8070	6,1837	0,0287	8,9209
3×3	80	Bipolar	3,7240	0,4647	5,6225	6,3585	-0,0902	9,4232
5×5	50	Bipolar	3,9445	0,4174	5,7765	5,7826	0,1343	8,7424
7×7	50	Bipolar	3,2685	0,6227	5,0157	5,9219	0,1730	8,6167
9×9	90	Bipolar	2,9285	0,7016	4,5898	6,3321	0,1116	9,0666
11×11	100	Bipolar	2,5984	0,8056	3,8519	6,6568	-0,0110	9,3085

Dari Tabel 2 nilai MAE, koefisien korelasi (COR), dan RMSE tiap domain *grid* berbeda-beda. Dari hasil pada Tabel 2 neuron yang lebih banyak cenderung menghasilkan MAE yang

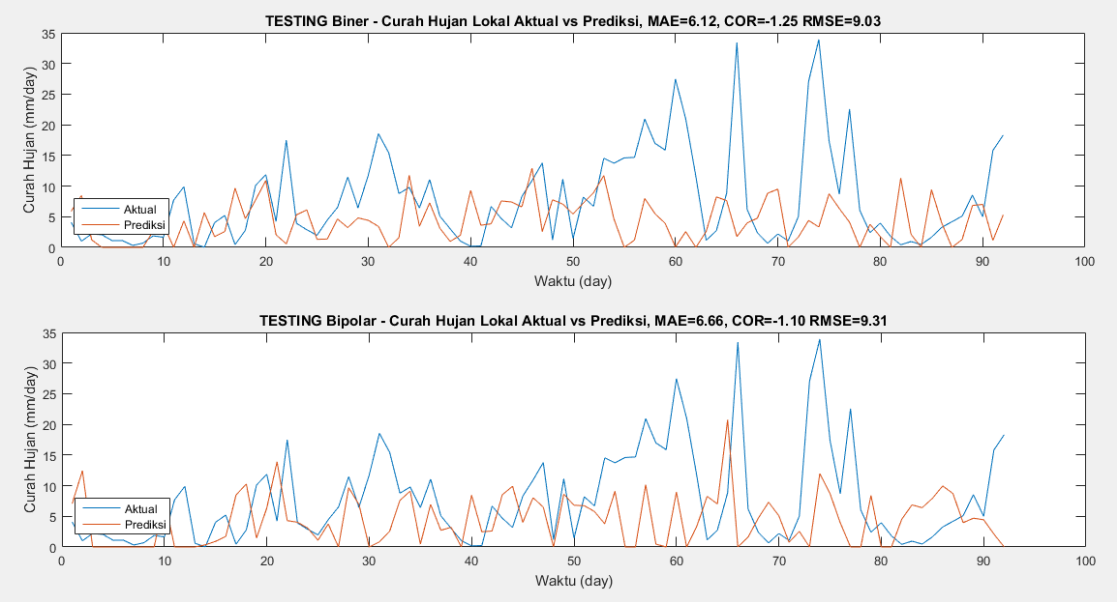
rendah pada pelatihan dan pengujian. Nilai terendah MAE dan RMSE pada domain grid 1 × 1 dengan 100 neuron untuk pelatihan sebesar 4,06 dan 5,85, akan tetapi dibandingkan 90 neuron, MAE dan RMSE yang dihasilkan menjadi lebih tinggi yaitu sebesar 6,1685 dan 8,9949. Sehingga model terbaik untuk domain *grid* 1 × 1 menggunakan model dengan neuron 90.

D. Analisis dan Pemilihan Model Terbaik

Pemilihan model JST pada fungsi aktivasi biner mempertimbangan nilai MAE dan RMSE pada pelatihan dan pengujian. Berdasarkan Tabel 2, nilai MAE dan RMSE pelatihan pada domain *grid* 11 × 11 menghasilkan nilai MAE dan RMSE yang terkecil. Pada pengujiannya diperoleh MAE dan RMSE dari fungsi aktivasi biner lebih kecil dari fungsi aktivasi bipolar. Nilai MAE dan RMSE fungsi aktivasi biner 6,1205 dan 9,0251 sementara pada fungsi aktivasi bipolar sebesar 6,6568 dan 9,3085. Grafik pelatihan dan pengujian model JST *grid* 11 × 11 dengan fungsi aktivasi biner dan bipolar dapat dilihat pada Gambar 2 dan Gambar 3. Pada kedua gambar tersebut terlihat bahwa hasil pelatihan curah hujan harian Kabupaten Jember menggunakan fungsi aktivasi biner maupun fungsi aktivasi bipolar memiliki pola curah hujan yang hampir sama. Sehingga model JST dengan *grid* 11 × 11 menjadi ukuran domain yang terpilih (model terbaik) untuk digunakan sebagai model peramalan, dengan fungsi aktivasi biner dan 100 neuron.



Gambar 2. Grafik Pelatihan Model dengan Fungsi Aktivasi Biner dan Bipolar



Gambar 3. Grafik Pengujian Model dengan Fungsi Aktivasi Biner dan Bipolar

E. Peramalan

Hasil pemodelan JST terbaik kemudian digunakan untuk meramalkan curah hujan harian di Kabupaten Jember selama 1 bulan penuh yaitu 1 Januari 2021 sampai dengan 31 Januari 2021. Parameter yang digunakan adalah 2 *hidden layer* dengan fungsi aktivasi biner, 1 *output layer* dengan fungsi aktivasi linier, 100 neuron pada hidden layer. Domain *grid* yang terbaik adalah *grid* dengan ukuran 11 × 11. Sebanyak 65 variabel yng mewakili 121 variabel dengan persentase varian kumulatif sebesar 98% sebagai variabel input. Hasil peramalan curah hujan harian Kabupaten Jember Januari 2021 dapat dilihat pada Tabel 3.

Tabel 3. Hasil Peramalan Curah Hujan Harian Kabupaten Jember Januari 2021

Hari	mm/hari	Hari	mm/hari	Hari	mm/hari
1	5.8603	11	8.5583	21	5.9703
2	1.3106	12	8.6513	22	8.9885
3	8.8921	13	6.9761	23	8.3052
4	2.6628	14	6.0941	24	6.6885
5	3.5648	15	5.9329	25	8.442
6	7.3438	16	3.7436	26	8.5982
7	2.5936	17	2.1324	27	7.9574
8	5.4087	18	6.2416	28	7.1832
9	6.8419	19	10.0471	29	6.0312
10	4.7952	20	9.0329	30	7.4895
				31	7.4289

SIMPULAN DAN SARAN

Model JST dengan fungsi aktivasi biner menggunakan dua *hidden layer* dan 100 neuron menjadi model terbaik untuk melakukan peramalan curah hujan harian Kabupaten Jember. Pemodelan tersebut diperoleh dengan menggunakan domain *grid* GCM 11 × 11 sebagai variabel input dan menghasilkan nilai MAE 6,1205, Koefisien Korelasi -0,0125, dan RMSE 9,0251.

UCAPAN TERIMAKASIH

Penelitian ini didanai oleh Lembaga Penelitian dan Pengabdian kepada Masyarakat (LP2M) Universitas Jember melalui Hibah Kelompok Riset dan Pengabdian Masyarakat Tahun 2023 No. 3287/UN25.3.1/LT/2023.

DAFTAR PUSTAKA

[1] A. H. Wigena, "Pemodelan statistical downscaling dengan regresi projection pursuit untuk peramalan curah hujan bulanan," *Disertasi. IPB*, 2006.

[2] A. R. Syakhala, D. Puspitaningrum, and E. P. Purwandari, "Perbandingan Metode Principal Component Analysis (PCA) Dengan Metode Hidden Markov Model (HMM) Dalam Pengenalan Identitas Seseorang Melalui Wajah," *Rekursif: Jurnal Informatika*, vol. 3, no. 2, Mar. 2016, doi: 10.33369/rekursif.v3i2.743.

[3] L. V. Fausett, *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*. London: Prentice-Hall, 1994.

[4] I. Sofian and Y. Apriani, "Metode Peramalan Jaringan Saraf Tiruan Menggunakan Algoritma Backpropagation (Studi Kasus Peramalan Curah Hujan Kota Palembang)," May 2017.

[5] A. Riski, A. F. Hadi, O. Tazkiyah, and D. Anggraeni, "Neural Network and Principal Component Analysis on Statistical Downscaling for Local Rainfall Forecasting," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 12, pp. 812–826, May 2020.

[6] K. E. Taylor, R. J. Stouffer, and G. A. Meehl, "An Overview of CMIP5 and the Experiment Design," *Bull Am Meteorol Soc*, vol. 93, no. 4, pp. 485–498, Apr. 2012, doi: 10.1175/BAMS-D-11-00094.1.

[7] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 6th ed. Upper Saddle River: Pearson Prentice Hall, 2007.

Analisis Regresi Logistik Biner Multilevel pada Status Kemiskinan di Pulau Jawa menggunakan Algoritma MCMC Metropolis-Hasting

Regita Putri Permata [(1)], Rifdatun Ni'mah[(2)]

[1]Institut Teknologi Telkom Surabaya

Ketintang 156, Surabaya

e-mail: regitapermata@ittelkom-sby.ac.id[(1)], rifdatun@ittelkom-sby.ac.id[(2)]

ABSTRAK

Pulau Jawa adalah pulau paling padat penduduk di Indonesia. Namun, ada beberapa provinsi di Pulau Jawa yang mengalami masalah kemiskinan. Provinsi Jawa Tengah memiliki tingkat kemiskinan sebesar 11,32% pada tahun 2018, lebih tinggi dari persentase kemiskinan pulau Jawa, yang merupakan akumulasi dari kemiskinan di semua kabupaten dan kota di provinsi tersebut. Model regresi logistik multilevel mempunyai struktur data hirarki yang terdiri dari satu variabel prediktor yang diukur pada level paling bawah (level 1) dan satu variabel penjelas yang diukur pada setiap level atau level selanjutnya. Struktur hirarki data kemiskinan mengindikasikan bahwa data tersebut berasal dari beberapa level, dimana level yang lebih rendah yaitu Kabupaten/Kota tersarang pada level yang lebih tinggi yaitu Provinsi. Data persentase kemiskinan daerah diubah ke dalam bentuk biner menjadi variabel status kemiskinan sehingga metode pendugaan parameter dilakukan dengan pendekatan model regresi logistik biner hirarki dengan algoritma *Metropolis-Hasting*. Pemodelan ini membantu pemerintah dalam mengambil kebijakan terhadap kelompok kabupaten/kota kategori miskin berdasarkan nilai Indeks Pembangunan Manusia (IPM). Analisa pemodelan menunjukkan bahwa variabel IPM memberikan dampak yang sama saja bagi kecenderungan status kemiskinan kabupaten/kota dengan asumsi parameter lain konstan. Variabel interaksi antara IPM dan dana program Bantuan Pangan Non Tunai memberikan dampak kecenderungan kabupaten/kota di Pulau Jawa berstatus tidak miskin sebesar 1,07 kali daripada miskin.

Kata kunci: Kemiskinan, Regresi Logistik, Hirarki, Metropolis-Hasting.

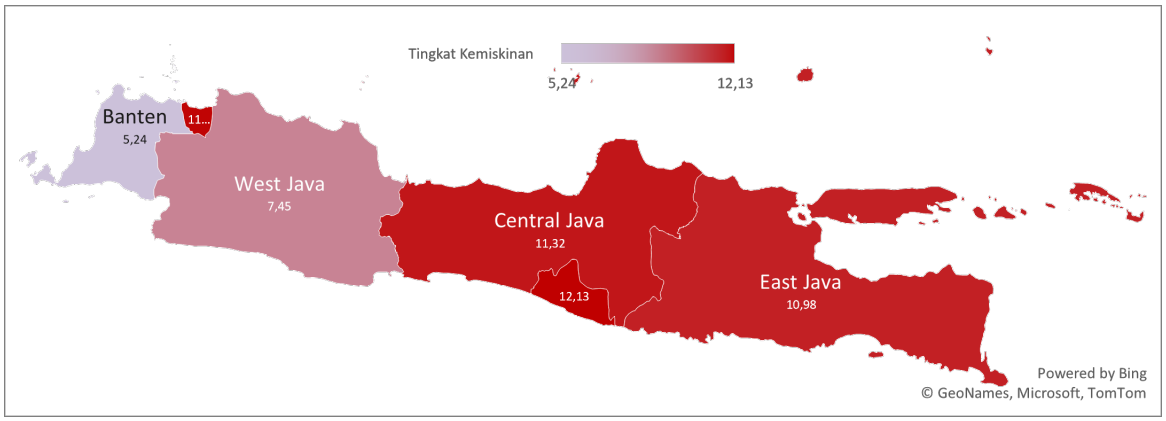
ABSTRACT

In Indonesia, Java Island has the highest density of inhabitants. However, socioeconomic issues are present in a number of Javan provinces. In 2018, the poverty rate in the Central Java province was 11.32%, which is greater than the overall Java island poverty rate due to the accumulation of poverty throughout all the province's cities and regencies. Multilevel logistic regression model has hierarchical data structure consisting one predictor variable measured at the lowest level (level 1) and one explanatory variable measured at each level or the next level. The hierarchical structure of poverty data indicates that the data comes from several levels, where the lower level is the Districts nested at the higher level, namely Province. Regional poverty percentage data is converted into binary form for poverty status variables so that the parameter estimation method is carried out using a hierarchical binary logistic regression model with the Metropolis-Hasting algorithm approach. This modeling helps the government in making policies towards the regional poverty status based on the Human Development Index (HDI). Analysis result shows that the HDI variable has the same impact on the trend of poverty status in districts/cities with the assumption that other parameters are constant. The interaction variable between HDI and Non-Cash Food Assistance program funds has an impact on the tendency of districts in Java Island to have a non-poor status of 1.07 times than poor.

Keywords: *Poverty, Logistics Regression, Hierarchy, Metropolis-Hasting.*

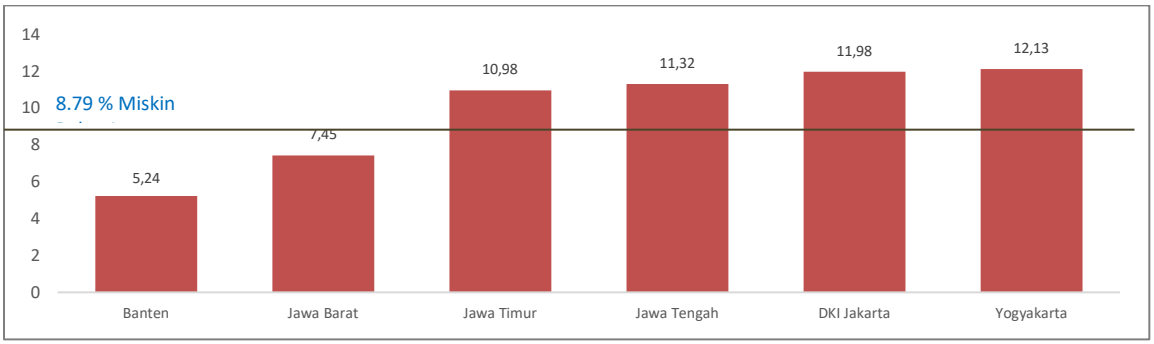
PENDAHULUAN

Menurut *World Bank,* kemiskinan adalah tingkat kesejahteraan yang rendah. Hampir setiap negara di dunia menghadapi fenomena kemiskinan yang dapat memengaruhi berbagai aspek masyarakat. Definisi kemiskinan menurut Badan Pusat Statistik (BPS) Indonesia adalah ketidakmampuan untuk memenuhi kebutuhan dasar yang diukur dari sisi pengeluaran [1]. Kemiskinan merupakan masalah yang sangat berbahaya bagi setiap daerah karena sering terjadi dan sulit untuk diatasi tanpa bantuan.



Gambar 1. Pemetaan tingkat kemiskinan di Pulau Jawa

Jumlah penduduk Pulau Jawa yang sebesar 13,19 juta orang adalah paling padat di Indonesia. Pulau Jawa yang terdiri dari enam provinsi Daerah Khusus Ibukota (DKI) Jakarta, Jawa Barat, Banten, Jawa Tengah, Daerah Istimewa Yogyakarta (DIY) dan Jawa Timur memiliki tingkat kemiskinan sebesar 8,79% [1]. Gambar 1 menunjukkan bahwa semakin merah, tingkat kemiskinan terus meningkat. Tingkat kemiskinan Jawa Tengah sebesar 11,32% pada tahun 2018 lebih tinggi dari persentase kemiskinan pulau Jawa. Provinsi DIY memiliki tingkat kemiskinan paling tinggi sebesar 12,13% disusul DKI Jakarta sebesar 11,98% dan Jawa Tengah sebesar 11,32%. Tingkat kemiskinan ini merupakan akumulasi dari semua kabupaten/kota di provinsi. Gambar 2 menunjukkan bahwa persentase kemiskinan di Jawa Timur, Jawa Tengah, DKI Jakarta, dan DIY masih lebih tinggi dari rata-rata Pulau Jawa. Banten dan Jawa Barat masih di bawah rata-rata.



Gambar 2. Persentase tingkat kemiskinan di Pulau Jawa

Indeks Pembangunan Manusia (IPM) adalah salah satu komponen yang dapat mempengaruhi tingkat kemiskinan. Jika IPM rendah, produktifitas kerja masyarakat juga akan rendah, yang berarti lebih banyak orang miskin. Penelitian yang dilakukan oleh Fadila menemukan bahwa IPM memiliki dampak negatif dan signifikan terhadap tingkat kemiskinan di Provinsi Sumatera Barat. Artinya, peningkatan IPM biasanya berkorelasi dengan penurunan tingkat kemiskinan [2]. Jumlah orang yang menerima Program Bantuan Pangan Non Tunai (BPNT) di setiap kabupaten/kota juga merupakan faktor yang dapat mempengaruhi tingkat kemiskinan. Pengaruh Program BPNT terhadap kemiskinan sangat negatif dan signifikan [3]. Keluarga Penerima Manfaat (KPM) yaitu keluarga dengan kondisi ekonomi 25% terendah di daerah pelaksanaan merupakan penerima BPNT. BPNT untuk setiap provinsi adalah 6.2 T dimana Jawa Tengah sebesar 6.9 T, Jawa Barat sebesar 4.7 T, DKI Jakarta sebesar 4.5 T, DIY sebesar 4 T dan Banten sebesar 6 T.

Indeks kedalaman kemiskinan digunakan untuk mengukur ketimpangan kemiskinan. Indeks ini cenderung tidak berubah dari tahun ke tahun. Oleh karena itu, penting untuk mengetahui faktor yang mempengaruhi tingkat kemiskinan saat ini di Pulau Jawa. Penelitian terhadap indeks kedalaman kemiskinan di Indonesia menunjukkan bahwa variabel IPM, persentase penduduk miskin, dan persentase rumah tangga yang memiliki akses air minum layak semuanya mempengaruhi klasifikasi daerah tertinggal di Indonesia [4]. Pada tahun 2020, Ni Putu dan Maulidia melakukan penelitian dengan analisis regresi logistik biner untuk menentukan apakah tingkat kedalaman kemiskinan provinsi-provinsi di Indonesia tepat untuk diklasifikasikan [5]. Respon dalam penelitian ini hanya terdiri dari dua kategori, yaitu tingkat kedalaman kemiskinan rendah dan tinggi serta adanya konsep hirarki, maka regresi logistik biner multilevel digunakan untuk menyelesaikan masalah tersebut. Regresi logistik yang tepat digunakan jika variabel respon kategorik dan variabel prediktornya diskrit, kontinu, atau kombinasi keduanya [6].

Penelitian ini menggunakan pemodelan multilevel dengan model regresi logistik yang mempunyai struktur data hirarki yaitu sebuah variabel prediktor yang diukur pada level paling bawah (level 1) dan satu variabel penjelas yang diukur pada setiap level atau level selanjutnya. Pada regresi biasa intersep dan kemiringan untuk setiap kelompok nilainya sama (*fixed*), sedangkan pada model multilevel intersep maupun kemiringan untuk setiap kelompok nilainya bisa berbeda (random), sehingga dapat dilihat variasi antar kelompok. Penelitian multilevel pernah digunakan untuk analisis pengaruh karakteristik individu, rumah tangga, dan wilayah terhadap status kemiskinan balita di kepulauan Maluku dan pulau Papua tahun 2019 [7]. Kekurangan dari penelitian tersebut tidak memberikan efek variasi dan distribusi dari setiap parameter yang estimasi. Pemodelan multi-level dan metode Markov Chain Monte Carlo (MCMC) adalah dua metode dengan peningkatan dalam kemampuan komputer baik dalam penyimpanan dan kecepatan operasi [8]. Multilevel MCMC dengan algoritma Metropolis Hastings dapat mengurangi waktu dan biaya komputasi estimator yang besar dari proses Markov dengan tolerasi 0,01 [9]. Pemodelan dengan multi-level MCMC lebih bersifat umum dan fleksibel sehingga sangat mudah diaplikasikan dengan regresi logistik biner [10]. Penelitian serupa menggunakan metode bayes pernah diterapakna untuk menganalisis faktor-faktor yang mempengaruhi bayi berat lahir rendah dengan model regresi logistik biner di RSUD Semarang [11]. Tujuan penelitian ini adalah untuk mengetahui pengaruh faktor indeks pembangunan manusia dan program bantuan non tunai terhadap status kemiskinan tinggi atau rendah di Pulau Jawa menggunakan Metode Regresi Logistik Hirarki dengan algoritma MCMC *Metropolis-Hasting*.

METODE

Data yang digunakan dalam penelitian ini adalah data indeks pembangunan manusia dan persentase kemiskinan tahun 2018 Kabupaten/kota Pulau Jawa. Data diakses melalui BPS pada tanggal 17 Mei 2020. Status kemiskinan ditransformasi menjadi data berskala biner dengan ketentuan jika persentase kemiskinan Kabupaten/kota berada di bawah persentase kemiskinan Pulau Jawa maka tersebut dikategorikan 0, sedangkan berada di atas persentase kemiskinan Pulau Jawa dikategorikan 1. Prediktor yang digunakan pada level pertama (Kabupaten) adalah IPM dan level kedua (Provinsi) adalah BNPT. Variabel yang digunakan untuk penelitian ini sebagai berikut:

Tabel 1. Variabel Status Kemiskinan

	Variabel	Keterangan	Skala
Level 1 (Kabupaten)			
Y	Status Kemiskinan	1= miskin (diatas 8.79) 0= tidak miskin (di bawah 8.79)	Nominal
X_1	Indeks Pembangunan manusia		Rasio
Level 2 (Provinsi)			
Z	Dana BPNT		Rasio

Struktur data penelitian mengindikasikan bahwa data yang dianalisis berasal dari beberapa level, dimana level yang lebih rendah tersarang pada level yang lebih tinggi. Struktur data tersebut dapat diselesaikan dengan model multilevel logistik biner. Analisa regresi logistik biner digunakan untuk mencari hubungan antara variabel respon (Y) yang bersifat biner atau dichotomous dengan variabel bebas (X) yang bersifat polychotomous [12].

Jika sekumpulan variabel $X_1, X_2, \dots, X_n$ adalah sampel acak yang berdistribusi bersyarat X dengan *probability density function* $f(x|\boldsymbol{\beta}), \boldsymbol{\beta} \in \Omega$ maka fungsi *likelihood* dimana $\mathbf{x} = (x_1, x_2, \dots, x_n)$ dapat ditulis:

$$L(\boldsymbol{\beta}; x) = f(x_1, \boldsymbol{\beta}).(x_2, \boldsymbol{\beta})\dots(x_n, \boldsymbol{\beta}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{\beta}) \tag{1}$$

dengan fungsi probabilitas binomial untuk setiap pasangan x adalah

$$f(x_i, \boldsymbol{\beta}) = P(x_i)^{y_i}[1 - P(x_i)]^{1-y_i};\ y_i = 0,1;\ P(x_i) = \frac{e^{\sum_{j=1}^{p} \beta_j x_{ij}}}{1+e^{\sum_{j=1}^{p} \beta_j x_{ij}}} \tag{2}$$

Untuk satu kejadian peubah respon Y yang mengikuti sebaran Bernoulli, fungsi penghubung regresi logistik biner yang digunakan adalah logit $P(X)$. Dari Persamaan 2, transformasi logit $P(X)$ diperoleh:

$$g(X) = \text{logit}P(X) = ln\left(\frac{P(X)}{1-P(X)}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k \tag{3}$$

Jika respon biner maka penaksiran parameternya memerlukan fungsi penghubung [13]. Secara umum rumusan matematis regresi logistik hirarki multilevel dengan satu prediktor pada level satu dan level kedua sebagai berikut.

$$Y_{ij} = P_{ij}; Y_{ij} \sim Binom(n_{ij}, P_{ij})$$
$$\text{Logit}(P_{ij}) = ln\left[\frac{P(X)}{1-P(X)}\right] = \beta_{0j} + \beta_{1j}X_{1j} \quad (4)$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \quad (5)$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \quad (6)$$

dimana indeks j adalah menunjukkan kelompok pada level 1, j=1,2,...,6. Dengan mensubstitusikan persamaan (5) dan (6) terhadap (4), maka menghasilkan persamaan model regresi logistik hirarki 2 level pada persamaan (7) :

$$\text{logit}(P_{ij}) = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}X_{1j} + \gamma_{11}X_{1j}W_j + u_{1j}X_{1j} \quad (7)$$

Metode penduga parameter model regresi biner hirarki adalah *Metropolis-Hasting.* Algoritma ini memerlukan *proposal distribution* (q) untuk membangkitkan sampel random yang mengikuti distribusi proposal yang akan menjadi kandidat parameter untuk diterima atau ditolak sebagai sampel dari distribusi posterior berupa fungsi likelihood. Distribusi posterior sangat penting digunakan dalam MCMC. Misal $f(\boldsymbol{\beta}|x)$ ialah distribusi posterior yang akan dibangkitkan dengan ukuran T, misalkan $\beta(t)$ ialah vektor dari nilai-nilai yang dibangkitkan di iterasi ke-t dari algoritma. Model distribusi posterior parameter β dinyatakan dalam persamaan (8) [14]:

$$f(\beta|y) = \frac{f(y|\beta)f(\beta)}{f(y)} \quad (8)$$

Langkah kerja *Metropolis-Hasting* dapat dibaca lebih lanjut pada [15].

Langkah yang dilakukan untuk mengestimasi parameter dalam membangun model gabungan regresi biner hirarki supaya dapat memprediksi status kemiskinan dan mengetahui besar faktor yang mempengaruhi status kemiskinan di Pulau Jawa adalah sebagai berikut:

a) Mendapatkan *link function* dari distribusi bernoulli $P(Y = y_i) = P_i^{y_i}(1 - P_i)^{1-y_i}$ dimana $log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$ maka didapatkan

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \quad (9)$$

b) Mendapatkan fungsi likelihood

$$L(\beta_0, \beta_1|y) = \prod_{i=1}^{n}\left(\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right)^{y_i}\left(\frac{1}{1+e^{\beta_0+\beta_1 x}}\right)^{1-y_i} \quad (10)$$

c) Mendapatkan inisialisasi awal parameter beta dari MLE.

d) Melakukan estimasi parameter regresi logistik biner level 1 pada status kemiskinan dan IPM dengan menjalankan algoritma *Metropolis-Hasting*
- Memilih distribusi proposal $g(.|X_i)$ sebagai target distribusi.
- Ambil nilai awal sebagai inisialisasi yaitu X_0 .
- Membangkitkan X_0 dari distribusi g dan parameter proposal sesuai distribusi likelihood *u*.
- Membangkitkan U berdistribusi uniform(0,1) jika $U \leq \frac{f(u)g(X_i|u)}{f(X_i),g(u|X_i)}$ iterasi jalan pada *i*.
- Kandidat parameter dari *u* diterima bila $R(X_i, Y) = min\left(1, \frac{f(u)g(X_i|u)}{f(X_i),g(u|X_i)}\right)$
- Melakukan iterasi N=100.000 untuk langkah diatas dengan dengan *burn in* B=10000 agar tidak bias hingga mendapatkan distribusi parameter yang konvergen.

e) Melakukan estimasi parameter regresi linier parameter level 2 yang didapatkan parameter level 1 sebagai respon dan anggaran dana BPNT prediktor baru seperti langkah d) pada fungsi likelihood linier

$$L(\beta_0, \beta_1|\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp(y_i - (\beta_0 + \beta_1 x_i))^2 \quad (11)$$

f) Mengintepretasi hasil dugaan parameter dan menarik kesimpulan.

HASIL DAN DISKUSI

Kriteria dalam mengkategorikan presentase penduduk miskin Kabupaten/Kota di Pulau Jawa berdasarkan rata-rata tingkat kemiskinan di Pulau Jawa sebesar 8.79%. Apabila persentase kemiskinan berada di bawah Pulau Jawa, maka Kabupaten/kota dalam provinsi tersebut dikategorikan 0 artinya tidak miskin, sedangkan jika presentase penduduk miskin berada di atas rata-rata tingkat kemiskinan Pulau Jawa, maka Kabupaten/kota dalam provinsi tersebut dikategorikan 1 artinya miskin. Kategori ini dilakukan agar pemerintah dapat mengambil kebijakan melihat kabupaten/kota yang dikategorikan miskin dengan meningkatkan nilai indeks pembangunan manusia dari segi pendidikan atau pendapatan perkapita. Dari Tabel 2 menunjukkan bahwa jumlah kabupaten/kota dengan indeks miskin tinggi di Pulau Jawa lebih banyak daripada daerah dengan indeks miskin rendah. Status kemiskinan terbanyak berada di Provinsi Jawa Timur sebesar 21,8%. Kemiskinan tinggi di Provinsi Jawa Timur berada pada daerah pedesaan, sedangkan di perkotaan cukup rendah. Tingkat kemiskinan ditentukan dengan pendapatan perkapita tiap rumah tangga, apabila di Jawa Timur kemiskinan tinggi, artinya orang kerja di Jawa Timur, upahnya masih rendah karena bekerja di sektor pertanian dengan Lahan garap pertanian rata-rata 0,5 hektar. Menurut BPS sepuluh Kabupaten/kota termiskin di Jawa Timur adalah Kab. Sampang, Sumenep, Bangkalan, Probolinggo, Tuban, Ngawi, Pamekasan, Pacitan, Bondowoso, dan Lamongan.

Tabel 2. Status Miskin tiap Provinsi

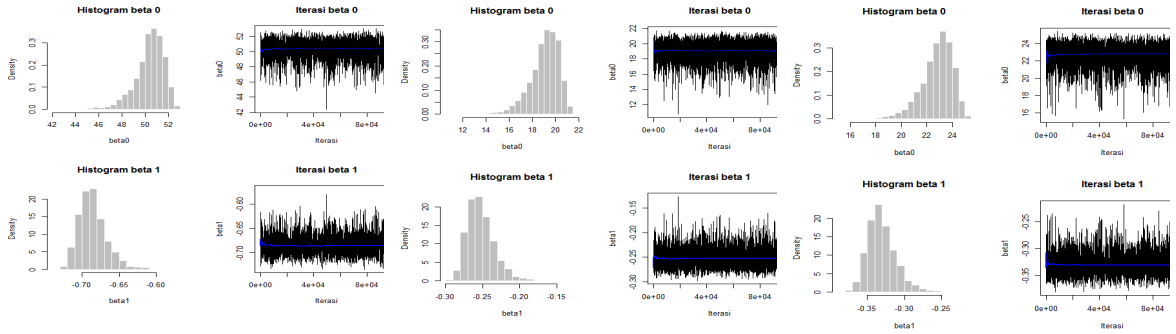
Provinsi		Banten	DIY	DKI Jakarta	Jabar	Jateng	Jatim	**Total**
Status Miskin	**Tidak Miskin**	7	2	4	16	11	12	**52**
	Miskin	1	3	2	11	24	26	**67**
Total		8	5	6	27	35	38	**119**

IPM menggambarkan indeks pengembangan manusia yang dilihat dari sisi perluasan, pemerataan, dan keadilan baik dalam bidang kesehatan, pendidikan, maupun kesejahteraan masyarakat. Rendahnya IPM akan mengakibatkan pada rendahnya produktivitas kerja dari penduduk. Produktivitas yang rendah mengakibatkan rendahnya perolehan pendapatan, sehingga menyebabkan tingginya jumlah penduduk miskin. IPM tertinggi berada di Provinsi DKI Jakarta mencapai 79,86. Artinya penduduk DKI Jakarta memiliki produktivitas kerja yang baik dan tingkat pendapatan yang relatif tinggi. Selain itu IPM terendah di Pulau Jawa adalah Jawa Timur dan Jawa Barat sebesar 70,97.

Tabel 3. Karakteristik IPM

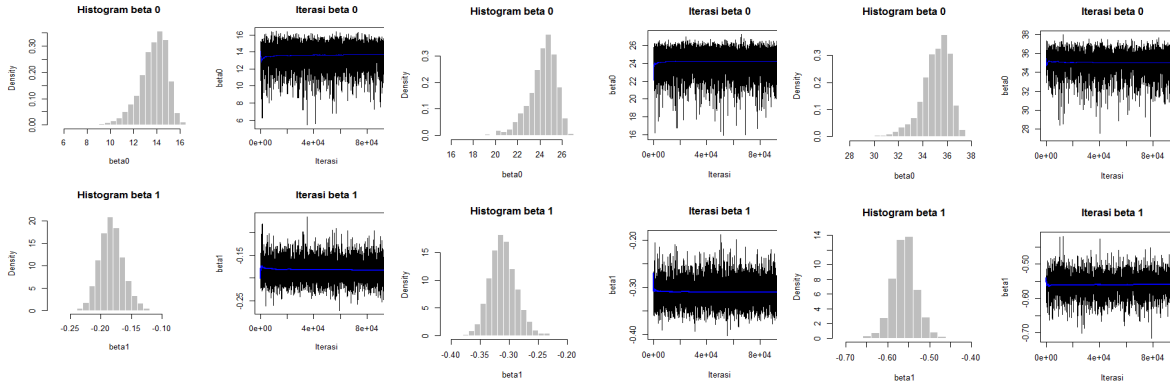
Provinsi	Banten	DIY	DKI Jakarta	Jabar	Jateng	Jatim
Rata-rata	71,08	75,80	79,86	70,97	71,79	70,97
Minimum	63,37	69,24	70,91	64,62	65,67	61,00
Maksimum	81,17	83,42	84,44	81,06	82,72	81,74

Level 1 berupa kabupaten/kota di setiap provinsi di pulau Jawa. Masing-masing iterasi estimasi parameter menggunakan *link function* biner dan secara keseluruhan bentuk histogram menunjukkan *bell shape* dan cenderung berpola normal, baik beta 0 maupun beta 1. Konvergensi iterasi N=100.000 dari masing-masing provinsi di Jawa disajikan dalam Gambar 3.



(a) (b) (c)

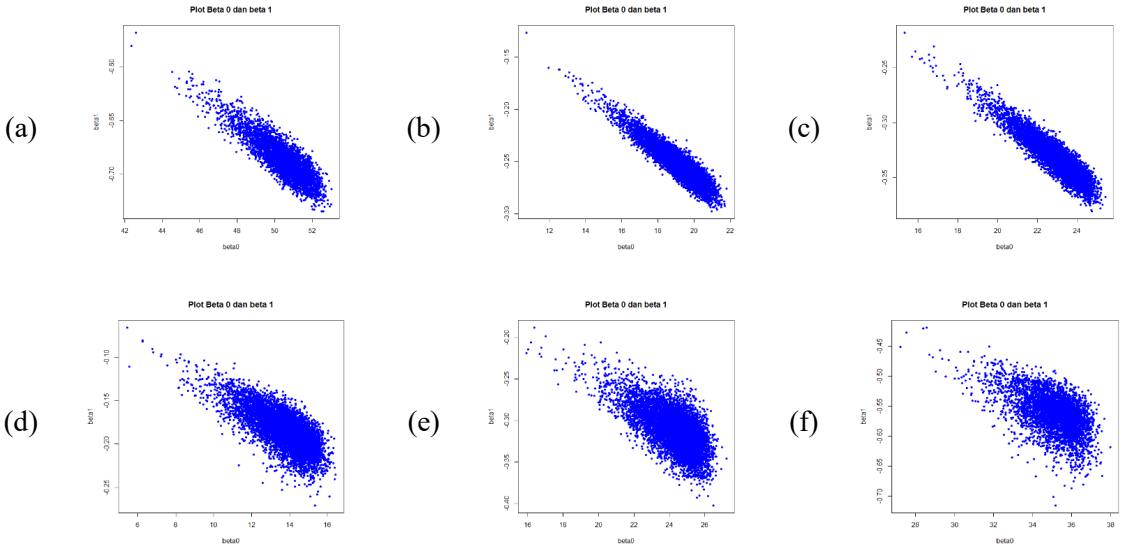
Gambar 3. Konvergensi regresi level 1 untuk: (a) Jawa Timur; (b) Jawa Tengah; (c) Jawa Barat;



(d) (e) (f)

Gambar 4. Konvergensi regresi level 1 untuk: (d) DKI Jakarta; (e) DIY; dan (f) Banten

Simulasi MCMC dilakukan dengan algoritma *Metropolis-Hasting* hingga mencapai konvergensi mencapai 100.000 iterasi dan kondisi 10.000 *burn-in. K*onvergensi algoritma diindikasikan melalui histogram dan *trace plot.* Gambar plot antara β_0 dan β_1 tiap provinsi di Pulau Jawa ditunjukkan pada Gambar 4. Hubungan β_0 dan β_1 mengumpul pada satu sisi dan membentuk pola tertentu. Hal ini menunjukkan adanya *bootstrap* dalam proses *Metropolis-Hasting sampling.*



Gambar 5. Plot $\beta_0$dan β_1 untuk a) Jatim, b) Jateng, c) Jabar, d) DKI Jakarta, e) DIY f) Banten

Pemodelan regresi logistik pada level 1 dimana terdiri dari kabupaten/kota di masing-masing provinsi sehingga didapatkan estimasi parameter regresi sehingga didapatkan estimasi parameter regresi pada Tabel 4. Hasil estimasi parameter ditunjukkan pada Tabel 4 bahwa indeks pembangunan manusia berbanding terbalik dengan status kemiskinan. Untuk lebih mudah dalam melakukan intepretasi maka dihitung nilai odds ratio parameter IPM dari masing-masing Provinsi. Semakin tinggi indeks pembanguna manusia di Provinsi Jawa Timur, maka kecenderungan kabupaten/kota berstatus indeks miskin rendah sebesar 2 kali lebih besar daripada berstatus miskin. Begitu pula dengan provinsi yang lainnya memiliki kecenderungan untuk kabupaten/kota berstatus tidak miskin lebih besar daripada berstatus miskin. Artinya indeks pembangunan manusia sangat berpengaruh terhadap status kemiskinan di Pulau Jawa. Pemerintah diharapkan dapat memberikan bantuan dana untuk meningkatan IPM dari segi pendidikan, atau segi pengeluaran rumah tangga berupa dana BPNT dari pemerintah provinsi yang dianalisis pada regresi hirarki level 2.

Tabel 4. Estimasi Parameter Level 1

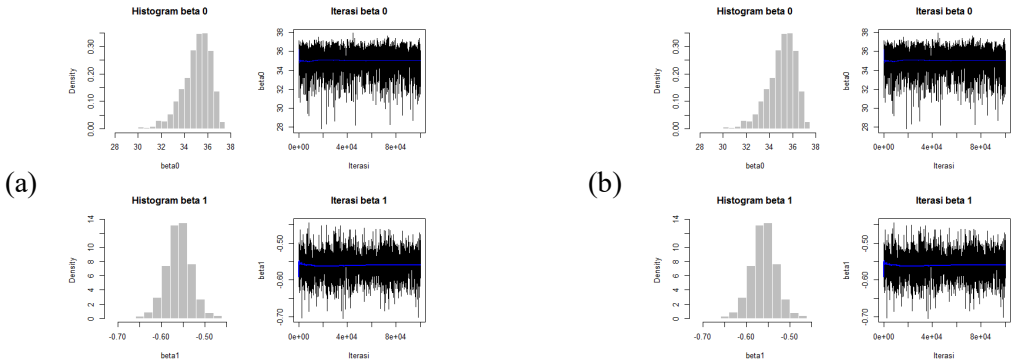
Parameter	Jawa Timur	Jawa Tengah	Jawa Barat	DKI Jakarta	DIY	Banten
$\boldsymbol{\beta_0}$	50,385	19,065	22,800	13,675	24,193	35,051
$\boldsymbol{\beta_1}$	-0,686	-0,253	-0,331	-0,183	-0,311	-0,561
$\boldsymbol{exp(\beta_1)}$	0,504	0,777	0,718	0,833	0,733	0,571
$\boldsymbol{1/exp(\beta_1)}$	1,985	1,287	1,392	1,201	1,364	1,985

Interval konfidensi 95% yang didapatkan dari estimasi parameter β_0, β_1 pada level 1 dirangkum pada Tabel 5. Tabel 5 menunjukkan bahwa konfidensi interval dari estimasi parameter setiap provinsi tidak memuat nilai nol, namun nilainya negative. Artinya dengan bertambahkan IPM, maka dapat menurunkan rata-rata persentase indeks kemiskinan.

Tabel 5. Interval Konfidensi 95% Parameter Level 1

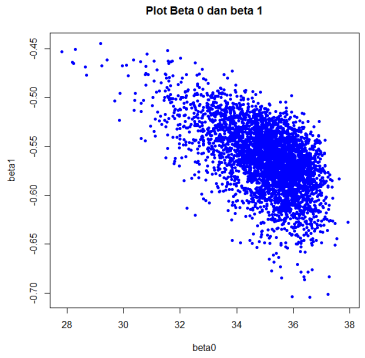
	Provinsi	Jawa Timur	Jawa Tengah	Jawa Barat	DKI Jakarta	DIY	Banten
CI $\boldsymbol{\beta_0}$	**Batas Bawah**	50,377	19,058	22,792	13,667	24,185	35,043
	Batas Atas	50,393	19,073	22,807	13,682	24,201	35,059
CI $\boldsymbol{\beta_1}$	**Batas Bawah**	-0,6859	-0,2528	-0,3309	-0,1829	-0,3107	-0,5608
	Batas Atas	-0,6857	-0,2525	-0,3306	-0,1826	-0,3104	-0,5605

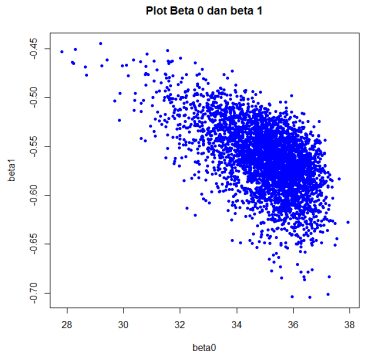
Untuk pemodelan regresi logistik biner level 2 berupa tingkat provinsi, terdapat enam parameter untuk mendapatkan persamaan regresi hirarkinya. Pada level 2 berbeda dengan level 1, level 2 parameter linier sehingga menggunakan regresi hirarki linier dengan estimasi parameter numerik *Metropolis-Hasting*. Pada level 2 ini digunakan distribusi proposal mengikuti distribusi normal baik di parameter beta 0 maupun beta 1, sehingga didapatkan konvergensi sampai N=100.000.



Gambar 6. Konvergensi regresi level 2 untuk (a) β_0 dan (b) β_1

Untuk mengetahui pola antara parameter w_0 dan w_1 ditunjukkan pada Gambar 6. Pada persamaan β_0, plot parameter w_0 dan w_1 membentuk pola tertentu, begitu pula dengan persamaan β_1. Sehingga dapat diduga bahwa adanya *bootstrap* dalam proses *Metropolis-Hasting*.

(a) 

(b) 

Gambar 6. Plot hubungan w_0 dan w_1 a) parameter β_0b) parameter β_1

Regresi hirarki level 2 dengan menambahkan variabel prediktor baru yaitu dana BPNT dari pemerintah provinsi sebagai upaya mengurangi penduduk miskin di Pulau Jawa sehingga tingkat kemiskinan kabupaten/kota di provinsi tertentu dapat berada di bawah rata-rata tingkat kemiskinan Pulau Jawa. Estimasi parameter level 2 ditunjukkan Tabel 6.

Tabel 6. Estimasi Parameter Level 2

Variabel	CI 95% untuk w_0			CI 95% untuk w_1		
	Batas Bawah	**Rata-rata**	**Batas Atas**	**Batas Bawah**	**Rata-rata**	**Batas Atas**
β_0	2,543	2,596	2,649	4,626	4,635	4,645
β_1	0,003	0,004	0,005	-0,073	-0,073	-0,073

Proses perhitungan odds rasio hirarki model berdasarkan persamaan (3) didapatkan berikut:

$$log\left(\frac{p(x)}{1-p(x)}\right) = 2{,}596 \text{ sehingga } \left(\frac{p(x)}{1-p(x)}\right) = exp(2{,}596) = 13{,}4 \quad (12)$$

Hasil estimasi level 2 menunjukkan bahwa kondisi dimana dana BPNT tidak berpengaruh pada status kemiskinan di Pulau Jawa. Meningkatnya dana BPNT pada tiap provinsi dapat meningkatkan status miskin sebesar 13,4 kali. Hal ini tidak sesuai dengan harapan pemerintah yang seharusnya dapat menurunkan kecenderungan kabupaten/kota miskin. Untuk parameter β_1 artinya dana BPNT dapat menurunkan kecenderungan status kemiskinan yang berkolaborasi dengan meningkatkan IPM. Kedua variabel ini berinteraksi sehingga mampu memberikan dampak yang signifikan pada kabupaten/kota supaya tidak miskin. Interval konfidensi 95% pada Tabel 7 yang didapatkan dari estimasi parameter w_0 dan w_1 pada level 2 menunjukkan bahwa setiap estimasi parameter w_0 dan $w_1$signifikan pada iterasi 100.000. Signifikansi terlihat dari nilai kuantil pada *credible interval* antara 5% dan 95% yang tidak memuat nilai nol. Artinya BNPT yang diberikan setiap provinsi berpengaruh signifikan terhadap indeks kemiskinan melalui peningkatan IPM. Estimasi parameter telah didapatkan maka dilakukan pemodelan regresi logistik biner hirarki 2 level sehingga didapatkan persamaan model sebagai berikut

$$Logit(P_{ij}) = 2{,}596 + 4{,}635W_j + 0{,}004X_{1j} - 0{,}073X_{1j}W_j + u_{0j} + u_{1j}X_{1j} \quad (13)$$

KESIMPULAN

Variabel IPM memberikan dampak yang sama saja bagi kecenderungan kabupaten/kota berstatus miskin atau tidak miskin, dengan asumsi parameter lain konstan. Sedangkan pada variabel interaksi antara IPM dan dana BPNT memberikan dampak kecenderungan kabupaten/kota di Pulau Jawa berstatus tidak miskin sebesar 1,07 kali daripada miskin. Dengan menggunakan interval konfidensi 95% bahwa dengan meningkatkan IPM, BNPT yang diberikan kepada setiap provinsi berpengaruh terhadap indeks kemiskinan.

DAFTAR PUSTAKA

[1] BPS, "Badan Pusat Statistik," 2017. [Online]. Available: https://www.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html. [Accessed 20 01 2022].

[2] R. Fadila and Marwan, "Pengaruh Indeks Pembangunan Manusia (IPM) dan Pertumbuhan Ekonomi terhadap Tingkat Kemiskinan di Provinsi Sumatera Barat periode tahun 2013-2018," *Jurnal Ecogen,* vol. III, pp. 120-133, 2020.

[3] L. R. Nadhifah and N. H. Mustofa, "Pengaruh PKH dan BPNT terhadap Kemiskinan dengan Pertumbuhan Ekonomi Sebagai Variabel Moderasi," *Journal of Islamic Economics and Banking,* vol. 3 No. 1, pp. 12-24, 2021.

[4] R. E. Fitri, E. Setiawan and D. Aziz, "Analisis Regresi Logistik Biner Terhadap Data Indeks Kedalaman Kemiskinan Di Indonesia Tahun 2020," *Jurnal Siger Matematika,* vol. 3, pp. 69-74, 2022.

[5] N. P. N. Hendayanti and M. Nurhayati, "Regresi Logistik Biner dalam Penentuan Ketepatan Klasifikasi Tingkat Kedalaman Kemiskinan Provinsi-Provinsi di Indonesia," *Jurnal Sains dan Teknologi,* pp. 63-70, 2020.

[6] A. Agresti, Categorical Data Analysis, Second Edition, New York: John Wiley and Sons, 2002.

[7] O. V. Amida and J. R. H. Sitorus, "Penerapan Regresi Logistik Biner Multilevel dalam Analisis Pengaruh Karakteristik Individu, Rumah Tangga, dan Wilayah terhadap Status Kemiskinan Balita di Kepulauan Maluku dan Pulau Papua (Analisis Hasil Survei Sosial Ekonomi 2019)," in *Seminar Nasional Official Statistics 2020:Pengembangan Official Statistics dalam mendukung Implementasi SDG's*, 2020.

[8] W. J. Browne, Applying MCMC Methods to Multi-level Models, United Kingdom: University of Bath, 1998.

[9] T. J. Dodwell, C. Ketelsen, R. Scheichl and A. L. Teckentrup, "Multilevel Markov Chain Monte Carlo," *Society for Industrial and Applied Mathematics and American Statistical Association,* vol. Vol. 3, pp. 1075-1108, 2019.

[10] F. Syafitri, R. Goejantoro and Wasono, "Regresi Logistik dengan Metode Bayes untuk Pemodelan Indek s Pembangunan Manusia Kabupaten/Kota di Pulau Kalimantan," *Jurnal EKSPONENSIAL,* pp. 103-110, 2021.

[11] L. Nadhifah, H. Yasin and Sugito, "Analisis Faktor-Faktor yang Mempengaruhi Bayi Berat Lahir Rendah dengan Model Regresi Logistik Biner menggunakan Metode Bayes," *Jurnal Gaussian,* pp. 125-134, 2012.

[12] D. Hosmer and S. Lemeshow, Applied Logistic Regression, New York: John Willey&Sons, 2000.

[13] H. Jox, Multilevel Analysis Techniques and Applications, New Jersey: Lawrence Erlbaum Associates, 2002.

[14] G. Box and G. Tiao, Bayesian Inference in Statistical Analysis, New York: John Wiley and Sons, 1973.

[15] D. Kurniawati, "Faktor-Faktor yang mempengaruhi Anemi Remaja Putri dengan menggunakan bayesian Regresi Logistik dan Algoritma Metropolis- Hasting," *Jurnal Ilmiah Matematika,* vol. Volume 7 No.1, pp. 1-6, 2019.

Peramalan Nilai Ekspor Migas di Indonesia dengan Model *Long Short Term Memory* (LSTM) dan *Gated Recurrent Unit* (GRU)

Prissy Nusaiba Yulisa [(1)], M. Al Haris [(2)*], Prizka Rismawati Arum [(3)]

Universitas Muhammadiyah Semarang,

Jl. Kedungmundu No.18, Kedungmundu, Kec. Tembalang, Kota Semarang, Jawa Tengah 50273

e-mail: prissynusaiba12@gmail.com(1), alharis@unimus.ac.id(2)*, prizka.rismawatiarum@unimus.ac.id(3)

ABSTRAK

Ekspor migas merupakan komoditas yang berperan penting dalam perekonomian negara dan pengelolaannya harus dimaksimalkan demi kemakmuran dan kesejahteraan rakyat. Namun realitanya, dalam kurun waktu 10 tahun terakhir, neraca perdagangan ekspor migas di Indonesia mengalami defisit sehingga berdampak pada pengeluaran negara lebih besar daripada pemasukan. Penelitian ini difokuskan pada peramalan yang dapat dijadikan sebagai bahan masukan bagi pemerintah dalam merencanakan arah kebijakan terkait ekspor migas pada masa mendatang. Metode peramalan yang digunakan dalam penelitian ini adalah *Long Short Term Memory* (LSTM) dan *Gated Recurrent Unit* (GRU) dengan optimasi *Nesterov Adam* (Nadam). LSTM mampu mengatasi masalah ketergantungan jangka panjang, sehingga dapat mengenali pola data dengan baik dan GRU merupakan variasi lain dari LSTM yang memiliki komputasi lebih sederhana. Sedangkan Nadam berperan dalam mempercepat proses training dan menurunkan nilai *error*. Berdasarkan hasil penelitian, diperoleh akurasi tertinggi dalam prediksi nilai ekspor migas menggunakan model terbaik LSTM dengan optimasi Nadam pada percobaan menggunakan nilai parameter α 0.001, jumlah neuron 20, epoch 100, dan nilai MAPE 12.8% dengan akurasi 87.2%.

Kata kunci : Ekspor migas, Gated Recurrent Unit, Long Short Term Memory, Peramalan.

ABSTRACT

Oil and gas exports were commodities that play an important role in the country's economy and their management must be maximized for the prosperity and welfare of the people. However, in reality, in the last 10 years, the trade balance of oil and gas exports in Indonesia has experienced a deficit so the impact on state expenditure was greater than income. This research was focused on forecasting that can be used as input for the government in planning policy directions related to oil and gas exports in the future. The forecasting methods used in this research are Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) with Nesterov Adam (Nadam) optimization. LSTM can overcome long-term dependency problems, so it can recognize data patterns well and GRU is another variation of LSTM which has simpler computations. Meanwhile, Nadam played a role in accelerating the training process and reducing error values. Based on the results of the study, the highest accuracy was obtained in predicting the value of oil and gas exports using the best LSTM model with Nadam optimization in the experiment using the parameter value 0.001, the number of neurons 20, epoch 100, and the MAPE value 12.8% with accuracy was 87.2%.

Keywords : Forecasting, Gated Recurrent Unit, Long Short Term Memory, Oil and gas exports.

PENDAHULUAN

Ekspor merupakan bentuk kegiatan ekonomi berupa menjual produk dari dalam negeri ke pasar di luar negeri [1]. Dampak dari adanya Globalisasi menyebabkan pengaruh diberbagai sektor salah satunya pertumbuhan yang meningkat dalam perdagangan internasional, dimana setiap negara dapat melakukan ekspor produknya ke negara lain dan impor kebutuhan mereka. Nilai ekspor di Indonesia merupakan perolehan hasil dalam kegiatan ekspor yang mengacu pada nilai *Free On Board* (FOB) dan diukur dalam satuan USD [2]. Terdapat beberapa manfaat dari adanya kegiatan ekspor, diantaranya menambah pemasukan kas negara dalam bentuk devisa, banyak tercipta lapangan pekerjaan dalam negeri, dan memajukan perekonomian dengan memperluas pasar [3], [4]. Dalam hal ini ekspor memiliki peranan yang sangat penting dalam mendorong pertumbuhan ekonomi, dimana pertumbuhan ekonomi yang baik adalah salah satu indikator keberhasilan yang menunjukkan tingkat produktivitas dari suatu negara. Selain itu, dengan meningkatnya kegiatan ekspor maka produksi yang dihasilkan oleh industri atau perusahaan akan semakin banyak, hal ini otomatis akan meningkatkan jumlah penyerapan tenaga kerja dan mengurangi angka pengangguran.

Di Indonesia, terdapat 2 bentuk kegiatan ekspor, yaitu ekspor migas dan non-migas [2]. Komoditi migas berupa minyak bumi dan gas alam. Sedangkan komoditi non-migas terdiri dari industri manufaktur, pertanian, hasil pertambangan, perkebunan dan sebagainya. Berdasarkan UU RI No. 22 tahun 2001, minyak dan gas bumi adalah komoditas vital yang berperan penting dalam perekonomian negara dan pengelolaannya harus dimaksimalkan demi kemakmuran dan kesejahteraan rakyat. Namun realitanya, dalam kurun waktu 10 tahun terakhir tepatnya tahun 2012 hingga 2021, neraca perdagangan ekspor migas di Indonesia mengalami defisit yang artinya pengeluaran negara lebih besar daripada pemasukan.

Defisit ekspor migas mengakibatkan terganggunya pembangunan nasional secara umum karena tidak tersedia cukup dana untuk melakukan investasi. Terbatasnya sumber dana dalam negeri membuat pemerintah memberlakukan kebijakan Utang Luar Negeri [5]. Ketika suatu negara sering melakukan utang luar negeri maka akan berdampak pada berkurangnya jumlah cadangan devisa. Dengan menurunnya cadangan devisa, mengakibatkan pertumbuhan ekonomi menurun sehingga nilai tukar rupiah melemah [6]. Maka untuk memenuhi kebutuhan dalam negeri dilakukan impor secara besar-besaran, sehingga ekspor akan semakin menyusut karena tingkat produktivitasnya rendah.

Dari permasalahan di atas, dapat dikatakan bahwa salah satu sebab defisit ekspor migas masih sering terjadi di Indonesia adalah perencanaan dan perbaikan yang belum matang. Perencanaan baik jangka pendek, maupun jangka panjang dapat dijadikan sebagai acuan kegiatan atau langkah apa yang harus diambil untuk mencegah minimnya ekspor migas di masa depan. Perencanaan ini dapat dilakukan dengan mengetahui informasi mengenai perkiraan nilai ekspor migas di masa depan [7]–[9]. Untuk memperkirakan nilai ekspor migas di masa yang akan datang, diperlukan sebuah metode peramalan (*forecasting*) yang dapat menghasilkan tingkat ketepatan yang tinggi, sehingga hasilnya bisa dijadikan dasar dalam merencanakan strategi bagi pihak-pihak yang berkepentingan, khususnya Badan Pengembangan Ekspor Nasional (BPEN) dan Kementerian Perdagangan [7].

Data ekspor migas sebagai data penelitian merupakan data *time series non-linier*. Salah satu metode peramalan berbasis data *time series non-linier* adalah Jaringan Saraf Tiruan (JST). Menurut

Aprilianto (2018), metode jaringan saraf tiruan lebih baik dalam hasil peramalannya jika dibanding dengan metode lainnya, karena proses komputasinya dilakukan secara berulang-ulang [10]. Dengan asumsi tersebut, penggunaan metode ini diharapan dapat menghasilkan peramalan yang tepat dengan tingkat kesalahan seminimal mungkin. Bentuk metode JST yang terbukti memiliki performa yang baik dalam beberapa kasus yaitu *Long Short Term Memory* (LSTM) dan *Gated Recurrent Unit* (GRU) [11]–[14].

Pada hakikatnya, metode JST memerlukan waktu kalkulasi yang lama untuk mencapai konvergen serta dapat mengalami masalah *overfitting* [15], termasuk metode LSTM dan GRU. Oleh karena itu, untuk mendapatkan hasil yang optimal diterapkan suatu algoritma optimasi dalam penyusunan jaringan saraf yang berperan dalam mempercepat proses *training* dan memperkecil nilai kesalahan [16]. Dalam penelitian ini, akan diterapkan optimasi Nesterov Adam (Nadam) yang merupakan pengembangan dari algoritma *Adaptive Moment Estimation* (Adam) dengan adanya penambahan momentum *Nesterov Accelerated Gradient* (NAG) [17]. Menurut Michael (2020), momentum Nesterov mencapai hasil yang lebih bagus dibandingkan momentum klasik, sehingga dihipotesiskan bahwa Nadam memiliki performa yang jauh lebih baik dari pada Adam maupun algoritma optimasi yang lain [18].

Berdasarkan observasi peneliti, hingga saat ini belum ada penelitian tentang peramalan ekspor migas yang menggunakan metode LSTM dan GRU. Salah satu contoh penelitian terkait ekspor migas dilakukan oleh [7], yang memprediksi nilai ekspor impor migas dan non-migas Indonesia Menggunakan *Extreme Learning Machine* (ELM), dan diperoleh hasil bahwa jumlah penerapan metode ELM dengan fitur data yang cukup dan jumlah hidden neuron yang cukup dapat memperbaik hasil prediksi [7]. Dari beberapa kasus, penelitian menggunakan JST mendapatkan hasil yang lebih optimal, sehingga sebagai bentuk pembeda dan perbaikan metode sebelumnya pada peramalan ekspor migas, penelitian ini digunakan metode LSTM dan GRU dengan optimasi Nadam.

METODE PENELITIAN

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari *website* resmi Badan Pusat Statistik Indonesia. Jumlah observasi sebanyak 348 amatan yang berupa nilai ekspor migas di Indonesia dalam satuan USD dari bulan Januari 1993 hingga Desember 2021. Data ini merupakan data *time series* yang akan dibagi menjadi 2 yaitu data *training* dan data *testing*. Proporsi masing-masing adalah 60% dan 40% [19]. Struktur data penelitian disajikan pada Tabel 1.

Tabel 1. Struktur data penelitian

Tanggal	Y	Keterangan

31/01/1993	Y_1	Training
28/02/1993	Y_2	Training
31/03/1993	Y_3	Training
⋮	⋮	⋮
⋮	⋮	⋮
31/05/2010	Y_{209}	Training
30/06/2010	Y_{210}	Testing
31/07/2010	Y_{211}	Testing
⋮	⋮	⋮
⋮	⋮	⋮
31/12/2021	Y_{348}	Testing

A. Langkah-Langkah Penelitian

langkah-langkah yang dilakukan pada penelitian ini adalah sebagai berikut:

1. Melakukan analisa statistika deskriptif untuk melihat struktur atau karakteristik dari data time series yang digunakan;
2. Melakukan *Preprocessing data*, dilakukan normalisasi dengan mengubah data aktual menjadi data yang berada pada rentang 0 hingga 1 menggunakan persamaan berikut [12]:

$$X_{sn}\frac{X-X_{min}}{X_{max}-X_{min}} \quad (1)$$

dengan X_{sn} data setelah dinormalisasi, X data input yang akan dinormalisasi, X_{min} nilai terkecil dari keseluruhan data, X_{max} nilai terbesar dari keseluruhan data.

3. Melakukan pembagian data menjadi 2 bagian, yaitu data training dan data testing dengan proporsi masing-masing 60% dan 40%.
4. Mendefinisikan algoritma Optimasi Nadam. Nadam bekerja dengan memperbarui bobot untuk menghaluskan gradien yang berdampak dalam mempercepat proses training dan meningkatkan akurasi. Persamaan Nadam dapat ditulis sebagai berikut [18]:

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{\hat{v}_t}+\epsilon}\left(\beta_1\hat{m}_t + \frac{1-\beta_1}{1-\beta_1^t}\right) \quad (2)$$

dengan

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

m_t dan $\hat{v}_t$ diinisialisasi ke-0, β_1 nilai konstan 0.9 untuk laju penurunan eksponensial pada estimasi momen pertama, β_2 nilai konstan 0.999 pada laju penurunan eksponensial pada estimasi momen kedua, konstanta yang diinisiasi kecil (1×10^{-7}) untuk presisi atau menstabilkan angka yang dihasilkan dari proses *training* [20], [21].

5. Membentuk jaringan arsitektur LSTM terbaik dengan mencobakan kombinasi jumlah *neuron* 10 dan 20 serta *epoch* pada proses *training* sebanyak 100, 200, dan 300 epoch. Gambaran skema arsitektur LSTM disajikan pada Gambar 1 [22].
6. Membentuk jaringan arsitektur GRU terbaik dengan mencobakan kombinasi jumlah neuron 10 dan 20 serta epoch pada proses training sebanyak 100, 200, dan 300 epoch.
7. Membandingkan akurasi model LSTM dengan GRU menggunakan nilai MAE dan MAPE. Persamaan MAPE diformulasikan berikut [23]:

$$MAE = \left|\frac{\sum_{i=1}^{n}(X_t - F_t)}{n}\right| \quad (3)$$

Persamaan nilai MAPE dapat dirumuskan sebagai berikut [24], [25]:

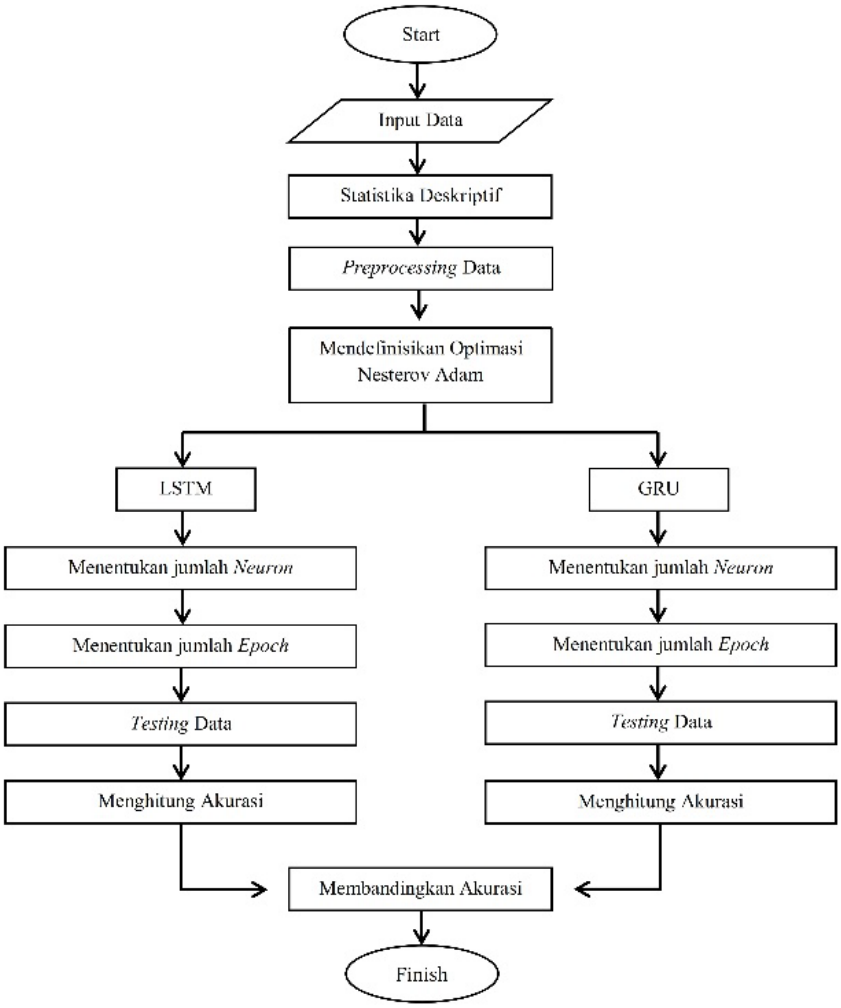
$$\text{MAPE} = \frac{\sum\frac{|F_t - X_t|}{X_t}}{n} \text{ x } 100\% \quad (4)$$

X_t merupakan nilai riil saat periode ke-t, F_t adalah hasil peramalan saat periode ke-t, dan n adalah jumlah observasi keseluruhan. Nilai MAPE dapat diiterpretasikan kedalam beberapa kriteria yang dapat dilihat pada Tabel 2 [12].

Tabel 2. Kriteria nilai MAPE

MAPE	Keterangan
< 10%	Hasil peramalan sangat baik
10-20%	Hasil peramalan baik
20-50%	Hasil peramalan cukup baik
>50%	Hasil peramalan buruk

Proses penelitian juga digambarkan pada diagram alir yang disajikan pada Gambar 1.

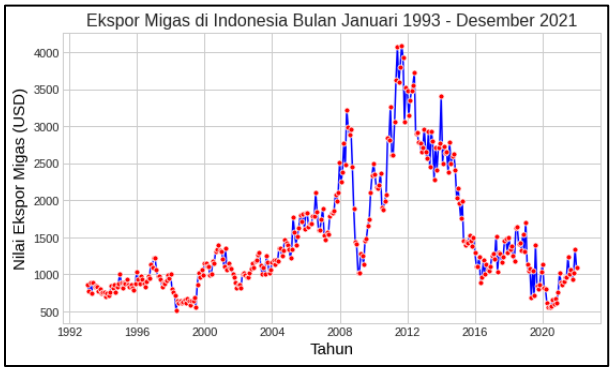


Gambar 1. Alur penelitian peramalan ekspor migas dengan metode LSTM dan GRU

HASIL DAN PEMBAHASAN

A. Analisis Deskriptif

Gambaran umum data nilai ekspor migas dari bulan Januari 1993 hingga Desember 2021 disajikan pada Gambar 2.



Gambar 2. Data nilai ekspor migas dari bulan Januari 1993 hingga Desember 2021

Berdasarkan data nilai ekspor migas di Indonesia dari tahun 1993 hingga 2021, diketahui rata-ratanya berkisar pada angka 1482.9 USD dengan nilai simpangan baku sebesar 775.8, artinya nilai ekspor migas cenderung beragam. Nilai ekspor migas mencapai nilai tertinggi sebesar 4091.6 USD tepatnya di bulan Agustus 2011 dan nilai terendahnya ada di bulan April 1998 dengan nominal sebesar 514 USD. Penelitian ini menggunakan 348 pengamatan yang kemudian dibagi menjadi 60% untuk data *training* sebanyak 209 pengamatan dan 40% data *testing* sebanyak 139 pengamatan.

B. *Preprocessing Data*

Hasil normalisasi seluruh data observasi menggunakan persamaan (1) ditunjukkan pada Tabel 3.

Tabel 3. Hasil normalisasi data penelitian

No	*Date*	Aktual	Normalisasi
1	31/01/1993	864.3	0.0979148
2	28/02/1993	767.5	0.07085756
⋮	⋮	⋮	⋮
347	29/02/2016	1332.4	0.22875671
348	31/03/2016	1093.4	0.16195215

C. Mendefinisikan Nesterov Adam (nadam)

Inisialisasi parameter optimasi Nadam yang digunakan pada penelitian adalah sebagi berikut:

$$\alpha = 0.001 \qquad \beta_1 = 0.9$$
$$\epsilon = 10^{-07} \qquad \beta_2 = 0.999$$

Berdasarkan *default* parameter di atas, penelitian ini dilakukan dengan mencobakan 3 nilai α yang berbeda yaitu 0.1, 0.01, dan 0.001 dan parameter yang lain bernilai tetap.

D. Membentuk jaringan arsitektur LSTM terbaik

Hasil pengolahan data *training* pada model LSTM menggunakan beberapa jumlah *neuron, epoch*, serta nilai α pada optimasi Nadam yang sudah didefinisikan disajikan pada Tabel 4.

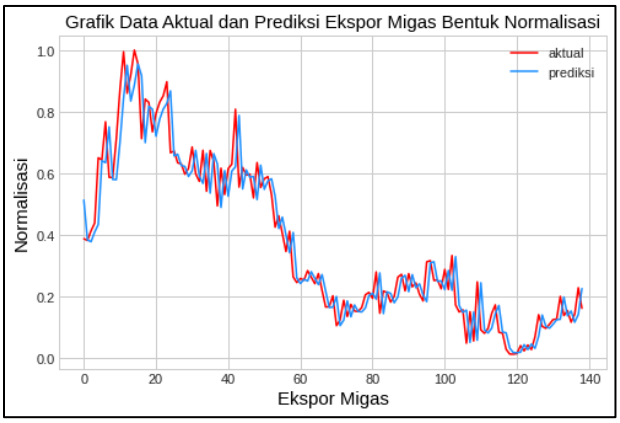
Tabel 4. Hasil *training* arsitektur model LSTM

No	$\boldsymbol{\alpha}$	***Neuron***	***Epoch***	**MAE**
1	0,1	10	100	253.5
2			200	273.2
3			300	249.0
4			100	250.8
5		20	200	263.8
6			300	256.1

7	0,01	10	100	218.1
8			200	226.4
9			300	236.0
10		20	100	244.1
11			200	247.5
12			300	230.3
13	**0,001**	10	100	217.1
14			200	218.2
15			300	217.6
16		**20**	**100**	**214.3**
17			200	216.6
18			300	221.1

Hasil pelatihan arsitektur model LSTM pada Tabel 4 diperoleh arsitektur ke-16 merupakan arsitektur terbaik. Arsitektur tersebut menghasilkan nilai MAE terkecil sebesar 214.3 dengan nilai parameter $\boldsymbol{\alpha}$ sebesar 0.001, jumlah neuron 20, dan epoch sebanyak 100.

Setelah mendapatkan arsitektur terbaik model LSTM pada proses *training*, maka model tersebut akan diuji menggunakan data *testing* 40%. Perbandingan antara data aktual dan hasil prediksi menggunakan arsitektur terbaik model LSTM ditunjukkan pada gambar 3.



Gambar 3. Grafik data aktual dan prediksi arsitektur terbaik model LSTM data normalisasi

Gambar 3 di atas menunjukkan bahwa pola data prediksi yang berwarna biru tidak jauh berbeda dengan pola data aktualnya. Hal ini menunjukkan model LSTM terbaik memiliki performa yang sesuai untuk memprediksi data ekspor migas di Indonesia. Selanjutnya nilai hasil prediksi yang masih dalam bentuk normalisasi dilakukan denormalisasi dengan persamaan (13). Hasil denormalisasi prediksi data *testing* nilai ekspor migas di Indonesia disajikan pada Tabel 5.

Tabel 5. Hasil prediksi data *testing* nilai ekspor migas di Indonesia dengan model LSTM terbaik setelah denormalisasi

No	Tanggal	Aktual	Prediksi

1	30-06-2010	1901.5	2350.21
2	31-07-2010	1881.4	1887.52
3	31-08-2010	1993.5	1867.26
⋮	⋮	⋮	⋮
138	30-11-2021	1332.4	1021.50
139	31-12-2021	1093.4	1322.37

Hasil denormalisasi prediksi data testing nilai ekspor migas di Indonesia kemudian dievaluasi menggunakan nilai MAPE. Hasil perhitungan nilai MAPE diperoleh nilai sebesar 12.8% atau akurasi prediksinya sebesar 87.2%. Berdasarkan hal tersebut, dapat disimpulkan bahwa hasil prediksi nilai ekspor migas di Indonesia dengan model LSTM terbaik dapat dikategorikan baik karena nilai MAPE memenuhi kriteria 10-20%.

E. Membentuk jaringan arsitektur GRU terbaik

Hasil pengolahan data *training* pada model GRU menggunakan beberapa jumlah *neuron, epoch*, serta nilai parameter α pada optimasi Nadam yang sudah didefinisikan disajikan pada Tabel 6.

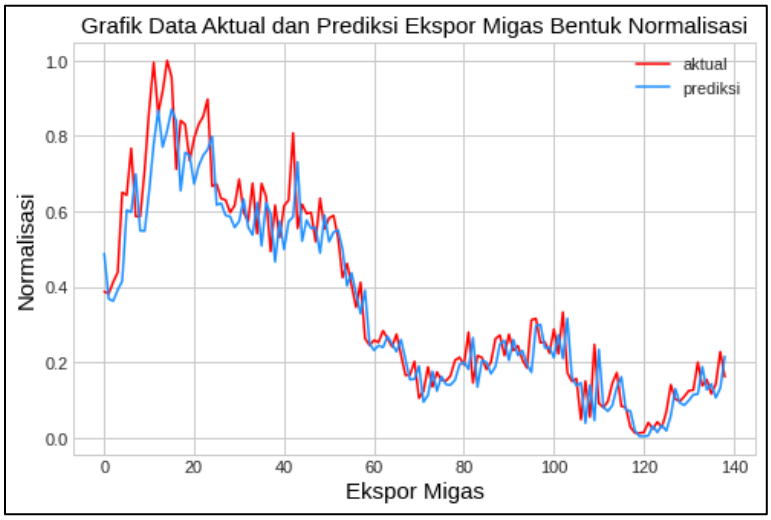
Tabel 6. Hasil *training* arsitektur model GRU

No	α	*Neuron*	*Epoch*	MAE
1	0.1	10	100	296.2
2			200	275.0
3			300	292.9
4		20	100	280.7
5			200	331.9
6			300	296.3
7	0.01	10	100	216.5
8			200	223.3
9			300	231.8
10		20	100	237.3
11			200	253.3
12			300	223.8
13	**0.001**	**10**	100	216.6
14			**200**	**215.6**
15			300	217.8
16		20	100	218.2

17		200	218.5
18		300	217.0

Hasil pelatihan arsitektur model GRU pada Tabel 6 diperoleh arsitektur ke-14 merupakan arsitektur terbaik. Arsitektur tersebut menghasilkan nilai MAE terkecil sebesar 214.3 dengan nilai parameter $\boldsymbol{\alpha}$ sebesar 0.001, jumlah *neuron* 20, dan *epoch* sebanyak 100.

Setelah mendapatkan arsitektur terbaik model GRU pada proses *training*, maka model tersebut akan diuji menggunakan data *testing* 40%. Perbandingan antara data aktual dan hasil prediksi menggunakan arsitektur terbaik model GRU ditunjukkan pada gambar 4.



Gambar 4. Grafik data aktual dan prediksi arsitektur terbaik model GRU data normalisasi

Pengujian arsitektur model GRU menghasilkan performa yang hampir sama dengan arsitektur model LSTM. Pola data prediksi arsitektur model GRU tidak jauh berbeda dengan pola data aktualnya. Selanjutnya nilai hasil prediksi data *testing* nilai ekspor migas di Indonesia dilakukan denormalisasi dan hasilnya ditunjukkan pada Tabel 7.

Tabel 7. Hasil prediksi data *testing* nilai ekspor migas di Indonesia dengan model GRU terbaik setelah denormalisasi

No	Tanggal	Aktual	Prediksi
1	30-06-2010	891.7	2261.65
2	31-07-2010	958.0	1831.60
3	31-08-2010	1187.4	1812.70
⋮	⋮	⋮	⋮

137	31-10-2021	1025.3	895.456
138	30-11-2021	1332.4	986.319
139	31-12-2021	1093.4	1286.57

Hasil denormalisasi tersebut dievaluasi menggunakan MAPE. Hasil perhitungan nilai MAPE diperoleh nilai sebesar 13.3% atau akurasi prediksinya sebesar 86.7%.

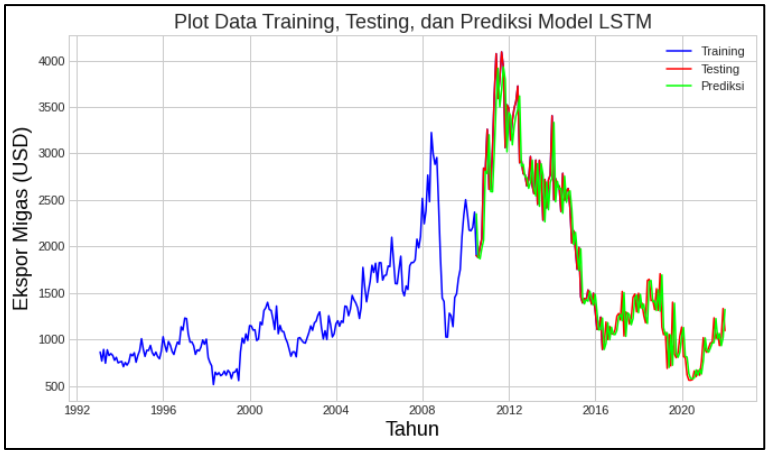
F. Perbandingan Akurasi Peramalan LSTM dan GRU

Setelah dihasilkan arsitektur terbaik model LSTM dan GRU, maka selanjutnya dilakukan evaluasi terhadap kedua model tersebut. Perbandingan akurasi peramalan antara model LSTM dan GRU dilakukan dengan memperhatikan nilai MAPE masing-masing model. Model dengan nilai MAPE terkecil merupakan model yang terbaik. Hasil perbandingan model LSTM dengan model GRU disajikan pada Tabel 8.

Tabel 8. Nilai MAPE prediksi data *testing* nilai ekspor migas di Indonesia model LSTM dan GRU

Metode	MAPE	Akurasi
LSTM	**12,8%**	**87,2%**
GRU	13,3%	86,7%

Hasil pada Tabel 8 menunjukkan bahwa peramalan Ekspor Migas di Indonesia menggunakan arsitektur model LSTM memiliki akurasi yang lebih bagus dibandingkan arsitektur model GRU karena menghasilkan nilai MAPE terkecil sebesar 12,8% dan akurasinya mencapai 87,2%. Berdasarkan arsitektur model terbaik tersebut, dilakukan peramalan untuk bulan berikutnya yaitu Januari 2022 dan diperoleh nilai ramalannya sebesar 1312,48 USD. Grafik performa arsitektur model LSTM terbaik dan hasil prediksinya disajikan pada Gambar 5.



Gambar 5. Grafik data aktual dan prediksi arsitektur terbaik model LSTM

KESIMPULAN

Nilai ekspor migas di Indonesia sepanjang tahun 1993 hingga 2021 mengalami kondisi yang fluktuatif. Ekspor migas mencapai nilai tertinggi pada bulan Agustus tahun 2011 sebesar 4091.6 USD dan terendah pada bulan April tahun 1998 dengan nilai sebesar 514 USD. Arsitektur model LSTM terbaik untuk memprediksi nilai ekspor migas di Indonesia adalah arsitekur dengan menggunakan parameter α 0.001, jumlah *neuron* sebanyak 20 dan *epoch* sebanyak 100. Sedangkan arsitektur model GRU terbaik adalah arsitekur dengan menggunakan parameter α 0.001, jumlah *neuron* sebanyak 10 dan *epoch* sebanyak 200.

Perbandingan arsitektur model LSTM dengan GRU berdasarkan nilai MAPE dihasilkan arsitektur model LSTM merupakan arsitektur terbaik untuk memprediksi nilai ekspor migas di Indonesia karena memiliki nilai MAPE terkecil 13.3% atau memiliki akurasi tertinggi sebesar 87.2%. Hasil peramalan pada bulan berikutnya, yaitu Januari 2022 dengan model terbaik diperoleh nilai ramalannya sebesar 1312,48 USD.

DAFTAR PUSTAKA

[1] F. Farina and A. Husaini, "Pengaruh Dampak Perkembangan Tingkat Ekspor dan Impor Terhadap Nilai Tukar Negara ASEAN Per Dollar Amerika Serikat (Studi pada International Trade Center Periode Tahun 2013-2015)," *J. Adm. Bisnis*, vol. 50, no. 6, pp. 44–50, 2017.

[2] Badan Pusat Statistik, *Buletin Statistik Perdagangan Luar Negeri*, vol., no. Mei. 2022.

[3] M. Sihombing, J. Sihotang, and M. L. Purba, "Analisis Pengaruh Ekspor Migas, Ekspor Non Migas dan Penanaman Modal Asing Terhadap Pertumbuhan Ekonomi Indonesia Tahun 2000-2019," *J. Econ. Bus.*, vol. 02, no. 02, pp. 40–51, 2021.

[4] M. U. M. Putra and S. Damanik, "Pengaruh Ekspor Migas dan Non Migas Terhadap Cadangan Devisa di Indonesia," *J. Wira Ekon. Mikroskil*, vol. 7, no. 2, pp. 245–254, 2017.

[5] B. A. Rahman, M. Al Musadieq, and S. Sulasmiyati, "Pengaruh Utang Luar Negeri dan Ekspor Terhadap Pertumbuhan Ekonomi (Studi pada Produk Domestik Bruto Indonesia Periode 2015-2019)," *J. Adm. Bisnis*, vol. 45, no. 1, pp. 55–62, 2017, doi: 10.54980/imkp.v4i1.116.

[6] G. Jalunggono, Y. T. Cahyani, and W. Juliprijanto, "Pengaruh Ekspor, Impor dan Kurs Terhadap Cadangan Devisa Indonesia Periode Tahun 2004 – 2018," *J. Ekon. Bisnis, dan Akunt.*, vol. 22, no. 2, pp. 171–181, 2020, doi: 10.32424/jeba.v22i2.1593.

[7] D. Kertayuga, E. Santoso, and N. Hidayat, "Prediksi Nilai Ekspor Impor Migas dan Non-Migas Indonesia Menggunakan Extreme Learning Machine (ELM)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 6, pp. 2792–2800, 2021.

[8] M. W. Pramana, I. Purnamasari, and S. Prangga, "Peramalan Data Ekspor Nonmigas Provinsi Kalimantan Timur Menggunakan Metode Weighted Fuzzy Time Series Lee," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 14, no. 1, 2021, doi: 10.36456/jstat.vol14.no1.a3747.

[9] F. N. Hayati, D. Nurlaily, and E. Pusporani, "Peramalan Data Ekspor Non Migas Provinsi Kalimantan Timur Menggunakan Univariate Time Series," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 14, no. 2, pp. 59–66, 2021.

[10] H. Aprilianto, S. Kumalaningsih, and I. Santoso, "Penerapan Jaringan Syaraf Tiruan untuk Peramalan Penjualan dalam Mendukung Pengembangan Agroindustri Coklat di Kabupaten

Blitar," *Habitat*, vol. 29, no. 3, pp. 129–137, 2018, doi: 10.21776/ub.habitat.2018.029.3.16.

[11] Y. Setiawan, Tarno, and P. Kartikasari, "Prediksi Harga Jual Kakao dengan Metode Long Short-Term Memory Menggunakan Metode Optimasi Root Mean Square Propagation dan Adaptive Moment Estimation Dilengkapi Gui Rshiny," vol. 11, no. 1, pp. 99–107, 2022.

[12] U. I. Arfianti, D. C. R. Novitasari, N. Widodo, M. Hafiyusholeh, and W. D. Utami, "Sunspot Number Prediction Using Gated Recurrent Unit (GRU) Algorithm," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 2, pp. 141–152, 2021, doi: 10.22146/ijccs.63676.

[13] R. A. Falah and M. Rachmaniah, "Price Prediction Model for Red and Curly Red Chilies using Long Short Term Memory Method," *Indones. J. Stat. Its Appl.*, vol. 6, no. 1, pp. 143–160, 2022, doi: 10.29244/ijsa.v6i1p143-160.

[14] A. Nilsen, "Perbandingan Model RNN , Model LSTM , dan Model GRU dalam Memprediksi Harga Saham-Saham LQ45," *J. Stat. dan Apl.*, vol. 6, no. 1, pp. 137–147, 2022.

[15] H. G. Nugraha and A. SN, "Optimasi Bobot Jaringan Syaraf Tiruan Mengunakan Particle Swarm Optimization," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 8, no. 1, p. 25, 2014, doi: 10.22146/ijccs.3492.

[16] A. Nasuha, T. A. Sardjono, and M. H. Purnomo, "Pengenalan Viseme Dinamis Bahasa Indonesia Menggunakan Convolutional Neural Network," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 7, no. 3, pp. 258–265, 2018, doi: 10.22146/jnteti.v7i3.433.

[17] T. Dozat, "Incorporating Nesterov Momentum into Adam," *ICLR Work.*, pp. 1–4, 2016.

[18] E. H. Michael, K. R. Prilianti, and M. Subianto, "Rancang Bangun Aplikasi Klasifikasi Tingkat Kematangan Sangrai Kopi Melalui Citra Digital Menggunakan CNN (Convolutional Neural Network) Berbasis Android," *J. Ilm. SAINSBERTEK*, vol. 1, no. 1, pp. 1–11, 2020.

[19] H. Purnomo, H. Suyono, and R. N. Hasanah, "Peramalan Beban Jangka Pendek Sistem Kelistrikan Kota Batu Menggunakan Deep Learning Long Short-Term Memory," *Transmisi*, vol. 23, no. 3, pp. 97–102, 2021, doi: 10.14710/transmisi.23.3.97-102.

[20] M. Yu, F. Xu, W. Hu, J. Sun, and G. Cervone, "Using Long Short-Term Memory (LSTM) and Internet of Things (IoT) for Localized Surface Temperature Forecasting in an Urban Environment," *IEEE Access*, vol. 9, pp. 137406–137418, 2021, doi: 10.1109/ACCESS.2021.3116809.

[21] T. B. Shahi, A. Shrestha, A. Neupane, and W. Guo, "Stock Price Forecasting with Deep Learning: A Comparative Study," *Mathematics*, vol. 8, no. 9, pp. 1–15, 2020, doi: 10.3390/math8091441.

[22] K. E. ArunKumar, D. V. Kalaga, C. Mohan Sai Kumar, M. Kawaji, and T. M. Brenza, "Comparative Analysis of Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM) Cells, Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA) for Forecasting COVID-19 Trends," *Alexandria Eng. J.*, vol. 61, no. 10, pp. 7585–7603, 2022, doi: 10.1016/j.aej.2022.01.011.

[23] S. Sautomo and H. F. Pardede, "Prediksi Belanja Pemerintah Indonesia Menggunakan Long Short-Term Memory (LSTM)," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 99–106, 2021, doi: 10.29207/resti.v5i1.2815.

[24] I. N. Hidayati, M. Al Haris, and T. W. Utami, "Metode Average Based Fuzzy Time Series Markov Chain pada Data Laju Inflasi di Indonesia," in *Seminar Nasional UNIMUS*, 2022,

pp. 581–597.

[25] N. A. Elmunim, M. Abdullah, A. M. Hasbi, and S. A. Bahari, "Short-term forecasting Ionospheric Delay Over UKM, Malaysia, using the Holt-Winter method," *Int. Conf. Sp. Sci. Commun. Iconsp.*, no. July, pp. 106–109, 2013, doi: 10.1109/IconSpace.2013.6599443.

Application of Agglomerative Hierarchical Clustering Method for Grouping Non-Cash Food Assistance Recipients in Ngambon Bojonegoro

Alif Yuanita Kartini [(1)], Abdul Manaf Jamiluddin [(2)]

Universitas Nahdlatul Ulama Sunan Giri,

Jalan Ahmad Yani No. 10 Sukorejo, Bojonegoro, 62115, Telp. (0353) 887341

e-mail: alifyuanita@unugiri.ac.id [(1)], abdulmanafjamiluddin.1992@gmail.com [(2)]

ABSTRAK

Salah satu kecamatan di kabupaten Bojonegoro yang mendapatkan bantuan pangan non tunai adalah kecamatan Ngambon. Bantuan pangan non tunai yang diberikan di kecamatan Ngambon belum tepat sasaran. Hal ini dikarenakan masyarakat yang kurang mampu tidak mendapatkan bantuan, sementara masyarakat yang mampu justru mendapatkan bantuan pangan non tunai. Sehingga diperlukan penelitian dengan tujuan agar bantuan pangan non tunai yang diberikan pemerintah dapat disalurkan sesuai prosedur. Metode yang digunakan dalam penelitian ini adalah agglomerative hieralchical clustering untuk mengelompokkan penerima bantuan pangan non tunai masyarakat kecamatan Ngambon Bojonegoro. Variabel yang digunakan adalah 12 indikator bantuan pangan non tunai yang ditetapkan oleh Dinas Sosial kabupaten Bojonegoro. Data yang digunakan adalah 131 penerima bantuan pangan non tunai yang tersebar di lima desa di kecamatan Ngambon. Hasil pengelompokan dengan metode single linkage kurang relevan. Sementara itu dengan metode average linkage dan complate linkage didapatkan masing-masing lima klaster, dan dengan ward linkage didapatkan tiga klaster. Berdasarkan elbow rule didapatkan bahwa ward linkage merupakan metode pengelompokan terbaik, dengan klaster 1 berjumlah 57 orang, klaster 2 berjumlah 53 orang dan klaster 3 berjumlah 21 orang.

Kata kunci : agglomerative hierarchical clustering, bantuan pangan non tunai, pengelompokan

ABSTRACT

One of the sub-districts in Bojonegoro that received non-cash food assistance was Ngambon sub-district. The non-cash food assistance provided in Ngambon sub-district has not been on target. This is because underprivileged people do not get assistance, while people who can afford it actually get non-cash food assistance. So, research is needed with the aim that non-cash food assistance provided by the government can be distributed according to procedures. The method used in this study is agglomerative hieralchical clustering to group recipients of non-cash food assistance from the people of Ngambon Bojonegoro. The variables used were 12 indicators of non-cash food assistance set by the Bojonegoro district Social Office. The data used were 131 recipients of non-cash food assistance spread across five villages in Ngambion sub-district. Grouping results with the single linkage method are less relevant. Meanwhile, with the average linkage and complate linkage methods, five clusters were obtained, and with ward linkage, three clusters were obtained. Based on the elbow rule, it was found that ward linkage is the best grouping method, with cluster 1 totaling 57 people, cluster 2 totaling 53 people and cluster 3 totaling 21 people.

Keywords: *agglomerative hierarchical clustering, non-cash food assistance, grouping*

INTRODUCTION

Poverty is a condition where people are below the standard of living, both in terms of livelihood, income, clothing and shelter [1]. One of the government's efforts to overcome the problem of poverty is provide social assistance to the community. With this social assistance, it is hoped that the community can improve the standard of living to be more prosperous and prosperous. There are several types of social assistance provided by the Indonesian government to the poor through the Ministry of Social Affairs, including Non-Cash Food Assistance (BPNT), Family Hope Program (PKH), National Health Insurance (JKN) and Social Cash Assistance (BST) [2].

Bojonegoro Regency is one of the districts receiving social assistance in Indonesia. One type of social assistance received by Bojonegoro district is Non-Cash Food Assistance (BPNT). In Bojonegoro district, Non-Cash Food Assistance is a continuation of the Poor Rice Program (Raskin) which has been renamed the Prosperous Rice Program (Rastra). Non-Cash Food Assistance is assistance originating from the government in the form of goods to be given to people who are entitled to receive. This assistance is given 14 times in one year. Non-Cash Food Assistance in the form are 15 kg of rice with an exchange price of Rp. 1,600.00/kg and cash of Rp. 150,000.00 [3]. In Bojonegoro district, Non-Cash Food Assistance is a new thing. Therefore, in the process of implementation, there are still many problems. In Bojonegoro district, Non-Cash Food Assistance is a new thing. Therefore, in the process of implementation, there are still many problems [4]. This is a problem, especially in Ngambon sub-district. Many people objected to the decision and protested to the Ngambon sub-district [5]. This problem still has not found a solution. This is because the data used by the Ngambon sub-district comes from the Bojonegoro Regency Social Office which is likely that the data has not been updated. Therefore, a study is needed to group recipients of Non-Cash Food Assistance communities, especially in Ngambon Bojonegoro, so that it is right on target for aid beneficiaries.

For the grouping recipients of non-cash food assistance, several indicators are needed in accordance with the criteria of beneficiaries, especially non-cash food assistance. Based on research conducted by Parhusip (2019), the criteria used for the eligibility selection process for Non-Cash Food Assistance recipients are non-monetary variable poverty criteria. Meanwhile, based on Sugianto & Maulana (2019), the indicators used for grouping recipients of Non-Cash Food Assistance are using indicators from the Central Statistics Agency. And based on research conducted by Saputra et al. (2021), to group recipients of Non-Cash Food Assistance using nine criteria, namely house area, income, floor type, wall type, lighting source, water type, source of medical expenses, fuel and savings. Meanwhile, in this study, the indicators used to group recipients of Non-Cash Food Assistance are the criteria used by the Social Office of Bojonegoro district in determining beneficiaries of assistance, especially for non-cash food assistance.

To group recipients of non-cash food assistance, one technique that can be used is clustering techniques. Clustering is one technique in machine learning to group different objects so that clusters with members who have the same characteristics will be obtained [9]. Clustering as an exploratory procedure consists of two methods, namely hierarchical clustering and non-hierarchical clustering. There are two methods in hierarchical clustering, namely the division method and the agglomerative method [10]. Divisional methods are computationally intensive and have limited applications in the social sciences, while agglomerative methods have been implemented in many standard software packages. There are several techniques in the agglomerative method including single linkage, average linkage, complate linkage and ward linkage.

This study will use the agglomerative hierarchical clustering method to group recipients of Non-Cash Food Assistance in Ngambon Bojonegoro. Previously, there have been many studies using the Agglomerative hierarchical clustering method, including by Randriamihamison et al., (2021), Rong (2020) dan Wu et al. (2021). From some of these studies, the results obtained in the use of the Agglomerative hierarchical clustering method are that the resulting levels more

structured and the desired cluster is not widely displayed in the dendogram. In addition, the use of the Agglomerative hierarchical clustering method is more efficient in the use of time. This is because the output produced in the form of levels or hierarchies to facilitate storage.

The purpose of this study is to obtain the results of grouping recipients of non-cash food assistance in Ngambon, Bojonegoro. From this research, hoped that it will be useful for Bojonegoro district government agencies in supervising the distribution of non-cash food assistance and targeting recipients of non-cash food assistance. It is hoped that the assistance provided by the government can be distributed correctly in accordance with established procedures. Not only that, the government can also evaluate the recipients of non-cash food assistance so that it can create a breakthrough program in the development of social assistance disbursed by the government.

METHOD

The source of data in this study came from the Bojonegoro Regency Social Office. The data used is in the form of data on aid beneficiaries in 2021 as many as 131 recipients along with indicators of non-cash food assistance in Ngambon Bojonegoro. Furthermore, from these data, a field survey was conducted for each beneficiary of non-cash food assistance in Ngambon sub-district. The variables used are shown in Table 1 below.

Table 1. Research variables

Variables	Measurement scale	Information
Residence (X_1)	Categorical	Beneficiaries have daily shelter
Number of family members (X_2)	Categorical	Beneficiaries live together with other family members or are alone in a household
Activities that make money (X_3)	Categorical	Beneficiaries have had wage-generating activities in the past week
Food needs (X_4)	Categorical	Beneficiaries are concerned about the food eaten
Food expenditure (X_5)	Categorical	The expenditure used by beneficiaries to buy food is more than 2/3 of the total expenditure
Clothing needs (X_6)	Categorical	Beneficiaries have purchased clothes for themselves or for other family members in the past year
House floor (X_7)	Categorical	Beneficiaries have residences whose floors are made of earth
House walls (X_8)	Categorical	Beneficiaries have residences whose walls are made of bamboo, wire, or wood
Toilet Washing Bath (MCK) (X_9)	Categorical	Beneficiaries have a place to bathe, wash and defecate or urinate
Lighting sources (X_{10})	Categorical	Beneficiaries have 450/900 Watt electricity from the State Electricity Company (PLN) or not (using a tromber or torch)
Work (X_{11})	Categorical	Livelihoods of aid beneficiaries
House building X_{12}	Categorical	The feasibility form of the entire house building seen from the front, back, side, right and left

In this study using the help of Python software for clustering using the Agglomerative Hierarchical Clustering method. The analysis steps to group recipients of non-cash food assistance in Ngambon, Bojonegoro using the Agglomerative Hierarchical Clustering method are as follows.

a. Obtain descriptive statistics for the number of recipients of non-cash food assistance per sub-district

b. Perform clustering using the Agglomerative Hierarchical Clustering method with the following steps [14]
   1) Calculates the distance matrix on all data pairs using the Euclidean Distance formula with the following equation

$$d_{ij} = d(x_i, x_j) = \left[\sum_{s=1}^{p}(x_{is} - x_{js})^2\right] \quad (1)$$

   2) Combine two nearby groups into one data group
      i) Single Linkage with equation

$$d(U,V) = \min\{d(U,V)\}; d(U,V) \in D \quad (2)$$

      ii) Average Linkage with equation

$$d(U,V) = \frac{1}{nU\ X\ nV}\sum d(U,V)\ ; d(U,V) \in D \quad (3)$$

      iii) Complate Linkage with equation

$$d(U,V) = \max\{d(U,V)\}; d(U,V) \in D \quad (4)$$

      iv) Ward Linkage with equation

$$SSE = \sum_{j=1}^{p}\left(\sum_{i=1}^{n} x^2ij\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x^2ij\right)^2 \quad (5)$$

   3) Update the distance matrix between data to represent between the new and remaining groups
   4) Repeat steps 2 through 3 until there is only one group left

c. Choose the best cluster result by looking inside SSE with the elbow rule (sum of square error) method
d. Interpretation of best cluster results

RESULT AND DISCUSSION

A. Descriptive Statistics

Descriptive statistics are used to determine the description of non-cash food assistance beneficiaries in each village in Ngambon sub-district as shown in Table 2 below.

Table 2. Number of non-food assistance recipients' cash in each village in Ngambon sub-district

Number	Village Name	Number of Recipients
1	Bondol	23
2	Ngambon	38
3	Sengon	17
4	Nglampin	31
5	Karangmangu	22

Sum	131

Based on Table 2, it can be seen that the total number recipients of non-cash food assistance in Ngambon sub-district is 131 recipients. The largest number of recipients of non-cash food assistance is Ngambon village, which is 38 recipients, and then Nglampin village as many as 31 recipients, Bondol village as many as 23 recipients, Karangmangu village as many as 22 recipients and at least Sengon village as many as 17 recipients.

Grouping Results Using Agglomerative Hierarchical Clustering Method

1) *Single Linkage*

The results of clustering using the single linkage method are obtained with the help of Python software. Dendogram obtained from the results of the single linkage method for clustering beneficiaries of Non-Cash Food Assistance in Ngambon District as shown in Figure 1 below.

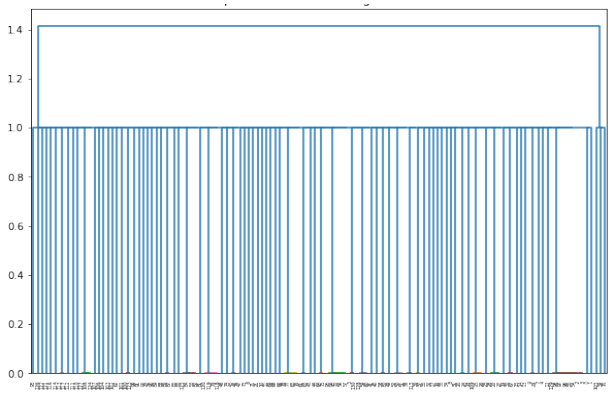


Figure 1. Results of dendogram with single linkage method

Based on the dendogram in Figure 1 shows that the results of clustering using the single linkage method produce 2 homogeneous clusters. The results of clustering using this single linkage method cannot describe the state of existing data. This is in contrast to a hierarchical agglomeration cluster validation where clusters should be heterogeneous between one cluster and another cluster member. Meanwhile, the results of clustering using the single linkage method are homogeneous.

In addition to generating a dendogram to obtain clustering results, with the help of Python software can also produce labeling array. This labeling array shows how many clusters have been formed and knows who the members of the cluster have formed. The results of the labeling array for clustering recipients of Non-Cash Food Assistance in Ngambon District are shown in Figure 2 below.

```
array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],
      dtype=int64)
```

Figure 2. Labeling array results using the single linkage method

Based on the labelization array in Figure 2 produces two labels, namely 0 and 1. This means that the clustering of Non-Cash Food Assistance recipients in Ngambon District is divided into two, namely cluster 1 and cluster 2. Cluster 1 consists of 3 recipients of Non-Cash Food Assistance, 1 person from Ngambion Village, 1 person from Karangmangu Village and 1 person from Nglampin Village. Meanwhile, cluster 2 consists of 128 recipients of Non-Cash Food Assistance. 3 people from cluster 1 do not carry out activities that make money, live alone, and meet all the criteria set by the Social Service. Therefore, cluster 1 is called the cluster that is eligible to receive non-cash food assistance and cluster 2 is called the cluster that is not eligible to receive non-cash food assistance. However, judging from the cluster tree that is formed, the cluster that is formed does not represent the existing data conditions. Therefore, use of the single linkage method in data on recipients of Non-Cash Food Assistance in Ngambon Regency is less relevant.

2) ***Average Linkage***

Just like in single linkage, the average linkage method also uses Python software. Dendogram for clustering recipients of Non-Cash Food Assistance using the average linkage method as shown in Figure 3 below.

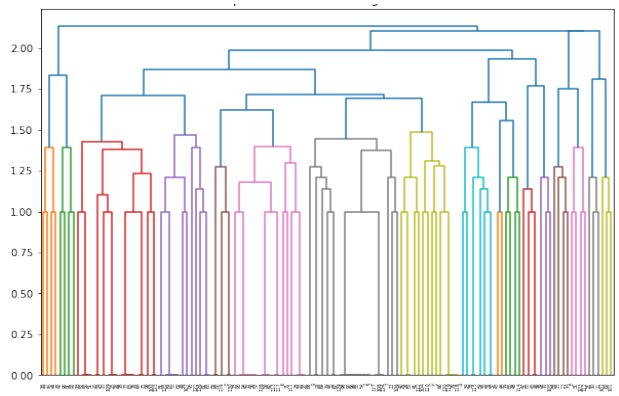


Figure 3. Dendogram results using the average linkage method

Based on the dendogram in Figure 3, the clustering of Non-Cash Food Assistance recipients in Ngambon District resulted in five clusters that were heterogeneous between one cluster and another. To find out the members of each cluster can be seen from the results of the labeling array as shown in Figure 4 below.

```
array([1, 1, 1, 1, 0, 1, 1, 1, 3, 1, 2, 2, 1, 3, 3, 3, 4, 4, 1, 1, 1, 2,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 4, 4, 0, 0, 0, 0, 0, 0, 1, 1, 4, 0,
       0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1,
       1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1,
       3, 0, 4, 4, 4, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 3, 1, 0, 1, 1, 1, 1,
       1, 1, 3, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],
      dtype=int64)
```

Figure 4. Labeling array results using the average linkage method

Based on the results of the labelization array using the average linkage method, five labels are obtained, namely 0, 1, 2, 3, and 4. This means that the cluster results formed from recipients of non-cash food assistance are five clusters, namely cluster 1, cluster 2, cluster 3, cluster 4 and cluster

5. The members of each cluster are cluster 1 with 21 people, cluster 2 with 88 people, cluster 3 with 6 people, cluster 4 with 8 people and cluster 5 with 8 people. From each cluster member, characteristics of non-cash food assistance recipients can be obtained. The characteristic of cluster 1 is that the recipients of Non-Cash Food Assistance in Ngambon District all meet the criteria set by the Social Service. Meanwhile, for cluster 2, the recipients of non-cash food assistance meet 80 percent of the criteria set by the Social Service and 20 percent of the criteria set by the Social Service were not meet. For cluster 3, most recipients of non-cash food assistance meet 60 percent of the criteria set by the Social Service and 40 percent did not meet the criteria set by the Social Service. Furthermore, for cluster 4, most of the recipients of Non-Cash Food Assistance meet 40 percent of the criteria set by the Social Service and 60 percent of the criteria set by the Social Service were not meet. And for cluster 5, almost all criteria set by the Social Service are not meet by the recipients of Non-Cash Food Assistance. Based on the characteristics of each cluster, cluster 1 is called a cluster that is very eligible to receive non-cash food assistance, for cluster 2 is called a cluster that is quite eligible to receive Non-Cash Food Assistance, cluster 3 is called a cluster that is worthy of receiving Non-Cash Food Assistance, cluster 4 is called a cluster that is not eligible to receive Non-Cash Food Assistance and cluster 5 is called a cluster that is not eligible to receive Non-Cash Food Assistance.

3) *Complate Linkage*

The results of the dendogram use the complete linkage method as shown in Figure 5. Based on Figure 5, it can be seen that the clustering results for Non-Cash Food Assistance beneficiaries in Ngambon District are five heterogeneous clusters. Dendograms generated using the complete linkage method have a neater arrangement and each cluster has a fairly tight error distance between one cluster and another.

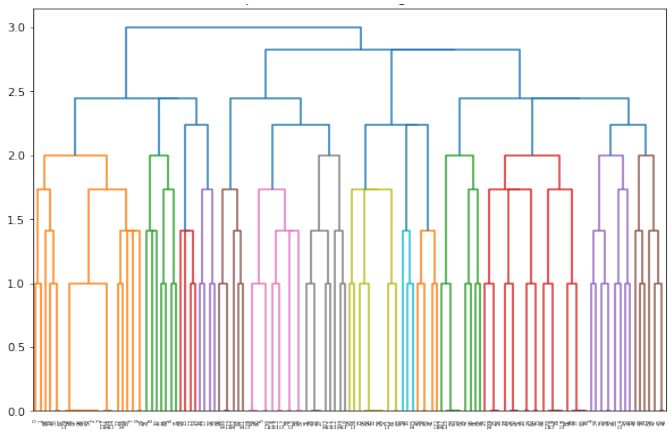


Figure 5. The results of the dendogram using the complate linkage method

The members of each cluster formed from data on recipients of Non-Cash Food Assistance in Ngambon District can be seen in the results of the labeling array as shown in Figure 6 below.

```
array([1, 1, 1, 1, 0, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1,
       2, 0, 0, 0, 0, 2, 2, 1, 1, 2, 1, 1, 0, 0, 0, 0, 2, 3, 3, 3, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 3, 3, 3, 2, 2, 3, 4, 3, 0, 0,
       0, 1, 1, 1, 3, 1, 3, 3, 0, 0, 0, 0, 0, 0, 3, 0, 0, 1, 0, 0, 2, 2,
       1, 0, 0, 0, 0, 3, 3, 3, 1, 1, 1, 4, 4, 4, 4, 2, 2, 3, 1, 0, 1, 0,
       2, 2, 2, 2, 0, 2, 2, 2, 0, 0, 0, 3, 3, 4, 1, 0, 3, 1, 2, 1, 2],
      dtype=int64)
```

Figure 6. Labeling array results using the complate linkage method

Based on the results of the labeling array using the complate linkage method as shown in Figure 6, it can be seen that there are four labels produced, namely label 0, label 1, label 2, label 3 and label 4. So that the results of clustering formed are 5, namely cluster 1, cluster 2, cluster 3, cluster 4 and cluster 5. Based on Figure 6, it can be seen that the members for cluster 1 are 46 people, cluster 2 is 38 people, cluster 3 is 21 people, cluster 4 is 19 people and cluster 5 is 7 people. When viewed from the results of the dendogram and labeling array, it can be seen that cluster 1 and cluster 2 dominate more, followed by cluster 3, cluster 4 and cluster 5. The recipients of Non-Cash Food Assistance in Ngambon District are mostly included in cluster 1. Based on the characteristics obtained, cluster 1 is called a cluster that is very eligible to receive Non-Cash Food Assistance. This is because recipients of Non-Cash Food Assistance meet the requirements set by the Social Service. However, there are still some recipients of Non-Cash Food Assistance who are members of clusters 4 and 5. Cluster 4 and cluster 5 are categorized as not eligible to receive Non-Cash Food Assistance. This is because it does not meet the criteria for beneficiaries set by the Social Service.

4) ***Ward Linkage***

To get the results of clustering along with the members of each cluster, it can be seen from the dendogram and labeling array. The results of dendogram using the ward linkage method for clustering recipients of Non-Cash Food Assistance in Ngambon District as shown in Figure 7 below.

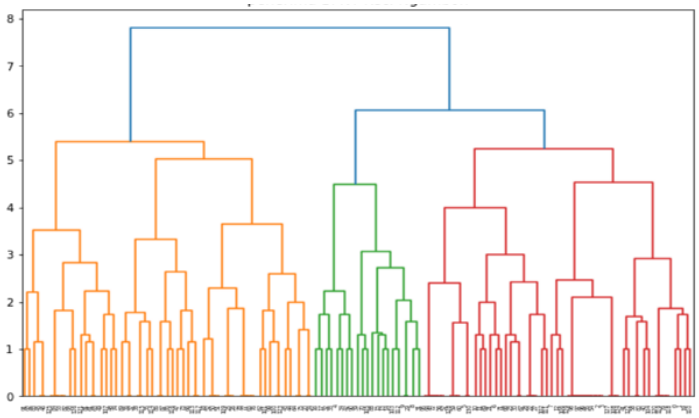


Figure 7. Dendogram results using the Ward Linkage method

Based on results of the dendogram as shown in Figure 7, it can be seen that the clustering results for recipients of Non-Cash Food Assistance in Ngambon District resulted in 3 heterogeneous clusters. The resulting arrangement of dendograms is very neat and each cluster has a fairly tight error distance between each other. In addition, the resulting cluster has a high heterogeneity between one cluster and another. This can be seen from the color of the resulting dendogram.

Furthermore, for members of each cluster formed using the ward linkage method, it can be seen from the results of the labeling array as shown in Figure 8 below.

```
Out[7]: array([1, 1, 1, 1, 2, 1, 1, 1, 2, 2, 1, 1, 1, 2, 2, 2, 2, 2, 0, 1, 0, 1,
               1, 1, 0, 0, 0, 1, 1, 2, 1, 2, 2, 2, 2, 0, 0, 0, 0, 2, 1, 1, 2, 0,
               0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0,
               0, 1, 1, 1, 2, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0,
               2, 0, 2, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 2, 1, 2, 1, 0, 1, 0,
               1, 1, 2, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1],
              dtype=int64)
```

Figure 8. Labeling array results using ward linkage method

Based on the labelization array in Figure 8 shows that the clustering results in three labels, namely 0, 1 and 2. This means that there are 3 clusters formed, namely cluster 1, cluster 2 and cluster 3. Clustering using the ward linkage method obtained results in the form of the first cluster, which is a cluster that is very eligible to receive Non-Cash Food Assistance totaling 57 people. Cluster 2 is a cluster that is quite eligible to receive Non-Cash Food Assistance totaling 53 people. Cluster 3 is a cluster that is not eligible to receive Non-Cash Food Assistance totaling 21 people.

Judging from the dendogram and labeling array using the ward lingkage method, the cluster that dominates is cluster 1. This is because the beneficiaries of Non-Cash Food Assistance in cluster 1 are very eligible to receive Non-Cash Food Assistance. Next followed by cluster 2. This is because the recipients still qualify as recipients of Non-Cash Food Assistance. There are 21 people or 16% of recipients of Non-Cash Food Assistance in Ngambon Sub-district who are categorized as not eligible to receive Non-Cash Food Assistance. This is because it no longer meet the criteria set by the Social Service. The results of clustering by four methods, namely single linkage, average linkage, complate linkage and ward linkage as shown in Table 3 below.

Table 3. Clustering results by four methods

Clustering method	Number of Clusters	Information
Single Linkage	2	Unable to describe the state of existing data.
Average Linkage	5	Cluster 1 (very eligible to receive assistance) amounted to 21 people, cluster 2 (moderately deserving of assistance) amounted to 88 people, cluster 3 (deserving of assistance) amounted to 6 people, cluster 4 (less eligible to receive assistance) amounted to 8 people and cluster 5 (not eligible to receive assistance) amounted to 8 people.
Complete Linkage	5	Cluster 1 (very deserving of assistance) as many as 46 people, cluster 2 (moderately deserving of assistance) as many as 38 people, cluster 3 (deserving of assistance) as many as 21 people, cluster 4 (less deserving of assistance) as many as 19 people and cluster 5 (not eligible to receive assistance) as many as 7 people
Ward Linkage	3	Cluster 1 (very eligible for assistance) totaled 57 people, cluster 2 (moderately deserving of assistance) numbered 53 people, cluster 3 (not eligible for assistance) totaled 21 people.

B. Results of Comparison of Methods Using Elbow Rule

Data on recipients of Non-Cash Food Assistance in Ngambon, Bojonegoro were clustered using agglomerative hierarchy clustering. The methods used are single linkage, average linkage, complete linkage, and ward linkage. Each of these methods produces 2 clusters, 5 clusters, 5 clusters, and 3 clusters. The best clustering results are measured using elbow rule method. The elbow rule is a right line where the best cluster number of data can be seen by the elbow level of the elbow rule. In the elbow rule, the more clusters that are formed and are heterogeneous between clisters, the better the accuracy obtained. The results of the elbow rule can be seen in figure 9.

From the elbow rule for agglomertive clustering as shown in Figure 9, the elbow point of the elbow rule is indicated by point 3 or 3 cluster. When viewed at the point of cluster 1, the distance value is still very hig. As well as the value of cluster 2. However, when at point 3 the black elbow rule line cluster begins to ramp up. This means that in cluster 3, the data on recipients of Non-Cash Food Assistance in Ngambon District is maximized. This is seen from the location of cluster 3 right at the elbow point (clustering ramp point). This is used to determine the best cluster results. The ramp points on the elbow rule show the best amount of clustering. The conclusion is that the best cluster from the data on recipients of Non-Cash Food Assistance in Ngambon District is 3 clusters. So the best and appropriate clustering method in this study is ward linkage method. The ward linkage method produces 3 clusters according to the measurements produced by the elbow rule.

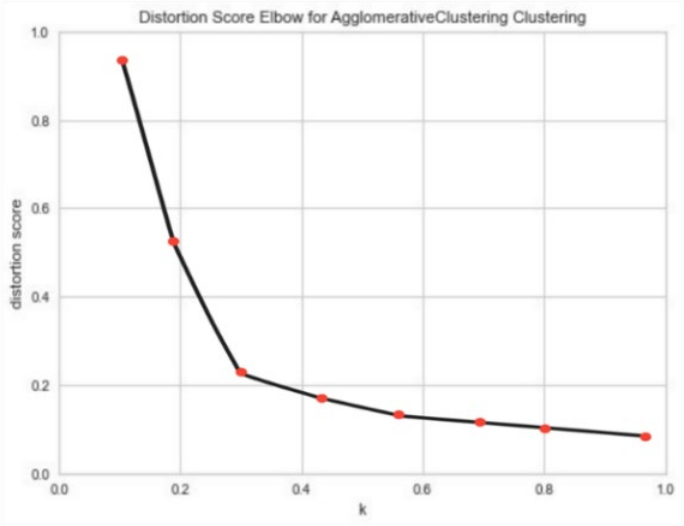


Figure 9. Elbow rule agglomerative hierarchical clustering

C. Interpretation of the Best Grouping Results

Based on the elbow role method, it was found that the best clustering method is to use ward linkage. The number of clusters generated using the ward linkage method is 3 clusters. The clusters formed have high heterogeneity between clusters 1, 2, and 3. A total of 131 recipients of Non-Cash Food Assistance in Ngambon District are divided into 3 clusters, namely cluster 1 category eligible to receive Non-Cash Food Assistance consisting of 57 people, cluster 2 categories moderately eligible to receive Non-Cash Food Assistance consisting of 53 people, and cluster 3 categories not eligible to receive Non-Cash Food Assistance consisting of 21 people. Overall, it can be concluded that the recipients of Non-Cash Food Assistance in Ngambon District are in accordance with the criteria set by the Social Service. This is because 84 percent of all recipients of Non-Cash Food Assistance in Ngambon District are eligible to receive the assistance. However, 16 percent of Non-Cash Food Assistance recipients in Ngambon District fall into the category of not eligible to receive Non-Cash Food Assistance.

Cluster 1 is a cluster that very eligible to receive Non-Cash Food Assistance. Cluster 1 has 57 people. 2 people from Bondol Village, 18 people from Ngambon Village, 14 people from Karangmangu, 15 people from Nglampin, and 8 people from Sengon. The beneficiaries are categorized as very eligible to receive Non-Cash Food Assistance because the beneficiaries meet the criteria set by the Social Services. This is because there are some recipients who are paralyzed and do not carry out daily activities, in the past week have not made money so they are worried about hunger in the next year, houses have dirt floors and do not have good MCK facilities. In addition, there are several beneficiaries whose houses are no longer habitable, and live in their relatives' houses.

Cluster 2 is a category eligible to receive Non-Cash Food Assistance. Cluster 2 has 53 people. From Bondol Village as many as 13 people, from Ngambon village as many as 13 people, from Karangmangu village as many as 7 people, from Nglampin village as many as 11 people, and from Sengon village as many as 9 people. The beneficiaries are categorized as eligible to receive Non-Cash Food Assistance. This is because the recipients meet the criteria set by the Social Services. The beneficiaries made money in the past week but the biggest expenditure was used to buy food, the floors of the house buildings were made of earth instead of cast or ceramics, there were some walls of houses made from red brick but there were some whose walls were half red brick and the top was covered with cassibot. In addition, the beneficiaries already have MCK facilities even though they are less crowded, and the overall house buildings are habitable.

Cluster 3 is a category that is not eligible to receive Non-Cash Food Assistance. Cluster 3 has 21 people. 8 people came from Bondol Village, 7 people came from Ngambion Village, 1 person came from Karangmangu Village, 5 people came from Nglampin Village, and none of them came from Sengon Village. The beneficiaries are categorized as not eligible to receive Non-Cash Food Assistance. This is because it does not meet the criteria set by the Social Service. The characteristics of the category that is not eligible to receive non-cash food assistance are staying at home with other family members, making money in the past week and the largest expenditure not being used to buy food. In addition, the beneficiaries are not worried about starvation in the next year, the floor of the house is made of earth but the walls of the house are made of red brick and cast concrete.

CONCLUSION

In Ngambon sub-district, there are 131 recipients of non-cash food assistance spread across five villages, namely Bondol village, Ngambion village, Sengon village, Nglampin village and Karangmangu village. Clustering using the single linkage method is less relevant because the resulting dendogram tends to be homogeneous. Clustering with average linkage and complate linkage obtained five clusters. Meanwhile, clustering with the ward linkage method obtained three clusters. Based on the Elbow Rule, it was found that ward linkage is the best clustering method. The results of clustering using ward linkage are cluster 1 is a cluster that is very eligible to receive non-cash food assistance benefits totaling 57 people, cluster 2 clusters that are quite eligible to receive non-cash food assistance benefits totaling 53 people, and cluster 3 is a cluster that is not eligible to receive non-cash food assistance benefits totaling 21 people. In future studies, it is expected to use other clustering methods to get better results.

REFERENCE

[1] H. Hardianto, "Determinasi Pemberdayaan Masyarakat Dan Pemberantasan Kemiskinan Desa: Analisis Dana Desa Dan Alokasi Dana Desa (Literature Review Manajemen Keuangan)," *J. Manaj. Pendidik. DAN ILMU Sos.*, vol. 3, no. 1, pp. 266–275, 2022.

[2] N. R. R. Dani, "Implementasi Program Bantuan Sosial Tunai (BST) Pada Masa Pandemi Covid-19 di Kecamatan Kedungadem Kabupaten Bojonegoro," *Publika*, pp. 1187–1200, 2022.

[3] D. M. Hasimi, "Analisis Program Bantuan Pangan Non Tunai (BPNT) guna meningkatkan kesejahteraan masyarakat dalam perspektif ekonomi Islam," *REVENUE J. Manaj. Bisnis Islam*, vol. 1, no. 01, pp. 61–72, 2020.

[4] N. A. Putri and H. Purnaweni, "Implementasi Program Keluarga Harapan (Pkh) Dalam Upaya Penanggulangan Kemiskinan Di Kecamatan Bojonegoro," *J. Public Policy Manag. Rev.*, vol. 10, no. 3, pp. 510–522, 2021.

[5] S. Rahayu and A. Y. Kartini, "Algoritma K-Means Dan K-Medoids Untuk Pengelompokan Kecamatan Penerima Bantuan Sosial Di Kabupaten Bojonegoro," *MEDIA BINA Ilm.*, vol. 16, no. 5, pp. 6815–6822, 2021.

[6] J. Parhusip, "Penerapan Metode Analytical Hierarchy Process (AHP) Pada Desain Sistem Pendukung Keputusan Pemilihan Calon Penerima Bantuan Pangan Non Tunai (BPNT) Di Kota Palangka Raya," *J. Teknol. Inf. J. Keilmuan dan Apl. Bid. Tek. Inform.*, vol. 13, no. 2, pp. 18–29, 2019.

[7] C. A. Sugianto and F. R. Maulana, "Algoritma NaÃ¯ ve Bayes Untuk Klasifikasi Penerima Bantuan Pangan Non Tunai (Studi Kasus Kelurahan Utama)," *Techno. Com*, vol. 18, no. 4, pp. 321–331, 2019.

[8] R. A. Saputra, S. Wasiyanti, and D. Pribadi, "Information Gain Pada Algoritma C4. 5 Untuk Klasifikasi Penerimaan Bantuan Pangan Non Tunai (BPNT)," *Indones. J. Bus. Intell.*, vol. 4, no. 1, pp. 25–30, 2021.

[9] P. R. Garikapati, K. Balamurugan, T. P. Latchoumi, and G. Shankar, "A quantitative study of small dataset machining by agglomerative hierarchical cluster and K-medoid," in *Emergent Converging Technologies and Biomedical Systems: Select Proceedings of ETBS 2021*, Springer, 2022, pp. 717–727.

[10] A. Naeem, M. Rehman, M. Anjum, and M. Asif, "Development of an efficient hierarchical clustering analysis using an agglomerative clustering algorithm," *Curr. Sci.*, vol. 117, no. 6, pp. 1045–1053, 2019.

[11] N. Randriamihamison, N. Vialaneix, and P. Neuvial, "Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints," *J. Classif.*, vol. 38, no. 2, pp. 363–389, 2021.

[12] Y. Rong, "Staged text clustering algorithm based on K-means and hierarchical agglomeration clustering," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2020, pp. 124–127.

[13] S. Wu, J. Lin, Z. Zhang, and Y. Yang, "Hesitant fuzzy linguistic agglomerative hierarchical clustering algorithm and its application in judicial practice," *Mathematics*, vol. 9, no. 4, pp. 1–16, 2021, doi: 10.3390/math9040370.

[14] C. Chatfield, *Introduction to multivariate analysis*. Routledge, 2018.

Multiperiod Logit on Survival Analysis of Financial Distress in Manufacturing Company

Wilda Yulia Rusyida(1) Anas Yoga Nugroho(2)

[1,2] UIN K.H. Abdurrahman Wahid Pekalongan

Kab. Pekalongan, Jawa Tengah, Indonesia

e-mail: wilda.yulia.rusyida@uingusdur.ac.id(1)

ABSTRAK

Perusahaan dituntut untuk dapat mempertahankan kelangsungan hidupnya agar tujuan perusahaan dapat tercapai dengan baik. Financial distress merupakan salah satu faktor yang menyebabkan perusahaan tidak dapat mempertahankan kelangsungan hidupnya sehingga tujuan perusahaan tidak tercapai. Faktor yang menyebabkan perusahaan berada dalam keadaan tertekan adalah faktor internal dan eksternal. Penelitian ini adalah penelitian deskriptif kuantitatif menggunakan 16 rasio keuangan, IHSG dan BI rate. Metode penelitian ini adalah metode kuantitatif dengan menggunakan data time series dengan model multiperiod logit. Penentuan sampel menggunakan purposive sampling sehingga terdapat 79 sampel yang digunakan dalam penelitian ini. Berdasarkan hasil deskripsi kurva Kaplan Meier, uji log rank, model logit multiperiod dengan pemilihan variabel, berarti perusahaan survive dan financial distress memiliki perbedaan yang menonjol pada rasio profitabilitas dan rasio market measure. Sedangkan berdasarkan hasil uji parsial 4 dari 5 rasio keuangan, hasil pemilihan variabel berpengaruh signifikan terhadap financial distress. Lima perusahaan terbaik untuk berinvestasi dengan nilai peluang hazard minimum adalah perusahaan dengan kode emiten SKBM, IGAR, PBRX, PSDN dan UNIC.

Kata kunci : Logit Multiperiod; Analisis Kelangsungan Hidup; Kesulitan keuangan; Perusahaan manufaktur

ABSTRACT

The company is required to be able to maintain its survival so that the company's goals can be achieved properly. Financial distress is one of the factors that causes the company to be unable to maintain its viability so that the company's goals are not achieved. The factors that cause the company to be in a state of distress are internal and external factors. This is descriptive quantitative research which used 16 financial ratios, IHSG and BI rate. This research method is a quantitative method using time series data with a multiperiod logit model. Determination of the sample using purposive sampling so that there are 79 samples used in this study. Based on the results of the description of the Kaplan Meier curve, log rank test, multiperiod logit model with variable selection, it means that companies survive and financial distress have prominent differences in profitability ratios and market measure ratios. Meanwhile, based on the results of the partial test 4 out of 5 financial ratios, the results of the selection of variables have a significant effect on financial distress. The five best companies to invest in with a minimum hazard opportunity value are companies with issuer codes SKBM, IGAR, PBRX, PSDN and UNIC.

Keywords : *Multiperiod Logit; Survival Analysis; Financial Distress; Manufacturing Company*

INTRODUCTION

Financial developments in Indonesia can be observed through the share of capital from the entire industry on the Indonesia Stock Exchange (IDX). The absolute number of organizations listed on the IDX is approximately 685 organizations which are separated into accompanying characterizations. The Indonesian Industry Service sees that the manufacturing business is one of the fields that basically contributes to the interests of the whole of Indonesia. According to the Minister of Industry, Airlangga Hartanto previously, "the manufacturing business area is the backbone for public finance development and is a pillar area in driving value towards a comprehensive turn of events and local government assistance" (www.kemenperin.go.id). In the attached image, the JCI information shows that of the eight areas listed on the IDX, only one area is in the green zone (positive zone), particularly the essential and synthetic business areas. While the other seven regions are in the red zone (negative zone). A particularly fragile area is the consumer goods industry.

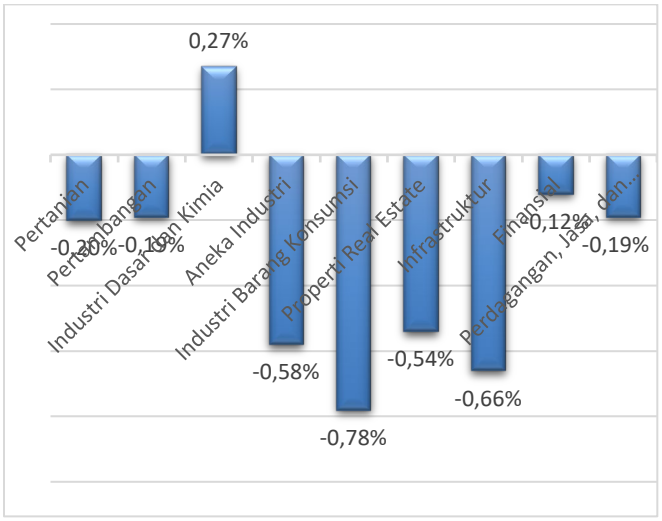


Figure 1. Stock Trading Index on the Indonesia Stock Exchange
Source: idx.com (Data processed by the author, 2020)

Based on Figure 1, it can be seen that only 1 sector rose and the rest fell. The fundamental and substance industry areas are the more down-to-earth areas, while the buyer's merchandise industry is the more vulnerable area. The decrease in stock costs can be used as an indication of a decrease in monetary execution in the regions[1]. Financial distress is a stage of decline in financial conditions that occurs before the company goes bankrupt or goes into liquidation. So this should be a concern for the company's management because it can affect sales and profits to be obtained so that later it will also have an impact on the company's operational activities.

The motivation behind establishing an organization is actually to generate benefits. These benefits are relied on to work on organizational presentations and follow the organization's business coherence in the long run. For that, the organization must compete with other competing organizations. In addition, the organization will also compete with changes in various parts of improvement in Indonesia. Similarly, high-benefit-generating organizations will attract

consideration of financial backers to contribute. The financial backer will see and examine the state of the organization before making a choice to contribute. Budget summary examination serves to decide the functional performance of the organization for a period and is used to forecast the organization's accounts at a later date. If bad things happen to the organization, for example facing financial problems or financial challenges, the organization definitely knows how to deal with those problems.

In 1968, Altman directed his exploration of the liquidation of manufacturing companies in America using discriminant analysis. The review reveals that there are five factors that significantly influence the forecast for bankruptcy including working capital per total asset, retained earnings per total asset, EBIT per total asset, market value of equity per total liability and sales per total asset. Besides that, [2]. also researched about bankruptcy with logistic regression. The test shows that the factors of net income per net sales, current liabilities per total assets, current assets per current liabilities, net income per asset growth are factors that significantly affect corporate chapter 11.

The two tests above focus on using a static model that relies on information for an indefinite period of time. This review means predicting the condition of the company that fails in the accompanying period by ignoring how the condition of the organization changes after some time. [3]. offers a multiperiod logit technique that can represent this progress and is admittedly more stable than the static model. In his exploration, Shumway showed that the multi-period logit model can predict bankruptcy better than using the discriminant analysis made by Altman to estimate corporate bankruptcy on the financial data of companies on the NYSE and AMEX from 1962 to 1992. In addition,[4] provides another experimental model on the use of the multiperiod logit technique on the insolvency data of commercial banks in America from the FDIC website in 1980-1992. In their research, Cole and Wu compared the single time frame probit model and the multiperiod logit model, the consequence of this study finding the results that the multiperiod logit model gives a great hope better than the static model with the consequence of forecasting accuracy of 93.12% compared to 72.34%. in the first decile.

Based on the above description in the form of a demonstration of the bankruptcy of a manufacturing company listed on the Indonesia Stock Exchange (IDX), a survival analysis will be carried out using the multiperiod logit technique. This technique can predict several companies in financial distress. Based on research [5][6] that the hazard function needs to be refined to get the best estimation model. Using eighteen financial ratios in the company's report and two macroeconomic indicators that are suspected to affect the company's liquidation. This exploration is expected to contribute to several groups such as investors, creditors, company management, and public authorities in providing an overview to take steps that are believed to be used to survive or prevent company bankruptcy.

LITERATURE REVIEW

A. Survival Analysis

Survival analysis is a statistical method where the variable to be observed is the time duration until the occurrence of an event[7], [8]. In this study, the event in question is a condition of Financial distress. The time that is the focus of survival analysis is called survival time (T) because it shows the time an individual "survive" in a certain observation period. While the event can be considered

as a failure (*d*). An event is denoted by the symbol d to define the status of the event whether it is failure or censored. The value of *d=1* indicates failure and *d=0* indicates censored. In general, the purpose of survival analysis is as follows.

a. Estimate and interpret survival function and/or hazard function from survival data.
b. Compare survival and/or hazard functions.
c. Knowing the effect of predictor variables on survival time.

B. Survival Function dan Hazard Function

In the endurance test, there are two basic quantities that are commonly used, namely the survival function referred to as S(t) and the hazard function indicated by h(t). The survival function is characterized as the probability that an individual can survive more than a certain time, while the hazard function is characterized as the rate at which an event occurs after the individual survives for a certain period of time. It tends to be numerically expressed as follows.

$$\boldsymbol{S(t) = P(T > t)} \tag{1}$$

Where T is the time of occurrence as an arbitrary nonstop variable, the survival function is the complement of the cumulative distribution function. Where the cumulative distribution function is characterized as the probability value of the random variable T that is incorrect or equivalent to time t written as $F(t) = P(T \leq t)$, so the survival function can be expressed as follows.

$$\boldsymbol{S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t)} \tag{2}$$

When expressed in a survival probability density function (PDF) the results are obtained as follows.

$$\boldsymbol{S(t) = P(T > t) = \int_{t}^{\infty} f(u)du} \tag{3}$$

The second fundamental quantity, especially the hazard function, is defined as the rate at which a single event experiences an event in the time span t to $t + \Delta t$ if it is realized that the individual is still alive up to time t. Mathematically the hazard function can be denoted as follows.

$$\boldsymbol{h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}} \tag{4}$$

The relationship between the survival function and the hazard function can use the conditional probability theory $P(A|B) = \frac{P(A \cap B)}{P(B)}$, where A is the hazard function and B is the survival function. In addition, $P(A \cap B)$ is the probability of a joint event between A and B. The conditional probability value of the meaning of the hazard function is as follows.

$$\frac{P(t \le T < t + \Delta t)}{P(T > t)} = \frac{F(t + \Delta t) - F(t)}{S(t)} \quad (5)$$

where F(t) is the distribution function of T, then we get

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{F(t + \Delta t) - F(t)}{\Delta t} \right\} \frac{1}{S(t)} \quad (6)$$

So, the relationship between the survival function and the hazard function is as follows.

$$h(t) = \frac{f(t)}{S(t)} \quad (7)$$

C. K-Nearest Neighbour Multiple Imputation

Missing data is one of the problems that are often encountered in the use of big data. Therefore, we need a special method to deal with missing data. McGraw Hill Finance (2015) in their work report states that the k-nearest neighbor (KNN) method is better for imputing financial ratio data. The KNN imputation method is one method to overcome missing data without the need for the formation of a prediction model for each item that experiences missing data, but only uses a distance measure[9].

1. The procedure for imputing missing data using the KNN method is as follows: Determine the value of K, which is the number of closest observations that will be used to estimate the missing data.
2. Calculate the distance between observations containing missing data on the jth variable and other observations that do not contain missing data on variables other than j (denoted by j') using the following formula.

$$d(x_a, x_b) = \sqrt{\sum_{\substack{j'=1 \\ j' \neq j}}^{m} \left(x_{aj'} - x_{bj'}\right)^2}$$

where:

$d(x_a, x_b)$ = distance between observations x_a and observations x_b on variables other than variable j

x_{aj} = the value of the j[th] variable in the target observation x_a

x_{bj} = the value of the j[th] variable on the target observation x_b

3. Find the closest K observations based on the smallest distance value. The value of the variable in the nearest K observations will be used for the imputation process in the observations that contain missing values.
4. Calculate the weight (weight) on each K closest observation. The closest observation will get the biggest weight.
5. Calculate the average value of the nearest K observations that do not contain missing values with the weighted mean estimation procedure, which is with the following formula.

$$\widehat{x_j} = \frac{1}{w}\sum_{k=1}^{K} w_k v_{kj}$$

where vkj is the value of the j[th] variable at the k[th] observation, k=1,2,…,K dan $W = \sum K \sum_{k=1}^{K} w_k v_{kj}$ is the k-th nearest neighbor observation weight, where $w_k = \frac{1}{(x,v_k)^2}$

6. Perform the imputation process of missing data on observations that contain missing values with the average value obtained in stage 5.

D. Kaplan Meier Curve and Log Rank Test

In the survival analysis, the Kaplan-Meier curve was used to estimate the survivor function[7]. The Kaplan-Meier curve is a curve that describes the relationship between survivor function estimation and survival time. If the probability of Kaplan-Meier is denoted by $S(t(j))$ then the general Kaplan-Meier equation is as follows.

$$S(t(j)) = S(t(j\text{-}1)) \times Pr(T > t(j) | T \geq t(j))$$

$$\hat{S}(t(j-1)) = \prod_{i=1}^{j-1} \hat{P}r(T > t(j) | T \geq t(j))$$

So that $S(t(j))$ can be formulated as follows.

$$\hat{S}(t(j)) = \prod_{i=1}^{j} \hat{P}r(T > t(j) | T \geq t(j))$$

The Kaplan-Meier survival curve can be illustrated through Figure 2.2 below.

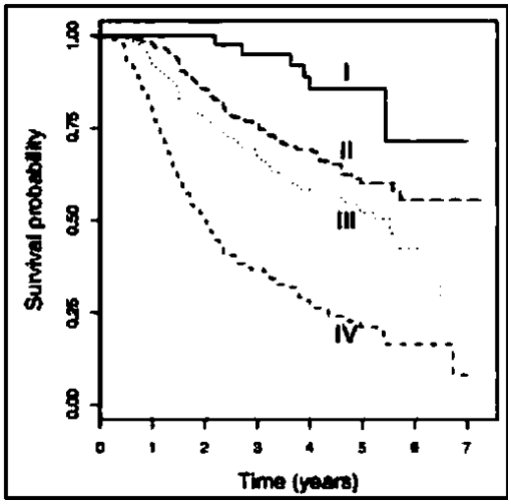


Figure 2 Illustration of Kaplan-Meier Curve

The Kaplan-Meier curve illustration in Figure 2 shows that in a period of 7 years, the survival curve of group I individuals is above the survival curve of group II, III and IV individuals. This shows that individuals belonging to group I have a higher probability of surviving for 7 years when compared to individuals of other groups. On the other hand, group IV individuals have the lowest probability of surviving for 7 years when compared to individuals belonging to other groups. In addition to the Kaplan-Meier curve, there is also a log rank test that is used to compare survival curves in different groups[7]. The hypothesis used in the log rank test for two or more is as follows.
H_0: there is no difference in the survival curve between different groups
H_1: there is at least one difference in the survival curve between different groups

E. Variable Selection

The hazard model using many independent variables will cause problems, namely the occurrence of multicollinearity cases. Therefore, a method is needed to select independent variables that can produce the best model and avoid multicollinearity cases. Variable selection methods that can be used in the hazard model are forward, backward and stepwise methods. The forward method is a variable selection method that works by inserting independent variables into the model gradually, the backward method is a variable selection method that works by inserting all independent variables first, then independent variables that do not have a significant effect in the model will be removed from the model, while the stepwise method is a method that combines forward and backward methods. In these three methods, it is important that a variable is not measured based on the AIC value (Akaike Information Criterion)[10].

$$AIC = -2L + 2(p + 1) \tag{8}$$

where:
L: log-likelihood model
p: many parameter estimates in the model

The steps of variable selection using the forward method are as follows.

1. Calculate the value of $AIC0$, where the value is the AIC value for the model with intercept only.
2. Select the independent variable included in the model by calculating the value of $AIC(0)$ for the model containing the variable x_j. Suppose x_j is an independent variable with $j = 1, 2, \ldots, p$. The selected independent variable is the variable in the model with the independent variable x_{e1} that has the smallest $AIC(0)$.
3. Test $AIC(0)$ with $AIC0$. If $AIC(0)$ is smaller than $AIC0$ then the independent variable x_{e1} can enter the model.
4. Calculate the value of $AIC(1)$ which expresses the AIC value of the model containing the variables x_{e1} and x_j. Suppose the model containing independent variables $xe1$ and $xe2$ is the model that has the smallest AIC value of $\text{AIC}_{e_{ij}}^{(1)}$ and $\text{AIC}_{e_{ij}}^{(1)}$ is smaller than $AIC(0)$ then the independent variable $xe2$ can enter the model.
5. The iteration stops when there is no model with the addition of a new variable that has an AIC value smaller than the AIC value of the previous model.

The steps of variable selection using the backward method are as follows

1. Enter all independent variables into the model and calculate the value of $AIC(0)$.
2. Calculate the value $\text{AIC}_{e_{ij}}^{(1)}$ which expresses the AIC value of the model containing the variable xej . The most suitable variables to leave the model are variables with the smallest value of $\text{AIC}_{e_{ij}}^{(1)}$. If this variable is expressed as x_{r1} .
3. Calculate the value of $\text{AIC}_{-e_{r1}}^{(2)}$ which expresses the AIC value of the model without variables x_{r1}. If the value of $\text{AIC}_{-e_{r1}}^{(2)}$ is less than the value of $AIC(0)$ then $-er1$ continue reducing the independent variable as in step 2.
4. The iteration stops when there is no model with the elimination of a new variable that has an AIC value smaller than the AIC value of the previous model.

While the stepwise method is a method that combines forward and backward. So the step to use the stepwise method is to perform selection using forward and backward in each stage to obtain the smallest AIC value.

F. Multiperiod Logit

Multiperiod Logit model is a logit model estimated by using survival data with autonomous perception between objects. The multiperiod logit model is comparable to the hazard function model in discrete time with a function form

$$h(t_i, x_i; \theta) = P(T \leq t | y_i = 1)$$

with y = {1; there is an event and 0; others}, so that the multiperiod logit model can be described as a hazard function model[3].

The relationship between the multiperiod logit model and the hazard model can be described as follows. Because the multiperiod logit model is estimation data taken from independent observations, the likelihood function is as follows[3].

$$L = \prod_{i=1}^{n} \left(F(t_i, x_i; \theta)^{y_i} \prod_{j<t_i} [1 - F(j, x_i; \theta)] \right)$$

As a probability distribution function, the value of F will be between zero and one $(0 \leq F \leq 1)$, with $F(0) = 0$ and $F(\infty) = 1$. The value of F always depends on t, so that F can be interpreted as a hazard function $h(t)$

$$L = \prod_{i=1}^{n} \left(h(t_i, x_i; \theta)^{y_i} \prod_{j<t_i} [1 - h(j, x_i; \theta)] \right)$$

defined the likelihood of the survival function as follows.

$$S(t,x;\theta) = \prod_{j<t_i}[1 - h(j,x_i;\theta)]$$

If the above survival function is substituted into the hazard function equation, the likelihood function is obtained as follows.

$$L = \prod_{i=1}^{n}\left(h(t_i,x_i;\theta)^{y_i}S(t,x;\theta)\right)$$

The likelihood function is equivalent to the likelihood function produced by the hazard model which was first introduced by Cox and Oakes in 1984 [11]. So that the model obtained from the Multiperiod Logit method is equivalent to be used as a hazard function.

METHODS

A. Data Types and Sources

This type of research is quantitative research. While the data used in this review is optional data obtained from the organization's quarterly fiscal reports on the Indonesia Stock Exchange website IDX and ICMD from the first quarter of 2001 to the third quarter of 2021. In the manufacturing sector, eight sub-areas were selected for observation, more specific, food, drink, and food. pets, ceramics, glass and porcelain, synthetic materials, footwear, plastics and bundling, mash and paper, and apparel materials and articles. This data is data from the organization's financial statements so that calculations are completed first to obtain financial proportion information that will be used as a predictor variable.

B. Variables

The response variable is in the form of a dummy variable, where the company's status will have a value of 1 if the company experiences financial distress and a value of 0 for others.
While the predictor variables consist of 16 company financial ratios and the JCI and the BI rate which are used as macroeconomic indicators.
The stages of data analysis to be carried out in this study are as follows.

1. Collect financial data for manufacturing companies listed on the Indonesia Stock Exchange (IDX) from the first quarter of 2001 to the third quarter of 2021. Financial data is obtained from the publication of quarterly issuer financial reports on www.idx.go.id (IDX website) as well as from www.icamel.id (IDX electronic library)
2. Pre-processing the company's financial data using the progress report, namely: working on the company's financial reports, converting data into discrete form, handling missing qualities and anomalies. The company's financial ratios are determined using the equation of financial ratios in general. Then, at that point, the missing data is associated using the k-Closest Neighbor technique. From there, exclusions are controlled using the 2.5% quantile for the upper and lower quantiles.
3. Describe the quality of financial ratio information by using a clear measurable strategy as mean, middle, minimum and maximum. As well as showing the difference in the

circumstances of the sample surviving and failing (facing bankruptcy) using the Kaplan-Meier curve.

4. Create a multiperiod log model with progress attached to it.
   a. Selecting variables by utilizing forward, backward and gradual strategies.
   b. Form a multiperiod logit model by utilizing indicator factors from the best model results from the three strategic choices above.
   c. Estimating the parameters of the multiperiod log model
   d. Performing a significance test of the parameters of the multiperiod logit model using simultaneous testing (likelihood ratio) and partial testing (Wald test)
   e. interpret the Multiperiod Logit model and then calculate the probability value of the hazard function, survival, and probability of financial distress in each sampled company.
5. Make a conclussions

RESULT AND DISSCUSSION

This chapter examines the characteristics of manufacturing companies listed on the IDX from 2001 to the third quarter of 2021. There are 79 companies in the manufacturing sector, out of these 79 companies 73 have survived, 4 financial distress companies and 2 Relisting companies. Companies that are categorized as experiencing financial distress are companies that survive and financial distress. The Relisting Company will not be used in the analysis because the survival model used is not a repeating model. The two Relisting companies will be discussed in a separate sub-chapter. The data consists of 16 financial ratios and 2 macroeconomic indicators. In addition, there are also EPS and PBV variables contained in the data, but these two variables are not financial ratios, so statistical descriptions will only be carried out and not included in the modeling.

In the next stage, multiperiod logit modeling is carried out. Modeling begins with selecting variables to get the best model. Then to find out the factors that influence the bankruptcy of companies, simultaneous and partial testing of the best model obtained from the selection of variables is carried out. The interpretation of the model will be carried out at the end of the discussion.

A. Pre-Processing Data

The risk of using big data in analysis is that there are outliers and missing data. In this sub-chapter, we will discuss methods for overcoming the problem of missing data and outliers.

B. Missing Data

The financial ratio data used is data that still contains missing data (missing value). Comparison of observations which are complete data and contain missing data is shown in the following figure.

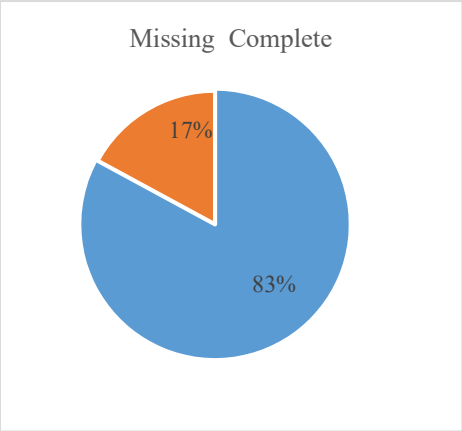


Figure 3. Complete Data Comparison with Missing Data

From the picture above it appears that 17% of the total data is in the form of incomplete time-observation or contains missing values. Eliminating missing data means deleting one observation (company) because the data used is panel data which are independent of one another. It should also be noted that 17% of the missing data comes from almost all companies. So deleting lost data is not a workable solution.

To handle missing data, imputation is carried out using k-nearest neighbor. This imputation method uses the data around the missing data as a reference to fill in the missing data.

C. Outliers in Financial Ratios

In addition to problems regarding missing data, outliers are also a problem encountered in the use of financial ratio data. This is due to the long observation interval which reaches 84 quarters or nearly 21 years, and the diversity of companies' financial conditions which can be very different from one another.

As an illustration, the following is a description of the data before handling the outliers.

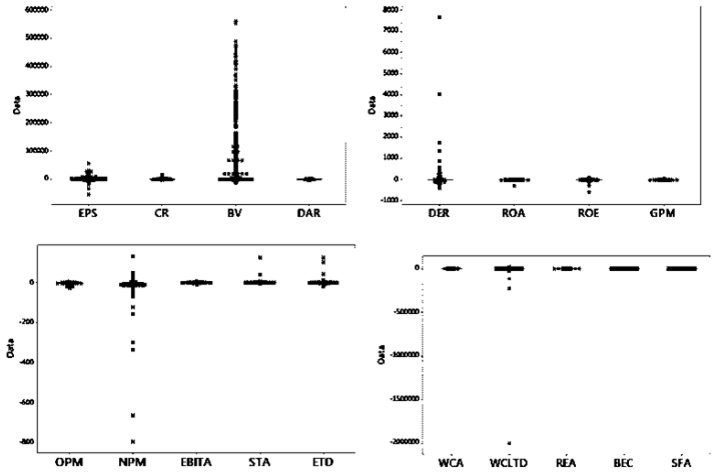


Figure 4. Box-Plot of Financial Ratios Variable

Based on the picture above, it can be seen how the outliers are in each financial statement variable. The distribution of financial ratio data is very wide. Even the shape of the box-plot is not very visible because the range compared to the quartiles is very large. Normally financial ratio data will spread between zero and one. However, in reality, several financial ratio variables are spread very wide.

D. Kaplan Meier Curve and Log Rank Test

The probability that a company can maintain its shares listed on the IDX is shown in the following figure.

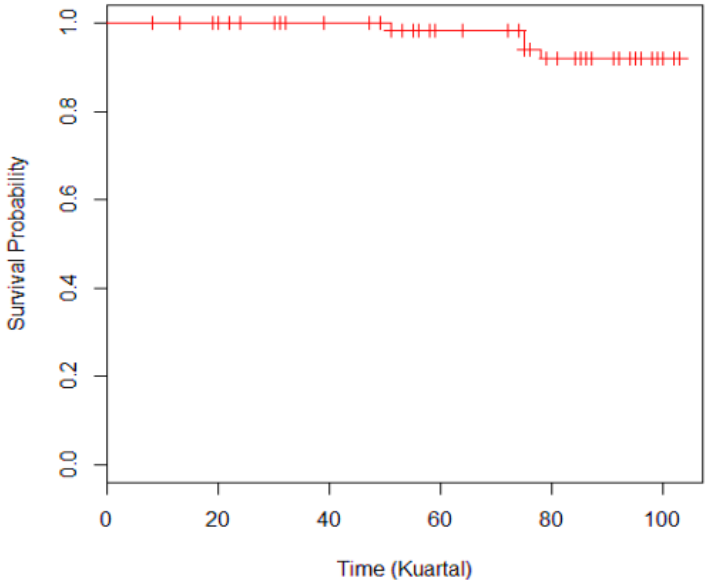


Figure 4. Kaplan-Meier Survival Curve for All Companies in the Manufacturing Sector Listed on the IDX

Figure 4 shows the survival curve of all manufacturing companies listed on the IDX. With the limited data on delisted companies used, there is no significant decrease in survival probability. Based on data used by manufacturing sector companies listed on the IDX, they were able to maintain their shares on the IDX for 103 quarters, which were relatively the same, namely above 80%. This also shows that the effectiveness of the company's business in the manufacturing sector is able to provide a sense of security to investors quite well.

While the probability that a company can maintain its shares listed on the IDX is grouped based on three sectors, namely the basic chemical industry sector, the various industrial sectors and the consumer goods industry sector as shown in Figure 4.6 as follows.

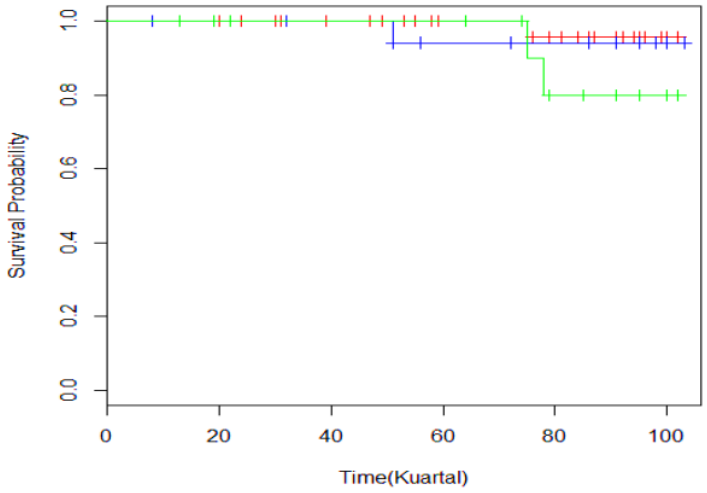


Figure 5. Kaplan-Meier Survival Curve of All Manufacturing Sector Companies Listed on IDX by Sub-Sector

Figure 5 shows that the red plot is the survival curve for companies in the basic chemical industry sector, the blue plot is the survival curve for various industrial companies and the green plot is the survival curve for companies in the consumer goods industry sector. The picture above can be interpreted that the probability of survival in the three sectors has a relatively equal probability of being able to maintain its shares on the IDX for 21 years, namely above 75% and still close together. The curves for the basic chemical industry and various industries coincide and are constant from the beginning of the observation to the end of the observation, but in the 50th quarter companies in the various industrial sectors experienced a decrease in the probability of survival, while the basic chemical industry sector experienced a decrease in the probability of survival in the 74th quarter. For the food and beverage industry sector, there was a sharper decline than the other two sectors, starting in the 74th quarter and dropping again in the 80th quarter.

To strengthen the conclusion that the survival chances of the three sub-sectors are not different, a log rank test was carried out which produced a log rank test statistical value of 2.2 and a p-value of 0.340. If a 90% confidence level is used, a failure to reject H0 is obtained, which means that there is no difference in the survival curves between the three manufacturing sub-sectors. So it can be concluded that companies in the three manufacturing sub-sectors have the same probability of staying on the stock exchange for 84 quarters.

E. Manufacturing Company Financial Distress Modeling on IDX Multiperiod Logit

With variables that always change over time. Then the static model will be difficult to be able to describe the status of the company will continue to survive or experience financial distress. The use of multiperiod logit is expected to give better results than the static model. Because the shape of the likelihood function is the same, the multiperiod logit estimator can be calculated using the logit program. The predictor variable used in modeling the financial distress of manufacturing companies on the IDX is a predictor variable that has gone through a variable selection process. Parameter estimation values and standard errors of the multiperiod logit method are shown in the following table.

Table 1. Estimator Value and Standard Error

Variable	Estimate	Std. Error	VIF
(Intercept)	-6,7649	5,2845	
CR	0,2784	0,129	1,1727
GPM	6,1418	3,1744	1,3412
SFA	-5,335	2,7433	1,0687
BI Rate	-1,9689	0,882	1,2223

Based on Table 4.1 it can be seen that there are no VIF values that are more than ten, so there is no longer multicollinearity in the model. The hazard model is also obtained which is shown in the following equation.

$$h(t_i, x_i) = \frac{a}{1+a}$$

with,

$$a = exp(-6{,}7649 + 0{,}2784CR_i + 6{,}1418GPM_i - 5{,}335SFA_i - 1{,}9689BI.Rate_i)$$

Furthermore, a simultaneous test was carried out to determine whether the predictor variables affect the rate of financial distress of manufacturing companies. Simultaneous testing was carried out using the likelihood ratio test and the $\chi 2$ value was 27.956 while the $\chi 2$ value was 7.78. Because the value of $\chi 2$ is greater than $\chi 2$, then reject H0, which means that there is at least one predictor variable that has a significant effect on the model. Furthermore, partial testing is carried out, the partial test values are shown in the following table.

Table 2. Partial Test Wald Value

Variable	Estimate	Std. Error	Z Value	Pr(>|z|)
(Intercept)	6.7649	5.2845	1.2800	0.2005
CR	0.2784	0.1290	2.1590	0.0309
GPM	6.1418	3.1744	1.9350	0.0530
SFA	-5.3350	2.7433	-1.9450	0.0518
BI.Rate	-1.9689	0.8820	-2.2320	0.0256

In Table 4.2, the four variables have a significant effect at the 90% confidence level on the occurrence of financial distress for manufacturing companies on the IDX, namely CR (Current Ratio), GPM (Gross Profit Margin), SFA (Sales to Fixed Assets) and BI rate. Based on the modeling that has been done, the parameter values for the Current Ratio and GPM variables are 0.2784 and 6.1418, respectively, these values indicate that the greater the value of the Current Ratio and Gross Profit Margin, the chances of a company experiencing financial distress will increase. at one time period. Meanwhile, the parameter value for the Sales to Fixed Assets (Fixed Asset Turnover Ratio) variable is -5.335, this value indicates that the greater the Sales to Fixed Assets

value, the less chance the company will experience financial distress during a period of time. The Sales to Fixed Asset Ratio measures the company's ability to make sales of the fixed assets used. The more efficiently a company uses its fixed assets, the smaller the chance the company will experience financial distress at one time.

Bank Indonesia interest rates also have a significant influence on the chances of financial distress for manufacturing companies on the IDX in a certain period. Bank Indonesia's interest rate reflects the government's response to Indonesia's economic conditions. In the multiperiod logit model obtained, the estimated value of the BI rate parameter is -1.924, this value indicates that the greater the value of the BI rate, the chances of a manufacturing company experiencing financial distress in a certain period will decrease.

F. Hazard, Survival, and Delisting Opportunities for Manufacturing Companies Based on the Multiperiod Logit Model

The hazard opportunity is obtained by adding up the hazard rate for each company in each quarter until the last determined quarter. Meanwhile, calculating the probability of survival is obtained by using the relationship between the hazard function and the survival function which has been described in equation (2.14). Opportunities for financial distress are obtained from the remaining survival opportunities. For example, it will calculate the value of the opportunity hazard, survival and financial distress of a glass producing company with the issuer code AMFG in the last quarter or the 103rd quarter. So it must be known in advance the value of the company's hazard rate from the start of the IPO in the 23rd quarter to the end. The hazard rate is calculated using the hazard function that has been obtained from the multiperiod logit model.

$$h(t_i, x_i) = \frac{a}{1+a}$$

with,

$$a = exp(-6{,}7649 + 0{,}2784CR_i + 6{,}1418GPM_i - 5{,}335SFA_i - 1{,}9689BI.Rate_i)$$

The hazard opportunity is obtained by adding up the hazard rate in the 23rd quarter to the end so that the hazard opportunity for AMFG issuers is 0.0334. After knowing the probability of hazard, the probability of survival is known to be 0.9672 so that the probability of financial distress for AMFG issuers is only 0.0328. The full value of hazard, survival and financial distress opportunities can be seen in the attachment. Descriptive value of hazard, survival and financial distress opportunities is shown in the following table.

Table 3. Description of Hazard Opportunity Statistics, Survival and Financial Distress

Probability	Statistik					
	Mean	Min	Q1	Median	Q3	Max
Hazard	0,051	0,000	0,002	0,013	0,042	0,861
Survival	0,955	0,422	0,959	0,987	0,998	100,0
Financial Distrees	0,044	0,000	0,002	0,013	0,041	0,577

Based on Table 3, it is known that there are companies with a large probability hazard value of up to 85% while on the contrary, there are also companies with the smallest hazard value of 0%. Companies with smaller hazard values will be safer to invest because the chances of financial distress occurring in these companies are smaller, and vice versa. The five companies with the smallest opportunity hazard values are companies with company codes SKBM, IGAR, PBRX, PSDN and UNIC.

While the companies with the greatest opportunity hazard values are IIKP, UNIT, SIMM, PWSI and DAVO. Companies with the codes SIMM, PWSI and DAVO are companies that have experienced financial distress, so it is only natural that these three companies have a high probability hazard value. Companies that need to be considered are IIKP and UNIT, these two companies have high hazard opportunities. A company rescue step is needed so that the two companies are not in financial distress from the stock exchange.

CONCLUSSION AND SUGGESTIONS

Descriptively, surviving and financial distress companies have a prominent difference in the profitability ratio and market measure ratio. In the variable profitability ratios the different financial ratios of the two groups of companies are Book Value, Debt on Equity Ratio, Return on Assets, Operating Profit Margin and Net Profit Margin. Meanwhile, in the market measure ratio, the variables that differ between the two groups of companies are EPS, Book Value per Share and Debt Equity Ratio. Meanwhile, a relisting company has a lower profitability ratio than during financial distress, but when it improves again on the stock exchange, the company's profitability ratio increases again. Differences in the company sector do not show significant differences in survival curves. This is evidenced by the insignificant Log Rank test.

The resulting model from modeling using multiperiod logit using the backward variable selection results produces the following model.

$$h(t_i, x_i) = \frac{a}{1+a}$$

with,

$$a = exp(-6{,}7649 + 0{,}2784CR_i + 6{,}1418GPM_i - 5{,}335SFA_i - 1{,}9689BI.Rate_i)$$

Based on the results of the simultaneous test of the multiperiod logit model, it is concluded that there is at least one predictor variable that has a significant effect on the financial distress of manufacturing companies on the IDX. From the partial test results it is known that of the five variables in the model there are four significant variables namely CR (Current Ratio), GPM (Gross Profit Margin), SFA (Sales to Fixed Assets) and BI rate. The five best companies to invest in with minimum hazard opportunity values are companies with issuer codes SKBM, IGAR, PBRX, PSDN and UNIC. Accuracy and geometric mean values are used to measure the accuracy of the model. The multiperiod logit model obtained has accuracy and geometric mean values of 0.9726 and 0.8541, respectively.

REFERENCES

[1] H. D. Platt and M. B. Platt, "Development of a class of stable predictive variables: the case of bankruptcy prediction," *J. Bus. Financ. Account.*, vol. 17, no. 1, pp. 31–51, 1990.
[2] L. S. Almilia and K. Kristijadi, "Analisis rasio keuangan untuk memprediksi kondisi financial distress perusahaan manufaktur yang terdaftar di bursa efek Jakarta," *J. Akunt. dan Audit. Indones.*, vol. 7, no. 2, 2003.
[3] T. Shumway, "Forecasting bankruptcy more accurately: A simple hazard model," *J. Bus.*, vol. 74, no. 1, pp. 101–124, 2001.
[4] R. A. Cole and Q. Wu, "Predicting bank failures using a simple dynamic hazard model," in *22nd Australasian Finance and Banking Conference*, 2009, pp. 16–18.
[5] F. T. Kristanti, S. Rahayu, and A. N. Huda, "The determinant of financial distress on Indonesian family firm," *Procedia-Social Behav. Sci.*, vol. 219, pp. 440–447, 2016.
[6] F. T. Kristanti, N. Effendi, A. Herwany, and E. Febrian, "Does corporate governance affect the financial distress of Indonesian company? A survival analysis using Cox hazard model with time-dependent covariates," *Adv. Sci. Lett.*, vol. 22, no. 12, pp. 4326–4329, 2016.
[7] F. Anderson, "Statistics by Example: Hands on Approach Using R and/or Excel." CreateSpace Independent Publishing Platform, 2016.
[8] G. Anuraga, A. Indrasetianingsih, and M. Athoillah, "Pelatihan Pengujian Hipotesis Statistika Dasar dengan Software R," *BUDIMAS J. Pengabdi. Masy.*, vol. 3, no. 2, 2021.
[9] M. Athoillah, I. Irawan, M., and M. Imah, Elly, "Study Comparison of SVM-, K-NN- and Backpropagation-Based Classifier for Image Retrieval," *J. Ilmu Komput. dan Inf. (Journal Comput. Sci. Information)*, 2015.
[10] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*, vol. 398. John Wiley & Sons, 2013.
[11] D. R. Cox, *Analysis of Survival Data*. CRC Press, 2018.

Optimasi Produk Plastik pendekatan Taguchi *Mixed Level* pada Faktor Interaksi Injeksi Molding

Muhammad Ahsan(1), Galuh Kusuma W(2), and Salman Alfarizi P A(3)

Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Jl. Teknik Mesin No.175, Keputih, Sukolilo Telp. (031) 5943352 Surabaya 60115

e-mail: muh.ahsan@its.ac.id(1), galuhkusuma14@gmail.com(2), dan s_alfarizi98@yahoo.com(3)

ABSTRAK

Penelitian parameter injeksi molding diperlukan untuk meminimalkan produk cacat penyusutan (*shrinkage)*. Diharapkan perbaikan penelitian yang mengintegrasikan metode Taguchi untuk mengoptimalkan parameter injeksi serta interaksi yang mungkin terjadi. Berdasarkan uraian tersebut permasalahan yang dibahas adalah bagaimana mendapatkan parameter terbaik untuk optimasi produk plastik menggunakan pendekatan taguchi *mixed level* dengan 6 faktor faktor interaksi injeksi molding dan variasi level. Metode Taguchi merupakan suatu metode pengendalian kualitas sebelum proses berlangsung, *orthogonal array* yang akan digunakan $L_{18}(2^1 3^5)$ sesuai dengan jumlah faktor sebanyak 6 dan menggunakan mixed level. Variabel respon *shrinkage* diukur dari plastik *tray* yang terbuat dari campuran polimer atau poliblend. Faktor yang digunakan yaitu dengan memvariasikan *injection speed* (kecepatan injeksi), *melting temperature* (suhu leleh), *injection pressure* (tekanan injeksi), *holding pressure* (tekanan penahanan), *holding time* (waktu penahanan) dan *cooling time* (waktu pendinginan). Parameter optimal tanpa efek interaksi dengan kondisi optimum adalah kecepatan injeksi (90 %rpm), suhu leleh (240 °C), tekanan injeksi (110 bar), tekanan penahanan (96 bar), waktu penahanan (5 detik), dan pendinginan waktu (10 detik). Parameter optimal dengan efek interaksi kondisi optimum yaitu suhu leleh (240 ° C), tekanan injeksi (110 bar), tekanan penahanan (80 bar), waktu penahanan (5 detik), dan waktu pendinginan (10 detik). Kontribusi parameter dengan interaksi adalah suhu leleh (56,65%), interaksi suhu leleh dan tekanan injeksi (15,75%), waktu pendinginan (11,12%), dan kekuatan tekanan (7,83%).

Kata kunci : Injeksi Molding, Mixed Level, Taguchi, Interaksi.

ABSTRACT

Injection molding parameter research is required to minimize product shrinkage defects. Expected improvement of research that integrates Taguchi method to optimize injection parameters as well as possible interactions. Based on this description, the problem discussed is how to get the best parameters for plastic product optimization on injection molding interaction factors with taguchi mixed level approach. Taguchi is a method of quality control before the process takes place, orthogonal array to be used $L_{18}(2^1 3^5)$ according to the number of factors as much as 6 and using mixed level. Shrinkage response variable measured from plastic tray made of a mixture of polymers or poliblend. Factors used are by varying the injection speed, melting temperature, injection pressure, holding pressure, holding time and cooling time. The optimal parameters without interaction effect with optimum conditions are injection speed (90% rpm), melting temperature (240 °C), injection pressure (110 bar), holding pressure (96 bar), holding time (5 seconds), and cooling time (10 seconds). Optimal parameters with the effect of the interaction of optimum conditions ie. melting temperature (240 °C), injection pressure (110 bar), holding pressure (80 bar), holding time (5 seconds), and cooling time (10 seconds). The contribution of parameter with interaction is melting temperature (56.65%), interaction of melting temperature and injection pressure (15.75%), cooling time (11.12%), and pressure strength (7.83%).

Keywords : *Injection Molding, Mixed Level, Taguchi, Interaction.*

PENDAHULUAN

Saat ini penggunaan plastik menjadi kebutuhan hampir setiap orang, plastik disukai karena mudah digunakan dan diproduksi secara masal. Hal ini diikuti peningkatan kualitas proses pembuatan plastik untuk meningkatkan produktivitas, pada industri manufaktur proses injeksi molding berperan besar dalam pengolahan plastik. Penelitian parameter injeksi molding diperlukan agar meminimalkan banyaknya produk cacat [1]. *Shrinkage, Warpage, sink marks* dan *weld lines* merupakan cacat yang banyak terjadi pada injeksi molding. Parameter proses meliputi kecepatan injeksi, tekanan injeksi, tekanan ketahanan, suhu leleh, waktu ketahanan, waktu pendinginan, dan lainnya. Penelitian sebelumnya telah banyak dilakukan untuk meningkatkan dan mengoptimalkan kualitas produk plastik pada mesin injeksi molding [2]. Terdapat juga penelitian mengenai kerangka general untuk mengoptimalkan injeksi molding plastik dengan parameter suhu leleh, suhu molding, tekanan injeksi, waktu injeksi, injeksi pengepakan, waktu pengepakan, dan sebagainya [3], Penelitian lainnya membahas mengenai seleksi parameter injeksi molding menggunakan metode taguchi dan ANOVA [4]. Penentuan parameter proses yang optimal secara rutin dilakukan di bidang industri injeksi molding plastic yang memiliki pengaruh langsung pada kualitas. Penelitian sebelumnya kurang memperhatikan interaksi parameter dan *mixed level* yang kemungkinan besar mempengaruhi kualitas produk dan biaya, seperti pada penelitian *Job Stress As A Predictor Of Employee Health* [5], serta pada penelitian [6] dengan judul *Optimization of plastic injection molding process parameters for manufacturing a brake booster valve body.*

Diharapkan perbaikan penelitian yang mengintegrasikan metode Taguchi untuk mengoptimalkan parameter injeksi serta interaksi yang mungkin terjadi. Pengaturan parameter pada proses injeksi berhubungan erat dengan hasil suatu produk plastik. Penentuan parameter proses yang kurang akan menyebabkan hasil akhir suatu produk plastik cacat. Salah satu cacat yang dihasilkan adalah cacat *sink marks*. Penelitian mengenai cacat *sink marks* dengan menggunakan metode taguchi telah dilakukan dengan parameter *mold surface temperature*, *melt temperature*, *mold open time dan injection pressure* [7].

Metode Taguchi merupakan metode desain yang telah distandarisasi agar dengan mudah digunakan peneliti, selain itu mengatasi limitasi faktorial dan percobaan fraksional factorial [8]. Pada metode taguchi biasanya akan mempertimbangkan definisi kualitas, standarisasi DOE, strategi *robust design, loss function,* analisis *signal to noise (S/N)* [9]. Berdasarkan uraian tersebut permasalahan yang akan dibahas adalah bagaimana mendapatkan parameter terbaik untuk optimasi produk plastik dengan pendekatan metode taguchi *mixed level* pada faktor interaksi injeksi molding.

KAJIAN PUSTAKA

A. Taguchi

Metode Taguchi ditemukan oleh Dr. Genichi Taguchi. Metode ini merupakan suatu metode pengendalian kualitas sebelum proses berlangsung atau sering juga dinamakan *off-line quality control*. Riset tersebut sering kali menggunakan teknik *Design of Experiment* (DOE).

Design of Experiment (DOE) menggunakan seperangkat matriks khusus yang disebut *orthogonal array*, langkah untuk mengelompokkan faktor-faktor yang berpengaruh pada proses dan level-level yang bervariasi sehingga dapat memperoleh kualitas yang baik, serta menentukan

jumlah eksperimen minimal yang dapat memberikan informasi-informasi dari semua faktor yang berpengaruh terhadap parameter.

B. Klasifikasi Karakeristik Kualitas

Menurut [10], terdapat lima klasifikasi karakteristik kualitas. Kelima klasifikasi karakteristik kualitas tersebut adalah :

a. *Nominal-the-best*
   Nominal-the-best adalah karakteristik kualitas yang dapat diukur dengan target yang spesifik. Nilai dari target tersebut dapat berupa bilangan positif maupun negatif. Contohnya adalah penyinaran lampu LED pada kendaraan bermotor tidak terlalu terang tidak terlalu redup.
b. *Smaller-the-better*
   Karakteristik kualitas *smaller-the-better* memiliki nilai non-negatif. Karakteristik kualitas ini dapat diukur dan memiliki nilai target nol (0). Contoh dari karakteristik kualitas *smaller-the-better* adalah nilai penyusutan pada busa spandex terkecil merupakan produk yang terbaik.
c. *Larger-the-better*
   Karakteristik kualitas *larger-the-better* memiliki nilai non-negatif. Karakteristik kualitas ini dapat diukur dan memiliki nilai target tidak terhingga (*infinity*, ∞). Contoh kekuatan produksi mobil pickup, dimana mobil yang chasis mampu menahan bobot terberat itu lebih baik.
d. *Signed-target*
   Signed-target adalah karakteristik kualitas yang dapat diukur. Karakteristik kualitas ini memiliki nilai target nol (0). Karakteristik kualitas ini berbeda dengan karakteristik kualitas *smaller-the-better*. *Signed-target* dapat memiliki nilai negatif. Contoh dari karakteristik kualitas ini adalah aliran arus listrik dalam suatu wahana permainan.
e. *Classified attribute*
   Karakteristik kualitas atribut klasifikasi (*classified attribute*) adalah karakteristik kualitas yang memiliki skala diskret. Biasanya penilaian kualitas pada karakteristik ini berdasarkan penilaian subjektif. Terdapat skala diskret seperti Baik-Buruk, Setuju-Tidak Setuju, maupun Puas-tidak puas.

C. Faktor Variasi Proses *Injection Molding*

Variasi parameter proses berguna untuk menghasilkan suatu produk plastik dengan kualitas yang paling optimal. Pendekatan untuk *setting* parameter guna pengoptimalan produk plastik disini adalah dengan pendekatan kecacatan *shrinkage* [11]. Kecacatan dapat dipengaruhi oleh beberapa setting parameter, yaitu:

a. Temperatur leleh (*melt temperature*) adalah batas temperatur dimana bahan plastik mulai meleleh apabila diberikan *setting* temperatur dinaikkan.
b. Waktu penekanan (*holding time*) adalah lamanya waktu yang diperlukan untuk memberikan tekanan untuk mendorong lalu menahan cairan plastik yang sudah meleleh didalam barrel sehingga plastik cair tidak kembali kedalam barrel.
c. Waktu pendinginan (*cooling time*) adalah waktu pendinginan setelah cairan plastik diinjeksikan ke dalam cetakan agar cepat menjadi produk.

d. Tekanan injeksi (*injection pressure*) merupakan tekanan yang digunakan untuk menginjeksi cairan plastik kedalam cetakan. Tekanan ini dipengaruhi oleh luas proyeksi benda dan gaya yang dibutuhkan
e. Kecepatan Injeksi (*Injection Speed*) merupakan kecepatan mesin dalam menyemprotkan plastik polimer ke cetakan yang berada di bawahnya.
f. Kekuatan Tekanan (*Holding Pressure*) adalah kekuatan tekanan untuk menekan biji plastik polimer yang berada pada wadah.

D. Orthogonal Array

Orthogonal array merupakan suatu "jalan pintas" dalam melakukan *design of experiment* (Roy, 1990). Faktor-faktor yang ada tersebut dapat berubah sesuai dengan level pada faktor tersebut. Lambang dari orthogonal array dapat dilihat pada contoh berikut ini.

$$L_n(m^p)$$

Lambang tersebut memiliki arti:
L :*Latin Square*. Notasi memberikan gambaran informasi *orthogonal array*.
n :Jumlah baris yang mewakilkan jumlah eksperimen yang akan dilakukan.
m : Level yang ada pada eksperimen.
p : Jumlah faktor yang mewakilkan jumlah kolom yang ada.

E. *Signal to Noise* (S/N Ratio)

Percobaan atau eksperimen *robust design* memiliki fungsi objektif. Fungsi objektif tersebut sering disebut dengan *signal to noise ratio* atau S/N Ratio. S/N Ratio digunakan untuk mengoptimalkan karakteristik kualitas yang ada [12]. Perhitungan S/N ratio bergantung dengan karakteristik kualitas yang diinginkan.

a. Mean Squared Deviation (MSD)

Beberapa tujuan dari eksperimen *robust design* adalah meminimalkan sensitivitas dari faktor *noise* yang ada terhadap karakteristik kualitasnya. Beberapa keputusan yang diambil mungkin berasal dari MSD.

b. S/N Ratio *smaller-the-better* (SNR S)

Karakteristik kualitas untuk *smaller-the-better* memiliki nilai tujuan nol (0). Contoh dari karakteristik kualitas untuk *smaller-the-better* adalah nilai polusi yang dihasilkan dari sebuah perusahaan. Maka yang harus dilakukan adalah maksimasi S/N Ratio *smaller-the-better*. Rasio S/N untuk karakteristik ini dirumuskan dengan persamaan sebagai berikut:

$$S/N = -log\left[\sum_{i=1}^{n}\frac{(y_i)^2}{n}\right]$$

c. S/N Ratio *nominal-the-best* (SNR N)

Karakteristik kualitas dari *nominal-the-best* memiliki nilai yang kontinyu dan *non-negative*. Nilai tersebut antara nol (0) hingga tidak terhingga (∞). Fungsi tujuan dari *nominal-the-*

best adalah maksimasi S/N Ratio *nominal-the-best.* Rasio S/N untuk karakteristik ini dirumuskan dengan persamaan sebagai berikut:

$$S/N = -log\left[\sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{n}\right]$$

d. S/N Ratio *larger-the-better* (SNR L)

Target dari karakteristik kualitas *larger-the-better* adalah memperoleh nilai sebesar mungkin (∞). Fungsi tujuan dari S/N Ratio *larger-the-better* adalah maksimasi S/N Ratio *larger-the-better*. . Rasio S/N untuk karakteristik ini dirumuskan dengan persamaan sebagai berikut:

$$S/N = -log\left[\sum_{i=1}^{n}\frac{(\frac{1}{y_i{}^2})}{n}\right]$$

F. *Analysis of Variance* (ANOVA)

Analisis ragam pada metode Taguchi digunakan sebagai metode statistik untuk menginterpretasikan data hasil percobaan. Tujuan dari analisis varians (ANOVA) adalah untuk temukan parameter mana yang secara signifikan memengaruhi kualitas. Analisis variansi adalah teknik yang digunakan untuk menganalisis data yang telah disusun dalam desain secara statistik [13].

Tabel 1 Parameter Injeksi Molding dan Level

Sumber Variasi	Sum of Square (SS)	Degrees of Freedom (df)	Mean Square (MS)	F Ratio
Variabel Proses A	$ss_A = n_A\sum_{i=1}^{n}(A_i-\bar{y})^2$	$d_A - 1$	$MSA = \frac{ss_A}{dfA}$	$\frac{MSA}{MSE}$
Variabel Proses B	$ss_B = n_B\sum_{i=1}^{n}(B_i-\bar{y})^2$	$d_B - 1$	$MSB = \frac{ss_B}{dfB}$	$\frac{MSB}{MSE}$
Interaksi AB	$ss_{AB} = \sum_{I=1}^{n}\sum_{j=1}^{n}\frac{x_{ij}^2}{n_A} - \frac{T}{n_{AB}}$	$(a-1)(\mathrm{b}-1)$	$MSAB = \frac{ss_{AB}}{dfAB}$	$\frac{MSAB}{MSE}$
Error	$ss_E = ss_T - ss_A - ss_B$	$ab(n-1)$	$MSB = \frac{ss_B}{dfE}$	
Total	$ss_A = \sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$		

Uji distribusi F, menunjukkan bukti adanya perbedaan pengaruh masing-masing faktor dalam eksperimen [13]. Pengujian ini dilakukan dengan cara membandingkan variansi yang disebabkan oleh masing-masing faktor dan variansi error.

$$y_{ij} = \mu_i + e_{ij},\ j = 1,2,\dots,n_i, dan\ i = 1,2,\dots,c,$$

Dimana μ_i merupakan mean dan e_{ij} independen berdistribusi $N(0,\sigma)$. Hipotesis yang digunakan dalam pengujian ini untuk faktor yang tidak diambil secara random (*fixed*) adalah:

H_0 : $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_c$

H_1 : sedikitnya ada satu pasangan μ yang tidak sama

Kegagalan menolak H_0 mengindikasikan tidak adanya perbedaan rata-rata dari nilai respon yang dihasilkan pada perlakuan yang berbeda. Tolak H_0 apabila $F\ ratio \geq F_\alpha$.

METODE PENELITIAN

A. *Analysis of Variance* (ANOVA)

Data yang digunakan dalam penelitian ini merupakan data sekunder yang diambil dari penelitian sebelumnya yang berjudul "*Application of Taguchi Method in the Optimization of Injection Moulding Parameters for Manufacturing Products from Plastic Blend*" [14]. Perbedaan dari jurnal sebelumnya adalah penggunaan *orthogonal array mixed level* untuk analisis ANOVA dan pemilihan variabel respon menggunakan rata-rata dari *shrinkage*.

Metode menggunakan inovasi dari Dr. Genichi Taguchi, Jepang yang telah sukses digunakan di bidang industry dan melengkapi aktivitas peningkatan kualitas [15]. Pendekatannya memberikan strategi eksperimental baru dengan *design of experiment* (DOE) yang dimodifikasi dan distandarisasi. Taguchi menyarankan desain fraksional faktorial dan *orthogonal array* dengan metode statistika yang sesuai, untuk penelitian ini menggunakan ANOVA. *Orthogonal array* yang akan digunakan $L_{18}(2^1 3^5)$ sesuai dengan jumlah faktor sebanyak 6 dan menggunakan *mixed level* yang terlihat pada Tabel 2.

Tabel 2. Parameter Injeksi Molding dan Level

Faktor	Parameter	Satuan	Level 1	Level 2	Level 3
A	*Injection Speed*	rpm (%)	80	90	-
B	*Melting Temperature*	◦C	220	230	240
C	*Injection Pressure*	Bar	100	110	120
D	*Holding Pressure*	Bar	80	88	96
E	*Holding Time*	sec.	5	8	10
F	*Cooling Time*	sec.	5	8	10

B. FAKTOR

Variabel respon penelitian ini menggunakan *shrinkage* dari plastik *tray* yang terbuat dari campuran polimer atau poliblend [14]. Plastik tray diproduksi oleh mesin injeksi molding Battenfeld TM750/210, secepat mungkin setelah tercetak setiap tray akan diukur panjang dan lebarnya. Setelah 24 jam diukur kembali untuk mengetahui *shrinkage*, variabel ini menunjukkan stabilitas produk.

Faktor yang digunakan seperti pada tabel 2 (Parameter Injeksi Molding dan Level) terdapat 6 faktor yang mempengaruhi variabel respon yaitu dengan memvariasikan beberapa faktor yakni:

injection speed (injeksi kecepatan) di kisaran 80 – 90 rpm (%), *melting temperature* (suhu leleh) dalam kisaran 220 – 240 °C, *injection pressure* (tekanan injeksi) dalam kisaran 100-120 bar, *holding pressure* (tekanan penahan) pada kisaran 80 – 96 bar, *holding time* (waktu penahanan) dalam kisaran 5 – 10 detik dan *cooling time* (waktu pendinginan) dalam kisaran 5 – 10 detik [14].

HASIL DAN PEMBAHASAN

A. Rancangan Desain Optimasi Parameter

Digunakan *orthogonal array* agar lebih ekonomis dan menentukan jumlah percobaan paling minimal yang diperlukan dari sekumpulan parameter. Pemilihan *orthogonal array* yang sesuai dengan percobaan yaitu derajat kebebasan *orthogonal array* standar harus lebih besar atau sama dengan perhitungan derajat kebebasan pada percobaan [15]. *Mixed level* merupakan percobaan yang memiliki jumlah level berbeda antar faktor, Taguchi memodifikasi standar *array* agar sesuai dengan kondisinya. Desain *orthogonal array* $L_{18}(2^1 3^5)$ penelitian ini disajikan pada Tabel 2, bermakna terdapat 18 percobaan berdasarkan kombinasi faktor dan level, terdapat 2 level pada 1 faktor dan 3 level untuk 5 faktor lainnya. Selanjutnya analisis ini akan melihat juga interaksi antara faktor A dengan B dan faktor B dengan C mengggunakan aplikasi Minitab 17. Interaksi *Injection Speed, Melting Temperature,* dan *Injection Pressure* didukung penelitian bahan plastic yang meleleh dipengaruhi faktor tersebut dan interaksinya secara simultan untuk menyelidiki karakter desain parameter yang optimal [16].

Table 3. *Orthogonal Array* $L_{18}(2^1 3^5)$ untuk Desain Percobaan Parameter Injeksi Molding

Percobaan	Faktor					
	A	**B**	**C**	**D**	**E**	**F**
1	1	1	1	1	1	1
2	1	1	2	2	2	2
3	1	1	3	3	3	3
4	1	2	1	1	2	2
5	1	2	2	2	3	3
6	1	2	3	3	1	1
7	1	3	1	2	1	3
8	1	3	2	3	2	1
9	1	3	3	1	3	2
10	2	1	1	3	3	2
11	2	1	2	1	1	3
12	2	1	3	2	2	1
13	2	2	1	2	3	1
14	2	2	2	3	1	2
15	2	2	3	1	2	3
16	2	3	1	3	2	3
17	2	3	2	1	3	1
18	2	3	3	2	1	2

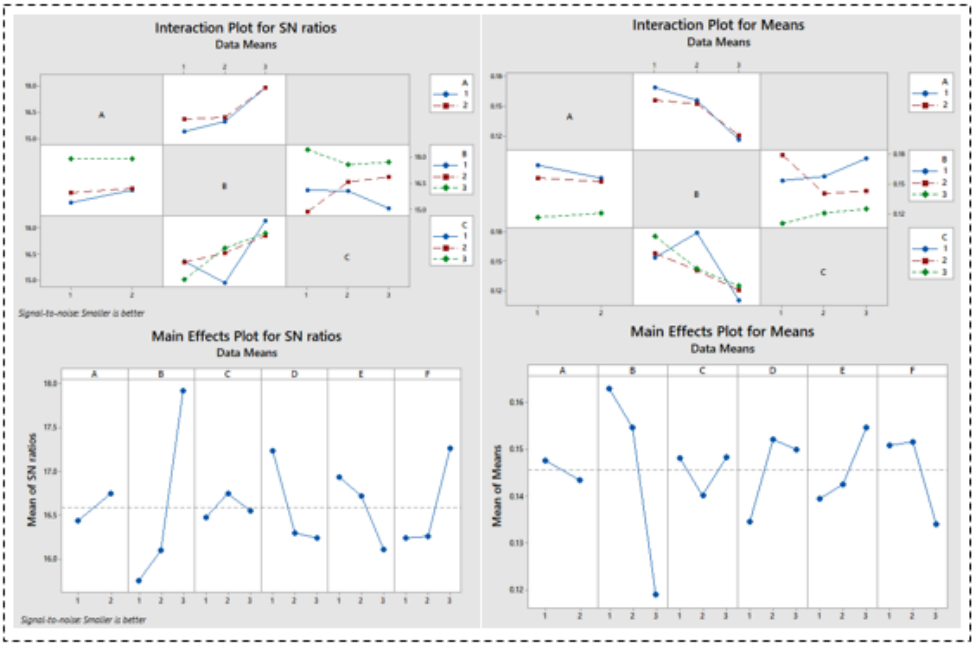
B. Hasil Analisis

Hasil pengamatan sesuai dengan desain percobaan *orthogonal array* didapatkan data pada Tabel 3 [14]. S*hrinkage* pada setiap *tray* akan diukur panjang dan lebarnya segera setelah tercetak. Setelah 24 jam diukur kembali untuk mengetahui perbedaannya, s*hrinkage* merupakan cacat produk yang sering terjadi pada proses injeksi molding.

Table 4. Hasil Pengamatan *Shrinkage*

Percobaan	*Shrinkage*		*Mean*
	Width	*Length*	
1	0.1054	0.1752	0.1403
2	0.1741	0.1813	0.1777
3	0.1845	0.1961	0.1903
4	0.1543	0.18	0.16715
5	0.0913	0.1801	0.1357
6	0.1736	0.1603	0.16695
7	0.071	0.1534	0.1122
8	0.0571	0.1822	0.11965
9	0.0735	0.1637	0.1186
10	0.1487	0.1874	0.16805
11	0.142	0.136	0.139
12	0.1761	0.1498	0.16295
13	0.2226	0.1629	0.19275
14	0.0966	0.196	0.1463
15	0.0899	0.1483	0.1191
16	0.0575	0.1588	0.10815
17	0.0836	0.1617	0.12265
18	0.1538	0.1099	0.13185

Berdasarkan hasil data pada Tabel 3 dilanjutkan analisis dengan karakteristik kualitas *smaller is better,* dipilih karena pada proses injeksi molding diinginkan *shrinkage* yang kecil karena menunjukkan proses tersebut telah berjalan stabil. *shrinkage* merupakan masalah yang banyak dipengaruhi faktor diantaranya *injection speed, melting temperature, injection pressure, holding pressure, holding time,* dan *cooling time*. Untuk mengetahui pengaruh level dan faktor terhadap rata-rata dapat ditunjukkan pada Gambar 1.



Gambar 1. Plot untuk S/N Ratio dan Rata-Rata Interaksi

Gambar 1 pada plot S/N rasio maupun Mean menunjukkan adanya interaksi pada faktor AB (*Injection Speed* dengan *Melting Temperature*) dan BC (*Melting Temperature* dengan *Injection Pressure*) terlihat dari garis yang saling berpotongan. Hal ini mendukung teori penelitian sebelumnya terdapat interaksi yang mempengaruhi *shrinkage* [14]. Pada Gambar *Main Effect Plot For Means* dan Tabel 4 disajikan *rank* pengaruh level dan faktor terhadap *shrinkage* yang muncul pada produk plastik *tray* yang selanjutnya digunakan untuk mencari *signal noise to ratio*. Delta menunjukkan selisih dalam tiap parameter, *rank* menunjukkan urutan dari selisih tiap parameter. Selanjutnya dapat dilihat pada Gambar *Main Effect Plot for SN ratios* dan Tabel 3 yang menunjukkan delta terbesar pada faktor B dengan *rank* 1, hal ini berarti bahwa faktor B (*Melting Temperature*) memiliki pengaruh paling besar terhadap hasil akhir. Sehingga *Melting Temperature* menjadi faktor paling besar mempengaruhi *shrinkage* dan *Injection Pressure* menjadi faktor paling lemah dalam mempengaruhi *shrinkage*. Kombinasi parameter paling optimum dapat dipilih dengan dengan level yang memiliki *signal to noise ratio* paling besar yaitu kombinasi *Injection Speed* 90 rpm (%), *Melting Temperature* 240◦C, *Injection Pressure* 110bar, *Holding Pressure* 80bar, *Holding Time* 5sec, dan *Cooling Time* 10sec. (*smaller the better)*

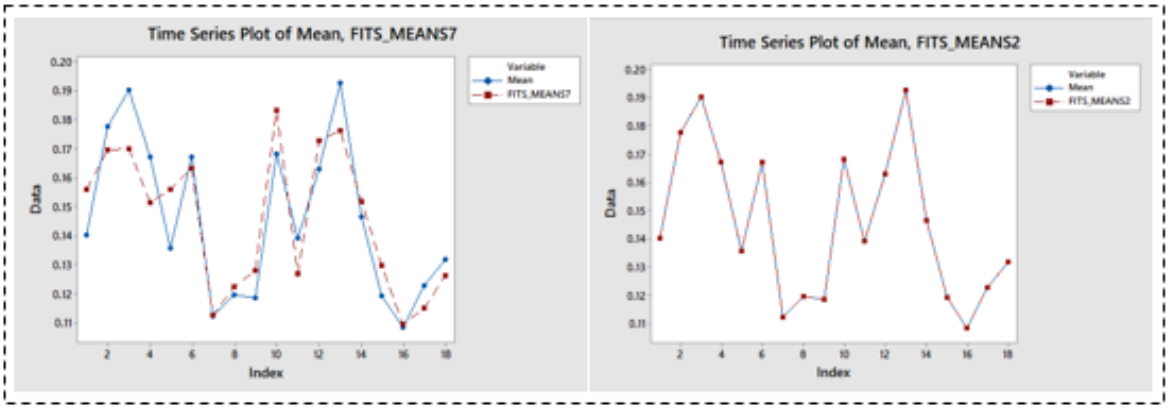
Table 5 Tabel Respon untuk *Signal to Noise Ratios*

Level	**A**	**B**	**C**	**D**	**E**	**F**
1	16.44	15.75	16.48	17.24	16.94	16.24
2	16.74	16.1	16.74	16.29	16.72	16.26
3		17.92	16.55	16.24	16.11	17.26
Delta	0.31	2.17	0.27	0.99	0.82	1.02
Rank	5	1	6	3	4	2

Table 6. Tabel Respon untuk *Mean*

Level	A	B	C	D	E	F
1	0.148	0.163	0.148	0.134	0.139	0.151
2	0.143	0.155	0.140	0.152	0.142	0.152
3		0.119	0.148	0.149	0.155	0.134
Delta	0.004	0.044	0.008	0.018	0.015	0.017
Rank	6	1	5	2	4	3

Untuk melihat apakah interaksi antar faktor AB dan faktor BC benar-benar mempengaruhi dapat juga dilihat dengan membuat plot berdasarkan data variabel respon hasil pengamatan dengan nilai fit yang dapat dilihat pada Gambar 2. Tampak desain sebelum menyertakan interaksi masih terdapat titik nilai fit yang yang seharusnya sama tapi berada dibawah ataupun diatas titik hasil pengamatan. Namun, setelah menggunakan interaksi data menjadi lebih baik berada pada titik yang sama antara hasil pengamatan dengan nilai fit. Sehingga interaksi faktor AB (*Injection Speed* dengan *Melting Temperature*) dan BC (*Melting Temperature* dengan *Injection Pressure*) perlu dipertimbangkan dalam analisa selanjutnya yang didukung juga oleh penelitian sebelumnya [14].



Gambar 2. Perbandingan Plot *Mean* dengan *Fit* sebelum dan setelah terdapat interkasi

Berikut adalah hipotesis pada penelitian ini:
H_0 : Tidak ada pengaruh faktor terhadap *Shrinkage*
H_1 : Terdapat pengaruh semen terhadap *Shrinkage*

Tabel 7. Tabel Respon untuk ANOVA tanpa interaksi

Sumber	df	Seq SS	Kontribusi	Adj SS	Adj MS	F-value	P-value
A	1	0.419	1.30%	0.419	0.419	0.47	0.518
B	2	16.261	50.39%	16.261	8.131	9.14	0.015
C	2	0.231	0.72%	0.231	0.116	0.13	0.880
D	2	3.751	11.62%	3.751	1.875	2.11	0.202
E	2	2.189	6.78%	2.189	1.095	1.23	0.356
F	2	4.087	12.66%	4.087	2.043	2.30	0.182

Error	6	5.335	16.53%	5.335	0.889
Total	17	32.274	100.00%		

Tabel 8. Nilai R-Square

S	R-Sq	R-Sq (adj)	PRESS	R-sq (pred)
0.9429	83.47%	53.16%	48.0166	0.00%

Dari hasil output anova tanpa interaksi diatas diapat didapat bahwa nilai kontribusi pada faktor B terbesar bernilai 50,39% artinya variabel faktor B dapat berpengauh terhadap *Shrinkage* sebesar 50,39% dan sisanya dijelaskan dari selain faktor B.

Nilai *R-Squared* adalah sebesar 83,47%, yang mengandung arti bahwa faktor A-F diatas beserta interaksi secara bersama-sama berpengaruh 83,47% terhadap shrinkage, sisanya 16,53% dijelaskan oleh faktor lain diluar penelitian.

Tabel 11. Tabel Respon ANOVA dengan interaksi

Sumber	Df	Seq SS	Kontribusi	Adj SS	Adj MS	F-value	P-value
A	1	0.254	0.58%	0.254	0.254	*	*
B	2	24.835	56.65%	24.835	12.4235	*	*
C	2	0.624	1.42%	0.624	0.312	*	*
D	2	3.432	7.83%	3.084	1.542	*	*
E	2	2.268	5.17%	2.474	1.237	*	*
F	2	4.874	11.12%	1.488	0.744	*	*
A*B	2	0.648	1.48%	0.648	0.324	*	*
B*C	4	6.904	15.75%	6.904	1.726	*	*
Error	0	*	*	*	*		
Total	17	43.839	100.00%				

Tabel 9. Tabel Respon ANOVA dengan Interaksi dan *Pooled*

Sumber	df	Seq SS	Kontribusi	Adj SS	Adj MS	F-value	P-value
B	2	24.835	56.65%	24.835	12.4173	41.29	0.007
C	2	0.624	1.42%	0.624	0.312	1.04	0.455
D	2	3.432	7.83%	3.084	1.542	5.13	0.108
E	2	2.268	5.17%	2.475	1.237	4.11	0.138
F	2	4.874	11.12%	1.488	0.744	2.47	0.232
B*C	4	6.904	15.75%	6.904	1.726	5.74	0.091
Error	3	0.902	2.06%	0.902	0.301		
Total	17	43.839	100.00%				

Tabel 10. Nilai R-Square dengan Interaksi

S	R-Sq	R-Sq (adj)	PRESS	R-sq (pred)

0.548412	97.94%	88.34%	32.4816	25.91%

Dari hasil output anova dengan interaksi diatas didapat bahwa *p-value* dan *F-value* tidak muncul dikarenakan ada salah satu faktor yang memiliki *contribution* terlalu kecil, sehingga faktor A di pooled atau dihilangkan. Setelah dihilangkan faktor A didapat bahwa nilai kontribusi pada faktor B (suhu leleh) terbesar bernilai 56,65% artinya variabel faktor B dapat berpengaruh terhadap *Shrinkage* sebesar 56,65% dan sisanya dijelaskan dari selain faktor B. Nilai *R-Squared* adalah sebesar 97,94%, yang mengandung arti bahwa faktor B-F diatas beserta interaksi secara bersama-sama berpengaruh 97,94% terhadap *shrinkage*, sisanya 2,06% dijelaskan oleh faktor lain diluar penelitian. Hasil ini berbeda dengan penelitian aslinya karena tidak ada faktor yang dihilangkan, dimana faktor waktu penahanan memiliki kontribusi terbesar [14].

KESIMPULAN DAN SARAN

Dari hasil analisa terhadap injeksi molding untuk *tray* plastik campuran 75% polypropylene (PP) dan 25% polietilen densitas rendah (LDPE) menggunakan pendekatan Taguchi, dapat disimpulkan berikut ini:

Parameter optimal tanpa efek interaksi:

- Kondisi optimum kecepatan injeksi (90 %rpm), suhu leleh (240 °C), tekanan injeksi (110 bar), tekanan penahanan (96 bar), waktu penahanan (5 detik), dan waktu pendinginan (10 detik).
- Suhu leleh adalah parameter yang paling signifikan sedangkan faktor yang lain adalah parameter yang tidak signifikan.
- Kontribusi parameter yaitu suhu leleh (50,39%), waktu pendinginan (12,66%), kekuatan tekanan (11,62%), dan waktu penahanan (6,78%).

Parameter optimal dengan efek interaksi

- Kondisi optimum adalah suhu leleh (240 ° C), tekanan injeksi (110 bar), tekanan tahan (80 bar), waktu penahanan (5 detik), dan waktu pendinginan (10 detik).
- Suhu leleh adalah parameter yang paling signifikan sedangkan faktor yang lain adalah parameter yang tidak signifikan.
- Kontribusi parameter adalah suhu leleh (56,65%),interaksi suhu leleh dan tekanan injeksi (15,75%) , waktu pendinginan (11,12%), dan kekuatan tekanan (7,83%).

Saran untuk penelitian pengembangan berikutnya, jika pada proses produksi plastik dengan metode injeksi molding terdapat bahan yang menarik dan faktor baru yang tentunya diperkirakan berpengaruh itu jauh lebih baik, untuk perbandingan dengan faktor lama dan bahan yang lama. Penelitian ini sebaiknya memperhatikan kembali untuk menggunakan pengulangan yang lebih banyak agar mendapatkan hasil yang lebih optimal, selain itu perlu diperhatikan juga faktor *noise* yang mungkin mempengaruhi seperti suhu dan kelembaban ruangan.

DAFTAR PUSTAKA

[1] G. Singh and A. Verma, “A Brief Review on injection moulding manufacturing process,” *Mater. Today Proc.*, vol. 4, no. 2, pp. 1423–1433, 2017.

[2] C. Budiyantoro and H. Sosiati, “Optimalisasi Parameter Proses Injeksi Pada Hdpe Recycle

Material Untuk Memperoleh Minimum Sink Marks Menggunakan Pendekatan Metode Taguchi," *JMPM (Jurnal Mater. dan Proses Manufaktur)*, vol. 1, no. 2, pp. 56–62, 2017.

[3] X.-P. Dang, "General frameworks for optimization of plastic injection molding process parameters," *Simul. Model. Pract. Theory*, vol. 41, pp. 15–27, 2014.

[4] R. Pareek and J. Bhamniya, "Optimization of injection moulding process using Taguchi and ANOVA," *Int. J. Sci. Eng. Res.*, vol. 4, no. 1, pp. 1–6, 2013.

[5] N. C. Fei, N. M. Mehat, and S. Kamaruddin, "Practical applications of Taguchi method for optimization of processing parameters for plastic injection moulding: a retrospective review," *Int. Sch. Res. Not.*, vol. 2013, 2013.

[6] Y. Wang, J. Kim, and J. Song, "Optimization of plastic injection molding process parameters for manufacturing a brake booster valve body," *Mater. Des.*, vol. 56, pp. 313–317, 2014.

[7] O. A. Mohamed, S. H. Masood, A. Saifullah, and J. L. Bhowmik, "Investigation on warpage and sink mark for injection moulded parts using Taguchi method," *SPE ANTEC Indianap.*, pp. 1723–1728, 2016.

[8] M. K. Damayanti, "Desain Parameter Eksperimen Untuk Optimasi Nilai Frangibility Factor Material Komposit Dengan Metode Taguchi dan Neural Network," *Inst. Teknol. Sepuluh Nop.*, 2017.

[9] P. Sidi and M. T. Wahyudi, "Aplikasi metoda taguchi untuk mengetahui optimasi kebulatan pada proses bubut CNC," *J. Rekayasa Mesin*, vol. 4, no. 2, pp. 101–108, 2013.

[10] A. A. Wulandari, T. Wuryandari, and D. Ispriyanti, "Penerapan Metode Taguchi Untuk Kasus Multirespon Menggunakan Pendekatan Grey Relational Analysis Dan Principal Component Analysis (Studi Kasus Proses Freis Komposit Gfrp)," *J. Gaussian*, vol. 5, no. 4, pp. 791–800, 2016.

[11] D. Rahmalina, E. Prayogi, A. S. Atmaja, S. Sudiro, A. Suhadi, and I. C. Setiawan, "Analisis Pengaruh Tekanan Injeksi pada Proses Injection Molding terhadap Kekerasan Komposit Polyurethane-15% Carbon Black," in *Prosiding Seminar Rekayasa Teknologi (SemResTek)*, 2018, pp. 711–715.

[12] I. O. Fagbolagun and S. A. Oke, "The optimization of packaging system process parameters using Taguchi method," *Int. J. Ind. Eng. Eng. Manag.*, vol. 2, no. 1, pp. 1–14, 2020.

[13] R. Risnawati, "Mengidentifikasi Faktor-Faktor yang mempengaruhi Kualitas Gula dengan Metode Taguchi pada Pabrik Gula." Universitas Islam Negeri Alauddin Makassar, 2018.

[14] S. Kamaruddin, Z. A. Khan, and S. H. Foong, "Application of Taguchi method in the optimization of injection moulding parameters for manufacturing products from plastic blend," *Int. J. Eng. Technol.*, vol. 2, no. 6, p. 574, 2010.

[15] P. R. Maulidia, N. Budiharti, and E. Adriantantri, "Analisis Pengendalian Kualitas Menggunakan Metode Taguchi pada UMKM Rubber Seal RM Products Genuine Parts Sukun, Malang," *Ind. Inov. Tek. Ind. ITN Malang,* pp. 83–91, 2020.

[16] R. H. Widyatmoko, "Optimalisasi Parameter Injeksi untuk Minimasi Shrinkage, Sink Marks dan Warpage pada Industri Mold Modern." UAJY, 2017.

Comparison of Support Vector Machine (SVM) and Autoregressive Integrated Moving Average (ARIMA) Methods for Predicting Air Quality Using Python and KNIME

Tiara Melati Putri Wiryawanto [(1)], Zuyyina Hawani [(2),] Muhammad Attar Ramadhani [(3)]

[1,2,3] Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, UIN Sunan Ampel Surabaya

Jalan Ir. H. Soekarno No. 682, Gunung Anyar, Surabaya

e-mail: tiarawir@gmail.com[(1)], zuyyinahawani@gmail.com[(2)], attarramadhani@gmail.com[(3)]

ABSTRAK

Udara merupakan komponen terpenting bagi makhluk hidup di bumi. Namun, perubahan-perubahan yang ada di bumi menyebabkan permasalahan yang salah satunya pencemaran udara. Aktivitas manusia menjadi salah satu penyebab pencemaran udara. Hal inilah yang menjadikan kualitas udara masa depan penting untuk diprediksi. Untuk melakukan prediksi kualitas udara, penelitian ini menggunakan metode Support Vector Machine (SVM) dan Autoregressive Integrated Moving Average (ARIMA). Untuk SVM sendiri mempresentasikan salah satu metode dari teknik machine learning. Sedangkan ARIMA mempresentasikan salah satu metode dari model statistik. Dengan menggunakan data dari situs Open Data Jakarta mengenai pengukuran Indeks Standar Pencemaran udara (ISPU) pada lima stasiun pemantau kualitas udara (SPKU) di Provinsi DKI Jakarta tahun 2021, kemudian dilakukan analisis untuk membandingkan performa dan keakuratan dua metode ini dalam memprediksi kualitas udara. Hasil dari penelitian ini menunjukkan bahwa antara pengujian ARIMA dan SVM, dapat dikatakan pengujian SVM memiliki hasil akurasi yang lebih tinggi. Hal ini dapat dilihat dari rata-rata hasil akurasi dengan beberapa perlakuan.
Kata kunci: Kualitas Udara, Support Vector Machine, ARIMA

ABSTRACT

Air is the most important component for living things on earth. However, the changes that exist on earth cause problems, one of which is air pollution. Human activity is one of the causes of air pollution. This is what makes future air quality feasible to predict. To predict air quality, this research use the Support Vector Machine (SVM) and Autoregressive Integrated Moving Average (ARIMA) methods. For SVM itself, it presents one of the methods of machine learning techniques. Meanwhile, ARIMA presents one of the methods of the statistical model. Using data from the Open Data Jakarta website regarding measurements of the Air Pollution Standard Index (ISPU) at five air quality monitoring stations (SPKU) in DKI Jakarta Province in 2021, an analysis was then carried out to compare the performance and accuracy of these two methods in predicting air quality. The results of this study indicate that between ARIMA and SVM testing, it can be said that SVM testing has higher accuracy results. This can be seen from the average accuracy results with several treatments.
Keywords: *Air Quality, Support Vector Machine, ARIMA*

INTRODUCTION

Over the last few decades, most of humanity transformed into city dwellers. However, this rapid change towards urbanization gave rise to several problems, including air pollution [1]. Air is one of the many components that are important in everyday life for living things on earth. Air is a mixture of gases, which is composed of 78 percent Nitrogen, 20 percent Oxygen, 0.3 percent Carbon Dioxide (CO2) and the remainder consists of Neon (Ne), Helium (He), Methane (CH4) and Hydrogen (H2) [2].

Air can be said to be polluted or polluted when it gets additional other gases that can cause disturbances and changes in composition. Air pollution results from a combination of high emissions and unfavorable weather [1]. Air pollution itself consists of a mixture of particulate matter (PM2.5 and PM10) and gas species (NO2, CO, O3, and SO2), which have acute and chronic effects on human health, especially among adolescents and the elderly. [3]. Emissions of air pollutants and their precursors determine regional air quality and can change climate [4]. Entering an era of rapidly changing climate, the impact on air quality needs to be better understood and reviewed. This is good for air quality management purposes and as one of the social impacts of climate change.

Human activity is one of the causes of air pollution. Air pollution is a process of entering substances, energy, and other components into the air which causes air quality to decrease and function improperly. [2]. This is why, in addition to monitoring, there is a demand to predict air quality in the future. One of the aims is to be able to inform the government or authorities in policy making, such as conducting traffic control when the air is heavily polluted.

One of the tools for prediction or classification is to use the support vector machine (SVM) method. SVM is one of the classification methods with a fairly high degree of accuracy in making predictions on air quality assessments. SVM itself has several advantages in classifying a pattern accurately even though the dataset has limitations. On the other hand, SVM has limitations when the number of attributes used tends to be large which results in heavy computational loads and reduced accuracy [5].

Several studies related to air quality prediction have been carried out in the last decade. Lu et al., in 2002 predicted air pollutant parameters by comparing the SVM and the classical radial basis function (RBF) network [6]. The results of this study indicate that SVM is superior to conventional RBF in predicting air quality parameters with different time series. Then, research related to air pollution prediction using the SVM method was developed by Dun et al., in 2020 which predicted short-term air quality using a hybrid model, which combines the fractional gray linear regression and SVM methods [7]. The results of this study indicate that the hybrid model is used to predict three air pollutants (PM10, PM2.5, and NO2) in Shijiazhuang and Chongqing, and it appears that the prediction accuracy of the hybrid model is significantly higher than that of the single model.

Besides SVM, a method that can be used to make predictions is ARIMA. This method has several stages including identification, estimation, diagnostic check and finally forecasting. The models in ARIMA itself are divided into three, namely the Autoregressive Model, the Moving Average, and the Mixed Model which have the characteristics of the two previous models. However, before carrying out these stages the data must be ascertained to be stationary first. That

is, the data does not increase or decrease. This is because the requirements for the formation of the ARIMA model are stationary data.

Research on air pollution prediction using the ARIMA method shows that the ARIMA model is suitable for short-term predictions if the data is stationary [8]. Subsequent research related to air quality prediction by comparing NO2 and SO2 using the ARIMA method [9]. The results of this study indicate that the ARIMA model is the most effective model for evaluating results through analysis and prediction. With this method it basically helps environmental technicians to study and analyze air pollution levels, and therefore helps the government to take preventive actions.

From these previous studies, in this study a comparison will be made of machine learning techniques using the support vector machine (SVM) method and statistical models using the autoregressive integrated moving average (ARIMA) method. These two methods are used to compare which method is more suitable for use in air quality datasets. Currently, there is a lot of information available on the internet about air quality in several places. In this study, a sample dataset was taken from the Open Data Jakarta website regarding air quality information for DKI Jakarta in 2021. From this dataset, the performance and accuracy of these two methods in predicting air quality will be compared.

METHOD

The research uses two methods, namely SVM and ARIMA. The SVM method itself represents one of the methods of machine learning techniques. Meanwhile ARIMA represents one of the methods of the statistical model. Actually, there are many other methods for making predictions both from statistical models and machine learning techniques. However, these two methods are considered to have significant advantages when used based on several previous studies.

In comparing the two methods, namely SVM and ARIMA to predict air quality, a flow design is needed to provide a reference or direction in conducting this research. The flow of this research consists of six steps that are done sequentially. The following in Figure 1 is a research flow design.

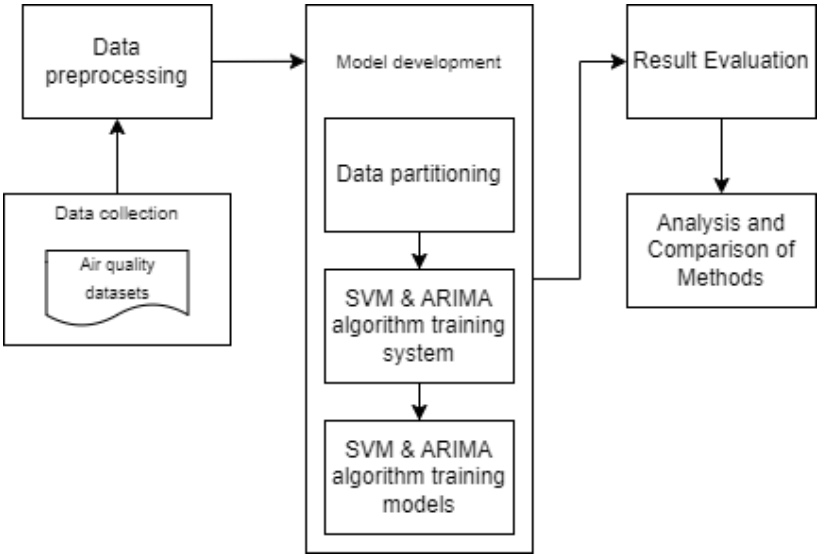


Figure 1. Research flow design

A. Data Collection

The data source used in this study is secondary data obtained from the Open Data Jakarta website regarding measurements of the Air Pollution Standard Index (ISPU) at five air quality monitoring stations (SPKU) in DKI Jakarta Province in 2021. From this dataset, one variable is then defined response or dependent and seven predictor or independent variables as shown in Table 1 below.

Table 1. Variabel dan indikator penelitian

Attribute	Indicator Variables
Y	Category
X1	PM_{10}
X2	$PM_{2.5}$
X3	SO_2
X4	CO
X5	O_3
X6	NO_2
X7	Max

B. Data Preprocessing

At this stage, the data that has been obtained before will be processed in stages which consist of several steps, including:

1. Data cleaning is performed on data that is null and irrelevant to the purpose of analysis.
2. Data transformation is performed, which is used when data needs to be converted to suit the purpose of analysis.
3. Normalization is carried out so that the data used in the two methods does not have large deviations. Analysis using Python with StandardScaler library.

C. Model Development

At the development stage of this model consists of several sub-steps. First, after the data has been processed and modeled, data partitioning is performed. This stage divides the data by dividing the data for training and testing in four experiments, namely with a ratio of 80:20, 70:30 and 60:40. Then from the training data, the algorithms of the two methods are implemented, namely SVM and ARIMA. Then, testing will be carried out, namely predicting the data.

D. Result Evaluation

The results of the data that have been predicted will be tested by model testing and assessing forecasting performance with indicators of percentage accuracy, RMSE and MAE.

1. RMSE

   Root Mean Square Error (RMSE) is a method for measuring the bias or difference in the prediction value of the model.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{I=1}^{n}(Yi - \underline{Yi})^2}{n}} \qquad (1)$$

2. MAE

Mean Absolute Error (MAE) is the average value of the absolute difference between the actual (actual) value and the predicted (forecasting) value.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | \text{Ai} - \text{Fi} | \quad (2)$$

E. Analysis and Comparison of Methods

From the evaluation results of RMSE and MAE for each method using several kinds of partitioning methods, then analysis and comparison are carried out to see the advantages and disadvantages of each method.

F. Support Vector Machine (SVM)

In machine learning, SVM is a supervised learning model that is usually used for pattern recognition, classification, and regression analysis [10]. SVM works based on the Structural Risk Minimization (SRM) principle to find the best hyperplane that separates the two classes in the input space. The performance and effectiveness of SVM itself is greatly affected by the type of kernel functions implemented. The use of the kernel aims to transform data into a high-dimensional space, namely by making non-linear data separate linearly [11]. There are several types of kernel functions that can be implemented based on data criteria. This study uses the two most popular and frequently used kernel functions, namely RBF (Radial Basis Function) and Polynomial.

a. RBF (Radial Basis Function)

Gaussian radial basis function (RBF) is one of the most widely used kernel functions. RBF itself is usually used when the data is not linearly separated. The following is the kernel's RBF equivalent.

$$K(x,z) = exp\left[-\gamma||x-z||^2\right] \quad (3)$$

b. Polynomial

The polynomial kernel function is a kernel function that is used when the data is not linearly separated. This kernel function is appropriate when the training dataset has been normalized. The following is the kernel polynomial equation.

$$K(x,z) = (x^T z)^d \; atau \; (1+x^T z)^d \quad (4)$$

With x^T is the training data, z is the testing data, and d is the degree of the polynomial

G. Autoregressive Integrated Moving Average (ARIMA)

The ARIMA model dominates in the field of time series forecasting. ARIMA stands for AutoRegressive (AR), Integrated (I), Moving Average (MA). Each of these phrases describes a different part of the mathematical model. Here are the equations ARIMA$(p,d,q)(P,D,Q)$ where (p,d,q) is the non-seasonal part of the model and (P,D,Q) adalah bagian *seasonal* dari model is the seasonal part of the model.

$$\phi_p(B)\Phi_P(B^x)\nabla^d\nabla_s^D Y_t = \theta(B)_Q(B^s)_{e_t} \quad (5)$$

RESULT AND DISCUSSION

The dataset used in this study was published in June, 2022. This dataset consists of 365 data with several attributes related to components in the air. Some of the attributes in the dataset are shown in Table 2 below.

Table 2. Attributes used in research

No	Attribute	Description	Data Type
1	Date	Air quality measurement date	datetime
2	PM_{10}	Particulate matter	integer
3	$PM_{2.5}$	Particulate matter	integer
4	SO_2	Sulfide (in SO_2 form)	integer
5	CO	Carbon monoxide	integer
6	O_3	Ozone	integer
7	NO_2	Nitrogen dioxide	integer
8	Max	The maximum measured value of all parameters measured simultaneously	integer
9	Critical	The maximum measured value of all parameters measured simultaneously	string / object
10	Category	The air quality category is based on the calculation of the air pollution standard index	string / object

From the dataset, data cleaning is then carried out by removing the Critical column because it is irrelevant to the predicted results and the data type does not match the method to be used. After cleaning the data, data transformation is then performed to change the Category data type to integer. This is done so that ARIMA reads the response variable and can make predictions about the Category.

At the last data preprocessing stage, namely normalization. Python uses the StandardScaler library and KNIME uses the Normalizer library. After passing through the data preprocessing stage, data partitioning is carried out using 3 scenarios, namely with a ratio of 80:20, 70:30 and 60:40. From the data partition, then each method is implemented with two tools, namely Python and KNIME. Figure 2 illustrates the flow of SVM and ARIMA modeling using KNIME. The nodes used in this workflow will also be explained in Table 3.

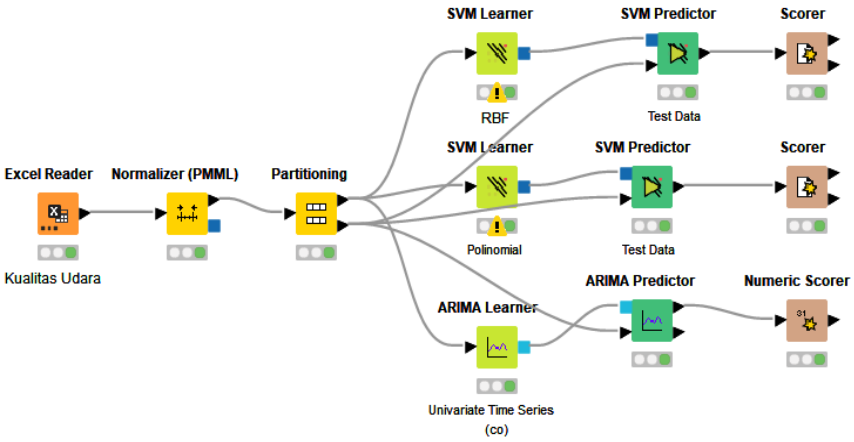


Figure 2. Workflow of SVM and ARIMA using KNIME

Figure 2 above shows the workflow in SVM and ARIMA modeling using KNIME. From the several nodes used [12], a description of each node will be explained in Table 3 below.

Table 3. Nodes used in SVM

No	Node	Description
1	Excel Reader	Reading excel files (xlsx, xlsm, xlsb, and xls formats)
2	Normalizer (PMML)	Normalizes all numeric columns
3	Partitioning	Split data by row to train and test data
4	SVM Learner	Forward partitioned input data using the kernel and parameters in the form of HyperTangent, Polynomial, and RBF
5	ARIMA Learner	Estimating time series parameters on the ARIMA model
6	ARIMA Predictor	Calculating predictions from estimated ARIMA models (forecast and In-sample production)
7	SVM Predictor	Predicting the output value issued by the SVM learner
8	Numeric Scorer	Calculates certain statistics from the predicted results
9	Scorer	Shows the results of the comparison of attribute columns and matrices

A. Support Machine Vector (SVM) Experiment Results

Testing the method to be carried out in this study is based on the distribution of training data and testing using the dataset division procedure. Figure 3 shows the results of data classification based on air quality after training data.

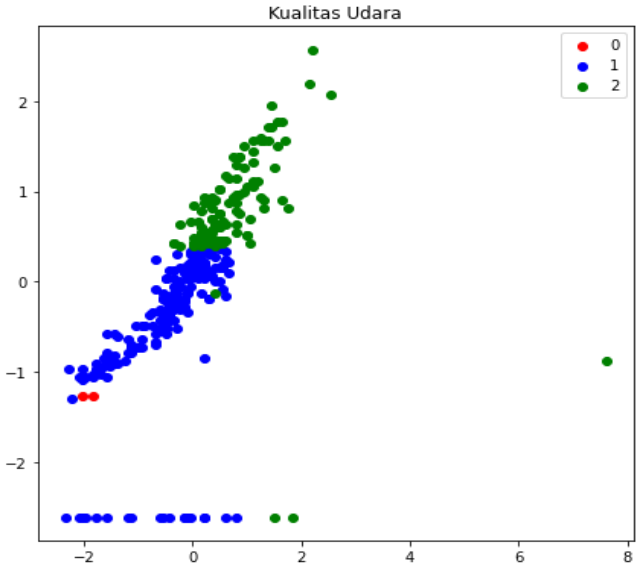


Figure 3. Classification results of dataset training

Figure 4 below shows the results of data classification based on air quality after testing data.

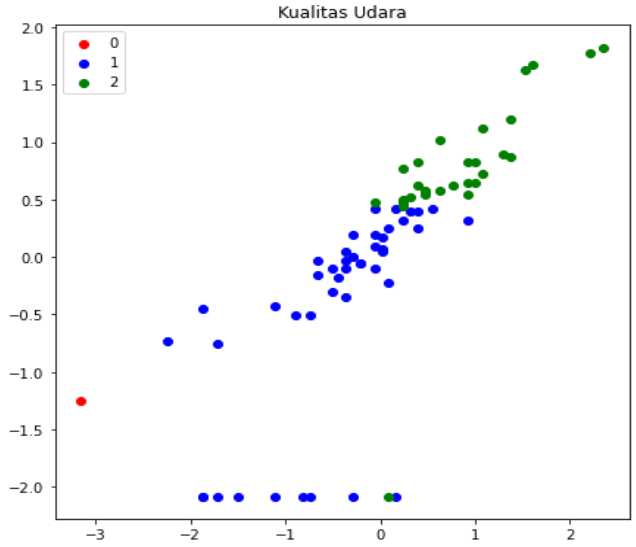


Figure 4. Classification results of dataset testing

At the partitioning stage (training and data testing) two treatments were given in the partitioning method, namely by using random sampling and stratified sampling. Then, the next step is to create an SVM classifier with two treatments in the kernel function, namely polynomial and

RBF. Table 4 below shows the results of the accuracy of the SVM method with some of these treatments.

Table 4. The level of accuracy in the SVM method using error analysis

Partition Method	Ratio	Polynomial		RBF	
		MAE	RMSE	MAE	RMSE
Random Sampling	80:20	0.19	0.44	0.09	0.31
	70:30	0.18	0.43	0.11	0.33
	60:40	0.14	0.37	0.09	0.29
Stratified Sampling	80:20	0.23	0.51	0.04	0.20
	70:30	0.21	0.49	0.05	0.23
	60:40	0.19	0.46	0.06	0.25

B. Autoregressive Integrated Moving Average (ARIMA) Experiment Results

Some of the attributes or variables in the dataset used in the study are not stationary, namely Max and Category. Then, differentiation is performed once on data that is still not stationary. Comparison of the state of the dataset before and after differentiation is shown in Figures 5 and 6 below.

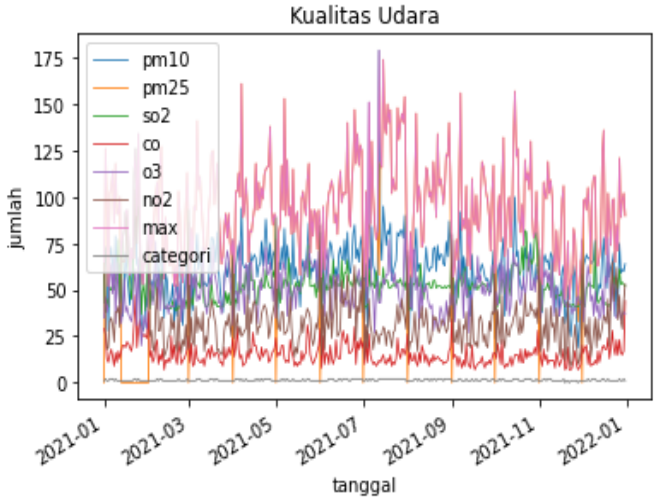


Figure 5. Dataset before differentiation (non-stationary)

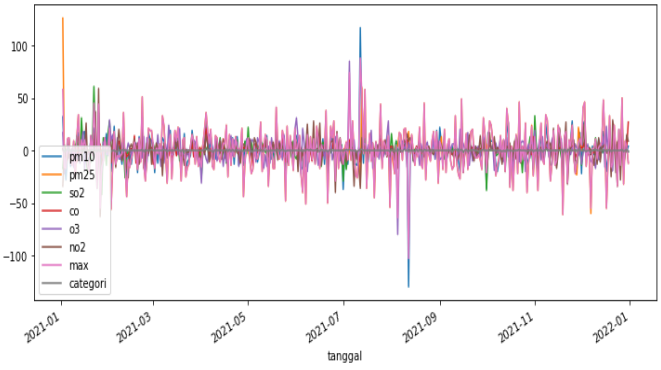


Figure 6. Dataset after differentiation (stationary)

After the data is stationary, ARIMA modeling is then carried out with parameters (1,1,1). The results of this parameter are obtained by using the auto_arima function in python. From the results of this modeling, forecasting is then carried out and the accuracy results are shown in Table 5.

Tabel 5. The level of accuracy in the ARIMA method using error analysis

Partition Method	Ratio	MAE	RMSE
Random Sampling	80:20	0.51	0.55
	70:30	0.58	0.61
	60:40	0.57	0.62

CONCLUSION

From the experiments result, several conclusions can be drawn:

1. In testing the dataset using SVM using random sampling and stratified sampling partitions produce different levels of accuracy as described above, that the random sampling partition with a ratio of 60:40 has the best accuracy, namely with RBF producing MAE 0.09 and RMSE 0.29 and with polynomial produces MAE 0.14 and RMSE.
2. In the SVM test, treatment using RBF produces a higher level of accuracy. This is because usually linear and polynomial kernels take less time and provide lower accuracy than rbf or Gaussian kernels.
3. In testing the dataset using ARIMA before differentiating several variables such as Max and Category it produces non-stationary data, but after differentiation with parameters (1,1,1) the dataset becomes stationary. The level of accuracy obtained in the use of random sampling partitions obtained a ratio of 80:20 which is the best, namely MAE of 0.51 and RMSE of 0.55.
4. From ARIMA and SVM testing, it can be said that SVM testing has higher accuracy results. This can be seen from the average accuracy results with several treatments.

From the testing, it is possible to do research again to test the dataset on ARIMA using a stratified sampling partition in the next test, so that the results of this study can be perfected.

REFERENCE

[1] S. Chapman, J. E. M. Watson, A. Salazar, M. Thatcher, and C. A. McAlpine, "The impact of urbanization and climate change on urban temperatures: a systematic review," *Landsc Ecol*, vol. 32, no. 10, pp. 1921–1935, Oct. 2017, doi: 10.1007/s10980-017-0561-4.

[2] L. Manisalidis, E. Stavropoulou, A. Stavropoulos, and E. Bezirtzoglou, "Environmental and Health Impacts of Air Pollution: A Review," *Front Public Health*, vol. 8, Feb. 2020, doi: 10.3389/fpubh.2020.00014.

[3] X. Yi, J. Zhang, Z. Wang, T. Li, and Y. Zheng, "Deep Distributed Fusion Network for Air Quality Prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA: ACM, Jul. 2018, pp. 965–973. doi: 10.1145/3219819.3219822.

[4] A. M. Fiore *et al.*, "Global air quality and climate," *Chem Soc Rev*, vol. 41, no. 19, p. 6663, 2012, doi: 10.1039/c2cs35095e.

[5] T. B. Sasongko, "Komparasi dan Analisis Kinerja Model Algoritma SVM dan PSO-SVM (Studi Kasus Klasifikasi Jalur Minat SMA)," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 2, no. 2, Aug. 2016, doi: 10.28932/jutisi.v2i2.476.

[6] W. Lu *et al.*, "Air pollutant parameter forecasting using support vector machines," in *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)*, IEEE, 2002, pp. 630–635. doi: 10.1109/IJCNN.2002.1005545.

[7] M. Dun, Z. Xu, Y. Chen, and L. Wu, "Short-Term Air Quality Prediction Based on Fractional Grey Linear Regression and Support Vector Machine," *Math Probl Eng*, vol. 2020, pp. 1–13, May 2020, doi: 10.1155/2020/8914501.

[8] A. Kaya, R. Ozturk, and C. Altin Gumussoy, "Usability Measurement of Mobile Applications with System Usability Scale (SUS)," 2019, pp. 389–400. doi: 10.1007/978-3-030-03317-0_32.

[9] Gourav, J. K. Rekhi, P. Nagrath, and R. Jain, "Forecasting Air Quality of Delhi Using ARIMA Model," 2020, pp. 315–325. doi: 10.1007/978-981-15-0372-6_25.

[10] Y. Zhang, H. Yang, H. Cui, and Q. Chen, "Comparison of the Ability of ARIMA, WNN and SVM Models for Drought Forecasting in the Sanjiang Plain, China," *Natural Resources Research*, vol. 29, no. 2, pp. 1447–1464, Apr. 2020, doi: 10.1007/s11053-019-09512-6.

[11] M. Awad and R. Khanna, *Efficient Learning Machines*. Berkeley, CA: Apress, 2015. doi: 10.1007/978-1-4302-5990-9.

[12] KNIME Community Hub, "KNIME Base nodes," *KNIME*. https://hub.knime.com/knime/extensions/org.knime.features.base/latest (accessed Jan. 01, 2023).

Klasifikasi Pengeluaran per Kapita di Tiga Provinsi Sulawesi menggunakan K-Nearest Neighbor

Ismi Rizqa Lina [(1)], Dia Cahya Wati [(2)]

[12]Prodi Sains Data, Universitas Insan Cita Indonesia

Graha Binakarsa, 10th floor, Jl. H.R Rasuna Said, Kota Jakarta Selatan, DKI Jakarta

e-mail: irizqalina@gmail.com[(1)], diacahyawati@gmail.com[(2)]

ABSTRAK

Klasifikasi pengeluaran per kapita merupakan analisis pasar yang penting bagi banyak perusahaan untuk menentukan Kabupaten/kota mana yang paling cocok untuk menjual suatu produk di perusahaan tersebut. KNN dapat digunakan untuk berbagai jenis data, termasuk data ekonomi seperti pengeluaran per kapita. Pada penelitian ini, 56 data pengeluaran per kapita di provinsi Sulawesi Selatan, Sulawesi Utara, dan Sulawesi Tenggara pada tahun 2022 diklasifikasi dengan algoritma KNN. Proses klasifikasi menggunakan algoritma KNN diawali dengan melakukan *pre-procecing* data dan menghitung jarak antara data pelatihan (data *training)* dengan data uji (data *testing*). Dalam perhitungan jarak, digunakan metrik *Euclidean* dan metrik *Manhattan*. Selanjutnya, dilakukan perhitungan nilai prediksi berdasarkan k data *training* terdekat. Hasil pengujian pada penelitian ini menunjukkan akurasi tertinggi pada $k = 9$ untuk jarak *Eucledian* sebesar 76,47% yang berarti klasifikasi cukup dan untuk jarak *Manhattan* sebesar 94,12% yang berarti klasifikasi sangat baik. Dari hasil akurasi kedua jarak tersebut dapat disimpulkan bahwa jarak *Manhattan* lebih baik daripada jarak *Eucledian*.

Kata kunci: K-Nearest Neighbor; Euclidean; Manhattan; Pengeluaran per Kapita; Sulawesi

ABSTRACT

Classification of expenditure per capita is an important market analysis for many companies to determine which district/city is most suitable for selling products in the company. KNN can be used for various types of data, including economic data such as expenditure per capita. In this study, 56 per capita expenditure data in the provinces of South Sulawesi, North Sulawesi and Southeast Sulawesi in 2022 were classified using the KNN algorithm. The classification process using the KNN algorithm begins with pre-processing the data and calculating the distance between the training data and the test data. In calculating distances, the Euclidean metric and the Manhattan metric are used. Furthermore, the predicted value is calculated based on the closest k training data. The test results in this study showed the highest accuracy at $k = 9$ for the Eucledian distance of 76.47% which means that the classification is sufficient and for the Manhattan distance it is 94.12% which means the classification is very good. Based on the accuracy results of both distances, the conclusion can be drawn that the Manhattan distance outperforms the Euclidean distance.

Keywords: *K-Nearet Neighbor; Eucledian; Manhattan; Expenditure per Capita; Sulawesi*

PENDAHULUAN

Pengeluaran per kapita adalah salah satu indikator ekonomi yang berguna dalam menilai tingkat produktivitas dan kesejahteraan masyarakat. Perbedaan antara desil terkaya dan termiskin dalam masyarakat tercermin dari distribusi pendapatan dan tingkat ketimpangan pendapatan, yang dipengaruhi oleh faktor-faktor struktural ekonomi dan situasi sosial dalam lingkungan masyarakat [1]. Pernyataan ini juga di dukung oleh penelitan yang berkaitan dengan pertumbuhan populasi, semakin banyak penduduk dalam suatu wilayah akan berdampak pada perubahan jumlah penduduk dan status ekonomi di wilayah tersebut [2].

Pertumbuhan ekonomi Sulawesi Selatan pada triwulan I tahun 2013 mencapai 7,79%, yang merupakan angka yang sedikit lebih rendah daripada pertumbuhan ekonomi pada triwulan I tahun 2012 yang mencapai 7,95%. Namun, pertumbuhan ekonomi Sulawesi Selatan tersebut lebih tinggi dibandingkan dengan pertumbuhan ekonomi secara nasional pada periode yang sama tahun 2013, yang hanya mencapai 0,02% [3]. Provinsi Sulawesi Selatan menjadi wilayah yang paling berkembang dan berperan sebagai pendorong utama pertumbuhan ekonomi, sehingga menyebabkan disparitas antara daerah-daerah yang lebih besar [4]. Dalam konteks indeks Gini, nilai indeks Gini yang berada antara 0 hingga 1 menunjukkan sejauh mana distribusi pendapatan yang merata. Nilai indeks Gini yang sama dengan 0 mengindikasikan distribusi pendapatan yang sangat merata, sedangkan nilai indeks Gini yang sama dengan 1 menunjukkan distribusi pendapatan yang sangat tidak merata [5]. Hal serupa dilakukan oleh penelitian [6] bahwa ketimpangan perekonomian dan pembangunan juga terjadi di Kawasan Timur seperti Sulawesi Selatan, Sulawesi Utara, dan Sulawesi Tenggara.

Penggunaan Algoritma *K-Neares Neighbor* (KNN) sesuai untuk melakukan klasifikasi pengeluaran per kapita di wilayah-wilayah Sulawesi Selatan, Sulawesi Utara, dan Sulawesi Tenggara. KNN merupakan suatu teknik klasifikasi yang digunakan dalam data mining dan termasuk dalam kelompok metode pembelajaran berbasis contoh (*instance-based learning*). Dalam KNN, pencarian dilakukan untuk menemukan k objek terdekat dari data pelatihan yang paling mirip dengan objek pada data uji [7]. Penelitian [8] melakukan prediksi nilai tanah menggunakan algoritma KNN yang menghasilkan tingkat akurasi prediksi data testing sebesar 80%. penggunaan Algoritma KNN dengan metrik jarak Manhattan dapat membantu penetapan dalam pemenangan lelang dengan tingkat akurasi sebesar 0,8 [9].

Berdasarkan paparan tersebut penulis ingin mengklasifikasikan pengeluaran per-kapita di Sulawesi Selatan, Sulawesi Utara, dan Sulawesi Tenggara dengan KNN. Metode klasifikasi K-NN dengan menggunakan metrik jarak Euclidean dan Manhattan telah diteliti oleh [10]. Hasil penelitiannya menunjukkan bahwa dalam klasifikasi transportasi bus, metrik jarak Manhattan lebih akurat dibandingkan dengan metrik jarak Euclidean. Penelitian yang dilakukan oleh [11] menunjukkan hasil yang berbeda. Dalam penelitian tersebut, metrik jarak *Euclidean* dan *Minkowski* pada algoritma KNN pada data dengan representasi vektor dari kalimat menunjukkan akurasi terbaik, yang lebih baik daripada menggunakan metrik jarak *Chebyshev* maupun *Manhattan*.

METODE

Penelitian ini menggunakan data sekunder yang diperoleh dari situs resmi Badan Pusat Statistik (BPS) Sulawesi Selatan (https://sulsel.bps.go.id/), Sulawesi Utara

(https://sulut.bps.go.id/), dan Sulawesi Tenggara (https://sultra.bps.go.id/) untuk tahun 2022. Adapunvariabel-variabel yang menjadi fokus dalam penelitian ini sebagai berikut.

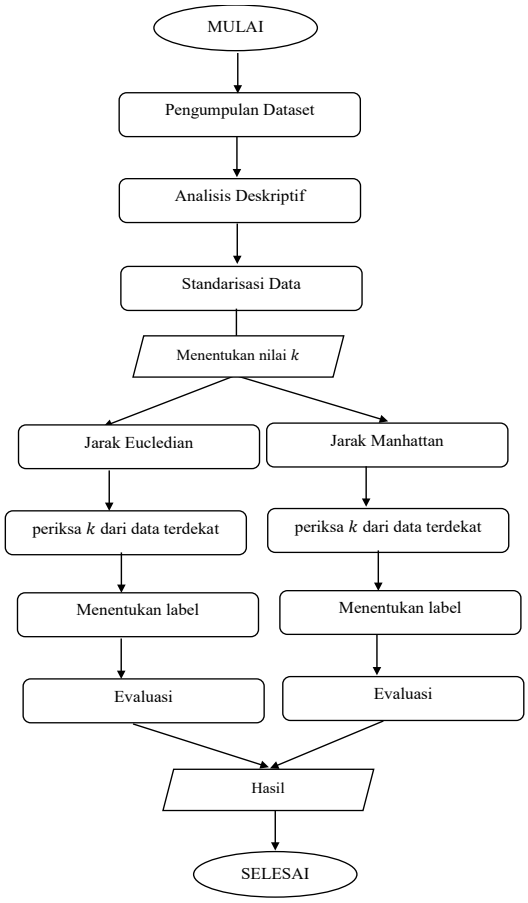
Tabel 1. Struktur data

Variabel	Notasi	Keterangan
Pengeluaran per kapita	y	Kategorik : 1 : Tinggi ($\geq$ 10610) 0 : Rendah (< 10610)
HLS	$x1$	Numerik
IPM	$x2$	Numerik
PDRB	$x3$	Numerik
Mi	$x4$	Numerik
RLS	$x5$	Numerik
UHH	$x6$	Numerik

Keterangan.
HLS : Harapan Lama Sekolah
IPM : Indeks Pembangunan manusia
PDRB : Produk Domestik Regional Bruto
Mi : Presentase penduduk miskin menurut kabupaten/kota
RLS : Rata-rata lama sekolah
UHH : Umur harapan hidup.

Penelitian ini direpresentasikan melalui sebuah flowchart yang menunjukkan rangkaian tahapan pelaksanaan secara visual dan terstruktur yang ditunjukkan pada **Gambar 1.**

Tahapan awal penelitian ini dimulai dengan pengumpulan dataset dari sumber-sumber yang relevan. Selanjutnya, dilakukan analisis deskriptif untuk memahami karakteristik data. Jika satuan data berbeda, standarisasi data digunakan untuk menyamakan skala variabel. Selanjutnya, dataset dibagi menjadi dua bagian, yaitu data training (data pelatihan) dan data testing (data uji), agar model KNN dapat diuji pada data yang tidak digunakan selama pelatihan. Pada tahap klasifikasi, algoritma KNN digunakan untuk melakukan klasifikasi data berdasarkan tetangga terdekat. KNN adalah salah satu algoritma machine learning yang digunakan untuk analisis klasifikasi dan regresi. Prinsip kerjanya sederhana, yaitu melakukan klasifikasi pada objek baru berdasarkan jumlah k tetangga terdekatnya [12]. Dalam ruang dua dimensi, perhitungan jarak antara dua titik digunakan untuk mengidentifikasi tetangga yang jauh atau dekat dari suatu titik tertentu. Salah satunya, yaitu jarak *Euclidean* dan jarak *Manhattan* [13].



Gambar 1. *Flowchart* Algoritma KNN

Jarak *Euclidean* merupakan metode perhitungan jarak yang digunakan untuk mengukur jarak antara dua buah titik dalam ruang *Euclidean*. Metode ini pertama kali diperkenalkan oleh Euclid, seorang matematikawan terkemuka dari Yunani kuno. Secara umum, jarak *Euclidean* didefinisikan sebagai panjang garis lurus yang menghubungkan dua titik tersebut. Jarak *Euclidean* antara dua titik (x_1, y_1) dan (x_2, y_2) dalam ruang dua dimensi dihitung menggunakan rumus berikut [12].

$$d_E(x_1, x_2) = \sqrt{\sum_{i=0}^{n} (x_{1i} - x_{2i})^2 + \left(y_{1i} - y_{2i}\right)^2 + \cdots}, \quad (1)$$

dengan

$d_E(x_1, x_2)$: jarak *Euclidean* dari data *trainig* ke-1 ke data *testing* ke-2
x_1 : koordinat data *training* ke-1
x_2 : koordinat data *testing* ke 2

Rumus ini menghitung panjang garis lurus (jarak *Euclidean*) antara dua titik dalam ruang dua dimensi yang diwakili oleh data training ke-1 (x_1) dan data testing ke-2 (x_2). Hasil perhitungan jarak *Euclidean* ini akan digunakan dalam algoritma KNN untuk menentukan k tetangga terdekat dari data testing (x_2) berdasarkan data training (x_1). Semakin kecil nilai jarak *Euclidean*

$d_E(x_1, x_2)$, semakin dekat data training (x_1) dengan data testing (x_2), dan semakin besar nilai jarak *Euclidean*, semakin jauh keduanya.

Jarak *Manhattan*, juga dikenal sebagai *City Distance*, adalah metode perhitungan jarak yang digunakan dalam algoritma KNN untuk mengidentifikasi kasus yang paling cocok dari basis kasus dengan mengukur jumlah bobot absolut dari perbedaan antara kasus yang sedang diuji dan kasus lain dalam basis kasus. Untuk menghitung bobot (jarak *Manhattan*) antara dua data x dan $y$dalam ruang dua dimensi, digunakan persamaan berikut:

$$d_M(x, y) = \sum_{i=0}^{n} |x_i - y_i|, \quad (2)$$

x dan y adalah dua data dalam ruang dua dimensi, dan n adalah jumlah dimensi dalam ruang tersebut. Perhitungan bobot menggunakan jarak Manhattan ini melibatkan penjumlahan dari selisih absolut antara koordinat data x_i dan y_i dalam masing-masing dimensi. Hasil bobot ini akan digunakan dalam algoritma KNN untuk menentukan k tetangga terdekat dari data testing x_i berdasarkan data training y_i. Semakin kecil nilai bobot (jarak *Manhattan*), semakin dekat data x_i dengan data y_i dalam ruang berdimensi dua, dan semakin besar nilai bobot, semakin jauh keduanya.

Setelah menghitung jarak *Euclidean* dan *Manhattan* serta melakukan prediksi pada data *testing* menggunakan algoritma *KNN*, langkah selanjutnya adalah melakukan evaluasi model untuk membandingkan nilai aktual dan nilai hasil prediksi. Evaluasi yang digunakan dalam penelitian ini adalah *confusion matrix* (matriks kebingungan). *Confusion matrix* adalah tabel yang digunakan untuk menggambarkan kinerja model klasifikasi. Berikut bentuk perhitungan dari *comfussion matrix* dan menghasilkan empat nilai.

Tabel 2. *Confussion matrix* pengklasifikasi biner

Predict	**Actual**	
	True	**False**
True	True Positif	False Negatif
False	False Positif	True Negatif

Keterangan :

- True Positive (TP): Jumlah data positif yang benar-benar diprediksi dengan benar, artinya model berhasil mengklasifikasikan data positif secara tepat.
- False Positive (FP): Jumlah data negatif yang salah diprediksi sebagai positif, artinya model keliru mengklasifikasikan data negatif sebagai positif.
- True Negative (TN): Jumlah data negatif yang benar-benar diprediksi dengan benar, artinya model berhasil mengklasifikasikan data negatif secara tepat.
- False Negative (FN): Jumlah data positif yang salah diprediksi sebagai negatif, artinya model keliru mengklasifikasikan data positif sebagai negatif.

Dari *confusion matrix* dapat dihitung berbagai metrik evaluasi seperti akurasi, presisi, *recall*, dan F1-score yang memberikan informasi lebih lanjut tentang performa model dalam melakukan

klasifikasi. Evaluasi model ini membantu dalam menilai sejauh mana model KNN telah berhasil dalam memprediksi data dengan benar.

Accuracy merupakan salah satu metrik evaluasi yang digunakan untuk mengukur kinerja model klasifikasi. *Accuracy* menggambarkan seberapa akurat model dalam mengidentifikasi kasus dengan benar dari seluruh kasus yang ada dalam dataset. Rumus untuk menghitung *Accuracy* adalah:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP'} \times 100\%. \quad (4)$$

Presisi merupakan salah satu metrik evaluasi yang digunakan untuk mengukur kinerja model klasifikasi, khususnya dalam kasus klasifikasi positif. Presisi mengukur proporsi dari kasus dengan hasil positif (positif yang diprediksi) yang benar-benar diidentifikasi dengan benar dari seluruh kasus yang diprediksi sebagai positif. Rumus untuk menghitung Presisi sebagi berikut.

$$Precision = \frac{TP}{(TP + FP)} \times 100\% \quad (5)$$

Presisi memberikan informasi tentang seberapa akurat model dalam mengidentifikasi kasus positif. Semakin tinggi nilai presisi, semakin sedikit data negatif yang salah diprediksi sebagai positif, dan semakin tepat model dalam mengklasifikasikan data sebagai positif.

Recall yang juga dikenal sebagai *Sensitivity* atau *True Positive Rate* (TPR), adalah salah satu metrik evaluasi yang digunakan untuk mengukur kinerja model klasifikasi, khususnya dalam kasus klasifikasi positif. *Recall* mengukur proporsi kasus positif yang benar-benar diidentifikasi dengan benar (True Positive) dari seluruh kasus positif yang ada dalam dataset.Berikut formula untuk menghitung Recall.

$$Recall = \frac{TP}{(TP + FN)} \times 100\% \quad (6)$$

Recall memberikan informasi tentang seberapa baik model dapat mengidentifikasi kasus positif dengan benar. Semakin tinggi nilai *Recall*, semakin sedikit kasus positif yang terlewatkan (*False Negative*), dan semakin baik model dalam mengenali data positif.

F-measure, juga dikenal sebagai F1-score, adalah metrik evaluasi yang mengkombinasikan pengukuran Precision dan Recall menjadi suatu nilai tunggal yang mencerminkan keseimbangan antara keduanya. F-measure dihitung dengan menggunakan rumus berikut:

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \times 100\% \quad (7)$$

F-measure mencakup *Precision* dan *Recall* dalam perhitungannya, sehingga memberikan gambaran tentang kinerja model klasifikasi yang lebih holistik. Metrik ini berguna dalam situasi di mana kita ingin menemukan keseimbangan antara ketepatan (*Precision*) dan kemampuan mengenali positif (*Recall*) dari model klasifikasi.

F-measure memberikan nilai dari 0 hingga 1, di mana nilai 1 menunjukkan performa yang sempurna (keseimbangan yang ideal antara *Precisio*n dan *Recall*) dan nilai 0 menunjukkan performa yang sangat buruk.

Dalam klasifikasi data mining, nilai akurasi umumnya dapat digolongkan ke dalam beberapa kategori sebagai berikut [14]:

- 0.90 – 1.00: Klasifikasi sangat baik, yang berarti model memiliki tingkat keakuratan yang sangat tinggi dalam melakukan klasifikasi. Model mampu memprediksi data dengan sangat tepat dan memiliki performa yang sangat baik.
- 0.80 – 0.90: Klasifikasi baik, menunjukkan bahwa model memiliki tingkat keakuratan yang baik dalam melakukan klasifikasi. Model mampu memprediksi data dengan baik dan memberikan hasil yang dapat diandalkan.
- 0.70 – 0.80: Klasifikasi cukup, yang berarti model memiliki tingkat keakuratan yang cukup dalam melakukan klasifikasi. Meskipun performanya cukup baik, masih ada ruang untuk perbaikan.
- 0.60 – 0.70: Klasifikasi buruk, menunjukkan bahwa model memiliki performa yang buruk dalam melakukan klasifikasi. Model mungkin mengalami kesulitan dalam memprediksi data dengan tepat.
- 0.50 – 0.60: Klasifikasi salah, yang berarti model memiliki performa yang sangat buruk dan hasil prediksinya hampir tidak lebih baik daripada tebakan acak.

HASIL DAN DISKUSI

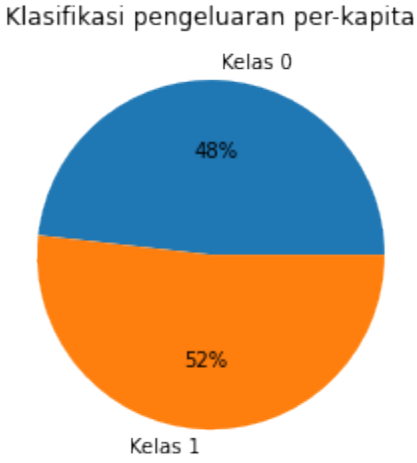
Perhitungan nilai prediksi dilakukan pada data sekunder berdasarkan waktu tahunan, yaitu tahun 2022. Data tersebut diprediksi menggunakan *forecasting.* Setelah data di prediksi kemudian dilakukan analisis deskriptif. Analisis Deskriptif dalam statistika digunakan untuk memperoleh sedikit gambaran umum tentang karakteristik data pengeluaran per-kapita di Sulawesi Selatan, Sulawesi Utara, dan Sulawesi Tenggara berjumlah 56 Kabupaten/Kota.

Tabel 3. Deskripsi Data

No	Variabel	Rata-rata	N	Min	Maks
1	HLS	13.20	56	11.61	16.90
2	IPM	71.26	56	65.13	84.51
3	PDRB	43.11	56	2.50	145.89
4	Mi	19.89	56	4.57	80.34
5	RLS	8.79	56	6.75	12.52
6	UHH	70.11	56	64.74	73.93
7	Pengeluaran per-kapita	10610.38	56	7192	17406

Berdasarkan **Tabel 3**, rata-rata angka pengeluaran per-kapita di Sulawesi Selatan, Sulawesi Utara, dan Sulawesi Tenggara sebesar 10610.38. Untuk HLS terendah pada Bolaang Mongondow sebesar 11,61 dan tertinggi pada Kota Kendari sebesar 16,9. Untuk IPM terendah pada Jeneponto sebesar 12.11 dan tertinggi pada Kota Kendari 84,51. Untuk PDRB terendah pada Bolaang

Mongondow sebesar 2,5 dan tertinggi pada Kota Makassar sebesar 145,89. Untuk Mi terendah pada Kota Kendari 4,57 dan tertinggi pada Bone sebesar 80,34. Untuk RLS terendah pada Jenepoto sebesar 6,75 dan tertinggi sebesar Kota Kendari 12,52. Untuk UHH terendah pada Bolaang Mongondow sebesar 64,74 dan tertinggi pada Kota Kendari sebesar 73,93. Untuk pengeluaran perkapita terendah pada Konawe Kepulauan sebesar 7192 dan tertinggi pada Kota Makassar sebesar 17406.



Gambar 2. Klasifikasi pengeluaran per-kapita menurut BPS

Berdasarkan **Gambar 2.** ada 29 Kabupaten/Kota atau 52% yang tergolong di kelas 1 dan 27 Kabupaten/Kota atau 48% tergolong di kelas 0. Hal ini menunjukkan bahwa angka pengeluaran per-kapita pada kelas 1 atau kategori tinggi sedikit lebih besar dibandingkan kelas 0 atau kategori rendah. Penerapan standarisasi menjadi penting ketika data dalam sebuah penelitian memiliki unit yang berbeda-beda. Standarisasi data bertujuan agar data tersebut mempunyai jarak yang tidak terlalu jauh dan membuat kisaran nilai data dengan format tertentu sehingga seluruh data menjadi standar. Proses standarisasi pada data asli dilakukan dengan menggunakan nilai Z-Score, yang dihitung menggunakan rumus berikut.

$$Z = \frac{x - \mu}{\sigma}$$

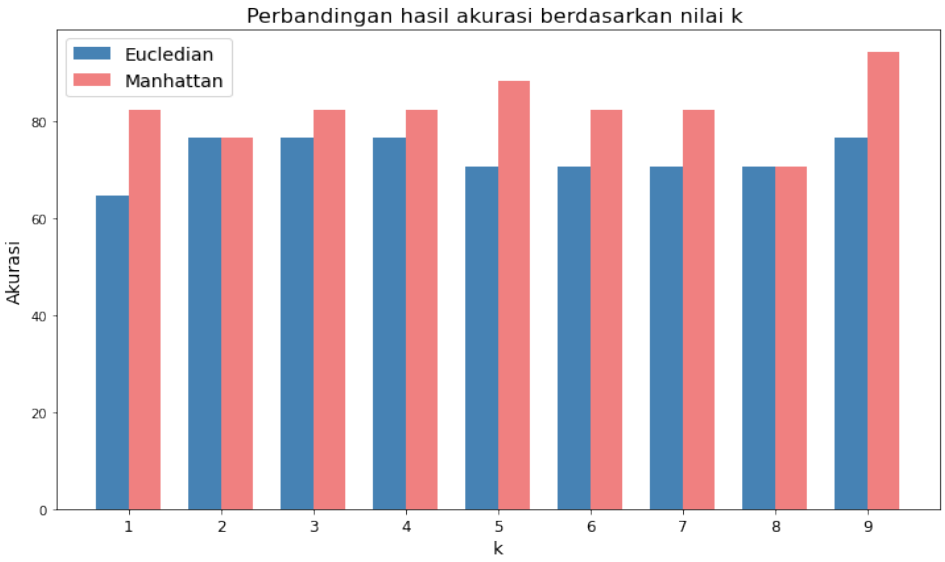
dengan
Z : Z-Score
x : data yang diamati
μ : rata-rata populasi
σ : standar deviasi.
Hasil standarisasi data sebagai berikut.

Tabel 4. Data yang telah distandarisasi

Kab/ Kota	$x1$	$x2$	$x3$	$x4$	$x5$	$x6$
Kep. Selayar	-5.5494	-6.9678	3.7029	-1.9329	-5.7573	-7.1276
Bulukumba	2.3114	-2.2005	-1.4712	6.9873	-6.4107	-8.7671
Bantaeng	-9.4275	-3.7577	6.2205	-1.6386	-1.6212	4.1844
⋮						
Bitung	-5.7590	8.2443	-7.8208	-8.2477	9.1092	6.9168
Tomohon	1.0801	1.4808	-1.3242	-8.7381	1.5970	1.2709
Kotamobagu	-3.2435	6.4715	-1.3457	-8.7504	1.2458	4.7309

Pada penelitian ini, split ratio yang digunakan adalah 8:2, yang berarti 80% dari dataset digunakan sebagai data training dan 20% dari dataset digunakan sebagai data testing. Penggunaan split ratio ini penting dalam proses pembagian dataset menjadi dua bagian yang berbeda untuk tujuan evaluasi model. Dengan menggunakan 80% data training, model akan belajar dari mayoritas data untuk melakukan klasifikasi. Sedangkan, data testing sebesar 20% digunakan untuk menguji performa model terhadap data yang belum pernah dilihat sebelumnya. Penggunaan split ratio yang tepat akan membantu menghindari overfitting atau underfitting, serta memberikan evaluasi yang lebih kredibel terhadap performa model dalam keadaan yang lebih realistis.

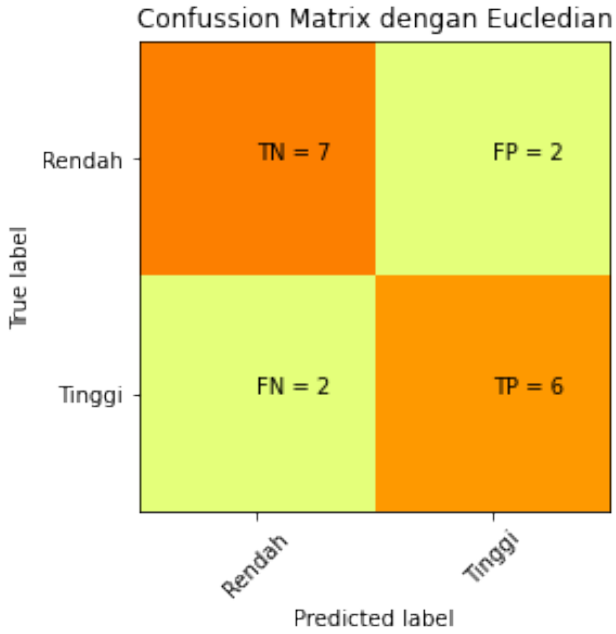
Pada penelitian ini, nilai k divariasikan dari 1 hingga 9 untuk menghasilkan prediksi berdasarkan k data terdekat yang memiliki jarak Euclidean dan Manhattan paling kecil dengan data yang akan diprediksi. Dengan melakukan variasi nilai k, penelitian ini bertujuan untuk mengeksplorasi pengaruh jumlah tetangga terdekat pada performa prediksi algoritmaKNN. Dengan cara ini, model dapat menyesuaikan diri dengan berbagai situasi dan memilih k yang optimal untuk mencapai hasil prediksi yang akurat dan konsisten.



Gambar 3. Hasil akurasi berdasarkan nilai k dengan *Eucledian* dan *Manhattan*

Berdasarkan Gambar 3. memperlihatkan hasil perbandingan dari beberapa percobaan nilai k dengan jarak Eucledian dan Manhattan. Nilai akurasi yang paling tinggi untuk kedua jarak tersebut terdapat pada $k = 9$. Akurasi pada jarak Eucledian dengan $k = 9$ sebesar 76,47% yang berarti hasil prediksi kelompok angka pengeluaran per kapita yang tergolong cukup dengan metode KNN sesuai dengan dataset sebelum diprediksi. Sedangkan akurasi dengan jarak Manhattan pada $k = 9$ sebesar 94,12%. Hal ini berarti hasil prediksi kelompok angka pengeluaran per kapita dengan metode KNN tergolong sangat baik sesuai dengan dataset sebelum diprediksi.

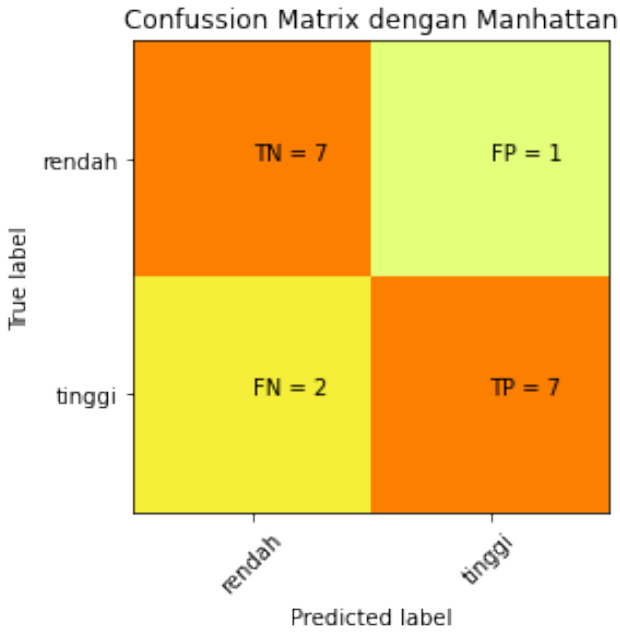
Setelah data diklasifikasi selanjutnya dievalusi dengan *Confusion matrix.* Empat indikator utama yang digunakan sebagai hasil evaluasi, yaitu akurasi, presisi, *recall,* dan F-1 score. Hasil akurasi *Confusion matrix* dengan *Eucledian* menggunakan nilai $k = 9$ ditunjukkan pada **Gambar 4**.



Gambar 4. Hasil *Confussion* dengan *Matrix Eucledian*

Berdasarkan **Gambar 4**, terdapat 7 data yang diprediksi dengan benar sebagai angka pengeluaran per kapita yang tergolong rendah, dan 2 data yang salah diprediksi sebagai rendah. Selanjutnya, terdapat 6 data yang diprediksi dengan benar sebagai angka pengeluaran per kapita yang tergolong tinggi, dan 2 data yang salah diprediksi sebagai tinggi.

Dengan demikian, hasil evaluasi model menunjukkan nilai precision, recall, dan F1-score yang sama, yaitu sebesar 0.78. Hal ini berarti model memiliki tingkat keakuratan yang cukup baik dalam mengidentifikasi kasus positif dan negatif secara benar.



Gambar 5. Hasil *Confussion Matrix* dengan *Manhattan*

Berdasarkan **Gambar 5**, terdapat 7 data yang diprediksi dengan benar sebagai angka pengeluaran per kapita yang tergolong rendah, dan 1 data yang salah diprediksi sebagai rendah. Selanjutnya, terdapat 7 data yang diprediksi dengan benar sebagai angka pengeluaran per kapita yang tergolong tinggi, dan 2 data yang salah diprediksi sebagai tinggi.

Dengan demikian, hasil evaluasi model dengan menggunakan jarak Manhattan dan nilai $k = 9$ menunjukkan adanya peningkatan dalam jumlah data yang diprediksi dengan benar pada kategori angka pengeluaran per kapita yang tergolong rendah (dari 6 menjadi 7) dan kategori yang tergolong tinggi (dari 6 menjadi 7) dibandingkan dengan hasil pada **Gambar 4**.

Berdasarkan hasil evaluasi model, nilai presisi, *recall*, dan F1-score berturut-turut adalah 0.78, 0.88, dan 0.82. Nilai-nilai ini menunjukkan bahwa model memiliki kinerja yang baik dalam mengidentifikasi kasus secara benar. Dari hasil evaluasi tersebut, dapat disimpulkan bahwa metode KNN dengan $k = 9$ dan jarak *Manhattan* tergolong klasifikasi sangat baik, karena memiliki nilai *precision*, *recall*, dan F1-score yang tinggi. Hal ini menunjukkan bahwa model mampu mengklasifikasikan data dengan baik dan akurat.

Sementara itu, jarak *Euclidean* tergolong klasifikasi cukup dengan nilai *precision, recall*, dan F1-score yang lebih rendah dibandingkan dengan jarak *Manhattan*. Meskipun masih memberikan hasil yang cukup baik, performa KNN dengan jarak *Manhattan* lebih unggul dalam mengklasifikasikan data.

KESIMPULAN

Dari hasil dan diskusi yang telah disajikan, klasifikasi angka pengeluaran per kapita kabupaten/kota di tiga provinsi Sulawesi pada tahun 2022, menggunakan metode KNN menghasilkan nilai k tertinggi, yaitu $k = 9$, dengan akurasi terbaik. Hasil akurasi yang diperoleh adalah sebesar 76,47% dengan jarak *Euclidean,* menunjukkan hasil klasifikasinya dapat dianggap

cukup baik. Selain itu, dengan jarak Manhattan, diperoleh akurasi sebesar 94,12%, yang berarti hasil klasifikasinya tergolong sangat baik.

Dengan menggunakan jarak *Manhattan*, model KNN mampu mengklasifikasikan data dengan akurasi yang lebih tinggi dan memberikan hasil prediksi yang lebih akurat. Hal ini menunjukkan bahwa jarak *Manhattan* lebih sesuai dengan karakteristik data pengeluaran per kapita di tiga provinsi Sulawesi pada tahun 2022, sehingga memberikan kinerja model yang lebih optimal dalam mengklasifikasikan data tersebut.

DAFTAR PUSTAKA

[1] U. Hasannah and H. Ahmadi, "Pengaruh Ketimpangan Pendapatan, Pendapatan Per Kapita, Dan Pengeluaran Pemerintah Di Bidang Kesehatan Terhadap Sektor Kesehatan Di Indonesia," *J. Ilmu Ekon. Terap.*, vol. 2, no. 1, pp. 1–18, 2017.

[2] R. F. K. Dewi, O. Obert, and R. Gusmana, "Implementasi Metode K-Nearest Neighbor (KNN) dalam Pengelompokan Status Ekonomi Warga," *J. Big Data Anal. Artif. Intell.*, vol. 4, no. 1, pp. 15–22, 2018.

[3] A. Hakib, "Pengaruh Konsumsi Rumah Tangga Dan Pengeluaran Pemerintah Terhadap Pertumbuhan Ekonomi Di Sulawesi Selatan Periode 2012-2016," *J. Ekon. Balanc. Fak. Ekon. Dan Bisnis*, vol. 15, no. 1, 2019.

[4] J. Zakaria, "Analisis Pertumbuhan Ekonomi dan Disparitas Pendapatan Terhadap Kesejahteraan Antar Daerah Kabupaten/Kota di Provinsi SulawesiSelatan Tahun 2005-2019," *Tata Kelola*, vol. 8, no. 1, pp. 2–24, 2021.

[5] S. I. S. Dai, S. Canon, and D. O. Bauty, "Analisis Pengaruh Rls, Pengeluaran Perkapita, Uhh, Dan Tingkat Kemiskinan Terhadap Ketimpangan Distribusi Pendapatan Di Kbi Dan Kti," *Jesya (Jurnal Ekon. dan Ekon. Syariah)*, vol. 6, no. 1, pp. 535–544, 2023.

[6] A. N. Insany, M. Fajri, and others, "Pemodelan IPM Di Kawasan Timur Indonesia Menggunakan Multivariate Adaptive Regression Spline (MARS)," *Nat. Sci. J. Sci. Technol.*, vol. 8, no. 2, pp. 94–98, 2019.

[7] M. Y. Darsyah, "Lasifikasi indeks pembangunan manusia (ipm) dengan pendekatan k-nearset neighbor (k-nn)," in *Prosiding Seminar Nasional \& Internasional*, 2017.

[8] W. Yustanti, "Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah," *J. Mat. Stat. dan Komputasi*, vol. 9, no. 1, pp. 57–68, 2012.

[9] K. Latifah, "Kombinasi Algorithma K-NN dan Manhattan Distance untuk Menentukan Pemenang Lelang," *J. Inform. Upgris*, vol. 1, no. 1 Juni, 2015.

[10] R. K. Dinata, H. Akbar, and N. Hasdyna, "Algoritma K-Nearest Neighbor dengan Euclidean Distance dan Manhattan Distance untuk Klasifikasi Transportasi Bus," *Ilk. J. Ilm.*, vol. 12, no. 2, pp. 104–111, 2020.

[11] W. Wahyono, I. N. P. Trisna, S. L. Sariwening, M. Fajar, and D. Wijayanto, "Perbandingan penghitungan jarak pada k-nearest neighbour dalam klasifikasi data tekstual," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 1, pp. 54–58, 2020.

[12] L. Hakim and A. Saefudin, *Introduction to Machine Learning Using R*, 1st ed. Bogor: IPB Press, 2022.

[13] T. Wahyono, "Fundamental of Python for Machine Learning," *Yoyakarta Penerbit Gava Media*, 2018.

[14] F. Gorunescu, *Data Mining: Concepts, models and techniques*, vol. 12. Springer Science \& Business Media, 2011.

Perbandingan Kinerja Hasil Luaran Model Jaringan Syaraf Tiruan dan SARIMA Untuk Prediksi Awal Musim Hujan Kota Pangkalpinang

Presli Panusunan Simanjuntak

Stasiun Klimatologi Bangka Belitung

Jalan Kartika I, Koba, Bangka Tengah, Kep. Bangka Belitung 33681
e-mail: presli.simanjuntak@bmkg.go.id

ABSTRAK

Bangka Belitung merupakan wilayah penghasil utama dari lada dan karet. Informasi prediksi curah hujan dan awal musim hujan (AMH) diperlukan untuk meningkatkan produksi komoditi ini. Tujuan penelitian untuk membangun model Seasonal ARIMA (SARIMA) dan Jaringan Syaraf Tiruan (JST) guna memprediksi curah hujan dan penentuan AMH dengan model terbaik. Model prediksi yang digunakan adalah SARIMA dengan menggunakan data curah hujan dari masa lalu dan JST propagasi balik (backpropagation) dengan menggunakan prediktor suhu muka laut, angin zonal dan precipitable water periode 1981-2010. SARIMA merupakan metode yang digunakan dalam peramalan data runtun waktu musiman sedangkan JST backpropagation merupakan pelatihan model untuk mempelajari set pola data masa lalu dan mengevaluasi serta membuat formula yang dihubungkan dengan keluaran yang diinginkan. Model dibandingkan kinerjanya dengan menghitung nilai koefisien korelasi dan akar rerata kuadrat kesalahan. Hasil penelitian menunjukkan model JST lebih baik daripada SARIMA untuk memprediksi curah hujan dengan nilai RMSE sebesar 48,70 dan korelasi sebesar 0,25. Akan tetapi, JST belum mampu untuk menduga curah hujan ekstrim. Penentuan AMH menggunakan prediksi JST menghasilkan kesesuaian model prediksi sebesar 42,86 % dengan nilai skill sebesar 0,05 lebih baik daripada penentuan AMH menggunakan prediksi SARIMA terpilih yang menghasilkan kesesuaian model prediksi sebesar 14,26 % dengan nilai skill sebesar 0,00.
Kata kunci: backpropagation, korelasi, prediksi AMH, RMSE, SARIMA

ABSTRACT

Bangka Belitung Islands are the main producing areas of pepper and rubber. Predictive information on rainfall and the start of the rainy season (AMH) is needed to increase the production of this commodity. The aim of the research is to build a Seasonal ARIMA model (SARIMA) and an Artificial Neural Network (ANN) to predict rainfall and determine AMH with the best model. The prediction model used is SARIMA using rainfall data from the past and ANN backpropagation using predictors of sea surface temperature, zonal wind and precipitable water for the period 1981-2010. SARIMA is a method used in forecasting seasonal time series data while ANN backpropagation is a training model to study sets of past data patterns and evaluate and create formulas related to the desired output. Model performance is compared by calculating the value of the correlation coefficient and the mean square root of the error. The results showed that the ANN model was better than SARIMA for predicting rainfall with an RMSE value of 48.70 and a correlation of 0.25. However, ANN has not been able to predict extreme rainfall. Determining AMH using ANN predictions resulted in a predictive model fit of 42.86% with a skill value of 0.05 which was better than the AMH determination using selected SARIMA predictions which resulted in a predictive model fit of 14.26% with a skill value of 0.00.
Keywords: *backpropagation, correlation, prediction of the onset of wet season,RMSE, SARIMA*

PENDAHULUAN

Pangkalpinang adalah ibu kota dari Provinsi Kepulauan Bangka Belitung. Pangkalpinang dan beberapa wilayah sekitarnya seperti Kabupaten Bangka Tengah dan Kabupaten Bangka merupakan salah satu wilayah penghasil utama lada dan karet [1]. Tanaman lada tumbuh optimal pada kondisi curah hujan 1.000 sampai dengan 3.000 mm/tahun, dengan rata-rata hari hujan 110 sampai dengan 170 hari hujan/tahun. Kondisi optimal juga dapat teercapai dengan kemarau hanya 2 sampai 3 bulan per tahun (Suprapto, 2008). Pada tanaman karet, kondisi optimal untuk pertumbuhan tanaman ini dengan curah hujan 1.500 sampai dengan 3.000 mm/tahun dan diperlukan 1 sampai 3 bulan kering per tahun [2]. Produktivitas lada dan karet ini sangat terpengaruh kondisi iklim terutama curah hujan serta polanya sehingga informasi AMH sangat diperlukan.

Banyak penelitian sebelumnya telah menganalisis faktor-faktor yang mempengaruhi AMH. Sebagai contoh, nilai anomali suhu muka laut (SML) Indonesia memilki peranan yang signifikan terhadap awal dan panjang musim hujan Indonesia. Keadaan SML Lookal dapat digunakan menjadi parameter dalam penentuan banyak sedikitnya kandungan uap air di atmosfer. Kondisi kandungan uap air di atmosfer berperan penting dalam pembentukan awan di Indonesia [3]. Selain itu, hal lain yang dapat dijadikan indicator untuk menjadi prediktor peluang terjadinya hujan disuatu wilayah adalah jumlah uap *precipitable water*. Kandungan jumlah uap air pada sebuah massa udara adalah indicator potensi terjadinya prespitasi. *Precipitable water* sebagai variabel tunggal persamaan regresi linier sederhana cukup baik digunakan untuk menghasilkan prediksi curah hujan di Pangkalpinang [4]

Terdapat beberapa cara yang dapat dilakukan dalam memprediksi AMH. Salah satu prediksi AMH adalah menggunakan teknologi *Artificial Intellegence* yaitu jaringan saraf tiruan (JST). Dengan mengunakan pemodelan JST, pengenalan pola data dari metode prediksi AMH dapat dilakukan dengan sistem pembelajaran. Berdasarkan hasil pembelajaran yang dimiliki, JST dapat dilatih untuk mengenali dan menganalisis pola data masa lalu dan mencari suatu fungsi yang menghubungkan pola data pada masa lalu dengan keluaran yang dikehendaki pada saat ini [5]. Saat ini metode JST telah banyak dipelajari dan diterapkan untuk berbagai masalah, termasuk prediksi.

Selain dengan pemodelan *Artificial Intelligence*, banyak penelitian sebelumnya menggunakan metode stastisik untuk memprediksi curah hujan maupun AMH. Salah satu metode statistik yang dapat digunakan dalam memprediksi curah hujan maupun AMH adalah ARIMA (*Autoregressive Integrated Moving Average*). Model yang dihasilkan dari metode ARIMA dapat meramalkan curah hujan dengn baik. Metode ARIMA mampu memberikan hasil yang baik dalam prediksi dengan menggunakan data dari masa lalu, dengan memenuhi beberapa kondisi diantaranya adalah *white noise,* stasioner dalam mean dan varians, berdistribusi normal. Pada data runtun waktu yang memiliki pola musiman model ARIMA dapat dikembangkan menjadi SARIMA (*Seasonal Autoregressive Integrated Moving Average*) [6]

Metode *Seasonal* ARIMA merupakan metode yang paling sering digunakan dalam peramalan data deret waktu [7]Metode lain yang juga dapat digunakan dalam melakukan peramalan adalah Jaringan Syaraf Tiruan Backpropagation. Jaringan Syaraf Tiruan (JST) merupakan suatu sistem pengolahan informasi yang terinspirasi dari sistem kerja syaraf biologis, seperti kinerja otak yang memproses informasi. Algoritma Backpropagation merupakan algoritma pembelajaran yang sering digunakan dalam melakukan pelatihan jaringan dalam berbagai aplikasi, seperti pemilihan lokasi, pengenalan pola maupun evaluasi kinerja [8]. Kota Pangkalpinang memerlukan

model yang handal untuk memprediksi curah hujan dan awal musim sebagai pedoman perencanaan tata Kelola kota dan Kelola pangan.

Berdasarkan uraian diatas, penelitian ini akan membahas lebih spesifik terkait perbandingan hasil model luaran JST dengan melibatkan prediktor SML, angin zonal dan *precipitable water* dengan hasil model luaran SARIMA untuk memprediksi AMH di Pangkalpinang.

METODE PENELITIAN

A. Data Penelitian

Data curah hujan harian diperoleh dari Stasiun Meteorologi Pangkalpinang. Data curah hujan harian diolah menjadi data curah hujan dasarian dengan cara menjumlahkan data curah hujan sepuluh harian tahun 1981-2017 untuk menentukan AMH.

Gambar 1. Pembagian Grid Prediktor Curah Hujan

Data prediktor curah hujan yang digunakan terdiri dari suhu muka laut (SML), angin zonal dan *precipitable water* pada lapisan 850 mb. Data harian prediktor tersebut diambil pada periode Januari 1981 – Desember 2017 dari situs *International Research Institute Data Library* (IRIDL). Data prediktor curah hujan tersebut dibatasi berdasarkan lintang dan bujur yang ditentukan yaitu 10^o LU - 10^o LS dan 90^o BT - 115^o BT. Daerah dengan luasan tersebut dibagi menjadi 20 grid dengan ukuran masing-masing grid 5^o x 5^o, seperti ditunjukkan pada Gambar 1. Data suhu muka laut (SML) harian diolah menjadi data SML dasarian dengan cara menghitung nilai rata-rata SML persepuluh hari.

$$SML_{dasarian} = \frac{\sum_{i=1}^{n} SML_1 + SML_1 + \cdots + SML_n}{n} \quad (1)$$

dimana:

$SML_{dasarian}$ = suhu muka laut dasarian (^{0}C)

n = banyaknya data (^{0}C)

SML = suhu muka laut (^{0}C)

Data angin zonal harian diubah menjadi data angin zonal dasarian dengan cara merata-ratakan data harian menjadi data persepuluh hari.

$$Mo = b + \left(\frac{b_1}{b_1+b_2}\right)P \quad (2)$$

dimana:

Mo = modus
B = batas bawah kelas interval dengan frekuensi terbanyak
P = panjang kelas interval
b_1 = frekuensi terbanyak dikurangi frekuensi kelas sebelumnya
b_2 = frekuensi terbanyak dikurangi frekuensi kelas sesudahnya

Data *precipitable water* harian diubah menjadi data *precipitable water* dasarian dengan cara menghitung menjumlahkan data *precipitable water* persepuluh hari.

B. Penetnuan Awal Musim Hujan (AMH)

Penentuan AMH didefenisikan menggunakan kriteria BMKG yaitu dengan menentukan AMH baru yang dihitung setelah normal dasarian AMH pada periode 1981-2010. Kriteria pertama yaitu AMH ditentukan dengan menjumlahkan curah hujan dasarian yang telah lebih dari 50 mm dan kemudiaan diharuskan diikuti dua dasarian berikutnya, apabila kriteria pertama tidak terpenuhi maka dapat menggunakan kriteria kedua. Kriteria kedua yaitu AMH ditentukan dengan jumlah curah hujan pada 3 (tiga) dasarian telah lebih dari 150 mm.

C. ARIMA

ARIMA adalah model pendekatan kuantitatif yang dipopulerkan oleh George Box dan Gwilym Jenkins[9]. Metode ARIMA menggunaka data masa lalu sebagai acuan dalam memprakirakan masa mendatang [10]. Metode ARIMA lebih mudah untuk mengikuti fluktuasi data dibandingkan metode statistik lainnya. ARIMA dibedakan menjadi model ARIMA nonmusiman, model ARIMA musiman (*Seasonal* ARIMA/ SARIMA) dan gabungan antara model ARIMA non-musiman dan musiman atau sering disebut sebagai ARIMA musiman multiplikatif. Secara umum model ARIMA non musiman terdiri dari model *autoregressive* (AR), *model moving average* (MA), model ARMA dan model ARIMA.

ARIMA merupakan metode runtut waktu yang tidak stasioner terhadap mean dan agar menjadi stasioner dibutuhkan proses *differencing* sebanyak *d.* Bentuk umum model ARIMA pada orde ke-p,q dengan *differencing* sebanyak d atau ARIMA(p,d,q) adalah sebagai [6]

$$\emptyset_p\,(B)(1-B)^d\,Z_t = \theta_0 + \theta_q(B)a_t \quad (3)$$

dengan

$$\emptyset_p\,(B) = (1 - \emptyset_1 B - \emptyset_2 B - \cdots - \emptyset_p B^p) \quad (4)$$
$$\theta_q(B) = (1 - \theta_1 B - \theta_2 B - \cdots - \theta_q B^q) \quad (5)$$

dimana $\emptyset_p\,(B)$ adalah operator dari AR, $\theta_q(B)$ adalah operator MA dan $(1-B)$ adalah operator *backward shift* dengan d adalah orde *differencing*. Ketika p = 0, model ARIMA (p,d,q) dapat disebut sebagai model *integrated moving average* . Begitu juga ketika q = 0, model ARIMA(p,d,q) dapat disebut sebagai model *autoregressive integrated*.

ARIMA musiman (SARIMA) adalah model yang dibuat dari komposisi data yang dipengaruhi faktor musiman, sehingga plot yang dihasilkan membentuk pola musiman. Model ARIMA dengan periode musiman s dapat dinotasikan ARIMA $(P,D,Q)^s$ dengan modelnya sebagai berikut[6]

$$\Phi_P\ (B^s)(1-B^s)^D\ Z_t = \Theta_Q(B^s)a_t \quad (6)$$

dengan

$$\Phi_P(B^s) = (1-\Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps} \quad (7)$$

$$\Theta_Q(B^s) = (1-\Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_Q B^{Qs}) \quad (8)$$

dimana $\Phi_P\ (B^s)$ adalah faktor dari AR musiman, $\Theta_Q(B^s)$ adalah faktor MA musiman dan $(1-B^s)^D$ adalah *differencing* musiman s dengan D adalah orde *differencing*.

1. Stasioneritas *Time Series*

Suatu data *time series* yang dapat analisis adalah data yang bersifat stasioner. Stasioner adalah keadaan dimana mean dan varians adalah konstan[11]. dengan demikian

Mean dari Z_t:

$$E\ (Z_t) = E\ (Z_{t+k}) = \mu \quad (9)$$

Varians dari Z_t:

$$E(Z_t-\mu)^2 = E(Z_{t+k}-\mu)^2 = \sigma^2 \quad (10)$$

Apabila nilai observasi sebanyak n berfluktuasi terhadap nilai varians dan *mean* secara konstan serta tidak terpangaruh waktu, maka dapat disimpulkan bahwa data *time series* tersebut adalah stasioner. Sebaliknya apabila nilai observasi sejumlah n tidak berfluktuasi terhadap means dan varian secara konstan, maka data disimpulkan bahwa *time series* tersebut tidak stasioner [11]Adapun cara untuk mengatasi ketidakstasioneran adalah dengan melakukan pembedaan (*differencing*) atau dengan transformasi *Box-cox*. Pembedaan (*differencing*) dibuat jika data tidak stasioner terhadap mean, sedangkan Transformasi *Box-cox* dibuat jika data tidak stasioner terhadap varians [12]

Stasioneritas data dalam mean dapat dilakukan dengan mengidentifikasi plot data dan bentuk ACF data. Apabila ACF menunjukkan kondisi pola yang turun lambat berarti data belum stasioner dalam mean. Dalam kondisi seperti maka dibutuhkan *differencing* agar datanya stasioner dalam mean. Namun apabila ACF menunjukkan kondisi pola yang menurun dengan cepat makan kondisi ini menunjukkan data sudah stasioner dalam mean. Adapun metode yang gunakan dalam mengatasi kondisi non-stasioner dalam mean adalah melakukan pembedaan (*differencing*) terhadap data dengan persamaan berikut [11]

$$W_t = Z_t - Z_{t-1} \quad (11)$$

dimana W_t merupakan nilai series Z_t setelah dilakukan *differencing*.

Sedangkan suatu deret waktu Z_t dikatakan tidak stasioner terhadap varians, apabila Z_t berubah sejalan dengan perubahan level $(Z_t) = cf(\mu_t)$, dimana c merupakan konstanta. *Box*

dan *Cox* pada teorinya memberikan pemahaman bahwa transformasi terhadap varians yang tidak konstan dapat menggunakan *power transformation* sebagai berikut [6]

$$T(Z_t) = \frac{{Z_t}^{\lambda}-1}{\lambda}, \text{ dimana } \lambda \neq 0 \quad (12)$$

λ adalah parameter transformasi, untuk λ = 0 dilakukan pendekatan berikut :

$$\lim_{\lambda \to 0} T(Z_t) = \lim_{\lambda \to 0} {Z_t}^{(\lambda)} = \lim_{\lambda \to 0} \frac{{Z_t}^{\lambda}-1}{\lambda} = ln(Z_t) \quad (13)$$

2. Identifikasi Model

Apabila data telah stasioner kemudian diiidentifikasi model dengan mengamati plot ACF dan PACF. Nilai p dilihat dengan mengamati plot PACF dan nilai q dilihat dengan mengamati plot ACF. Kondisi data yang tidak terdapat faktor musiman, maka untuk menduga model dapat dilakukan dengan memperhatikan kriteria seperti pada tabel berikut

Tabel 1. Struktur ACF dan PACF Pada Model ARIMA Reguler

Model	ACF	PACF
Autoregressive (p)	Turun eksponensial (*dies – down*)	Terpotong setelah lag-p (*cut off after lag*-p)
Moving Average (q)	Terpotong setelah lag-q (*cut off after lag*-q)	Turun eksponensial (*dies – down*)
Autoregressive-Moving Average (p,q)	Turun eksponensial (*dies – down*)	Turun eksponensial (*dies – down*)
Autoregressive (p) atau *Moving Average* (q)	Terpotong setelah lag-q (*cut off after lag*-q)	Terpotong setelah lag-p (*cut off after lag*-p)

Sedangkan pendugaan model yang terdapat faktor musiman dilakukan dengan memperhatikan kriteria seperti tabel berikut [11]

Tabel 2. Struktur ACF dan PACF Pada Model ARIMA Reguler

Model	ACF	PACF
Autoregressive (p)	Turun eksponensial (*dies – down*)	Terpotong setelah lag s, 2s,…,Ps (*cut off after lag* Ps)
Moving Average (q)	Terpotong setelah lag s, 2s,…,qs (*cut off after lag Qs*)	Turun eksponensial (*dies – down*)
Autoregressive-Moving Average (p,q)	Turun eksponensial (*dies – down*)	Turun eksponensial (*dies – down*)
Autoregressive (p) atau *Moving Average* (q)	Terpotong setelah lag s, 2s,…,qs (*cut off after lag Qs*)	Terpotong setelah lag s, 2s,…,Ps (*cut off after lag* Ps)

3. Uji Signifikansi Model

Pengujian parameter dilakukan untuk mengetahui apakah parameter signifikan terhadap model. pengujian signifikansi parameter dapat dinyatakan sebagai [11]

Hipotesis :

$H_0 : \beta = 0$ (parameter tidak signifikan)

$H_1 : \beta \neq 0$ (parameter signifikan)

dimana β adalah parameter pada model ARIMA

Statistik uji :

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} \quad (14)$$

Daerah penolakan : Tolak H_0 jika $|t| > t_{\alpha/2\,;n-m}$, dengan

$SE(\hat{\beta})$: standar error dari nilai taksiran β

m : banyaknya parameter yang ditaksir

4. Pengujian Asumsi

Asumsi yang harus dipenuhi pada model ARIMA meliputi asumsi residual *white noise* dan residual berdistribusi normal (Wei, 2006). Berikut adalah pengujian asumsi white noise dan uji kenormalan. *White noise* adalah suatu proses apabila tidak terdapat korelasi pada suatu deret residual. Untuk menguji kondisi residual yang telah memenuhi asumsi *white noise* maka dapat digunakan statistik uji yang dikemukakan oleh *Ljung Box*. Hipotesisnya adalah sebagai berikut :

$H_0 : \rho_1 = \rho_2 = \rho_K = 0$ (residual tidak saling berkorelasi)

H_1 : minimal ada satu $\rho_k \neq 0$ (residual saling berkolerasi), dengan $k = 1,2,3, \dots, K$.

Statistik uji :

$$Q = n(n+2)\sum_{k=1}^{k}(n-k)^{-1}\,\hat{\rho}_k{}^2 \quad (15)$$

Daerah Penolakan : tolak H_0 , jika nilai dari $Q > \chi^2{}_{(\alpha;K-p-q)}$ atau *P- value* < α

dimana : p dan q adalah orde ARIMA

n : jumlah observasi dari *time series*

$\hat{\rho}_k$: taksiran autokorelasi residual *lag k*

Untuk pengujian kenormalan dengan memeriksa kenormalan data residual denga melihat kecenderung membentuk garis lurus (linier) pada *normal probability plot*.

5. Prediksi

Dilakukan mengguakan model terbaik yang memenuhi syarat estimasi parameter, pemeriksaan asumsi residual *white noise* dan berdistribusi normal.

D. Jaringan Syaraf Tiruan *Backpropagation*

JST dibuat untuk memberikan kemampuan jaringan mengenai pola. Model JST dapat dibuat apabila sudah menentukan data latih dan data uji. Metode JST yang digunakan adalah *backpropagation*. Pelatihan *backpropagation* meliputi 3 fase, yaitu fase propagasi maju (*feedforward*), fase propasi mundur (*backpropagation*), dan fase modifikasi bobot. Selain itu,

beberapa hal yang perlu ditentukan guna optimalisasi arsitektur backpropagation adalah sebagai berikut.

a. Pemilihan bobot dan bias awal
   Inisialisasi bobot adalah dengan cara mengambil nilai random yang cukup kecil [13]
b. Jumlah lapisan dan unit tersembunyi
   Jumlah lapisan tersembunyi (*hidden layer*) yang digunakan di dalam penelitian ini adalah 1. 1 *hidden layer* sudah cukup dan tidak membutuhkan banyak komputasi saat pelatihan [14]Sedangkan, jumlah unit tersembunyi (*hidden neuron*) yang digunakan berdasarkan aturan Heaton. Aturan Heaton engemukakan beberapa aturan untuk menentukan jumlah unit tersembunyi (*hidden neuron*) yang tepat untuk digunakan dalam lapisan tersembunyi (*hidden layer*) ([15]
   1. Jumlah *hidden neuron* harus berada diantara ukuran *input layer* dan *output layer*
   2. Jumlah *hidden neuron* harus 2/3 ukuran *input layer*, ditambah ukuran *output layer*.
   3. Jumlah *hidden neuron* harus kurang dari dua kali ukuran *input layer.*
c. Jumlah pola pelatihan (training)
   Training dilakukan terhadap setiap model prediksi pada masing-masing hidden neuron-nya. Jumlah pola pelatihan (training*)* dilakukan sampai menghasilkan nilai error yang stabil.
d. Lama iterasi (*epoch*)
   Banyaknya epoch yang digunakan pada setiap training model yaitu 1000. dalam penelitian terhadap prediksi curah hujan bulanan di Stasiun Tenggarong, Kalimantan Timur menunjukkan 1000 epoch menghasilkan nilai *Mean Square Error* terbaik [16].
e. Parameter laju pelatihan (*learning rate*)
   Pada penelitian ini menggunkanan *adaptive learning rate*. Nilai learning rate pertama yaitu 0.01. Kemudian untuk nilai *learning rate* berikutnya bernilai 1.05 kali lebih besar dari yang sebelumnya hingga mencapai hasil konvergen. Namun, jika hasil yang diperoleh menunjukkan divergen, maka nilai *learning rate* berikutnya dikali 0.7.

E. Validasi dan Verifikasi Hasil Prediksi

a. *Root Mean Square*

Root Mean Square Error (RMSE) merupakan nilai rerata akar kuadrat dari residu antara prediksi dengan observasi. Hal ini untuk mengetahui besaran nilai simpangannya (*error*). RMSE digunakan untuk mengetahui besarnya penyimpangan yang terjadi antara nilai prakiraan curah hujah dasarian yang dihasilkan model JST dan ARIMA dengan nilai observasi. Hasil validasi RMSE dikatakan baik apabila nilainya mendekati 0 (semakin kecil). Secara matematis formula RMSE sebagai berikut [17].

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(X_t - F_t)^2}{n}} \qquad (16)$$

dengan

X_t : nilai aktual pada waktu ke-t
F_t : nilai dugaan pada waktu ke-t
n : jumlah data

b. Metode Kontingensi

Metode Kontingensi yang digunakan untuk mengetahui apakah model dapat menunjukkan kinerja yang dapat diterima dalam memprediksi AMH. Dalam buku verifikasi prediksi iklim Indonesia, verifikasi terhadap suatu hasil prediksi sangat penting dilakukan, tujuannya adalah untuk memantau kualitas dari hasil prediksi, tujuannya adalah untuk mengetahui sejauhmana ketepatan prediksi yang dibuat serta mengetahui perbaikan dari waktu ke waktu. Dalam Kajian ini menggunakan metode verifikasi dengan metode kontingensi. Dalam buku Verifikasi Prediksi Iklim Indonesia Pada Tahun 2012, keakuratan suatu prediksi dinyatakan dalam persentase kesesuaian dengan istilah "Sesuai Prediksi" dan "Menyimpang Prediksi". Apabila (x_i,y_i) merupakan data series pasangan pprediksi dengandata observasinya maka dapat dibentuk table kontingensi untuk setiap pasangan (x_i,y_i). Jumlah pasangan (x_i,y_i) kemudian dijabarkan dalam notasi n_i.

Table 3. Kontingensi Jumlah Pasangan Prediksi dan Observasi di Setiap Kategori

ij		**Observasi**				
		1	2	…j…	K	**Jml**
Prediksi	1	P_{11}	P_{12}	P_{1j}	P_{1k}	$\sum P_{1j}$
	2	P_{21}	P_{22}	P_{2j}	P_{2k}	$\sum P_{2j}$
	…i…	P_{i1}	P_{i2}	P_{ij}	P_{ik}	$\sum P_{2j}$
	K	P_{k1}	P_{k2}	P_{kj}	P_{kk}	$\sum P_{2j}$

Dari tabel diatas dibentuk tabel kontingensi berikutnya yang berisi nilai *frekuensi relative.* Apabila jumlah data adalah ni, maka *frekuensi relative* dari jumlah pasangan prediksi tersebut kategori i dan observasi kategori j adalah P_{ij}.

$$P_{ij} = {}^{n_{ij}}/_{n} \tag{17}$$

c. *Skill Score*

Model prediksi AMH dikatakan memiliki skill yang lebih baik dibandingkan dengan prediksi menggunakan data normalnya jika skill bernilai antara 0 hingga sama dengan 1. Menurut WMO (2000) nilai skill dapat dihitung menggunakan rumus sebagai berikut :

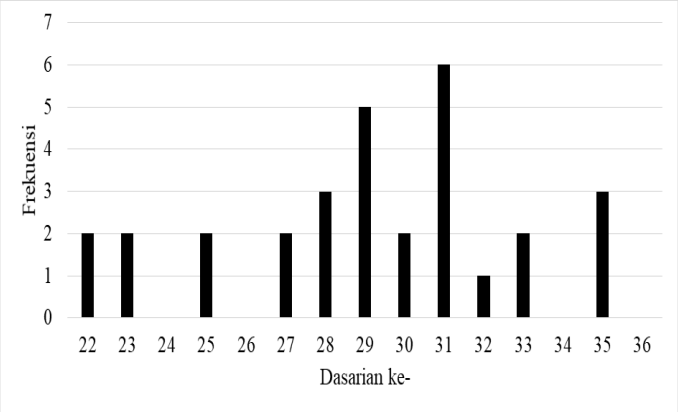
$$Skill\ Score\ = 1 - \frac{MAE_f}{MAE_n} \tag{18}$$

dengan MAE_f adalah selisih data observasi dengan prediksinya sedangkan MAE_n adalah selisih data observasi dengan data normalnya.

HASIL DAN PEMBAHASAN

A. Pola Curah Hujan dan Awal Musim Hujan Kota Pangkalpinang

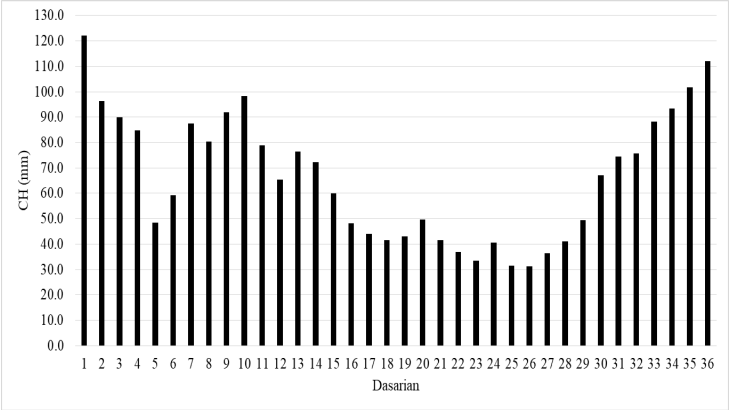
Berdasarkan pembagian pola hujan di Indonesia, Pangkalpinang termasuk dalam wilayah yang memiliki pola hujan monsunal yang mana musim hujan dan musim kemarau dapat dibedakan secara jelas [18]. Berikut adalah karakteristik pola curah hujan di Pangkalpinang.



Gambar 2. Pola Curah Hujan Pangkalpinang Periode 1981-2010

Berdasarkan gambar 2 menunjukkan bahwa secara umum curah hujan tertinggi terjadi pada dasarian ke-1 yaitu 122,2 mm (Januari dasarian pertama), sedangkan curah hujan terendah terjadi pada pada dasarian ke-26 dengan 31,2 mm (September dasarian kedua).

Penentuan awal musim hujan (AMH) kemudian dilakukan dengan metode kriteria yang telah ditetapkan oleh BMKG yaitu AMH dilihat dengan jumlah curah hujan dasarian telah lebih 50 mm dan diikuti duaa dasarian berikutnya, apabila tidak memenuhi kriteria pertama maka AMH dimulai dengan jumlah curah hujan dalam tiga (3) dasarian telah lebih dari 150 mm.



Gambar 3. Distribusi frekuensi AMH Pangkalpinang periode 1981-2010

Berdasarkan gambar 3, awal musim hujan sering terjadi pada dasarian ke-31 atau terjadi pada bulan November dasarian pertama dengan frekuensi kejadian 6 kali. Awal musim hujan tercepat terjadi pada dasarian ke-22 (pada bulan Agustus dasarian pertama) dan awal musim hujan paling akhir terjadi pada dasarian ke-35 (pada bulan Desember dasarian kedua). Selama periode tersebut

juga tidak pernah terjadi AMH pada dasarian 26 (pada bulan September dasarian kedua), dasarian 34 (Desember dasarian pertama) dan dasarian 36 (Desember dasarian ketiga).
Jumlah data AMH maju pada periode 1981-2010 sebanyak 11 kali, sedangkan jumlah data AMH mundur pada periode 1981-2010 sebanyak 14 kali sedangkan sesuai dengan normalnya sebanyak 5 kali. Hal ini menunjukkan bahwa AMH di Pangkalpinang cenderung mundur dari normalnya selama periode 1981-2010

B. Model Prediksi Curah Hujan dengan Jaringan Syaraf Tiruan

Pemodelan Jaringan Saraf Tiruan (JST) terdiri atas 1 lapisan input, *1 hidden layer* dan 1 lapisan output. Input pada pemodelan JST adalah data SML, angin zonal dan *precipitable water* hasil dari pemilihan prediktor. Pemodelan JST menggunakan pembelajaran *Backpropagation*. Karakteristik dan spesifik yang digunakan pada arsitektur jaringan syaraf tiruan dapat dilihat pada Tabel 4.

Tabel 4. Karakteristik dan Spesifikasi JST Untuk Prediksi Curah Hujan

Karakteristik	**Spesifikasi**
Arsitektur	1 *hidden layer*
Neuron Input	3 neuron (suhu muka laut, angin zonal,
Neuron Output	*precipitable water*)
Hidden Neuron	1 target data
Learning Rate	2
Fungsi Aktivasi	*adaptive learning rate*
Algoritma JST	Sigmoid biner *backpropagation*

Simulasi prediksi dilakukan dengan menggunakan skema *lag time* terbaik pada setiap periode musiman untuk tahun 2011 sampai dengan 2017. Simulasi prediksi untuk 9 dasarian kedepan (simulasi sesuai periode musiman).

C. Model Dugaan SARIMA Terbaik

Setelah menemukan beberpa model untuk dugaan yang elah memenuhi asumsi dan signifikansi, selanjutnya memilih model terbaik. Pemilihan model terbaik digunakan untuk mendapatkan model yang paling akurat diantara model-model lainnya. 2 model dugaan yang telah signifikan dan memenuhi asumsi adalah SARIMA (1,0,1) $(2,1,0)^{36}$ dan SARIMA (1,0,1) $(0,1,1)^{36}$. Selanjutnya dari kedua model ini akan dilihat nilai RMSE terendah dan korelasi tertinggi untuk dipilih sebagai model terbaik diantara keduanya.

Tabel 5. Hasil Perhitungan RMSE dan Korelasi Model SARIMA Terbaik Periode Musiman

Periode	**Model SARIMA (1,0,1) $(2,1,0)^{36}$**		**Model SARIMA (1,0,1) $(0,1,1)^{36}$**	
	RMSE	Korelasi	RMSE	Korelasi
DJF	78.80	-0.20	68,20	-0.03
MAM	62.30	-0.18	51,30	0.10
JJA	42.20	0.001	34,90	0.10

SON	57.10	0.2	50,50	0.32

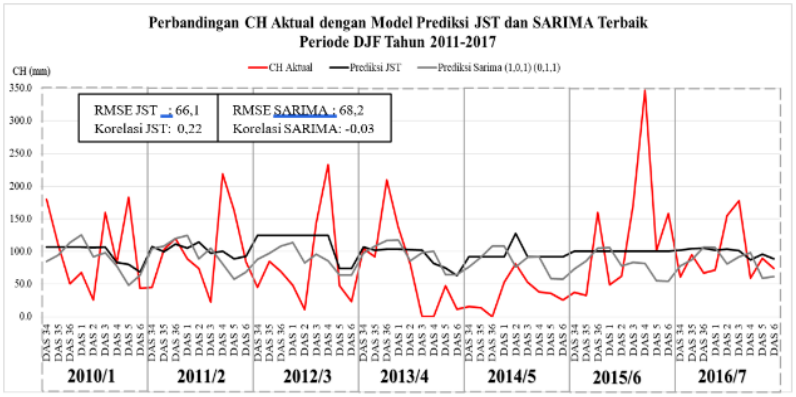
Berdasarkan dari penjabaran diatas dapat dilihat bahwa nilai RMSE terkecil terdapat pada model SARIMA (1,0,1) $(0,1,1)^{36}$ dengan 51,225 dan korelasi sebesar 0,12, sehingga model tersebut dipilih sebagai model ARIMA untuk prediksi curah hujan di Pangkalpinang.
Dengan mengkombinasikan persamaan (3) dan persamaan (6) maka Model SARIMA (1,0,1) $(0,1,1)^{36}$ secara matematis dapat dituliskan sebagai berikut.

- $(1-\emptyset_1 B)(1-B)(1-B^{36})\mathrm{Z}_t = (1-\theta_1 B)(1-\Theta_1 B^{36})\mathrm{a}_t$
- $(1-B^{36}-B+B^{37}-\emptyset_1 B+\emptyset_1 B^{37}+\emptyset_1 B^2-\emptyset_1 B^{38})\mathrm{Z}_t = (1-\theta_1 B-\Theta_1 B^{36}+\theta_1\Theta_1 B^{36})\mathrm{a}_t$
- $\mathrm{Z}_t = \mathrm{Z}_{t-1}+\mathrm{Z}_{t-36}-\mathrm{Z}_{t-37}+\emptyset_1\mathrm{Z}_{t-1}-\emptyset_1\mathrm{Z}_{t-2}-\emptyset_1\mathrm{Z}_{t-37}+\emptyset_1\mathrm{Z}_{t-38}-\theta_1\mathrm{a}_{t-1}-\Theta_1\mathrm{a}_{t-36}+\theta_1\Theta_1\mathrm{a}_{t-37}+\mathrm{a}_t$
- $\mathrm{Z}_t = \mathrm{Z}_{t-1}+\mathrm{Z}_{t-36}-\mathrm{Z}_{t-37}+0{,}8205\mathrm{Z}_{t-1}-0{,}8205\mathrm{Z}_{t-2}-0{,}8205\mathrm{Z}_{t-37}+0{,}8205\mathrm{Z}_{t-38}-0{,}77135\mathrm{a}_{t-1}-0{,}9468\mathrm{a}_{t-36}+0{,}7769\mathrm{a}_{t-37}+\mathrm{a}_t$

Dimana nilai Z_t merupakan nilai transformasi ln, untuk mengembalikan ke nilai aslinya (curah hujan) harus kembalikan dengan exp (Z_t). Model tersebut menunjukkan bahwa curah hujan di Pangkalpinang pada dasarian ke-t dipengaruhi oleh curah hujan pada 1 dasarian sebulmnya, curah hujan pada 36 dasarian sebelumnya, curah hujan pada 37 dasarian sebelumnya, curah hujan pada 38 dasarian sebelumnya, kesalahan prediksi pada 1 dasarian sebelumnya, kesalahan prediksi pada 36 dasarian sebelumnya, kesalahan peramalan pada 37 dasarian sebelumnya dan kesalahan prediksi pada waktu ke-t.

D. Perbandingan Prediksi Curah Hujan Model JST dan SARIMA Terbaik

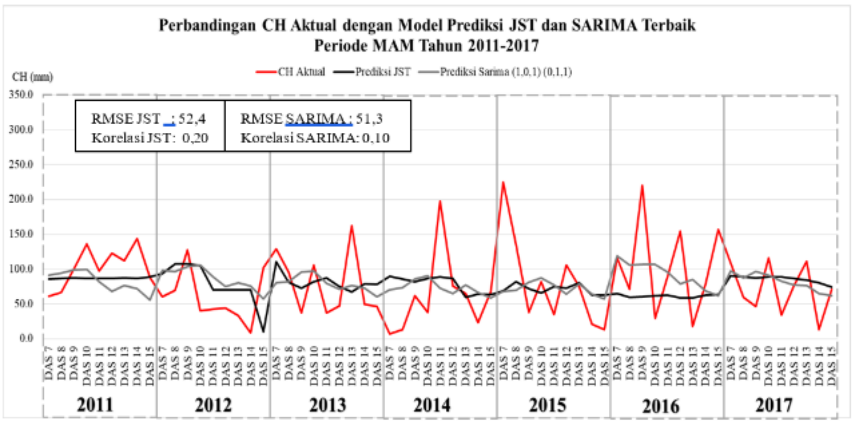
Selanjutnya hasil prediksi curah hujan dari dua model yaitu JST dan SARIMA akan dibandingkan dengan curah hujan aktual periode musiman tahun 2011-2017.



Gambar 4. Perbandingan curah hujan actual dengan model prediksi JST dan SARIMA terbeik periode DJF

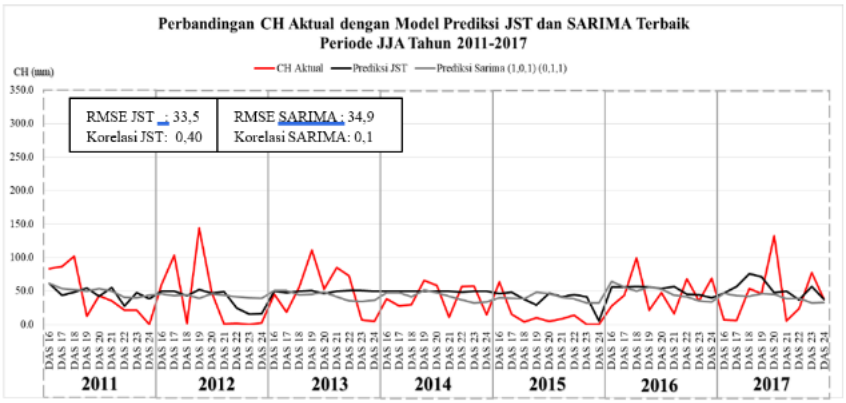
Perbandingan CH aktual dengan model prediksi JST dan SARIMA pada periode DJF menunjukkan bahwa persamaan model JST adalah model persamaan terbaik dalam memprakirakan curah hujan

dasarian di Pangkalpinang, pada periode DJF ditunjukkan dengan nilai korelasi paling tinggi dengan nilai 0,22 dan RMSE paling rendah dengan nilai 66,1.



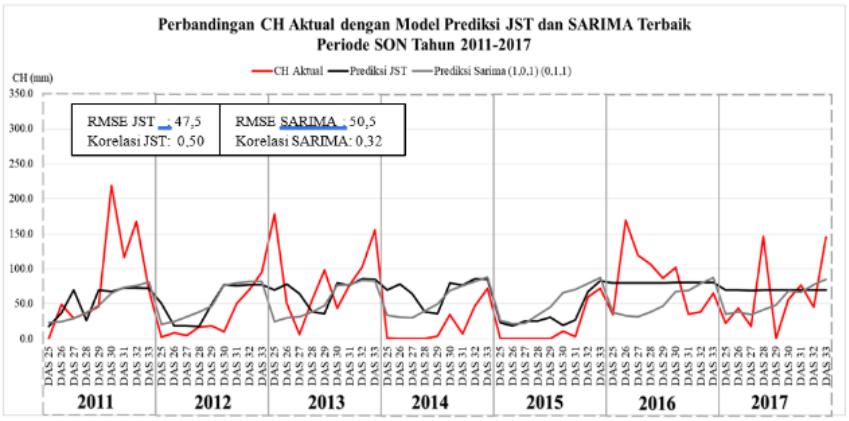
Gambar 5. Perbandingan curah hujan aktual dengan model prediksi JST dan SARIMA terbeik periode MAM

Perbandingan CH aktual dengan model prediksi JST dan SARIMA pada periode MAM menunjukkan bahwa persamaan model JST adalah model persamaan terbaik dalam memprakirakan curah hujan dasarian di Pangkalpinang, pada periode MAM ditunjukkan dengan nilai korelasi paling tinggi yaitu 0,20 dan RMSE yang hampir sama dengan SARIMA yaitu 52,4.



Gambar 6. Perbandingan curah hujan aktual dengan model prediksi JST dan SARIMA terbaik periode JJA

Perbandingan CH aktual dengan model prediksi JST dan SARIMA pada periode JJA (musim kemarau) menunjukkan bahwa persamaan model JST adalah model persamaan terbaik dalam memprakirakan curah hujan dasarian periode JJA di Pangkalpinang, pada periode JJA ditunjukkan dengan nilai korelasi paling tinggi yaitu 0,4 dan RMSE paling rendah yaitu 33,5.



Gambar 7. Perbandingan curah hujan aktual dengan model prediksi JST dan SARIMA terbeik periode SON

Perbandingan CH aktual dengan model prediksi JST dan SARIMA pada periode SON menunjukkan bahwa persamaan model JST adalah model persamaan terbaik dalam memprakirakan curah hujan dasarian periode SON di Pangkalpinang, pada periode SON ditunjukkan dengan nilai korelasi paling tinggi yaitu 0,50 dan RMSE paling rendah yaitu 47,5.

Curah hujan Pangkalpinang dipengaruhi oleh parameter iklim lain seperti suhu muka laut, angin zonal dan *precipitable water* dengan *time lag* yang sesuai dan saling mempengaruhi, sehingga model JST menghasilkan hasil prediksi curah hujan yang lebih baik dibandingkan model persamaan SARIMA terbaik yang hanya menghandalkan data *timeseries* curah hujan masa lampau untuk memprediksi curah hujan dasarian kedepan.

E. Prediksi Awal Musim Hujan (AMH) Berdasarkan Prediksi Model Terbaik

Berdasarkan hasil prediksi curah hujan dengan model JST dan SARIMA $(1,0,1)\ (0,1,1)^{36}$, kemudian dilakukan prediksi AMH tahun 2011-2017 di Pangkalpinang, sebagai berikut:

Tabel 6. Hasil Prediksi AMH Berdasarkan Prediksi Curah Hujan Model Terbaik

Tahun	Prediksi AMH JST	Prediksi SARIMA $(1,0,1)\ (0,1,1)^{36}$	AMH Aktual	Normal AMH
2011	27	28	30	29
2012	30	28	31	29
2013	30	28	25	29
2014	30	28	33	29
2015	31	29	32	29
2016	25	28	24	29
2017	23	28	28	29

Tabel 6 menunjukkan hasil prediksi AMH dengan JST di Pangkalpinang periode tahun 2011-2017, paling awal terjadi pada tahun 2017 yaitu AMH jatuh di dasarian ke-23 (Agustus dasarian ke-2), sedangkan AMH paling lambat terjadi pada tahun 2015 yaitu AMH jatuh di dasarian ke-31

(November dasarian-1). Hasil prediksi AMH dengan SARIMA (1,0,1) $(0,1,1)^{36}$ di Pangkalpinang periode tahun 2011-2017, secara umum jatuh di dasarian 28 (Oktober dasarian ke-1).

Pola antara AMH observasi dan AMH hasil prediksi model JST cenderung mempunyai fluktuasi yang serupa dan terlihat hasil prediksi AMH dengan JST cenderung mengikuti AMH observasi. Akan tetapi, pada terdapat tahun-tahun tertentu yang hasil prediksi dengan JST tidak mengikuti pola observasi. Penyimpangan terbesar dari hasil prediksi AMH dengan JST jika dibandingkan dengan observasinya sebesar 5 dasarian yang terjadi pada tahun 2013 dan 2017. Sedangkan, prediksi AMH dengan SARIMA tidak memiliki pola.

F. Verifikasi Awal Musim Hujan (AMH)

Salah satu upaya untuk mengetahui apakah hasil prediksi awal musim hujan yang telah dilakukan memiliki kualitas baik atau tidak maka dilakukan verifikasi. Untuk memverifikasi apakah hasil suatu prediksi yan telah dilakukan baik atau tidak ialah dengan membandingkan data hasil prediksi dengan data observasi. Metode verifikasi yang digunakan yaitu metode kontingensi. Nilai akurasi dari satu prediksi dinyatakan dalam bentuk persentase dengan istilah "Sesuai Prakiraan" dan "Menyimpang Prakiraan". Berdasarkan hasil verifikasi AMH dengan kontingensi dan definisi kesesuaian prakiraan awal musim dari BMKG dengan rentang kesesuaian +/- 1 dasarian (10 hari) dapat dihitung nilai akurasi untuk "Sesuai Prakiraan" dan "Menyimpang Prakiraan"

Sesuai Prakiraan : $\frac{3}{7} x\ 100\% = 42{,}86\ \%$ dan

Menyimpang Prakiraan : $100\% - 42{,}86\% = 57{,}14\%$

Berdasarkan hasil kesesuaian prakiraan, dapat terlihat bahwa hasil prediksi AMH dengan JST dari 7 kejadian yang mempunyai rentang kesesuaian dengan AMH aktual adalah sebanyak 3 kejadian. Tingkat akurasi hasil verifikasi yang tergolong "sesuai prakiraan" adalah sebanyak 3 kejadian dan yang "menyimpang prakiraan" sebanyak 4 kejadian. Persentase hasil prediksi yang "sesuai prakiraan" sebesar 42,86%. Adapun penjelasan dari sedikitnya kelompok jumlah tahun yang menyatakan "sesuai prakiraan" disebabkan oleh beberapa factor seperti pemilihan prediktor yang mempengaruhi AMH di Pangkalpinang, penggunaan metode untuk memprediksi awal musim hujan yang belum cukup baik untuk memprediksi jika adanya fenomena lokal maupun global yang mempengaruhi awal musim hujan di Pangkalpinang. Selain itu, menggunakan Skill Score untuk mengetahui AMH dikatakan memiliki skill yang lebih baik dibandingkan dengan prediksi menggunakan data normalnya. Hasilnya menunjukkan skill prediksi AMH periode 2011-2017 menggunakan prediksi CH model JST adalah 0,05. (dimana mendekati skill score 1, prediksi mendekati sempurna).

Berdasarkan hasil verifikasi AMH dengan kontingensi dan definisi kesesuaian prakiraan awal musim dari BMKG dengan rentang kesesuaian +/- 1 dasarian (10 hari) dapat dihitung nilai akurasi untuk "Sesuai Prakiraan" dan "Menyimpang Prakiraan"

Sesuai Prakiraan : $\frac{1}{7} x\ 100\% = 14{,}26\ \%$ dan

Menyimpang Prakiraan : $100\% - 14{,}26\% = 85{,}74\%$

Berdasarkan hasil kesesuaian prakiraan dan perhitungan *skill score*, dapat terlihat bahwa hasil prediksi AMH dengan SARIMA (1,0,1) $(0,1,1)^{36}$ dari 7 kejadian yang mempunyai rentang kesesuaian dengan AMH aktual adalah sebanyak 1 kejadian. Tingkat akurasi hasil verifikasi yang tergolong "sesuai prakiraan" adalah sebanyak 1 kejadian dan yang "menyimpang prakiraan"

sebanyak 6 kejadian. Persentase hasil prediksi yang "sesuai prakiraan" sebesar 14,26%. Selain itu, menggunakan *Skill Score* untuk mengetahui AMH dikatakan memiliki *skill* yang lebih baik dibandingkan dengan prediksi menggunakan data normalnya. Hasilnya menunjukkan *skill* prediksi AMH periode 2011-2017 menggunakan prediksi CH model SARIMA $(1,0,1)\ (0,1,1)^{36}$ adalah 0. (dimana mendekati *skill score* 1, prediksi mendekati sempurna).

Berdasarkan kesesuaian prakiraan dan perhitungan *skill score* prediksi AMH dari hasil prediksi curah hujan model JST dan SARIMA $(1,0,1)\ (0,1,1)^{36}$ menunjukkan bahwa hasil prakiraan AMH menggunakan model JST lebih baik daripada SARIMA di kota Pangkalpinang. Hal ini dikarenakan model JST menggunakan prediktor seperti *precipitable water*, angin zonal dan suhu muka laut dibandingkan dengan SARIMA yang hanya menghadalkan data masa lalu untuk melakukan prediksi.

KESIMPULAN

Berdasarkan hasil validasi model menunjukkan model prediksi JST lebih baik daripada model prediksi SARIMA yang hanya menggunakan data *series* curah hujan masa lampau. Akan tetapi model ini belum mampu untuk menduga curah hujan ekstrim. Menggunakan prediksi curah hujan model JST, hasil prediksi AMH di Pangkalpinang tahun 2011-2014 menunjukkan menyimpang dari normal AMH paling kecil adalah 1 dasarian dan paling besar adalah 6 dasarian. Sedangkan menggunakan prediksi curah hujan model SARIMA, hasil prediksi AMH di Pangkalpinang tahun 2011-2014 menunjukkan menyimpang dari normal AMH paling kecil adalah sesuai normal AMH (0 dasarian) dan paling besar adalah 1 dasarian. Verifikasi dari prediksi AMH menggunakan JST di Pangkalpinang menghasilkan kesesuaian model prediksi sebesar 42,86 % dengan nilai skill sebesar 0,05, sedangkan verifikasi dari prediksi AMH menggunakan SARIMA di Pangkalpinnag menghasilkan kesesuaian model prediksi sebesar 14,26 % dengan nilai skill sebesar 0,00.

DAFTAR PUSTAKA

[1] Kementerian Pertanian, *Statistik Perkebunan Indonesia 2015-2017 Lada*. Jakarta: Direktorat Jenderal Perkebunan, 2016.

[2] T. H. S. Siregar and I. Suhendry, *Budidaya dan Teknologi Karet*. Bogor: Kanisius, 2013.

[3] Badan Meteorologi Klimatologi dan Geofisika, *Prakiraan Musim Hujan 2013/2014 di Indonesia*. Jakarta: BMKG, 2013.

[4] A. N. Khoir, "Prediksi Curah Hujan Bulanan di Pangkalpinang dengan Prediktor Precitable Water dan Angin Zonal/ Meridional," Sekolah Tinggi Meteorologi Klimatologi dan Geofisika, Tanggerang Selatan, 2018.

[5] L. Lubis and A. Buono, "Pemodelan Jaringan Syaraf Tiruan untuk Memprediksi Awal Musim Hujan Berdasarkan Suhu Permukaan Laut," *Jurnal Ilmu Komputer dan Agri-Informatika*, vol. 1, p. 52, Nov. 2012, doi: 10.29244/jika.1.2.52-61.

[6] W. Wei, *Time Series Analysis: Univariate and Multivariate Methods, 2nd edition, 2006*. 2006.

[7] R. Ristiana, "Perbandingan Arima Dan Jaringan Syaraf Tiruan Propagasi Balik Dalam Peramalan Tingkat Inflasi Nasional," Institut Pertanian Bogor, Bogor, 2015.

[8] Y. Andriani, H. Silitonga, and A. Wanto, "Analisis Jaringan Syaraf Tiruan untuk prediksi volume ekspor dan impor migas di Indonesia," *Register: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 4, p. 30, Nov. 2018, doi: 10.26594/register.v4i1.1157.

[9] A. Singh and G. C. Mishra, "Application of Box-Jenkins Method and Artificial Neural Network Procedure for Time Series Forecasting of Prices," *Statistics in Transition*, vol. 16, pp. 83–96, Mar. 2015, doi: 10.21307/stattrans-2015-005.

[10] S. Rumagit and A. SN, "Prediksi Pemakaian Listrik Kelompok Tarif Menggunakan Jaringan Syaraf Tiruan dan ARIMA," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 7, p. 189, Jul. 2013, doi: 10.22146/ijccs.3359.

[11] Bowerman B. L. and O'Connell R. T., *Forecasting and time series : an applied approach (3rd ed.)*. Duxbury Press, 1993.

[12] J. D. Cryer and K.-S. Chan, *Time Series Analysis*. New York, NY: Springer New York, 2008. doi: 10.1007/978-0-387-75959-3.

[13] T. Sutojo, E. Mulyanto, and V. Suhartono, *Kecerdasan Buatan*. Yogyakarta: ANDI, 2010.

[14] L. Fausett, *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*. USA: Prentice-Hall, Inc., 1994.

[15] J. Heaton, "Programming Neural Networks with Encog 2 in Java," 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:64015064

[16] M. Mislan, H. H., S. Hardwinarto, S. Soeparto, and M. Aipassa, *Rainfall Monthly Prediction Based on Artificial Neural Network – A Case Study Tenggarong Station, East Kalimantan – Indonesia*, vol. 59. 2015. doi: 10.1016/j.procs.2015.07.528.

[17] Walpole, *Pengantar Statistika*, 3rd ed. Jakarta: PT Gramedia Pustaka Utama, 1995.

[18] E. Aldrian and R. Susanto, "Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature," *Int. J. Climatol.*, vol. 23, pp. 1435–1452, Oct. 2003, doi: 10.1002/joc.950.

Analisis Faktor-Faktor Yang Mempengaruhi Indeks Pembangunan Manusia Berdasarkan Kabupaten/Kota Di Jawa Tengah

Nur Huriyatullah Rona Nabila(1), Yulia Fitri(2), Prizka Rismawati Arum(3)*

[1,2,3]Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam,

Universitas Muhammadiyah Semarang

Jl. Kedungmundu No.18, Kedungmundu, Kec. Tembalang, Kota Semarang, Jawa Tengah 50273

e-mail: nurhuriyatullah@gmail.com(1), yuliafitri0307@gmail.com(2),
prizka.rismawatiarum@unimus.ac.id(3)*

ABSTRAK

Pembangunan merupakan indikator yang penting disuatu negara terutama negara berkembang seperti Indonesia. Pembangunan manusia merupakan salah satu upaya yang dilakukan oleh pemerintah guna mewujudkan masyarakat yang Makmur dan sejahtera. Salah satu cara untuk mengukur kesejahteraan suatu daerah yaitu dengan mengukur Indeks Pembangunan Manusia (IPM) daerah tersebut. Dimensi dari IPM sendiri yaitu umur Panjang dan hidup sehat, Pendidikan, dan kehidupan yang layak. Data yang digunakan pada penelitian ini yaitu data sekunder yang bersumber dari Badan Pusat Statistik berupa data Indeks Pembangunan Manusia di kabupaten/kota di Jawa Tengah. Melihat pentingnya IPM pada suatu daerah khusunya Jawa Tengah, maka perlu dilakukannya analisis mengenai faktor-faktor yang mempengaruhi IPM. Analisis regresi linier berganda merupakan salah satu metode penelitian yang dapat diterapkan dalam penelitian ini karena metode regresi adalah teknik analisis statistik yang digunakan untuk menentukan bagaimana dua atau lebih variabel berinteraksi. Sehingga, analisis regresi linear berganda dapat digunakan untuk melihat faktor-faktor yang mempengaruhi IPM kabupaten/kota di Jawa Tengah pada tahun 2022. Dari hasil penelitian, diketahui bahwa variabel harapan hidup, harapan lama sekolah, rata-rata lama sekolah, dan pengeluaran perkapita berpengaruh signifikan terhadap indeks pembangunan manusia pada tahun 2022. Hasil pengujian koefisien determinasi atau R-Square didapatkan nilai sebesar 99,9%.

Kata kunci : Indeks Pembangunan Manusia, Jawa Tengah, Regresi Linear Berganda

ABSTRACT

Development is an important indicator in a country, especially a developing country like Indonesia. Human development is one of the efforts made by the government to create a prosperous and prosperous society. One way to measure the welfare of an area is by measuring the Human Development Index (IPM) of that area. The dimensions of HDI itself are longevity and healthy living, education and a decent life. The data used in this study are secondary data sourced from the Central Bureau of Statistics in the form of Human Development Index data in regencies/cities in Central Java. Seeing the importance of HDI in an area especially Central Java, it is necessary to do an analysis of the factors that influence HDI. Multiple linear regression analysis is one of the research methods that can be applied in this study because the regression method is a statistical analysis technique used to determine how two or more variables interact. Thus, multiple linear regression analysis can be used to look at the factors that influence district/city HDI in Central Java in 2022. From the research results, it is known that the variables life expectancy, expected length of schooling, average length of schooling, and per capita expenditure have an effect significant to the human development index in 2022. The results of the test for the coefficient of determination or R-Square obtained a value of 99.9%.

Keywords: *Human Development Index, Central Java, Multiple Linear Regression*

PENDAHULUAN

Pembangunan ekonomi merupakan upaya untuk memperbaiki struktur ekonomi dengan tujuan meningkatkan lapangan kerja dan produktivitas, yang pada akhirnya akan meningkatkan pendapatan rata-rata penduduk. Selain mengejar pertumbuhan ekonomi yang lebih cepat, mengatasi ketimpangan regional, mengurangi kemiskinan, dan menurunkan tingkat pengangguran, pembangunan ekonomi adalah proses multifaset yang memasukkan beberapa perubahan mendasar pada struktur sosial, sikap masyarakat, dan institusi nasional [1]. Pembangunan marupakan indikator yang penting disuatu negara terutama negara berkembang seperti Indonesia. Pembangunan manusia merupakan salah satu upaya yang dilakukan oleh pemerintah guna mewujudkan masyarakat yang Makmur dan sejahtera [2]. Salah satu cara untuk mengukur kesejahteraan suatu daerah yaitu dengan mengukur Indeks Pembangunan Manusia (IPM) daerah tersebut. Indeks Pembangunan Manusia dapat menentukan level capaian pembangunan suatu daerah/negara. Indeks Pembangunan Manusia merupakan informasi penting bagi Indonesia, yang digunakan untuk mengukur kinerja pemerintah dan mengalokasikan sumber daya untuk Dana Alokasi Umum (DAU).

Indek Pembangunan Manusia merupakan indeks yang digunakan untuk pencapaian hasil pembangunan suatu wilayah yang terdiri dari tiga dimensi dasar pembangunan diantaranya yaitu lama hidup, tingkat Pendidikan dan standar hidup yang layak (Badan Pusat Statistik, 2022a). Indeks pembangunan manusia merupakan indeks fundamental yang mencakup komponen-komponen sebagai berikut: (1) Umur panjang dan kesehatan yang diukur dengan indikator angka harapan hidup; (2) pengetahuan, yang diukur dengan Harapan Lama Sekolah (HLS) dan Rata-Rata Lama Sekolah (RLS); dan (3) taraf hidup yang layak, yang diukur dengan indikator pengeluaran per kapita yang disesuaikan [3]. Dibandingkan dengan negara lain, Indeks Pembangunan Manusia Indonesia masih cukup rendah. Data Badan Pusat Statistik (2018) menunjukkan bahwa, terdapat perlambatan pertumbuhan Indeks Pembangunan Manusia (IPM) pada periode 2016-2018 dibandingkan tahun-tahun sebelumnya meskipun secara nominal tetap tumbuh secara positif. Indeks Pembangunan Manusia (IPM) Indonesia tahun 2021 mencapai 72,29, meningkat 0,35 poin (0,49%) dibandingkan capaian tahun sebelumnya (71,94). Selama 2010–2021, IPM Indonesia rata-rata meningkat sebesar 0,76% [4]

Berdasarkan data yang dikeluarkan oleh Badan Pusat Statistik Jawa Tengah, nilai Indeks Pembangunan Manusia di Jawa Tengah pada tahun 2021 yaitu sebesar 72,16 naik menjadi 72,79 pada tahun 2022. Badan Pusat Statistik mencatat Indeks Pembangunan di Jawa Tengah terus mengalami peningkatan setiap tahunnya. Meskipun IPM Jawa Tengah sudah tergolong tinggi, namun IPM Jawa Tengah masih tertinggal jika dibandingkan dengan provinsi lain seperti Jakarta yang memiliki nilai IPM 81,65 dan DI Yogyakarta dengan nilai IPM 80,64 [5]. Melihat pentingnya Indeks Pembangunan Manusia pada suatu daerah khusunya Jawa Tengah, maka perlu dilakukannya analisis mengenai Indeks Pembangunan Manusia. Analisis dapat dilakukan secara statistika dengan melihat faktor-faktor apa saja yang berpengaruh secara signifikan terhadap IPM. Salah satu metode statistika yang dapat digunakan sesuai tujuan penelitian ini adalah analisis regresi linier berganda. Regresi adalah teknik analisis statistik yang digunakan untuk menentukan bagaimana dua atau lebih variabel berinteraksi. hubungan antara variabel dependen dan satu atau lebih variabel independen, model regresi merupakan bagian penting dari analisis data. Untuk mengetahui seberapa besar

pengaruh variabel independen terhadap variabel dependen merupakan tujuan dari analisis regresi [6]

Analisis regresi berganda adalah pengembangan dari analisis regresi sederhana dengan lebih dari satu variabel bebas X. Pada regresi sederhana jumlah variabel bebas yang digunakan untuk memprediksi variabel terikat hanya satu, sedangkan pada regresi linear berganda variabel bebas yang digunakan lebih dari satu hal inilah yang membedakan antara regresi linear sederhana dengan regresi linear berganda [7]. Dalam regresi linear berganda, uji hipotesis berperan penting dalam mengevaluasi signifikansi hubungan antara variabel-variabel independen dan variabel dependen dalam model. Uji hipotesis membantu kita untuk memahami apakah hubungan yang diamati dalam model regresi memiliki dasar yang kuat secara statistik atau hanya merupakan hasil kebetulan. Uji Hipotesis sendiri yaitu bagian dari Statistika Inferensial yang berfokus pada pengujian validitas suatu pernyataan melalui pendekatan statistik, dengan tujuan untuk menyimpulkan apakah pernyataan tersebut dapat diterima atau ditolak secara statistik [8]. Secara keseluruhan, uji hipotesis dalam regresi linear berganda membantu kita dalam memvalidasi model dan interpretasi hasilnya. Ini memberikan dasar statistik yang kuat untuk menyimpulkan apakah variabel independen dalam model secara signifikan mempengaruhi variabel dependen. Dengan melakukan analisis regresi linear berganda maka akan dapat diketahui faktor apa saja yang mempengaruhi Indeks Pembangunan Manusia berdasarkan Kabupaten/Kota di Jawa Tengah pada tahun 2022.

METODE PENELITIAN

Struktur Data

Data yang digunakan dalam penelitian ini adalah data sekunder. Variabel yang digunakan adalah Indeks Pembangunan Manusia (Y), Angka Harapan Hidup (X1), Harapan Lama Sekolah (X2), Rata-Rata Lama Sekolah (X3), dan Pengeluaran Per Kapita (X4) menurut kabupaten/kota di Jawa Tengah Provinsi pada tahun 2022. Variabel-variabel tersebut terdiri dari satu variabel bebas dan empat variabel terikat.

Sumber Data

Sumber data dalam penelitian ini adalah data yang berasal dari Badan Pusat Statistik Provinsi Jawa Tengah.

Tahapan melakukan Regresi Linear Berganda

a. Uji asumsi yang harus terpenuhi dalam melakukan regresi linear berganda [9]:
   1. Uji Normalitas Residual

      Untuk mengetahui apakah nilai residual pada regresi yang diteliti berdistribusi normal atau tidak, digunakan uji normalitas. Nilai signifikansi dapat digunakan untuk menentukan apakah data terdistribusi secara normal atau tidak, jika lebih besar atau sama dengan 0,05 maka data dianggap berdistribusi normal. Adapun hipotesis ujinya yaitu sebagai berikut [10]:

      H_0 : Residual berdistribusi normal

      H_1 : Residual tidak berdistribusi normal
   2. Uji Autokorelasi

Uji autokorelasi digunakan untuk menilai apakah adanya korelasi antar anggota serangkaian dan observasi yang diuraikan menurut waktu (time series) atau ruang (cross section). Cara untuk mengetahui suatu data terjadi korelasi, salah satunya dengan melakukan uji durbin Watson. Adapun hipotesis ujinya yaitu sebagai berikut [10]:

H_0 : Tidak terdapat autokorelasi

H_1 : Terdapat autokorelasi

3. Uji Heteroskedastisitas

Tujuan dari uji heteroskedastisitas adalah untuk mengidentifikasi adanya perbedaan varian residual antar observasi. Uji heteroskedastisitas memeriksa varians yang tidak merata antara setiap residual dan setiap pengamatan. Homoskedastisitas mengacu pada konsistensi varians dari residual satu pengamatan ke residual pengamatan lain, sedangkan heteroskedastisitas mengacu pada perbedaan varians dari residual satu pengamatan ke residual pengamatan lain [11]

4. Uji Multikoloniaritas

Uji multikoloniaritas bertujuan untuk menegetahui apakah model regresi terdapat korelasi antar variabel independent atau tidak.

$$VIF_j = \frac{1}{(1 - {R_j}^2)}$$

Adapun hipotesis ujinya yaitu sebagai berikut [10]:

H_0 : Tidak terdapat multikolinieritas

H_1 : Terdapat multikolinieritas

Kriteria pengambilan keputusan pada uji multikolinearitas, jika nilai VIF<10, maka tidak terjadi multikolinearitas, sementara jika nilai VIF > 10 menunjukan bahwa terjadi multikolineritas yang kuat.

b. Pengujian kelayakan model terdiri dari [12]:

1. Uji F (secara simultan)

Uji F adalah alat statistik yang terkait erat dengan analisis varians (ANOVA), dimana ANOVA adalah salah satu teknik pengujian yang sering digunakan dalam statistik, karena pendekatan ini memungkinkan para peneliti untuk mengevaluasi perbedaan antara beberapa rata-rata populasi dengan membandingkan variasi di antaranya [13]. Uji f bertujuan untuk memastikan apakah masing-masing variabel independen mempengaruhi variabel dependen secara bersama-sama atau secara simultan. Jika F hitung> F tabel H0 ditolak dan Ha diterima artinya mempunyai pengaruh yang signifikan dan begitu sebaliknya.

2. Uji T (secara Parsial)

Uji t digunakan untuk menilai apakah variabel independen secara parsial mempengaruhi variabel dependen dengan cara yang signifikan secara statistik. Hal ini berdampak besar jika t hitung > t tabel, maka H0 ditolak dan Ha disetujui, begitu pula sebaliknya.

3. Koefisien Determinasi (R2)

Sejauh mana suatu model dapat menjelaskan variabel dependen diukur dengan koefisien determinasi atau goodness of fit. Nilai koefisien determinasi antara 0 dan 1.

$$R^2 \frac{SSR}{Syy}$$

HASIL DAN PEMBAHASAN

Data

Data yang digunakan dalam penelitian ini adalah data IPM (Y), Harapan Hidup (X1), Harapan lama sekolah (X2), Rata-rata Lama Sekolah (X3), dan Pengeluaran Perkapita (X4) Jawa Tengah tahun 2022 [14]. Dalam melakukan penelitian ini, penulis menggunakan data sekunder dari Badan Pusat Statistik Jawa Tengah. Data sekunder adalah data yang dikumpulkan oleh pihak lain atau perantara yang mengumpulkan data tersebut. Penelitian ini menggunakan metode regresi linear berganda.

Proses Regresi Linear Berganda

Proses regresi linear berganda, yaitu sebagai berikut:

Tabel 1 Data Penelitian

No	Kab/Kota	Y	X1	X2	X3	X4
1	Kabupaten Cilacap	70.99	74.07	12.66	7.18	10904
2	Kabupaten Banyumas	73.17	73.88	13.21	7.78	11905
3	Kabupaten Purbalingga	69.54	73.28	12.01	7.33	10277
4	Kabupaten Banjarnegara	68.61	74.37	11.81	6.84	9776
5	Kabupaten Kebumen	70.79	73.7	13.36	7.85	9282
6	Kabupaten Purworejo	73.6	75.03	13.52	8.32	10671
7	Kabupaten Wonosobo	68.89	72.05	11.78	6.88	11108
8	Kabupaten Magelang	70.85	74.03	12.58	7.81	10011
9	Kabupaten Boyolali	74.97	76.12	12.62	8.08	13250
10	Kabupaten Klaten	76.95	76.95	13.4	9.09	12522
11	Kabupaten Sukoharjo	77.94	77.82	13.9	9.62	11841
12	Kabupaten Wonogiri	71.04	76.41	12.51	7.42	9780
13	Kabupaten Karanganyar	76.58	77.64	13.7	8.79	11798
14	Kabupaten Sragen	74.65	75.87	12.91	7.79	13052
15	Kabupaten Grobogan	70.97	74.93	12.45	7.26	10610
16	Kabupaten Blora	69.95	74.6	12.44	7.01	10067
17	Kabupaten Rembang	71	74.68	12.13	7.41	10937
18	Kabupaten Pati	73.14	76.32	12.95	7.79	10948
19	Kabupaten Kudus	75.89	76.76	13.25	9.06	11609
20	Kabupaten Jepara	73.15	75.97	12.77	8.09	10913
21	Kabupaten Demak	73.36	75.52	13.33	8.1	10698
22	Kabupaten Semarang	74.67	75.86	13.04	8.05	12448
23	Kabupaten Temanggung	70.77	75.7	12.55	7.41	9773
24	Kabupaten Kendal	73.19	74.53	12.97	7.71	11999
25	Kabupaten Batang	69.45	74.79	12.14	6.9	9972

26	Kabupaten Pekalongan	70.81	73.8	12.43	7.46	10707
27	Kabupaten Pemalang	67.19	73.65	11.98	6.5	8994
28	Kabupaten Tegal	69.53	71.85	12.91	7.25	10020
29	Kabupaten Brebes	67.03	69.74	12.15	6.35	10514
30	Kota Magelang	80.39	77.02	14.31	10.94	12816
31	Kota Surakarta	83.08	77.43	14.89	10.92	15463
32	Kota Salatiga	84.35	77.72	15.43	10.95	16351
33	Kota Semarang	84.08	77.69	15.54	10.8	16047
34	Kota Pekalongan	75.9	74.51	12.86	9.2	13158
35	Kota Tegal	76.15	74.64	13.08	9	13455

Statistik Deskriptif

Analisis deskriptif adalah metode analisis yang melibatkan pengumpulan, pengolahan, penyajian, dan analisis data kuantitatif dengan tujuan memberikan gambaran yang terperinci tentang suatu peristiwa. Dalam proses ini, data disajikan dalam bentuk tabel dan grafik, menghasilkan pemahaman yang sistematis. Analisis deskriptif bertujuan untuk memberikan informasi yang rinci tentang data yang ada, tanpa niat untuk menguji atau merumuskan kesimpulan [15].

Tabel 2 Statistik Deskriptif

Statistik	**IPM**	**AHH**	**HLS**	**RLS**	**PPP**
Minimum	67,03	69,74	11,78	6,35	8994
Maximum	84,35	77,82	15,54	10,95	16351
Mean	73,50	75,11	13,016	8,14	11533,6
Std. Deviation	4,46	1,82	0,92	1,27	1813,54
N	35	35	35	35	

Berdasarkan Tabel 2. Dapat dilihat bahwa Mean Indeks Pembangunan Manusia yaitu sebesar 73,5% dengan nilai minimum sebesar 67,03% dan nilai maksimum sebesar 84,35%. Mean Angka Harapan Hidup yaitu sebesar 75,11% dengan nilai minimum sebesar 69,74% dan nilai maksimum sebesar 77,82%. Mean Harapan Lama Sekolah yaitu sebesar 13,016% dengan nilai minimum sebesar 11,78% dan nilai maksimum sebesar 15,54%. Mean Rata-Rata Lama Sekolah yaitu sebesar 8,14% dengan nilai minimum sebesar 6,35% dan nilai maksimum sebesar 10,95%. Mean Pengeluaran Perkapita yaitu sebesar 11533,6% dengan nilai minimum sebesar 8994% dan nilai maksimum sebesar 16351%.

Uji Asumsi

a. Normalitas

Tabel 3 Uji Normalitas

	Unstandardized Residual
Test Statistic	.121

Asymp. Sig. (2-tailed)	.200[c]

Asimp. Sig. (2-tailed) sebesar 0,200 ditemukan berdasarkan hasil uji Normalitas dengan menggunakan One-Sample Kolmogorov Smirnov. Hal ini menunjukkan bahwa nilai probabilitas atau signifikansi lebih besar dari 0,05 yang menunjukkan distribusi data normal.

b. Autokorelasi

Tabel 4 Uji Autokorelasi

Std.Error of the Estimate	DW
0,16350	1.894

Hasil output menunjukkan bahwa nilai DW adalah 1,750, berada dalam kisaran 1,65 < DW < 2,37, menunjukkan tidak ada autokorelasi pada data.

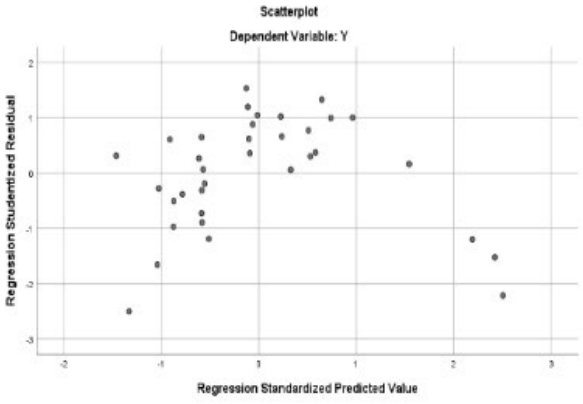
c. Multikolinieritas

Tabel 5 Uji Multikolinieritas

	X1	X2	X3	X4
VIF	2.206	6.257	9.260	3.341
Tolerance	0.453	0.160	0.108	0,299

Dari hasil output multikolinieritas didapatkan nilai VIF pada variabel X1 sebesar 2,206; X2 sebesar X3 sebesar 9,260; dan X4 sebesar 3,341. Karena nilai VIF untuk masing-masing variabel tersebut lebih kecil dari 10, maka dapat disimpulkan bahwa data tersebut tidak terjadi multikolinieritas.

d. Heteroskedastisitas



Gambar 1. Scatterplote Heterokedastisitas

Berdasarkan Grafik output Scatterplot tersebut, terlihat titik-titik data penyebar diatas dan di bawah atau di sekitar angka 0 dan tidak membentuk suatu pola tertentu dengan jelas. Hal ini berarti tidak terjadi heteroskedastisitas pada model regresi sehingga model regresi ini layak dipakai untuk memprediksi pengaruh variabel bedasarkan masukan variabel independennya

Uji Kelayakan Model

a. Uji F (uji secara simultan)

Tabel 6 Uji F

F	p-value	df
63.819	0,000	5

Bersadarkan output, didapatkan nilai F sebesar 63.819 dengan nilai P-Value sebesar 0,000 (p<0,05) artinya variabel harapan hidup (X1), harapan lama sekolah (X2), rata-rata lama sekolah (X3), dan pengeluaran perkapita (X4) berpengaruh secara signifikan terhadap variabel indeks pembangunan manusia (Y).

b. Uji T (uji secara parsial)

Tabel 7 Uji T

	X1	X2	X3	X4
T	20.227	11.767	19.949	29.472
sig	0.000	0.000	0.000	0,000

Bersadarkan output uji t, didapatkan nilai signifikansi untuk variabel harapan hidup (X1), harapan lama sekolah (X2), rata-rata lama sekolah (X3), dan pengeluaran perkapita (X4) sebesar 0,000 (p<0,05) artinya variabel tersebut secara negative berpengaruh signifikan terhadap indeks pembangunan manusia pada tahun 2022.

c. Koefisien Determinasi (R-Square)

Tabel 8 Uji Koefisien Determinasi

R Square	Adjusted R square
0.999	0. 999

Berdasarkan hasil pengujian koefisien determinasi atau R-Square adalah 0,999 yang berarti bahwa variabel harapan hidup (X1), harapan lama sekolah (X2), rata-rata lama sekolah (X3), dan pengeluaran perkapita (X4) mampu menjelaskan data sebesar 99,9% sedangkan 0,01% sisanya dijelaskan oleh variabel lain. Karena hasilnya sangat mendekati 1, maka dapat diasumsikan bahwa

variabel independen memiliki pengaruh yang signifikan terhadap variabel dependen, yang menunjukkan bahwa model tersebut efektif dalam menjelaskan pengaruh variabel tersebut.

Uji Regresi Linier Berganda

Karena dalam penelitian ini terdapat lima variabel (termasuk Y), digunakan regresi linier berganda. Dalam regresi linier berganda, X adalah variabel bebas dan Y adalah variabel terikat. Model regresi dapat dibuat sebagai berikut dengan menggunakan hasil uji regresi yang telah dilakukan.

$$\mathbf{Y= 6{,}550+0{,}463\ X_1+0{,}897\ X_2 +1{,}341\ X_3 + 0{,}001\ X_4}$$

Nilai konstanta β0= 6,550 artinya bila seluruh variabel independent yaitu X1, X2, X3, X4 disesuaikan diamsumsikan memiliki koefidien nol (konstan) maka nilai kemiskinan sebesar 6,550. Nilai koefisien regresi variabel harapan hidup β1=0,463 artinya jika IPM (Y) mengalami kenaikan 1% maka harapan hidup mengalami peningkatan 0,463. Koefisien bernilai positif artinya terjadinya hubungan positif antara harapan hidup dengan IPM. Nilai Koefisien regresi variabel Angka Harapan Lama sekolah β2= 0,897 artinya jika harapan lama sekolah mengalami kenaikan 1% maka harapan lama sekolah meningkat 0,897. Nilai koefisien regresi variabel Rata-rata Lama Sekolah disesuaikan β3= 1,341 artinya jika Rata-rata Lama Sekolah disesuiakan mengalami kenaikan 1% maka Rata-rata Lama sekolah 1,341. koefisien bernilai positif artinya terjadi hubungan antra rata-rata lama sekolah dan IP. Nilai koefisien regresi variabel pengeluaran perkapita disesuaikan β4= 0,001 artinya jika pengeluaran perkapita disesuiakan mengalami kenaikan 1% maka pengeluaran perkapita 0,001. koefisien bernilai positif artinya terjadi hubungan positif antara pengeluaran perkapita dan Indeks Pembangunan Manusia.

KESIMPULAN

Variabel harapan hidup (X_1), harapan lama sekolah (X_2), rata-rata lama sekolah (X_3), dan pengeluaran perkapita (X_4) berpengaruh signifikan terhadap indeks pembangunan manusia pada tahun 2022. Pada hasil pengujian koefisien determinasi atau R-Square didapatkan hasil sebesar 0,999 yang berarti bahwa variabel-variabel independen yang berpengaruh signifikan mampu menjelaskan variabel dependen sebesar 99,9% sedangkan 0,01% sisanya dijelaskan oleh variabel lain. Karena hasilnya sangat mendekati 1, maka dapat diasumsikan bahwa model tersebut efektif dalam menjelaskan pengaruh variabel independen terhadap variabel dependen

DAFTAR PUSTAKA

[1] D. Mahroji and I. Nurkhasanah, "Pengaruh Indeks Pembangunan Manusia Terhadap Tingkat Pengangguran Di Provinsi Banten," *J. Ekon.*, vol. 9, no. 1, 2019, doi: 10.35448/jequ.v9i1.5436.

[2] A. Melliana and I. Zain, "Indeks Pembangunan Manusia di Kabupaten / Kota Provinsi Jawa Timur dengan Menggunakan Regresi Panel," *J. Sains Dan Seni Pomits*, vol. 2, no. 2, pp. 237–242, 2013, [Online]. Available: http://dx.doi.org/10.12962/j23373520.v2i2.4844

[3] E. Yektiningsih, "Analisis Indeks Pembangunan Manusia (Ipm) Kabupaten Pacitan Tahun 2018," *J. Ilm. Sosio Agribis*, vol. 18, no. 2, pp. 32–50, 2018, doi: 10.30742/jisa1822018528.

[4] Badan Pusat Statistik, "Berita Resmi Statistik," *Badan Pus. Stat.*, 2021.
[5] Badan Pusat Statistik, "Indeks Pembangunan Manusia," *Badan Pusat Statistik*, 2022. https://jateng.bps.go.id/site/resultTab (accessed Dec. 25, 2022).
[6] Galton, "Regresi linier berganda 1.," pp. 1–6, 2004.
[7] B. A. Wisudaningsi, I. Arofah, and K. A. Belang, "Pengaruh Kualitas Pelayanan Dan Kualitas Produk Terhadap Kepuasan Konsumen Dengan Menggunakan Metode Analisis Regresi Linear Berganda," *Statmat J. Stat. Dan Mat.*, vol. 1, no. 1, pp. 103–117, 2019, doi: 10.32493/sm.v1i1.2377.
[8] G. Anuraga, A. Indrasetianingsih, and M. Athoillah, "Pelatihan Pengujian Hipotesis Statistika Dasar dengan Software R," *Budimas J. Pengabdi. Masy.*, vol. 3, no. 2, pp. 327–334, 2021, doi: 10.29040/budimas.v3i2.2412.
[9] G.- Mardiatmoko, "Pentingnya Uji Asumsi Klasik Pada Analisis Regresi Linier Berganda," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 14, no. 3, pp. 333–342, 2020, doi: 10.30598/barekengvol14iss3pp333-342.
[10] S. Alfiani and P. R. Arum, "Pemodelan Pertumbuhan Ekonomi di Jawa Barat Menggunakan Metode Geographically Weighted," *J Stat.*, vol. 15, no. 2, pp. 219–227, 2022.
[11] Setiawati, "Analisis Pengaruh Kebijakan Deviden Terhadap Nilai Perusahaan Pada Perusahaan Farmasi di BEI," *J. Inov. Penelit.*, vol. 1, no. 8, pp. 1581–1590, 2021.
[12] Y. Fitri, P. R. Arum, and A. Imron, "Pengaruh Rata-Rata Lama Sekolah, Angka Harapan Hidup Dan Pengangguran Terhadap Kemiskinan Di Kabupaten Solok Selatan," vol. 1, no. 1, pp. 27–33, 2023, [Online]. Available: http://journalnew.unimus.ac.id/index.php/jodi
[13] E. M. P. Hermanto, M. Athoilllah, W. N. Hamidah, and D. P. Putra, "Pelatihan Penggunaan Software R Untuk Menguji Perbandingan Berganda dan Asumsi Residual pada Rancangan Percobaan," vol. 71, no. 1, pp. 63–71, 2021.
[14] Badan Pusat Statistik, "Indeks Pembangunan Manusia," *Badan Pusat Statistik*, 2022. https://jateng.bps.go.id/subject/26/indeks-pembangunan-manusia.html#subjekViewTab1 (accessed Dec. 25, 2022).
[15] I. N. Azizah, P. R. Arum, and R. Wasono, "Model Terbaik Uji Multikolinearitas untuk Analisis Faktor-Faktor yang Mempengaruhi Produksi Padi di Kabupaten Blora Tahun 2020," *Pros. Semin. Nas. UNIMUS*, vol. 4, p. 63, 2021.

Analisis Faktor-Faktor Yang Mempengaruhi Anemia Pada Ibu Hamil Menggunakan CART

Atika Nurani Ambarwati [(1)], Naulia Fadilah [(2)], Safa'at Yulianto [(3)]

Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang.

Jl. Prof. Dr. Hamka Km 01 No. 17 Tambakaji Ngaliyan

e-mail: atika.nurani@gmail.com[(1)], naulia.fadilah.std@itesa.ac.id[(2)], safaatyulianto@yahoo.com[(3)]

ABSTRAK

Anemia selama kehamilan berdampak buruk bagi kesehatan ibu dan bayinya dan merupakan penyebab penting yang melatarbelakangi kejadian morbiditas dan mortalitas. Cikedal merupakan salah satu kecamatan yang ada di Kabupaten Pandeglang Provinsi Banten dimana angka KEK atau Kekurangan Energi Kronik pada ibu hamil terus meningkat dari tahun ke tahun. Metode yang digunakan untuk meneliti ini adalah metode CART. Dengan perolehan sampel menggunakan metode Cluster Random Sampling dan penentuan jumlah sampel dengan metode slovin, hingga diperoleh 106 responden yang mewaikili setiap desa yang ada di Kecamatan Cikedal. Berdasarkan analisis dan pembahasan, didapatkan kesimpulan bahwa variabel-variabel yang mempengaruhi anemia pada ibu hamil meliputi variabel (X_1) usia ibu ketika mengandung atau hamil, (X_2) usia kehamilan, (X_3) paritas, (X_4) jarak kehamilan antar anak, (X_6) status KEK, (X_7) pendidikan ibu hamil, (X_8) pemahaman ibu hamil, dan variabel (X_9) kondisi ibu sebelum mengandung atau pra kehamilan dengan ketepatan prediksi hasil klasifikasi sebesar 90,57%.

Kata kunci : Anemia, *Classification and Regression Tree*, Ibu Hamil

ABSTRACT

Anemia during pregnancy has a negative impact on the health of the mother and her baby and is an important underlying cause of morbidity and mortality. Cikedal is one of the sub-districts in Pandeglang Regency, Banten Province, where the number of KEK or Chronic Energy Deficiency in pregnant women continues to increase from year to year. The method used to research this is the CART method. By obtaining samples using the Cluster Random Sampling method and determining the number of samples using the slovin method, 106 respondents were obtained representing each village in Cikedal District. Based on the analysis and discussion, it was concluded that the variables that affect anemia in pregnant women include variables (X_1) the age of the mother when pregnant or pregnant, (X_2) gestational age, (X_3) parity, (X_4) the distance between pregnancies between children, (X_6) SEZ status, (X_7) education of pregnant women, (X_8) understanding of pregnant women, and variable (X_9) condition of the mother before pregnancy or pre-pregnancy with a prediction accuracy of the classification results of 90.57%.

Keywords: *Anemia, Classification and Regression Tree, Pregnant Women*

PENDAHULUAN

Kesehatan merupakan kegiatan yang dilakukan secara terpadu, terintegrasi, dan berkesinambungan untuk memelihara serta meningkatkan derajat kesehatan masyarakat dalam bentuk pencegahan penyakit, dan pemulihan kesehatan oleh pemerintah dan masyarakat. Indikator derajat kesehatan antara lain Angka Kematian Ibu (AKI) dan Angka Kematian Bayi (AKB) [1]. Angka Kematian Ibu merupakan salah satu tolak ukur untuk menilai sejauh mana ketercapaian kesejahteraan rakyat sebagai hasil dari pelaksanaan pembangunan bidang kesehatan [2]. Anemia dalam kehamilan merupakan masalah kesehatan nasional yang mencerminkan derajat kesehatan masyarakat, perkembangan sosial dan ekonomi masyarakat, serta kualitas sumber daya manusia suatu negara. Salah satu penyebab tidak langsung pada kematian maternal yang sangat penting adalah anemia dalam kehamilan [3]. Anemia selama kehamilan diketahui berdampak buruk bagi kesehatan ibu maupun bayinya dan merupakan penyebab penting yang melatarbelakangi kejadian morbiditas dan mortalitas [4].

Data Angka Kematian Ibu di Provinsi Banten tahun 2019 sebesar 135 per 100.000 kelahiran hidup, sementara target nasional adalah 125 per 100.000 kelahiran hidup. Prevalensi anemia pada ibu hamil di Provinsi Banten tahun 2018 adalah sebesar 35,2% dan pada tahun 2019 meningkat menjadi 37,7% [5].

Kejadian anemia pada ibu hamil sebagian besar disebabkan oleh defesiensi zat besi, untuk itu pemerintah memberikan kebijakan pemberian tablet zat besi pada ibu hamil di Puskesmas dan Posyandu secara gratis, setiap ibu hamil dianjurkan minum tablet tambah darah dengan dosis satu tablet setiap hari selama masa kehamilannya sampai 40 hari setelah melahirkan. Jumlah tablet zat besi yang dikonsumsi ibu hamil adalah minimal 90 tablet selama kehamilan [6] – [7].

Kabupaten Pandeglang untuk proporsi perolehan tablet tambah darah yang merupakan salah satu cara pencegah penyakit ini dapat diperoleh hingga diminum oleh ibu hamil hanya sebesar 22,24% dari anjuran sebenarnya atau lebih dari 90 tablet. Nilai yang sangat jauh dan masih tertinggal dengan ke enam kabupaten lain yang ada di provinsi Banten [8].

Cikedal merupakan salah satu kecamatan yang ada di Kabupaten Pandeglang dimana angka KEK atau Kekurangan Energi Kronik pada ibu hamil terus meningkat dari tahun ke tahun. Selain itu penyaluran tablet tambah darah untuk ibu hamil masih belum maksimal [9] – [10]. Selain kurangnya obat penambah darah, faktor-faktor lain yang mempengaruhi anamia pada ibu hamil yaitu usia ibu, usia kehamilan, paritas, jarak kehamilan anak, status KEK, pendidikan, pemahaman ibu hamil, kondisi ibu sebelum mengandung atau pra kehamilan, status bekerja dan frekuensi kunjungan ibu hamil dengan tenaga kesehatan professional [11].

Tujuan dari peneliti ini adalah mengetahui faktor-faktor resiko yang mempengaruhi terjadinya anemia pada Ibu hamil di wilayah kerja Puskesmas Kecamatan Cikedal dan mengklasifikasikannya berdasarkan faktor-faktor risiko tersebut. Salah satu metode yang dapat digunakan untuk menyelesaikan permasalahan ini adalah dengan menggunakan metode CART. Dengan adanya pengklasifikasian tersebut maka diharapkan jumlah anemia pada Ibu hamil dapat diturunkan sehingga akan berdampak pula pada penurunan angka kematian Ibu terutama pada kecamatan Cikedal.

METODE

Penelitian ini menggunakan data primer dengan cara menyebarkan kuesioner pada ibu hamil yang terdaftar di Puskesmas wilayah kerja Kecamatan Cikedal. Populasi yang terhitung sejak

September 2021 yaitu 145 ibu hamil sehingga didapat sampel menggunakan slovin yaitu sebanyak 106 ibu hamil. Teknik Sampling yang digunakan yaitu *Cluster Sampling* maka dengan begitu ibu hamil yang terdaftar datanya di puskesmas dan posyandu masing-masing desa itulah yang terpilih menjadi responden. Variabel yang digunakan yaitu variabel respon status anemia pada ibu hamil dan variabel prediktor yang digunakan dalam penelitian ini meliputi variabel (X_1) usia ibu, (X_2) usia kehamilan, (X_3) paritas, (X_4) jarak kehamilan anak, (X_5) pola konsumsi FE (X_6) status KEK, (X_7) pendidikan, (X_8) pemahaman ibu hamil, (X_9) kondisi ibu sebelum mengandung atau pra kehamilan, (X_{10}) status bekerja dan (X_{11}) frekuensi kunjungan ibu hamil dengan tenaga kesehatan professional [12].

Metode yang digunakan adalah analisis deskriptif untuk mengetahui status anemia pada ibu hamil di kecamatan Cikedal. Selanjutnya, mendapatkan faktor yang mempengaruhi anemia pada ibu hamil di kecamatan Cikedal dengan pendekatan klasifikasi pohon *(Classification Tree)*. Tahapan analisis klasifikasi pohon yakni menentukan kemungkinan pemilah pada setiap variabel prediktor, penentuan simpul terminal, dan penandaan label kelas. Selanjutnya pemangkasan pohon dan mendapatkan pohon klasifikasi yang optimal.

1. Pembentukan Pohon Klasifikasi
   a. Pemilihan Pemilah

   Menggunakan fungsi Indeks Gini dituliskan dalam persamaan berikut :

$$i(t) = \sum_{i \neq j} p(j|t)\, p(i|t) \quad (1)$$

dimana
$i(t)$: Fungsi keheterogenan indeks gini
$p(j|t)$: Proporsi kelas j pada simpul t
$p(i|t)$: Proporsi kelas i pada simpul t

Penurunan nilai impuritas dirumuskan sebagai berikut

$$\Delta i(s,t) = i(t) - P_L\, i(t_L) - P_R\, i(t_R) \quad (2)$$

Dimana
$\Delta i(s,t)$: Penurunan nilai impuritas kelas ke-s simpul ke-t
$i(t)$: Fungsi keheterogenan
P_L : Peluang observasi pada simpul kiri
$i(t_L)$: Nilai impuritas simpul ke-t kiri
P_R : Peluang observasi pada simpul kanan
$i(t_R)$: Nilai impuritas simpul ke-t kanan

   b. Penentuan Simpul Terminal

   Pengembangan pohon akan berhenti apabila pada simpul terdapat pengamatan berjumlah kurang dari atau sama dengan 5 ($n \leq 5$). Selain itu, proses pembentukan pohon akan berhenti

jika sudah mencapai batasan jumlah level yang telah ditentukan atau tingkat kedalaman (*depth*) dalam pohon maksimal [13].

c. Penandaan Label Kelas
Penentuan label kelas pada simpul terminal berdasarkan aturan jumlah terbanyak, yaitu jika

$$p\,(j_0|\mathrm{t}) = max_j\,(j|t)\,\frac{\mathrm{N}_j(\mathrm{t})}{\mathrm{N(t)}} \quad (3)$$

dimana
$p\,(j|t)$: proporsi kelas j pada simpul t
$\mathrm{N_j\,(t)}$: jumlah pengamatan kelas j pada terminal node t
N (t) : jumlah total pengamatan pada terminal node t
Label kelas untuk simpul terminal t adalah j_0 yang memberikan nilai dugaan kesalahan pengklasifikasian pada simpul t yang paling kecil sebesar $r(t) = 1 - maxj\ p(j|t)$.

2. Pemangkasan Pohon Klasifikasi
Pemangkasan dilakukan menggunakan *cost complexity prunning*.

$$R_a(T) = R(T) - \alpha|\tilde{T}| \quad (4)$$

Dimana
$R_a(T)$: *Resubtitution* suatu pohon T pada kompleksitas α
$R(T)$: *resubtitusion estimate*
a : *parameter cost-complexity*
$|\tilde{T}|$: banyaknya simpul terminal pohon t

Cost complexity pruning menentukan suatu pohon bagian $T(a)$ yang meminimumkan $R_a(T)$ pada seluruh pohon bagian untuk setiap nilai a, dicari pohon bagian $T(a) < T_{max}$ yang meminimumkan $R_a(T)$ yaitu :

$$R_a(T(a)) = min_{T<Tmax} R_a(T) \quad (5)$$

3. Penentuan Pohon Klasifikasi Optimal
Pohon klasifikasi optimal yang dipilih adalah pohon optimal yang berukuran tepat dan mempunyai nilai penduga pengganti yang cukup kecil.. Penaksir pengganti dilakukan menggunakan penaksir atau penduga sampel uji (*test sample estimate*).

$$R(T) = \frac{1}{N}\sum_n^N X(d(X_n) \neq j_n \quad (6)$$

dengan $X_n$adalah suatu fungsi indikator berbentuk :

$$X_n = \begin{cases} 1 \text{ jika pernyataan di dalam tanda kurung benar.} \\ 0 \text{ jika pernyataan di dalam tanda kurung salah} \end{cases}$$

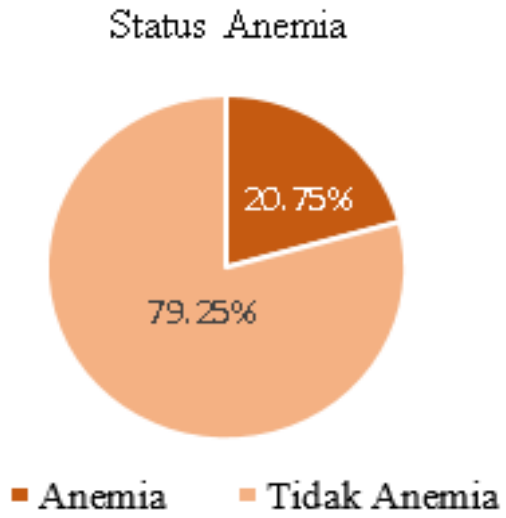
Pada penduga sampel, L dibagi menjadi 2 himpunan secara acak, yaitu L_1 untuk membentuk pohon T dan L_2 untuk menduga $R(T)$. Jika N_2 adalah banyaknya amatan dalam L_2, maka penduga sampel uji :

$$R^{ts}(T_k) = \frac{1}{N_2} \sum_{(xn,jn) \in L_2} X(d(X_n) \neq j_n \quad (7)$$

Pohon optimal yang terpilih adalah T^* dengan kriteria
$R^{ts}(T^*) = min_k R^{ts}(T_k)$

HASIL DAN PEMBAHASAN

Berdasarkan analisis deskriptif dari 106 sampel, status ibu hamil yang mengidap anemia yaitu 20,75%, angka itu lebih kecil dibandingkan ibu hamil yang tidak mengindap anemia (79,25%).



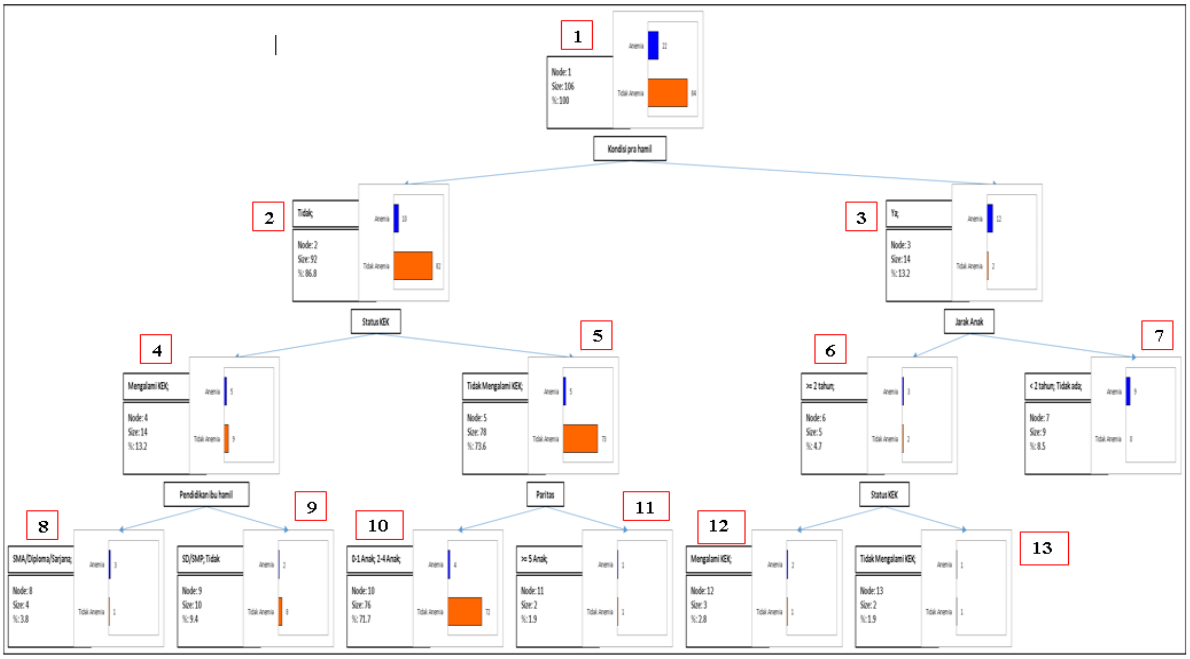
Gambar 1. Status Anemia Ibu Hamil di Kecamatan Cikedal

Faktor yang mempengaruhi resiko anemia pada ibu hamil didapatkan dengan pendekatan klasifikasi pohon kemudian didapat hasil pemilah terbaik pada simpul 1 (pemilah utama) pada penelitian ini adalah variabel kondisi pra hamil (X_9) karena memiliki nilai penurunan keheterogenan tertinggi pada simpul 1 yaitu 16,76 terlihat pada Tabel 1 di bawah ini.

Tabel 1. Struktur Pohon Klasifikasi

Nodes (1)	Objects (2)	% (3)	Improvement (4)	Split variable (5)	Values (6)	Parent node (7)	Sons (8)	Predicted values (9)
Node 1	106	100%	16.76				2; 3	Tidak Anemia
Node 2	92	86.79%	3.93	Kondisi pra hamil	Tidak;	1	4; 5	Tidak Anemia
Node 3	14	13.21%	2.38	Kondisi pra hamil	Ya;	1	6; 7	Anemia
Node 4	14	13.21%	1.87	Status KEK	Mengalami KEK;	2	8; 9	Tidak Anemia
Node 5	78	73.58%	1.96	Status KEK	Tidak Mengalami KEK;	2	10; 11	Tidak Anemia
Node 6	5	4.72%	0.07	Jarak Anak	≥2 tahun;	3	12; 13	Anemia
Node 7	9	8.49%		Jarak Anak	< 2 tahun; Tidak ada;	3		Anemia
Node 8	4	3.77%		Pendidikan ibu hamil	SMA/Diploma/Sarjana;	4		Anemia
Node 9	10	9.43%		Pendidikan ibu hamil	SD/SMP; Tidak sekolah/Tidak lulus SD;	4		Tidak Anemia
Node 10	71	66.98%		Usia ibu hamil	20-35 tahun;	5		Tidak Anemia
Node 11	7	6.60%		Usia ibu hamil	<20 tahun; >35 tahun;	5		Tidak Anemia
Node 12	3	2.83%		Status KEK	Mengalami KEK;	6		Anemia
Node 13	2	1.89%		Status KEK	Tidak Mengalami KEK;	6		Anemia

Terbentuk 13 simpul dengan 6 split variabel yang akan menjadi parents note pada pohon klasifikasi yaitu variabel kondisi pra-hamil, Status KEK, Jarak Anak, Pendidikan Ibu Hamil, dan Usia Ibu Hamil.



Gambar 2. Model CART yang terbentuk

Penjelasan dari simpul-simpul terminal yang terbentuk yaitu:

- Simpul terminal 1atau simpul 7 diprediksi ibu hamil mengalami anemia jika kondisi ibu sebelum hamil mengalami anemia, jarak kelahiran antar anak kurang dari 2 tahun (< 2 tahun) dan mengalami anemia ketika hamil. Hal ini terjadi 8,5% kasus dalam penelitian ini atau diverifikasi oleh 9 pengamatan.
- Simpul terminal 2 atau simpul 8 diprediksi ibu hamil mengalami anemia jika kondisi ibu ketika sebelum hamil tidak mengalami anemia, mengalami Kekurangan Energi Kronis (KEK) ketika

hamil, pernah mengenyam pendidikan SMA/Diploma/Sarjana dan mengalami anemia ketika hamil. Hal ini terjadi 3,8% kasus dalam penelitian ini atau diverifikasi oleh 4 pengamatan.

- Simpul terminal 3 atau simpul 9 diprediksi ibu hamil tidak mengalami anemia jika kondisi ibu ketika sebelum hamil tidak mengalami anemia, mengalami Kekurangan Energi Kronis (KEK) ketika hamil, pernah mengenyam pendidikan SD/SMP dan tidak mengalami anemia ketika hamil. Hal ini terjadi 9,4% kasus dalam penelitian ini atau diverifikasi oleh 10 pengamatan.
- Simpul terminal 4 atau simpul 10 diprediksi ibu hamil tidak mengalami anemia jika kondisi ibu ketika sebelum hamil tidak mengalami anemia, tidak mengalami Kekurangan Energi Kronis (KEK) ketika hamil, memiliki riwayat kelahiran 0-1 anak atau 2-4 anak dan tidak mengalami anemia ketika hamil. Hal ini terjadi 71,7% kasus dalam penelitian ini atau diverifikasi oleh 76 pengamatan.
- Simpul terminal 5 atau simpul 11 diprediksi ibu hamil mengalami anemia jika kondisi ibu ketika sebelum hamil tidak mengalami anemia, tidak mengalami Kekurangan Energi Kronis (KEK) ketika hamil, memiliki riwayat kelahiran 5 atau lebih dari 5 anak (≥ 5 anak) dan mengalami anemia ketika hamil. Hal ini terjadi 1,9% kasus dalam penelitian ini atau diverifikasi oleh 2 pengamatan.
- Simpul terminal 6 atau simpul 12 diprediksi ibu hamil mengalami anemia jika kondisi ibu ketika sebelum hamil mengalami anemia, memiliki riwayat kelahiran 2 atau lebih dari 2 anak (≥ 2 anak), mengalami Kekurangan Energi Kronis (KEK) ketika hamil dan mengalami anemia ketika hamil. Hal ini terjadi 2,8% kasus dalam penelitian ini atau diverifikasi oleh 3 pengamatan.
- Simpul terminal 7 atau simpul 13 diprediksi ibu hamil mengalami anemia jika kondisi ibu ketika sebelum hamil mengalami anemia, memiliki riwayat kelahiran 2 atau lebih dari 2 anak (≥ 2 anak), tidak mengalami Kekurangan Energi Kronis (KEK) ketika hamil dan mengalami anemia ketika hamil. Hal ini terjadi 1,9% kasus dalam penelitian ini atau diverifikasi oleh 3 pengamatan.

Tabel 2. Ketepatan Prediksi Hasil Klasifikasi

Kondisi Aktual	Prediksi			
	Anemia	Tidak Anemia	Total	% Correct
Anemia	16	6	22	72,73 %
Tidak Anemia	4	80	84	95,24 %
Total	20	86	106	90,57 %

Ketepatan prediksi hasil klasifikasi yang terbentuk secara keseluruhan yaitu 90,57%. Sedangkan ketepatan prediksi ibu hamil yang anemia yakni 72,73% dan ketepatan prediksi ibu hamil yang tidak anemia yakni 95,24%.

KESIMPULAN

Kesimpulan yang didapatkan yaitu variabel-variabel yang mempengaruhi ibu hamil terkena anemia meliputi variabel (X_1) usia ibu ketika mengandung atau hamil, variabel (X_2) usia kehamilan, variabel (X_3) paritas atau jumlah kelahiran anak baik yang lahir hidup ataupun lahir mati, variabel (X_4) jarak kehamilan antar anak, (X_6) status KEK (Kekurangan Energi Kronis), (X_7) pendidikan ibu hamil, variabel (X_8) pemahaman ibu hamil, dan variabel (X_9) kondisi ibu sebelum mengandung atau pra kehamilan.

Hasil model dari metode CART dari delapan varibel menghasilkan prediksi ibu hamil yang anemia pada Puskesmas Kecamatan Cikedal hanya 20,75% lebih rendah dibandingkan dengan jumlah ibu hamil yang diprediksi tidak anemia yaitu sebanyak 79,25%. Ketepatan prediksi hasil klasifikasi menggunakan metode CART ini secara general yaitu 90,57%.

DAFTAR PUSTAKA

[1] Tohari, A., "Pemodelan Derajat Kesehatan Menggunakan Structural Equation Modelingdi Kabupaten Kediri," Jurnal Imiah dan Aplikasi StatistikaVol. 10, No. 2, pp. 1-6, 2017.

[2] Fatima, M., Pramesti W. "Pemodelan Regresi Spline pada Studi Kasus Angka Kematian Bayi di Jawa Timur Tahun 2015," Jurnal Imiah dan Aplikasi StatistikaVol. 11, No. 2, pp. 9-16, 2018

[3] Anfiksyar, K. S. S., Aryana, M. B. D., Surya, I. G. N. H. W., & Manuaba, I. B. G. F., "Karakteristik Anemia pada Kehamilan di Poliklinik Kebidanan PSUP Sanglah Tahun 2016-2017," *Jurnal Medika Udayana*, vol. *8*, no. 7, pp. 1–7, 2019

[4] Rasmaliah. (2004). *Anemia kurang besi dalam hubungannya dengan infeksi cacing pada ibu hamil rasmaliah*.

[5] Serang, D. K. K., *Profil Dinas Kesehatan Kabupaten Serang Tahun 2020*, 2020

[6] Kesehatan, K., & Indonesia, R. (n.d.). *PROFIL KESEHATAN INDONESIA 2020*.

[7] Leny. "Faktor-faktor yang Berhubungan dengan Kejadian Anemia Pada Ibu Hamil," *Jurnal Kebidanan : Jurnal Medical Science Ilmu Kesehatan Akademi Kebidanan Budi Mulia Palembang*, vol. *9*, no. 2, pp. 161–167. 2019

[8] Riskesdas Banten, *Laporan Provinsi Banten RISKESDAS 2018*. Badan Penelitian Dan Pengembangan Kesehatan, 2018

[9] *Data ibu hamil cikedal September 2021*. (n.d.).

[10] Rukmaini, "Model Rencana Upaya Keluarga Menangani Anemia (RUKMA) Pada Kabupaten Pandeglang Provinsi Banten," *E-Skripsi Universitas Andalas*, vol. *7, no.* 1, pp. 37–72, 2018

[11] Nurhayati, I., Hidayat, A. R., & Hartati, T., "Gambaran Pengetahuan Ibu Hamil Tentang Kekurangan Energi Kronis (Kek) Di Klaten," *Jurnal Riset Gizi*, Vol.*8*, no. 1, pp. 48–51. 2020

[12] Hartati, A., Zain, I., & Suprih, S, "Analisis CART (Classification And Regression Trees) pada Faktor-Faktor yang Mempengaruhi Kepala Rumah Tangga di Jawa Timur Melakukan Urbanisasi," *Jurnal Sains Dan Seni Its*, vol. 1, no. 1, pp. 101–105, 2012

[13] Sjahrani, T., & Faridah, V., "Faktor-Faktor Yang Berhubungan Dengan Kejadian Anemia," *Jurnal Kebidanan*, vol. *5*, no. 2, pp.106–115, 2019

Pengelompokkan Kecamatan di Kabupaten Bima Berdasarkan Jumlah Produksi dan Luas Panen Bawang Merah Tahun 2021 Menggunakan *K-Means Clustering*

Ashabul Akbar Maulana [1], Athallah Widyatama Rafii [2], Yulia Anggi Anjelina [3], Edy Widodo [4]

Universitas Islam Indonesia, Fakultas Matematika dan Ilmu Pengetahuan Alam

Jl. Kaliurang No.Km. 14,5, Krawitan, Umbulmartani, Kec. Ngemplak, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55584

e-mail: 19611101@students.uii.ac.id(1), 19611106@students.uii.ac.id (2), 19611104@students.uii.ac.id(3), edywidodo@uii.ac.id(4)

ABSTRAK

Bawang merah adalah komoditas pertanian signifikan di Indonesia dengan potensi ekonomi yang menjanjikan. Kabupaten Bima mendominasi produksi bawang merah di Indonesia. Secara geografis, Kabupaten Bima berada pada kisaran ketinggian 0 hingga 477.5 mdpl dengan mayoritas wilayahnya memiliki iklim panas dan kering, menciptakan kondisi yang ideal untuk budidaya bawang merah. Penelitian ini bertujuan menggambarkan serta menganalisis Cluster hasil produksi bawang merah di Kabupaten Bima tahun 2021 menggunakan metode Clustering K-Means. K-Means Clustering merupakan analisis klaster non-hierarkis yang membagi data ke dalam kelompok berdasarkan karakteristik tertentu. Hasil analisis menunjukkan bahwa Kecamatan Lambu mencatat produksi dan luas panen bawang merah tertinggi di Kabupaten Bima, sementara Kecamatan Langgudu dan Donggo tidak menghasilkan bawang merah. Analisis klaster menghasilkan tiga kelompok dengan karakteristik yang berbeda, di mana Cluster 1 memiliki kategori produksi rendah, Cluster 2 memiliki kategori produksi tinggi, dan Cluster 3 memiliki kategori produksi sedang.

Kata kunci : Bawang Merah; *Clustering; K-Means.*

ABSTRACT

Shallots are a significant agricultural commodity in Indonesia with promising economic potential. Bima Regency dominates shallot production in Indonesia. Geographically, Bima Regency spans an elevation range of 0 to 477.5 meters above sea level, with the majority of its area experiencing a hot and dry climate, creating ideal conditions for shallot cultivation. This study aims to depict and analyze the clusters of shallot production outcomes in Bima Regency for the year 2021 using the K-Means Clustering method. K-Means Clustering is a non-hierarchical cluster analysis that groups data based on specific characteristics. The analysis results indicate that the Lambu Sub-district records the highest shallot production and harvest area in Bima Regency, while the Langgudu and Donggo Sub-districts do not yield shallots. The cluster analysis yields three distinct groups with varying characteristics, where Cluster 1 has a low production category, Cluster 2 has a high production category, and Cluster 3 falls into the moderate production category.

Keywords : *Red Onion, Clustering, K-Means.*

PENDAHULUAN

Bawang merah adalah salah satu komoditas pertanian yang menjanjikan dan memiliki nilai jual yang cukup tinggi. Kabupaten Bima menjadi salah satu daerah penghasil bawang merah terbesar di Indonesia. Dengan letak geografis yang berada pada 0 sampai dengan 477.5 mdpl memiliki iklim yang kering dan suhu udara yang tergolong panas sangat cocok untuk ditanami bawang merah, menjadikan Kabupaten Bima mempunya potensi yang sangat tinggi dalam menghasilkan bawang merah [1].

Salah satu metode yang dapat digunakan untuk melakukan analisis pada data tersebut adalah analisis clustering. Secara umum, Clustering didefinisikan sebagai pengelompokan data yang menunjukkan kesamaan berdasarkan standar khusus [2]. Sama seperti kelompok profil yang terbagi menjadi beberapa klaster yang saling serupa satu sama lain. Klasterisasi menghasilkan jumlah klaster yang kurang atau setara dengan jumlah data profil yang dimasukkan [3].

Metode k-Means merupakan salah satu metode yang populer digunakan untuk clustering data. *K-Means Clustering* adalah salah satu jenis analisis klaster non-hirarki. Metode ini bertujuan untuk membagi suatu objek data kedalam satu atau lebih jumlah *Cluster* berdasarkan karakteristik tertentu [4]. Terdapat beberapa keunggulan *Clustering K-Means* dibandingkan dengan *Clustering* yang lain diantaranya waktu komputasi yang relatif cepat, dan hasil yang mudah di representasikan [5].

Penelitian yang dilakukan oleh Febianto & Palarasa (2018) [6] yang meneliti tentang pengelompokkan daerah di Provinsi Jawa Timur berdasarkan 13 indikator kemiskinan. Penelitian ini mengasilkan 5 cluster dengan pemetaan karakteristik dari setiap kelompok yang terbentuk berdasarkan nilai rata-rata tertinggi dan terendah dari setiap indicator kemiskinan provinsi Jawa Barat tahun 2018.

Penelitian ini bertujuan untuk mengetahui hasil dari pengelompokkan daerah di Kabupaten Bima berdasarkan data Jumlah Panen dan Luas Panen Bawang Merah tahun 2021, serta mengetahui karakterstik yang ada dari setiap karakteristik yang terbentuk. Untuk selanjutnya dapat diambil keputusan dalam meningkatkan produksi Bawang Merah di Kabupaten Bima sehingga diharapkan mampu meningkatkan pendapatan petani Bawang Merah di Kabupaten Bima.

Banyak penelitian yang menggunakan metode *clustering K-Means* akan tetapi belum ada penelitian yang mengelompokkan data Jumlah Produksi dan luas panen bawang merah terkhusus pada daerah Kabupaten Bima.

Mengacu pada penjelasan sebelumnya, peneliti menjalankan sebuah penelitian yang memiliki potensi sebagai panduan bagi Pemerintah Kabupaten Bima dalam melaksanakan evaluasi atas hasil kelompok yang terbentuk serta ciri-ciri yang ada melalui judul penelitian "Pengelompokkan Kecamatan di Kabupaten Bima Berdasarkan Jumlah Produksi dan Luas Panen Bawang Merah Tahun 2021 Menggunakan *K-Means Clusterin*".

METODE PENELITIAN

Statistika Deskriptif

Statistika merupakan ilmu yang paling berguna untuk mengumpulkan dan mendefinisikan data kuantitatif sedemikian rupa sehingga kemungkinan kesalahan dalam kesimpulan serta estimasi dapat diperkirakan dengan menggunakan penalaran induktif berdasarkan pada probabilitas

matematika. Statistika Deskriptif merupakan kumpulan teknik yang meliputi teknik pengumpulan, penyajian, dan peringkasan data [7].

Statistik deskriptif adalah bidang studi yang berfokus pada cara-cara dan teknik-teknik untuk mengatur dan menyajikan data yang telah terkumpul dalam sebuah penelitian. Misalnya, ini melibatkan penyusunan data dalam bentuk tabel frekuensi atau dalam bentuk grafik, serta menghitung nilai statistik seperti rata-rata, median, variasi, simpangan baku, dan sejenisnya [8].

Statistika deskriptif dapat digunakan untuk melakukanan analisa dan menggambarkan sejumlah data hasil penelitian yang tidak digunakan dalam pengambilan kesimpulan yang sifatnya luas. Statistika deskriptif dapat berupa penyajian data dalam bentuk tabel distribusi frekuensi, grafik atau dalam bentuk lainnya yang tidak bertujuan sebagai pengambilan kesimpulan berkenaan dengan generalisasi [9].

Uji Multikolinearitas

Multikolinearitas artinya terdapat hubungan linier yang tinggi antara variabel yang menjelaskan model. Terjadinya Multikolinearitas dapat diketahui dari nilai koefisien korelasi dari masing-masing variabel independen. Tujuan dari Uji multikolinearitas adalah untuk mengetahui dalam suatu model regresi terdapat korelasi antar variabel independen atau variabel bebas yang signifikan. Hal tersebut berarti standar *error* besar, akibatnya ketika koefisien diuji, T-Hitung akan bernilai lebih kecil dari T-Tabel. Hal ini menunjukkan tidak adanya hubungan linier antara variabel independen atau variabel bebas yang dipengaruhi dengan variabel dependen atau variabel terikat [10].

Adanya multikolinearitas dapat mengakibatkan ketidaktepatan penggunaan metode regresi karena perkiraan regresinya menjadi tidak stabil dan koefisien variabel regresinya cenderung besar. Beberapa faktor yang dapat diamati untuk mengidentifikasi keberadaan multikolinearitas [11].

Untuk mengetahui ada atau tidaknya multikolinearitas pada suatu model regresi dapat dilihat dari nilai *tolerance* dan nilai *Variance Inflation Factor* (VIF). Nilai *tolerance* mengukur variabilitas dari variabel bebas yang terpilih dan tidak dapat didefinisikan oleh variabel bebas lainnya, jadi nilai *tolerance* yang rendah sama dengan nilai VIF tinggi, dikarenakan perhitungan VIF = 1/*tolerance* dan menunjukkan terdapat korelasi yang tinggi. Nilai *cut off* yang digunakan adalah untuk nilai *tolerance* 0,10 atau nilai VIF di atas angka 10. Rumus VIF untuk koefisien dijabaran sebagai berikut [12].

$$VIF_j = \frac{1}{1 - R_j^2} \tag{1}$$

keterangan:

VIF : *Variance Inflation Factor*

R^2 : Koefisien determinasi

j : Jumlah Sampel

Berdasarkan rumus di atas, jika nilai VIF < 10 atau nilai *tolerance* > 0,01 maka dinyatakan tidak terjadi multikolinearitas. Jika nilai VIF > 10 atau nilai *tolerance* < 0,01 maka dinyatakan terjadi multikolinearitas.

Analisis *Cluster*ing

Analisis *Cluster* adalah salah satu dari metode analisis multivariat yang memiliki tujuan untuk mengelompokkan objek-objek berdasarkan suatu karakteristik yang dimiliki. Analisis *Cluster* meengelompokkan individu atau objek suatu penelitian, sehingga setiap objek yang paling dekat kesamaannya dengan objek lain berada dalam satu *Cluster* yang sama. *Cluster*-*Cluster* yang terbentuk dalam suatu *Cluster* mempunyai ciri yang relatif mirip atau bahkan sama (homogen), sedangkan antar *Cluster* mempunyai ciri yang berbeda (heterogen). Pengelompokan ini dilakukan berdasarkan variabel yang diamati [13]. Analisis klaster adalah salah satu teknik dalam pembelajaran mesin yang memiliki tujuan utama dalam menemukan pola berinformasi atau pola yang bernilai pada kumpulan *big data* [14].

Untuk mendapatkan kelompok yang homogen, maka menggunakan dasar untuk pengelompokkan yaitu kesamaan nilai yang dianalisis. Semakin kecil nilai jarak pada suatu individu terhadap individu lain, maka semakin besar pula kemiripan individu tersebut. Data mengenai ukuran kesamaan tersebut dikelompokkan sehingga dapat ditentukan individu mana yang termasuk ke kelompok yang sesuai [13].

Algoritma *K-Means*

K-Means Clustering adalah salah satu jenis analisis klaster non-hirarki. Metode ini bertujuan untuk membagi suatu objek data kedalam satu atau lebih jumlah *Cluster* berdasarkan karakteristik tertentu [4]. Algoritma K-means merupakan salah satu metode algoritma partisional, karena K-means didasarkan pada pemilihan jumlah kelompok awal dengan menetapkan nilai awal pusatannya. Algoritma K-means mengadopsi langkah-langkah berulang untuk menghasilkan kelompok data yang konsisten [15].

Berikut adalah langkah-langkah dalam melakukan analisis *Cluster*ing *K-Means* [16]:

1. Menentukan secara acak jumlah data yang digunakan sebagai pusat *Cluster*
2. Menghitung jarak antara data dan pusat *Cluster*. Untuk menghitung semua jarak antara data dengan pusat *Cluster* dapat menggunakan rumus pada toeri jarak euclidian yang dirumuskan sebagai berikut :

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + \ldots + (X_{ki} - X_{kj})^2} \quad (2)$$

Keterangan:

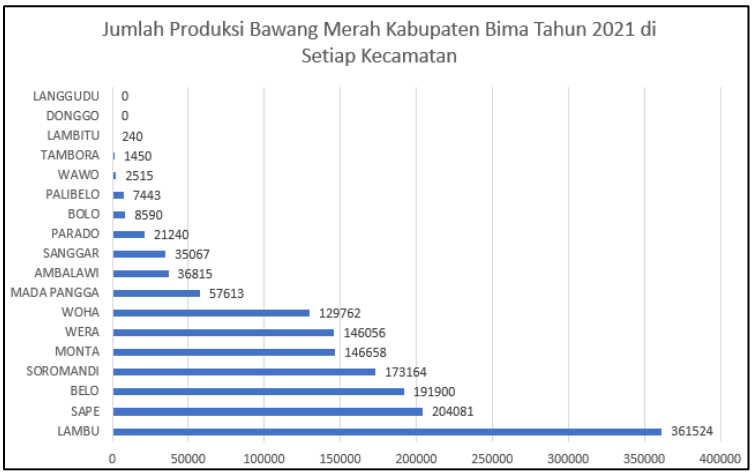
$D(i,j)$: Jarak data ke-i ke pusat *Cluster* j
X_{ki} : Data ke-i pada atribut data ke-k
X_{kj} : Titik pusat ke-j pada atribut data ke-k

3. Data ditempatkan *Cluster* terdekat, dihitung dari pusat *Cluster*.
4. Pusat *Cluster* akan dihitung kembali setelah semua data ditempatkan pada *Cluster*-nya masing-masing.
5. Proses pembagian *Cluster* selesai apabila tidak terdapat perubahan pada nilai *centroid*.

HASIL DAN PEMBAHASAN

Analisis Deskriptif

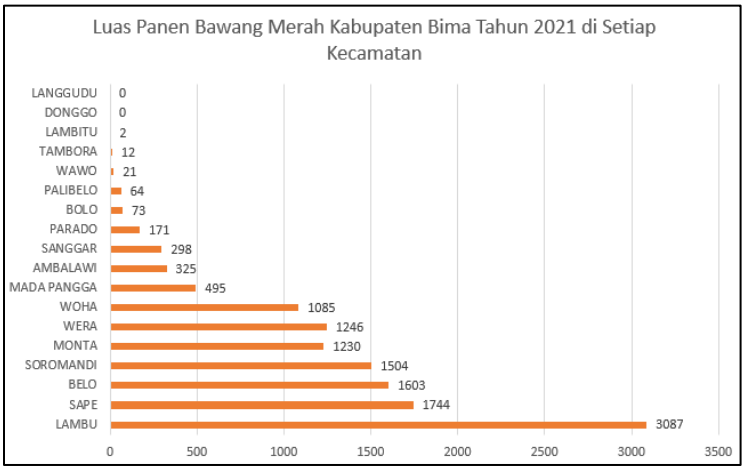
Tujuan dari analisis deskriptif adalah menyajikan data kedalam bentuk yang mudah dipahami dan menarik untuk diambil informasinya oleh pembaca [9]. *Bar chart* digunakan untuk melakukan visualisasi perbandingan jumlah produksi bawang merah untuk setiap Kecamatan yang ada di Kabupaten Bima.



Gambar 1. Jumlah Produksi Bwang Merah Kabupaten Bima Tahun 2021.

Diperoleh bahwa Kecamatan Lambu menjadi daerah dengan hasil produksi bawang merah terbanyak dengan jumlah produksi sebanyak 361524 ton, sedangkan Kecamatan Langgudu dan Donggo adalah daerah yang tidak memproduksi bawang merah pada tahun 2021. Karena daerah tersebut tidak cocok digunakan untuk menanam bawang merah karena keadaan georafis dan memiliki suhu terlalu dingin untuk pertanian bawang merah.

Kemudian peneliti membandingkan luas panen bawang merah kabupaten Bima tahun 2021 untuk setiap Kecamatan akan terlihat seperti pada **Gambar 2.**



Gambar 2. Luas Panen Bwang Merah Kabupaten Bima Tahun 2021.

Dapat dilihat pada **Gambar 2** bahwa Kecamatan Lambu adalah daerah dengan luas panen bawang merah terbesar dengan luas 3087 ha, sedangkan Kecamatan Langgudu dan Donggo menjadi daerah tanpa luas panen bawang Merah. Karena daerah tersebut tidak cocok digunakan untuk menanam bawang merah karena keadaan georafis dan memiliki suhu terlalu dingin untuk pertanian bawang merah.

Analisis *Clustering K-Means*

Peneliti melakukan proses Klasterisasi dengan tujuan untuk mengidentifikasi ciri khas dari masing-masing kelompok Klasterisasi yang terbentuk. Penelitian ini menggunakan dataset mengenai Jumlah Produksi Bawang Merah di Kabupaten Bima pada Tahun 2021. Sebelum menjalankan analisis Klasterisasi, terlebih dahulu dilakukan pengujian terhadap asumsi multikolinearitas melalui evaluasi nilai Toleransi dan Faktor Inflasi Varians (VIF). Hasil Klasterisasi diharapkan mampu memberikan arahan bagi pemerintah dalam meningkatkan produktifitas serta kesejahteraan petani bawang merah.

Untuk mengidentifikasi adanya multikolinearitas dalam suatu model regresi, dapat dianalisis melalui nilai toleransi dan Variance Inflation Factor (VIF). Toleransi mengukur variasi dari variabel independen yang tidak dapat dijelaskan oleh variabel independen lainnya, sehingga nilai toleransi yang rendah berhubungan dengan nilai VIF yang tinggi. Ini terjadi karena perhitungan VIF = 1/toleransi, yang mengindikasikan adanya korelasi yang kuat. Nilai ambang yang digunakan adalah toleransi 0,10 atau VIF melebihi 10.

Tabel 1. Uji Multikolinearitas.

Variabel	*VIF*	*Tolerance*
Luas Panen	1	1

Berdasarkan *output* pada **Tabel 1** didapatkan nilai *VIF* lebih kecil dari 10 dan nilai *Tolerance* lebih besar dari 0,1 dengan demikian menolak H_0 sehingga dinyatakan tidak terjadi Multikolinearitas antara variabel Jumlah Produksi dan Luas Panen.

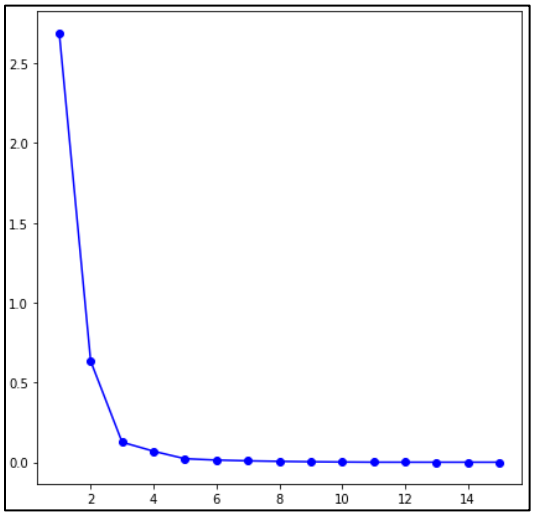
Setelah asumsi multikolinearitas terpenuhi, kemudian peneliti melakukan standarisasi data untuk membuat data memiliki rentang nilai yang sama hal ini disebabkan karena data memiliki satuan yang berbeda serta range data yang berbeda. Selain itu juga, standarisasi data dapat mempermudah dalam melakukan perhitungan. Hasil Standarisasi data dapat dilihat pada Tabel 2 berikut:

Tabel 2. *Data* Standarisasi.

Kecamatan	Jumlah Produksi	Luas Panen
Ambalawi	0.1018	0.1052
Belo	0.5308	0.5193
Bolo	0.0238	0.0236
Donggo	0.0000	0.0000
Lambitu	0.0007	0.0006
Lambu	1.0000	1.0000
Langgudu	0.0000	0.0000

Kecamatan	Jumlah Produksi	Luas Panen
Mada Pangga	0.1594	0.1603
Monta	0.4057	0.3984
Palibelo	0.0206	0.0207
Parado	0.0588	0.0554
Sanggar	0.0970	0.0965
Sape	0.5645	0.5649
Soromandi	0.4790	0.4872
Tambora	0.0040	0.0039
Wawo	0.0070	0.0068
Wera	0.4040	0.4036
Woha	0.3589	0.3514

Dari data yang telah di standarisasi, kemudian akan ditentukan *Cluster* yang terbentuk dengan menggunakan diagram *Elbow*.



Gambar 3. Diagram *Elbow* Untuk Jumlah *Cluster* Optimal.

Pada **Gambar 3** didapatkan bahwa jumlah *Cluster* optimal yang didapatkan dengan menggunakan metode *elbow* adalah sebanyak 3 *Cluster*. Sehingga, *data* Jumlah Produksi Bawang Merah Kabupaten Bima Tahun 2021 akan dibagi ke dalam 3 *Cluster*.

Setelah jumlah *Cluster* optimal didapatkan, peneliti menentukan pusat *Cluster* pada masing-masing variabel.

Tabel 3. Pusat *Cluster*.

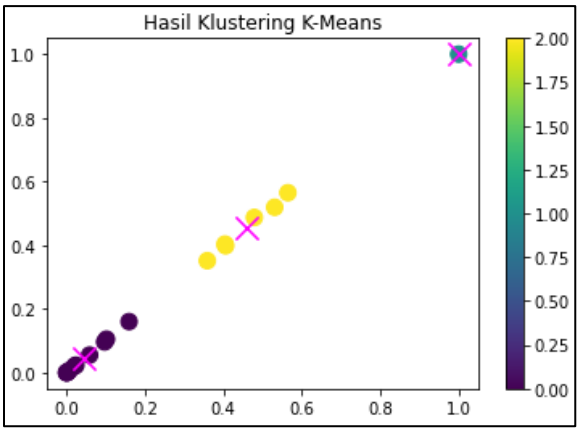
K	Pusat *Cluster*	
	Jumlah Produksi	Luas Panen
1	0.0430	0.0430
2	1	1
3	0.4571	0.4542

Didapatkan pada **Tabel 3** dengan menggunakan data yang di standarisasi bahwa pada variabel Jumlah Produksi, *Cluster* 1 memiliki pusat pada titik 0.0430, *Cluster* 2 berpusat pada titik 1, dan *Cluster* 3 berpusat pada titik 0.4571. sedangkan, pada variabel Luas Panen, *Cluster* 1 memiliki pusat pada titik 0.0430, *Cluster* 2 berpusat pada titik 1, dan *Cluster* 3 berpusat pada titik 0.4542. Dari hasil yang didapatkan, maka akan dilakukan pelabelan pada data asli sesuai dengan jarak terdekat dari pusat *Cluster.*

Tabel 4. Pelabelan *Data*.

Kecamatan	Jumlah Produksi	Luas Panen	*Cluster*
Ambalawi	36815	325	1
Bolo	8590	73	1
Donggo	0	0	1
Lambitu	240	2	1
Langgudu	0	0	1
Mada Pangga	57613	495	1
Palibelo	7443	64	1
Parado	21240	171	1
Sanggar	35067	298	1
Tambora	1450	12	1
Wawo	2515	21	1
Lambu	361524	3087	2
Belo	191900	1603	3
Monta	146658	1230	3
Sape	204081	1744	3
Soromandi	173164	1504	3
Wera	146056	1246	3
Woha	129762	1085	3

Dari *data* yang telah diberikan label dapat dilihat bahwa pada *Cluster* 1 memiliki 11 anggota, untuk *Cluster* 2 terdapat satu anggota, dan untuk *Cluster* 3 terdapat 6 anggota. Kemudian peneliti akan menampilkan penyebaran data dari setiap *Cluster* yang terbentuk dalam *scatterplot.*



Gambar 4. Penyebaran *Data* Anggota *Cluster*.

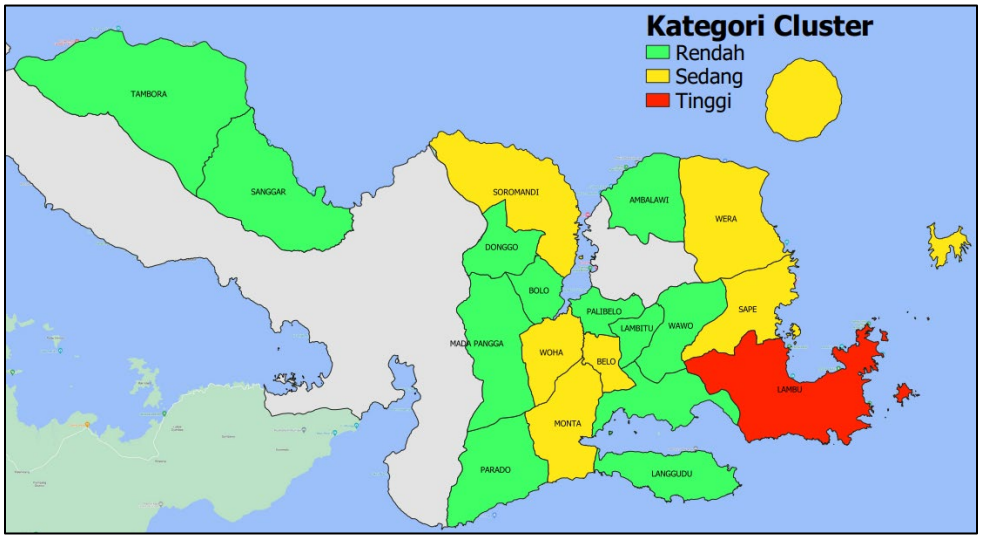
Pada **Gambar 4** dapat dilihat bahwa setiap data yang telah di kelompokkan dalam *Cluster*nya masing-masing membentuk hubungan yang linear, dan jarak dari setiap data terhadap pusat *Cluster* yang tidak berjauhan, artinya karakteristik dari setiap anggota cluster memiliki kesamaan. Untuk melihat persamaan atau karakteristik dari setiap anggota *cluster*, yang dapat dilihat berdasarkan rata-rata dari setiap variabel berdasarkan *Cluster*nya masing-masing.

Tabel 5. Karakteristik *Cluster*.

Cluster	Rata-rata	
	Jumlah Produksi	Luas Panen
1	15543.00	132.82
2	361524.00	3087.00
3	165270.67	1402.00

Pada **Tabel 5** dapat dilihat bahwa *Cluster* 1 adalah *Cluster* yang memiliki anggota dengan rata-rata jumlah produksi dan luas panen bawang merah yang rendah, *Cluster* 2 adalah *cluster* yang memiliki anggota dengan rata-rata jumlah produksi dan luas panen yang tinggi, dan *Cluster* 3 adalah *Cluster* dengan anggota yang memiliki rata-rata jumlah produksi dan luas panen yang sedang.

Dari hasil tersebut, maka akan dilakukan visualisasi dalam bentuk peta untuk melihat penyebaran wilayah setiap *Cluster* berdasarkan karakteristiknya masing-masing yang dapat dilihat pada **Gambar 5**.



Gambar 5. Visualisasi Penyebaran Anggota Cluster.

KESIMPULAN

Dari hasil statistik deskriptif yang didapatkan, Kecamatan Lambu adalah daerah yang mempunyai Jumlah Produksi dan Luas Panen Bawang merah tertinggi dari seluruh Kecamatan yang ada di Kabupaten Bima. Sedangkan, Kecamatan Langgudu dan Donggo adalah daerah yang tidak memproduksi Bawang Merah karena letak geografis dan suhu di daerah tersebut tidak cocok

untuk proses produksi bawang merah. Berdasarkan hasil Profilisasi *Cluster*, didapatkan hasil bahwa *Cluster* 1 dengan 11 anggota adalah *Cluster* dengan kategori Rendah karena memiliki rata-rata jumlah produksi dan luas panen terendah dari tiga *cluster* yang ada, *Cluster* 2 adalah *Cluster* dengan Kategori Tinggi karena Kecamatan Lambu adalah daerah yang penduduknya didominasi oleh petani bawang merah, dan *Cluster* 3 adalah Clsuter dengan kategori sedang.

DAFTAR PUSTAKA

[1] Kementrian Perdagangan RI, 22 Juni 2015. [Online]. Available: http://bppp.kemendag.go.id/media_content/2017/08/Potensi_Bawang_Merah_di_Kabupaten_Bima.pdf.

[2] R. O. Duda and P. E. Hart, Pattern classification and scene analysis, New York: Wiley, 1973.

[3] A. Al-Wakeel and J. Wu, "K-means based cluster analysis of residential smart meter measurements," *Energy Procedia,* vol. 88, pp. 754-760, 2016.

[4] A. N. Khomarudin, "Teknik Data Mining : Algoritma K-Means Clustering," *IlmuKomputer.com,* pp. 7-8, 2016.

[5] Syafnidawaty, 19 April 2020. [Online]. Available: https://raharja.ac.id/2020/04/19/k-means-clustering/#:~:text=Ada%20beberapa%20kelebihan%20pada%20algoritma,Mudah%20untuk%20diadaptasi..

[6] N. I. Febianto and N. D. Palasara, "Analisis Clustering K-Means Pada Data Informasi Kemiskinan Di Jawa Barat Tahun 2018," *SISFOKOM,* vol. 8, no. 2, pp. 130-140, 2019.

[7] H. H, Statistika Deskriptif, Banjarmasin: Poliban Press, 2019.

[8] Maizar, I. Mustika and S. D. Nabella, PENGANTAR STATISTIK 1, Media Sains Indonesia, 2022.

[9] B. R. Kustituanto B, Statistika 1 (Deskriptif), Depok, 1994.

[10] Ghozali, Aplikasi Multivariete Dengan Program IBM SPSS 23, Semarang: Badan Penerbit Universitas Diponegoro, 2016.

[11] V. I. Anggeyeny, Fear of Floating:Studi Empiris Sistem Nilai Tukar secara de facto di Indonesia dalam Periode 1994-2003, Depok: Universitas Indonesia, 2009.

[12] Roswirman and Elazhari, "Pengaruh Implementasi Manajemen Mutu Terpadu dan Disiplin Kerja Terhadap Kinerja Guru pada Era New Normal di SMK Swasta PAB 2 Helvetia," *All Fields of Science Journal Liaison Academia and Sosiety,* vol. 1, no. 4, pp. 316-333, 2021.

[13] S. Hussein, "Analisis CLuster," Oktober 2021. [Online].

[14] L. H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques with Java implementations," *Acm Sigmod Record ,* vol. 31, no. 1, pp. 76-77, 2002.

[15] N. T. Hartanti, "Metode Elbow dan K-Means Guna Mengukur Kesiapan Siswa SMK Dalam Ujian Nasional," . *Nas. Teknol. dan Sist. Inf ,* vol. 6, no. 2, pp. 82-89, 2020.

[16] A. Muhidin and I. Baragigiratri, "Pemetaan Penduduk Calon Penerima Bantuan Renovasi Rumah Desa Pesangkalan Menggunakan Algoritma Clustering K-Means," *Jurnal SIGMA,* vol. 9, no. 3, pp. 75-82, 2019.

Estimation of Survival Function in Head and Neck Cancer Patients Using the Kaplan-Meier Method

Ardi Kurniawan[(1)], Adelia Frielady Yosifa[(2)], Azizatul Kholidiyah[(3)], Vrisca Natalia Putri Wardhani[(4)]

[1,2,3,4]Program Studi Statistika, Fakultas Sains dan Teknologi, Universitas Airlangga

Jalan Dr. Ir. H. Soekarno, Mulyorejo, Kec. Mulyorejo, Surabaya, Jawa Timur

e-mail: ardi-k@fst.unair.ac.id[(1)], adelia.frielady.yosifa-2021@fst.unair.ac.id[(2)] , azizatul.kholidiyah-2021@fst.unair.ac.id[(3)] , vrisca.natalia.putri-2021@fst.unair.ac.id[(4)]

ABSTRAK

Analisis data uji hidup atau analisis survival adalah kumpulan metode statistik yang digunakan untuk menjawab pertanyaan yang berkaitan dengan apakah dan kapan suatu peristiwa terjadi. Data daya tahan hidup yang diperoleh berupa data tersensor tipe I yang dianalisis menggunakan Metode Kaplan-Meier atau Metode Produk Limit. Sampel tersensor tipe I adalah cara pengambilan sampel secara statistik didasarkan pada lama pengamatan yang telah ditentukan sebelumnya. Metode Kaplan-Meier merupakan salah satu metode yang digunakan untuk mengestimasi fungsi survival dalam analisis data uji hidup. Penelitian ini bertujuan untuk mengestimasi daya tahan hidup seorang penderita kanker kepala dan leher. Kanker kepala dan leher adalah jenis kanker yang paling umum di dunia dan terdiri dari berbagai kelompok tumor yang mempengaruhi saluran aerodigestive bagian atas. Penelitian dimulai dengan menentukan data tersensor dan tipe sensor, kemudian menghitung estimasi fungsi survivor. Fungsi survivor merupakan probabilitas individu atau suatu objek tertentu untuk tetap bertahan setelah melewati suatu satuan waktu. Data pada penelitian ini adalah Data Kanker Kepala dan Leher Northern California Oncology Group. Kelompok pasien kanker kepala leher diobati dengan terapi radiasi plus kemoterapi (RCT) dengan 45 pasien. Hasil penelitian menunjukkan bahwa pasien kanker kepala dan leher yang menjalani pengobatan RCT memiliki peluang bertahan hidup semakin kecil seiring berjalannya waktu.

Kata kunci: Kanker Kepala; Kanker Leher; Analisis Survival; Metode Kaplan-Meier; Tersensor Tipe I

ABSTRACT

Life test data analysis or survival analysis is a collection of statistical methods used to answer questions relating to whether and when an event occurred. The survival data obtained were type I censored data which were analyzed using the Kaplan-Meier Method or the Product Limit Method. Type I censored sample is a statistical sampling method based on a predetermined length of observation. The Kaplan-Meier method is one of the methods used to estimate the survival function in the analysis of survival test data. This study aims to estimate the survival of a patient with head and neck cancer. Head and neck cancer is the most common type of cancer worldwide and comprises a diverse group of tumors that affect the upper aerodigestive tract. The research begins by determining the censored data and the type of sensor, then calculating the estimated survivor function. The survivor function is the probability of an individual or a certain object to survive after passing a certain time. The data in this study is the Head and Neck Cancer Data from the Northern California Oncology Group. The head and neck cancer patient group was treated with radiation therapy plus chemotherapy (RCT) with 45 patients. The results showed that head and neck cancer patients undergoing RCT treatment had a smaller chance of survival over time.

Keywords: *Head Cancer; Neck Cancer; Survival Analysis; Kaplan-Meier method; Censored Type I*

INTRODUCTION

Analysis of survival test data is one of the statistical analyzes that is useful for testing the survival of a component or measuring the survival of a patient in the treatment of a disease [1]. In analyzing the survival test data, survival data is needed which includes the survival time and survival time status of the object or individual to be studied. Survival time is recorded as data about the duration of an event from start to finish [2]. Life time data obtained can be in the form of complete samples and incomplete samples (censored). The sample is said to be complete if all the individuals observed during a certain research period experience the desired event (failure). Meanwhile, the sample is said to be incomplete or censored if not all observation units observed during a certain period fail so that the actual survival time of some observations is unknown.

There are two types commonly used of censored samples, namely type I censored samples and type II censored samples [3]. Type I censored sample is a statistical sampling method based on a predetermined length of observation. If the life span of the sample exceeds the time allotted, the observation of the sample will be stopped or a censored sample will be obtained. Survival modeling is divided into two, namely the parametric model and the nonparametric model. The parametric survival model is used if the population distribution assumption is made first, while the nonparametric model is a model that does not depend on the population distribution assumption. There are several ways to estimate the survival function with nonparametric models, including life tables, Kaplan-Meier estimation or the Product Limit method. The Kaplan-Meier method is a method that can be used to estimate the survival function in the analysis of survival test data. The benefit of using this method is that it can provide an estimate of the chances of survival. The Kaplan-Meier method provides a graphical representation of the survival distribution [4].

One of the problems related to survival in the health sector is cancer. Cancer is a disease caused by abnormal growth of body tissue cells that turn into cancer cells [5]. There are several types of cancer including lung cancer, breast cancer, cervical cancer, head and neck cancer and so on. There are several factors that cause cancer including heredity, unhealthy lifestyle, radiation and others. In a previous study conducted by Pradika and Priatna (2019), explained the probability of survival for breast cancer patients with the Kaplan Meier method and it was concluded that the survival time of breast cancer patients who receive chemotherapy treatment has a greater chance of survival compared to those who do not receive chemotherapy [6].

Head and neck cancer is the most common type of cancer worldwide and comprises a diverse group of tumors that affect the upper aerodigestive tract [7]. The risk of this cancer is strongly associated with smoking and alcohol consumption. Head and neck cancer patients have been reported to show high rates of depression.

According to Ashraf-Ul-Alam (2021) in a previous study, head and neck cancer patients were given 2 different treatments, namely radiation therapy (RT) and radiation and chemotherapy (RTC) [8]. The research focuses on comparing the most suitable Bayesian models. It was concluded that the log-normal model fits the data better than the log-logistics and Weibull models.

From previous study, there has been no explanation regarding the estimation of survival chances. Therefore, it is necessary to know the factors that influence the survival of head and neck cancer patients. Adaptations in statistics about the factors that influence the survival time of head

and neck cancer patients can be analyzed using type I censored survival estimation using the Kaplan-Meier method.

Based on the explanation above, the authors are interested in estimating the survival of a patient with head and neck cancer. Estimating the survival of head and neck cancer patients begins by calculating the survivor function or S(t) and continues with the hazard function or h(t). The survivor function is the probability of an individual or a particular object surviving after a certain amount of time has passed [9]. The hazard function is the probability that an individual will reach a particular event at time t, provided that he has survived until that time. This function is used to express momentary opportunities then, it is known that a system has aged t [10]. After calculating the two functions, the results will be interpreted. Based on the background above, the authors wanted to conduct research with the title Estimation of Functional Survival in Head and Neck Cancer Patients Using the Kaplan-Meier Method.

METHOD

This type of research is quantitative research, namely a systematic research on a phenomenon by collecting data that can be measured using statistical, mathematical, or computational techniques [11]. with secondary data. The secondary data used is Type 1 Censored Head and Neck Cancer Data from the Northern California Oncology Group. The method used in this article is the Kaplan-Meier method, which is a nonparametric method used to estimate the survival function in the analysis of time to event data or survival data. The Kaplan-Meier method in head and neck cancer patients is very useful in estimating the probability of survival. This method requires research steps, including:

1. Collection of head and neck cancer patient data that involves information about the time of occurrence (such as death or disease progression) and the patient's status at any given point in time ensuring the data is complete and accurate.
2. Dataset preparation, by compiling a dataset that includes variable time of event, status (event or censoring), and other relevant variables such as patient characteristics or risk factors by ensuring the dataset is ready for analysis in the right format.
3. Conducted a descriptive analysis of type 1 censored head and neck cancer patient data from the Northern California Oncology Group.
4. Calculation of the survival function and hazard function, using statistical software that supports survival analysis to estimate the survival function using the Kaplan-Meier method where this software usually has a special function to calculate the survival function and produce estimates at each observation time point.
5. Visualization of the survival curve, by creating a graph of the survival curve that shows the development of the survival function over time where this graph will visualize the survival rate of head and neck cancer patients during the observation period.
6. Interpretation of the results, by analyzing the Kaplan-Meier and survival curves will provide an estimate of the survival function at each time point by interpreting these results with care, describing the survival rate of head and neck cancer patients and the factors that might affect survival.

RESULT AND DISCUSSION

The head and neck cancer patient group was treated with radiation therapy plus chemotherapy (RCT) with 45 patients. The patient's survival data is presented in Table 1 below.

Table 1. Survival Data (in days) of Head and Neck Cancer Patients treated with RCTs

Patient	Lifespan (Days)	Patient	Lifespan (Days)	Patient	Lifespan (Days)
1	37	16	173	31	633
2	84	17	179	32	725
3	92	18	194	33	759*
4	94	19	195	34	817
5	110	20	209	35	1092*
6	112	21	249	36	1245*
7	119	22	281	37	1331*
8	127	23	319	38	1557*
9	130	24	339	39	1642*
10	133	25	432	40	1771*
11	140	26	469	41	1776
12	146	27	519	42	1897*
13	155	28	528*	43	2023*
14	159	29	547*	44	2146*
15	169*	30	613*	45	2297*

Descriptive Statistics

Based on the data in table 1, the descriptive statistics are obtained in table 2 as follows.

Table 2. Descriptive Statistics

N Total	Mean	StDev	Min.	Median	Max.
45	639,20	668,859	37	319	2297

Based on Table 2, an average value of 639.2 is obtained with a standard deviation of 668.859. The basis for calculating the standard deviation is desire to know a diversity data group. In addition, the lowest (minimum) data obtained was 37 and the highest (maximum) data was 2297. The median value of 45 head and neck cancer patient data was 319. Data is classified as censored and observed (uncensored) data. There were 45 data with 30 (67%) being observed data and 15 (33%) being censored data. The proportion of censored and observed data is presented graphically in Figure 1 as follows.

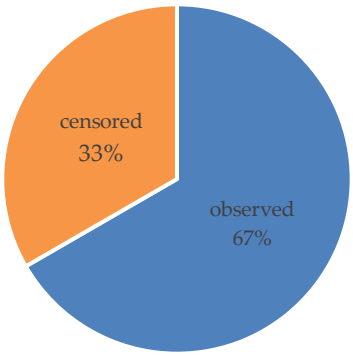


Figure 1. Graph of the Proportion of Senior and Observed Data

Normality Assumption Test

To continue the analysis of type 1 censored data using the Kaplan-Meier method, a normality test is performed first. The Kaplan-Meier method is a type of nonparametric method that can be used to estimate the survival function in the analysis of survival test data. One of the characteristics of the data that can be analyzed using nonparametric methods is that the data is not normally distributed [12]. The results of the normality test on the survival data of head and neck cancer patients are shown in table 3 as follows.

Table 3. Kolmogorov-Smirnov Normality Test

N		45
Normal Parameters[a,b]	Mean	639,20
	Std. Deviation	668,859
Most Extreme Differences	Absolute	0,207
	Positive	0,207
	Negative	-0,184
Test Statistic		0,207
Asymp. Sig. (2-tailed)		0,000[c]

Based on the normality test in table 3 above, it is found that the Asymp. Sig. 0,000 < α = 5%. It can be concluded that the data is not normally distributed, so it can be continued with type 1 censored data analysis using the Kaplan-Meier method.

Estimation of Survival and Hazard Functions with the Kaplan-Meier Method

To estimate the product limit or Kaplan-Meier estimator, it can be calculated in the following way.

1. Column ti is the survival of head and neck cancer patients undergoing treatment with radiation plus chemotherapy which includes censored and uncensored data.
2. Column Ri is the number of living individuals in the ith observation.
3. Column δi is the number of individuals that died in the i-th observation.
4. Column Ci is the number of individuals censored in the i-th observation.

5. Column qi is an estimate of the individual's probability of death in the i-th observation. The qi column can also be called a hazard function. The calculation of qi is as follows. The calculation of qi is as follows.

$$qi = \frac{\delta i}{Ri} \quad (1)$$

6. Column pi is an estimate of the probability of individual survival in the i-th observation. The calculation of pi is as follows.

$$pi = 1 - \frac{\delta i}{Ri} \quad (2)$$

7. Column S(t) is the estimated survival function for each individual included in the observation. The calculation of S(t) is as follows.

$$S(t) = \prod_{i=1}^{37} p_i \quad (3)$$

Some of Kaplan Meier's estimation calculations are as follows.

1. For a life time of 37 days
   a) Hazard function (qi)
   It is known that the number of individuals who died on the 37th day of observation (δi) was 1 patient and the number of individuals who lived on the 37th day of observation (Ri) was 45 patients, so the hazard function calculation can be done as follows.

$$q_1 = \frac{\delta_1}{R_1} = \frac{1}{45} = 0{,}022222$$

   The estimated probability of individual death (hazard function) on the 37th day is 0,022222.
   b) Estimated Chance of Individual Resilience (pi)
   It is known that the value of the hazard function is 0,022222, so the estimated chance of individual survival on the 37th day of observation is as follows.

$$p_1 = \left(1 - \frac{\delta_1}{R_1}\right) = (1 - q_1) = 1 - 0.022222 = 0{,}977778$$

   The estimated probability of individual survival on the 37th day of observation was 0,977778.
   c) Survival Function Estimation (S(t))

$$S(37) = \prod_{i=1}^{1} p_i = 0.977778$$

The estimated survival function for each individual included on the 37th day of observation was 0,977778.

2. For a life time of 155 days
   a) Hazard function (qi)
      It is known that the number of individuals who died on the 155th day of observation (δi) was 1 patient and the number of individuals who lived on the 155th day of observation (Ri) was 33 patients, so the hazard function calculation can be done as follows.

$$q_{13} = \frac{\delta_{13}}{R_{13}} = \frac{1}{33} = 0{,}030303$$

      The estimated probability of individual death (hazard function) on the 155th day is 0,030303.
   b) Estimated Chance of Individual Resilience (pi)
      It is known that the value of the hazard function is 0,030303, so the estimated chance of individual survival on the 155th day of observation is as follows.

$$p_{13} = \left(1 - \frac{\delta_{13}}{R_{13}}\right) = (1 - q_{13}) = 1 - 0{,}030303 = 0{,}969697$$

      The estimated probability of individual survival on the 155th day of observation was 0,969697.
   c) Survival Function Estimation (S(t))

$$S(155) = \prod_{i=1}^{13} p_i = 0{,}711111$$

      The estimated survival function for each individual included on the 155th day of observation was 0,711111.

3. For a life time of 1776 days
   a) Hazard function (qi)
      It is known that the number of individuals who died on the 1776th day of observation (δi) was 1 patient and the number of individuals who lived on the 1776th observation day (Ri) was 5 patients, so the hazard function calculation can be done as follows.

$$q_{41} = \frac{\delta_{41}}{R_{41}} = \frac{1}{5} = 0{,}2$$

      The estimated probability of individual death (hazard function) on day 1776 was 0,2.
   b) Estimated Chance of Individual Resilience (pi)

It is known that the value of the hazard function is 0.2, so the estimated probability of individual survival on the 1776th day of observation is as follows.

$$p_{41} = \left(1 - \frac{\delta_{41}}{R_{41}}\right) = (1 - q_{41}) = 1 - 0{,}2 = 0{,}8$$

The estimated probability of individual survival on the 1776th day of observation was 0,8.

c) Survival Function Estimation (S(t))

$$S(1776) = \prod_{i=1}^{41} p_i = 0{,}262696$$

It was found that the estimated survival function for each individual included on the 1776th day of observation was 0,262696.

Using the Kaplan-Meier method, the results of the product limit estimator or Kaplan-Meier survival data for type 1 censored head and neck cancer patients can be seen in table 4 as follows.

Table 4. Table of Estimation of the Survival Function of the Kaplan-Meier Method

ti	Ri	δi	Ci	qi	pi	S(t)
37	45	1	0	0,022222	0,977778	0,977778
84	44	1	0	0,022727	0,977273	0,955556
92	43	1	0	0,023256	0,976744	0,933333
94	42	1	0	0,02381	0,97619	0,911111
110	41	1	0	0,02439	0,97561	0,888889
112	40	1	0	0,025	0,975	0,866667
119	39	1	0	0,025641	0,974359	0,844444
127	38	1	0	0,026316	0,973684	0,822222
130	37	1	0	0,027027	0,972973	0,8
133	36	1	0	0,027778	0,972222	0,777778
140	35	1	0	0,028571	0,971429	0,755556
146	34	1	0	0,029412	0,970588	0,733333
155	33	1	0	0,030303	0,969697	0,711111
159	32	1	0	0,03125	0,96875	0,688889
169	31	0	1	0	1	0,688889
173	30	1	0	0,033333	0,966667	0,665926
179	29	1	0	0,034483	0,965517	0,642963
194	28	1	0	0,035714	0,964286	0,62
195	27	1	0	0,037037	0,962963	0,597037
209	26	1	0	0,038462	0,961538	0,574074
249	25	1	0	0,04	0,96	0,551111
281	24	1	0	0,041667	0,958333	0,528148
319	23	1	0	0,043478	0,956522	0,505185

339	22	1	0	0,045455	0,954545	0,482222
432	21	1	0	0,047619	0,952381	0,459259
469	20	1	0	0,05	0,95	0,436296
519	19	1	0	0,052632	0,947368	0,413333
528	18	0	1	0	1	0,413333
547	17	0	1	0	1	0,413333
613	16	0	1	0	1	0,413333
633	15	1	0	0,066667	0,933333	0,385778
725	14	1	0	0,071429	0,928571	0,358222
759	13	0	1	0	1	0,358222
817	12	1	0	0,083333	0,916667	0,32837
1092	11	0	1	0	1	0,32837
1245	10	0	1	0	1	0,32837
1331	9	0	1	0	1	0,32837
1557	8	0	1	0	1	0,32837
1642	7	0	1	0	1	0,32837
1771	6	0	1	0	1	0,32837
1776	5	1	0	0,2	0,8	0,262696
1897	4	0	1	0	1	0,262696
2023	3	0	1	0	1	0,262696
2146	2	0	1	0	1	0,262696
2297	1	0	1	0	1	0,262696

Table 4, shows that patients with head and neck cancer who are undergoing treatment with an RCT have a 37-day chance of survival of 0,977778, while the chance for patients with head and neck cancer who are undergoing treatment with an RCT is 0,2206, and so are so on. The value of the hazard function (failure rate) of head and neck cancer patients undergoing treatment with RCTs has a 37-day survival chance of 0,022222, while the patient's survival for 84 days is 0,044444, and so on.

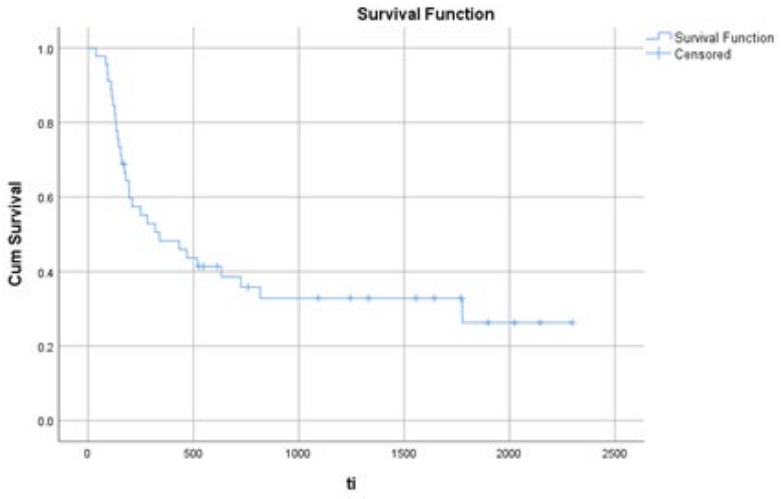


Figure 2. Survival Data Plot of Head and Neck Cancer Patients

Figure 2 shows that the survival time of head and neck cancer patients undergoing RCT treatment in the first 500 days decreased dramatically, after 500 days the plot slowly decreased, which means that the chances of a patient surviving are getting smaller as time goes by. Based on this figure, the survival chance of head and neck cancer patients undergoing RCT treatment for 130 days is 0,8.

CONCLUSION

Based on the results and discussion, head and neck cancer patients have an average life span of 639,2 days. The results of the calculation of the product limit or Kaplan-Meier estimator and the survival plot of head and neck cancer patients tend to decrease, which means that the chances of a patient surviving are getting smaller as time goes by. By using the Kaplan-Meier method, it can be seen that the chance of head and neck cancer patients undergoing RCT treatment surviving 37 days is 0,977778, while the failure rate for head and neck cancer patients surviving 37 days is 0,022222.

REFERENCE

[1] R. Nia and A. Agoestanto, "Estimator Bayes untuk Rata-Rata Tahan Hidup dari Distribusi Rayleigh pada Data Disensor Tipe II," *Unnes Journal of Mathematics*, vol. 3, no. 2, 2014.
[2] S. Mulyani and A. Mafthukhin, "Analysis Survival Therapy Hemodialysis pada Pasien Gagal Ginjal Kronik di RS Kelas B Dr. Sosodoro Djatikoesoemo Bojonegoro," *Asuhan Kesehatan Jurnal Ilmiah Ilmu Kebidanan dan Keperawatan*, vol. 7, no. 1, 2016.
[3] A. S. Yadav, S. K. Singh, and U. Singh, "Bayes estimator of the parameter and reliability function of Marshall-Olkin extended exponential distribution using hybrid Type-II censored data," *Journal of Statistics and Management Systems*, vol. 19, no. 3, pp. 325-344, 2016.
[4] B. Audina and M. Fatekurohman, "Analisis Survival pada Data Pasien Covid 19 di Kabupaten Jember," *Berkala Sainstek*, vol. 8, no. 4, pp. 118, 2020.
[5] S. S. Utami and Mustikasari, "Aspek Psikososial Pada Penderita Kanker Payudara: Studi Pendahuluan," *Jurnal Keperawatan Indonesia*, vol. 20, no. 2, pp. 65–74, 2017.
[6] R. Pradika, and B. A. Priatna, "Aplikasi Metode Kaplan Meier Sebagai Penduga Ketahanan Hidup Penderita Kanker Payudara," *Jurnal EurekaMatika*, vol. 9, no.1, pp. 35-44, 2019.
[7] M. D. Mody, J. W. Rocco, S. S. Yom, R. I. Haddad, and N. F. Saba, "Head and neck cancer," *The Lancet*, vol. 398 no. 10318, pp. 2289-2299, 2021.
[8] M. Ashraf-Ul-Alam and A. A. Khan, "Comparison of accelerated failure time models: A Bayesian study on head and neck cancer data," *Journal of Statistics Applications and Probability*, vol. 10, no. 3, pp. 715-738, 2021.
[9] R. Hidayat, M. Sam, R. Y. Wardi, and M. I. Iskandar, "Pemodelan Survival Pasien Covid-19 dengan Hazard Non-Proporsional," *Euler: Jurnal Ilmiah Matematika, Sains dan Teknologi*, vol. 10, no. 1, pp. 120–130, 2022.
[10] M. Gustina, "Identifikasi Karakteristik Hazard Rate Distribusi Generalized Exponential," Thesis, Universitas Lampung, Lampung, 2016.
[11] M. Ramdhan, *Metode penelitian*. Cipta Media Nusantara, 2021.
[12] S. Santoso, *Mahir statistik parametrik*. Elex Media Komputindo, 2019.

Analisis Tingkat Kepuasan Masyarakat Terhadap Pelayanan BPJS Kesehatan Cabang Utama Surabaya Dengan Metode *Customer Satisfaction Indeks* dan *Importance Performance Analysis*

Putri Nur Farida[(1)], Ardi Kurniawan[(2)], Sediono[(3)], Dita Amelia[(4)]

[1,2,3,4]Departement of Mathematics, Faculty of Science and Technologi, Airlangga University

Jl. Dr. Ir. H. Soekarno, Kec. Mulyorejo, Surabaya, Jawa Timur

e-mail: putri.nfarida@gmail.com[(1)], ardi-k@fst.unair.ac.id[(2)], sediono101@gmail.com[(3)]
dita.amelia@fst.unair.ac.id[(4)]

ABSTRAK

Kesehatan merupakan hak asasi manusia dan salah satu unsur kesejahteraan yang harus diwujudkan sesuai cita-cita bangsa Indonesia. Diperlukan pembentukan badan penyelenggara yang berbentuk badan hukum pelayanan publik salah satunya BPJS Kesehatan yang bertujuan untuk memberikan jaminan berupa perlindungan kesehatan kepada masyarakat, serta diharapkan dapat memberikan pelayanan kualitas yang bermutu baik, karena kepuasan merupakan aspek yang paling penting bagi instansi pemerintahan. Penelitian ini bertujuan untuk mengetahui tingkat kepuasan masyarakat terhadap Kantor BPJS Kesehatan dengan data yang digunakan sejumlah 100 responden. Metode yang digunakan yaitu *Customer Satisfaction Indeks* yang merupakan indeks untuk menentukan tingkat kepuasan pelanggan secara menyeluruh menggunakan pendekatan yang mempertimbangkan tingkat kepentingan dari atribut atribut produk dan jasa yang diukur dan *Importance Performance Analysis* yang merupakan merupakan salah satu teknik penerapan yang mudah untuk mengatur atribut dari tingkat kepentingan dan tingkat pelaksanaan itu sendiri. Hasil penelitian diperoleh bahwa metode CSI sebesar 82,038 % yang berarti masyarakat merasa sangat puas terhadap pelayanan di Kantor BPJS Kesehatan Cabang Utama Surabaya, pada metode IPA menunjukkan pada kuadran I,II,III,IV masing masing terdapat 11, 1, 9, dan 1 variabel, dan variabel yang pelu ditingkatkan adalah kuadran II, yaitu kecepatan dalam menangani keluhan secara online melalui saluran telepon dan email.
Kata kunci: BPJS Kesehatan; *Customer Satisfaction Indeks*; *Importance Performance Analysis*

ABSTRACT

Health is a human right and one of the elements of welfare that must be realized according to the ideals of the Indonesian nation. It is necessary to establish an organizing body in the form of a public service legal entity, one of which is BPJS Health which aims to provide guarantees in the form of health protection to the public, and is expected to be able to provide good quality services, because satisfaction is the most important aspect for government agencies. This study aims to determine the level of public satisfaction with the Health BPJS Office with the data used by 100 respondents. The method used is the Customer Satisfaction Index which is an index to determine the overall level of customer satisfaction using an approach that considers the level of importance of the attributes of products and services being measured and Importance Performance Analysis which is an easy application technique to adjust the attributes of the level of importance and the level of implementation itself. The results showed that the CSI method was 82,038%, which means that the community was very satisfied with the service at the BPJS Kesehatan Office, Surabaya Main Branch. The IPA method showed that in quadrants I, II, III, IV, there were 11, 1, 9, and 1 variables respectively, and the variable that needs to be improved is quadrant II, namely the speed in handling complaints online via telephone and e-mail.
Keywords: BPJS Kesehatan; *Customer Satisfaction Indeks*; *Importance Performance Analysis*

PENDAHULULAN

Kesehatan merupakan hak asasi manusia dan salah satu unsur kesejahteraan yang harus diwujudkan sesuai dengan cita cita bangsa Indonesia. Diperlukan pembentukan badan penyelenggara yang berbentuk badan hukum pelayanan publik salah satunya Badan Penyelenggara Jaminan Sosial (BPJS) yang bertujuan untuk memberikan jaminan perlindungan kesehatan agar masyarakat memperoleh manfaat pemeliharaan kesehatan. Data dari BPJS Kesehatan, jumlah peserta Jaminan Kesehatan Nasional di Indonesia per 31 Agustus 2022 sejumlah 243.282.029 jiwa [1]. Salah satu daerah di Indonesia yang wajib memiliki kantor BPJS Kesehatan yang memadai yaitu Kota Surabaya. Dikarenakan banyaknya masyarakat di Surabaya yang membutuhkan pelayanan BPJS Kesehatan, Kantor BPJS Kesehatan Cabang Utama Surabaya diharapkan dapat memberikan pelayanan kualitas yang bermutu baik dan berkualitas pada jasanya. Kualitas yang bermutu baik dalam pelayanannya pasti berpengaruh terhadap kepuasan dari masyarakat. Namun pada kenyataannya masih banyak kekurangan terkait pelayanan di kantor BPJS Kesehatan Cabang Utama Surabaya.

Dalam analisis ini, peneliti menggunakan metode *Customer Satisfaction Indeks* dan *Importance Performance Analysis*. *Customer Satisfaction Indeks* merupakan indeks untuk menentukan tingkat kepuasan pelanggan secara menyeluruh menggunakan pendekatan yang mempertimbangkan tingkat kepentingan dari atribut atribut produk dan jasa yang diukur [2]. Metode CSI efisien digunakan karena tidak hanya memperoleh indeks kepuasan tetapi sekaligus memperoleh informasi tentang dimensi atau atribut mana saja yang diperlukan. Namun, terdapat kekurangan dari metode CSI yaitu pada metode ini tidak dapat menentukan atribut prioritas dari atribut atribut yang dihasilkan [3]. Untuk menutupi kekurangan dari metode CSI ini maka diperlukan metode lain, yaitu metode *Importance Performance Analysis*. Metode *Importance Performance Analysis* merupakan salah satu teknik penerapan yang mudah untuk mengatur atribut dari tingkat kepentingan dan tingkat pelaksanaan itu sendiri yang digunakan untuk pengembangan program pemasaran yang efektif dan hasil dari analisis metode IPA akan disampaikan dalam bentuk kuadran yang tiap tiap kuadrannya telah memiliki kategori tertentu [4].

Penelitian terkait kepuasan pelanggan terhadap BPJS Kesehatan sebelumnya pernah dilakukan oleh Atma, dkk [5]. yang bertujuan untuk menganalisis kepuasan pasien BPJS Rawat Jalan dengan metode *Servqual*, CSI dan IPA di klinik Dr. M. Suherman. Hasil penelitiannya berdasarkan metode *Servqual* kualitas mutu pelayanan belum dapat memenuhi harapan pasien, metode CSI didapatkan hasil bahwa tingkat kepuasan pasien secara menyeluruh berada pada kriteria “puas”, dan metode IPA didapatkan hasil terdapat lima atribut pelayanan dengan prioritas tertinggi waktu tunggu pasien mendapatkan obat kurang dari 30 menit. Dari penelitian sebelumnya, penulis ingin mengembangkan penelitian lebih lanjut terkait kepuasan masyarakat terhadap pelayanan BPJS Kesehatan di Kantor BPJS Kesehatan Cabang Utama Surabaya. Metode yang digunakan adalah *Customer Satisfaction Indeks* dan *Importance Performance Analysis.*

METODE

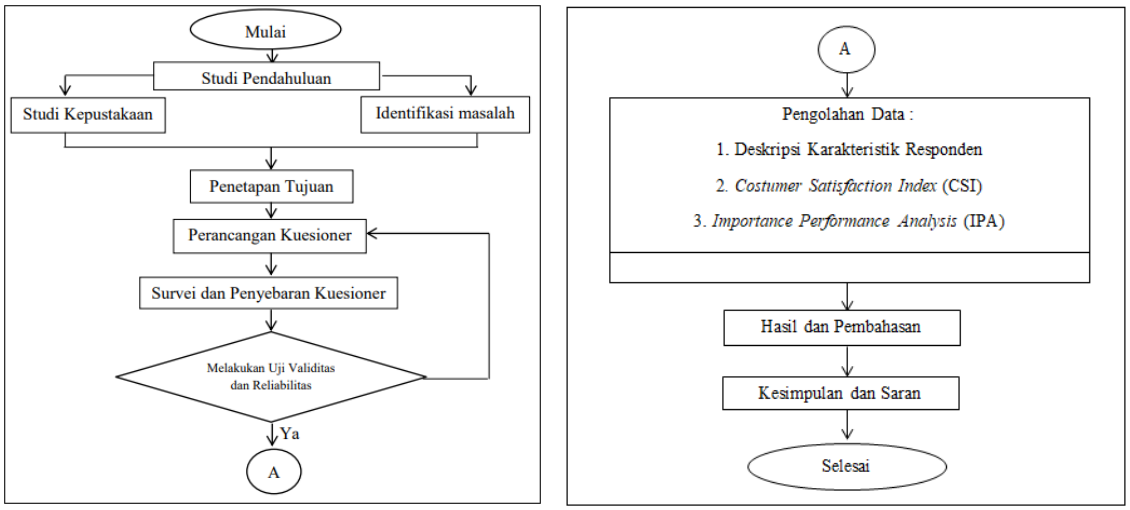
Sumber data yang digunakan dalam penelitian ini adalah data primer yang diperoleh dengan melakukan survei yang dilakukan dengan membagikan kuesioner secara langsung pada tanggal 12 Desember sampai 26 Desember 2022. Teknik pengumpulan data dalam penelitian ini menggunakan

purposive sampling yaitu menentukan responden dengan pertimbangan dan tujuan tertentu. Penentuan ukuran sampel menggunakan rumus [6]:

$$n = \frac{Z^2_{1-\frac{\alpha}{2}} P(1-P)}{d^2} \quad (1)$$

$$n = \frac{1{,}96^2(0{,}5)(0{,}5)}{0{,}1^2} = 96{,}04$$

Secara garis besar langkah langkah dalam penelitian ini dapat digambarkan melalui diagram alur *(flow chart)* sebagai berikut :



Gambar 1. Diagram Alir Penelitian

Berikut langkah yang digunakan untuk menganalisis data dalam penelitian sebagai berikut :

1. Melakukan Uji Validitas dan Uji Reliabilitas terhadap data hasil kuesioner.
2. Melakukan analisis statistika deskriptif untuk mendeskripsikan jawaban responden
3. Menganalisis tingkat kepuasan masyarakat di Kantor BPJS Kesehatan Surabaya berdasarkan *Customer Satisfaction Indeks* dengan langkah sebagai berikut :
   a. Menentukan *Mean Important Score* (MIS)
   b. Menghitung *Wight Factor* (WF)
   c. Menghitung *Satisfaction Score* (MSS)
   d. Menghitung *Weight Score* (WS)
   e. Menghitung *Customer Satisfaction Indeks*
4. Menganalisis faktor yang perlu ditingkatkan dan dipertahankan sebagai berikut :
   a. Menghitung rata rata perfaktor dari kolom kenyataan (X) dan harapan (Y)sehingga didapat $(\bar{X})$ dan $(\bar{Y})$.
   b. Menghitung rata rata $(\bar{X})$ dan $(\bar{Y})$ untuk mendapatkan batas $(\bar{\bar{X}})$ dan $(\bar{\bar{Y}})$.
   c. Membuat plot dalam diagram kartesius
   d. Membuat intepretasi diagram kartesius berdasarkan hasil dengan cara melihat variabel yang masuk pada kuadran I,II,III,dan IV, sehingga dapat ditarik kesimpulan.

HASIL DAN DISKUSI

Berdasarkan hasil perhitungan ukuran sampel, jumlah sampel minimal yang harus diambil dalam penelitian ini sebanyak 96 responden. Penelitian ini menggunakan data primer yang diperoleh dengan melakukan survei secara langsung dengan membagikan kuesioner dan melakukan wawancara kepada masyarakat yang berkunjung ke kantor BPJS Kesehatan Cabang Utama Surabaya.

1. Uji Validitas

Uji validitas merupakan uji yang digunakan untuk mengetahui serta menguji ketepatan dan ketetapan suatu alat ukur [7]. Hipotesis yang digunakan adalah :

H_0: $\rho = 0$ (tidak terdapat korelasi antara skor pengamatan dengan skor total)

H_1: $\rho \neq 0$ (terdapat korelasi antara skor pengamatan dengan skor total)

Dalam Uji Validitas, statistik uji yang digunakan pada Persamaan dengan daerah penolakan H_0 jika nilai *p-value* ≤ α (0,05).

Tabel 1. Uji Validitas pada Seluruh Dimensi

Variabel	P-Value	Keputusan	Kesimpulan	Variabel	P-Value	Keputusan	Kesimpulan
$X_{1.1}$	0,000	Tolak H_0	Valid	$X_{3.3}$	0,000	Tolak H_0	Valid
$X_{1.2}$	0,000	Tolak H_0	Valid	$X_{2.4}$	0,000	Tolak H_0	Valid
$X_{1.3}$	0,000	Tolak H_0	Valid	$X_{4.1}$	0,000	Tolak H_0	Valid
$X_{1.4}$	0,000	Tolak H_0	Valid	$X_{4.2}$	0,000	Tolak H_0	Valid
$X_{1.5}$	0,000	Tolak H_0	Valid	$X_{4.3}$	0,000	Tolak H_0	Valid
$X_{2.1}$	0,000	Tolak H_0	Valid	$X_{4.4}$	0,000	Tolak H_0	Valid
$X_{2.2}$	0,000	Tolak H_0	Valid	$X_{5.1}$	0,000	Tolak H_0	Valid
$X_{2.3}$	0,000	Tolak H_0	Valid	$X_{5.2}$	0,000	Tolak H_0	Valid
$X_{2.4}$	0,000	Tolak H_0	Valid	$X_{5.3}$	0,000	Tolak H_0	Valid
$X_{3.1}$	0,000	Tolak H_0	Valid	$X_{5.4}$	0,000	Tolak H_0	Valid
$X_{3.2}$	0,000	Tolak H_0	Valid	$X_{5.5}$	0,000	Tolak H_0	Valid

Berdasarkan hasil uji validitas, didapatkan bahwa semua variabel pertanyaan memiliki nilai *p-value* lebih kecil dari α=0,05 yang berarti Tolak H_0. Dapat disimpulkan semua pertanyaan untuk seluruh variabel pada kuesioner telah valid sehingga mampu mengukur apa yang diinginkan.

2. Uji Reliabilitas

Setiap instrumen dari pertanyaan memiliki tingkat reabilitas yang sangat tinggi jika nilai koefisien *Cronbach's Alpha* lebih besar dari 0,8, lalu dikatakan tinggi jika nilai koefisien *Cronbach's Alpha* jika memiliki nilai diantara 0,6 sampai 0,8, kemudian dapat pula dikatakan sedang jika nilai koefisien *Croncach's Alpha* diantara 0,4 sampai 0,6, lalu dikatakan rendah jika nilai koefisien *Cronbach's Alpha* diantara 0,2 sampai 0,4.

Tabel 2. Uji Reliabilitas

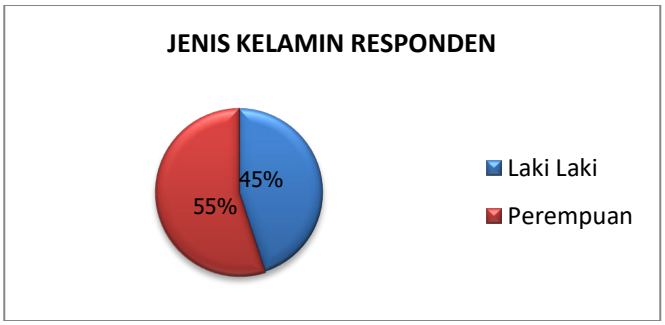
Variabel	*Cronbach's Alpha*	Kesimpulan	Variabel	*Cronbach's Alpha*	Kesimpulan

Dimensi Keandalan (X_1)	0,780	Reliabilitas Tinggi	Dimensi Empati (X_4)	0,639	Reliabilitas Tinggi
Dimensi Daya Tanggap (X_2)	0,608	Reliabilitas Tinggi	Dimensi Tampilan Fisik (X_5)	0,618	Reliabilitas Tinggi
Dimensi Jaminan (X_3)	0,794	Reliabilitas Tinggi			

Berdasarkan Tabel 2, diketahui bahwa hasil analisis dengan nilai *cronbach's alpha* memiliki reliabilitas tinggi. Uji reliabilitas dapat diukur dengan *Cronbach's alpha* (α) [8], dengan diperoleh nilai *Cronbach's Alpha* diantara 0,6-0,8 yang berarti keseluruhan jawaban dari responden terhadap pertanyaan yang diberikan konsisten stabil dari waktu ke waktu.

3. Statistika Deskriptif

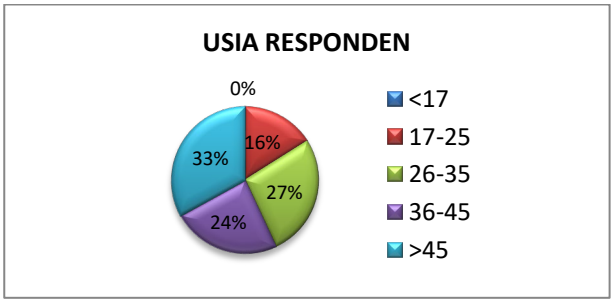
a. Hasil Analisis Karakteristik masyarakat Berdasarkan Jenis Kelamin



Gambar 2. Persentase Responden Berdasarkan Jenis Kelamin

Berdasarkan Gambar 2 dari 100 responden, dapat diketahui bahwa jenis kelamin masyarakat pengguna BPJS Kesehatan sebesar 55% atau 55 responden laki-laki dan sebesar 45% atau 45 responden berjenis kelamin perempuan. Hal tersebut selaras dengan jumlah data dari Badan Pusat Statistik Provinsi Jawa Timur bahwa jumlah masyarakat perempuan di Jawa Timur lebih banyak dibandingkan dengan jumlah masyarakat berjenis kelamin laki laki, dengan presentase perempuan 50,20% dan laki laki 49,80% [9].

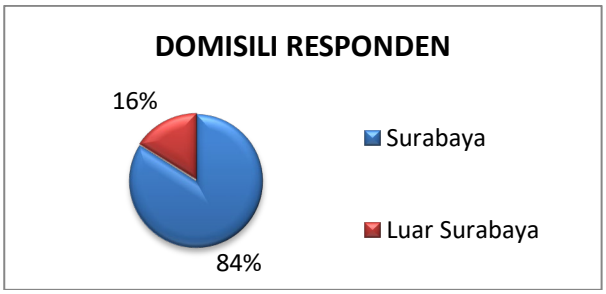
b. Hasil Analisis Karakteristik masyarakat Berdasarkan Usia



Gambar 3. Persentase Responden Berdasarkan Usia

Berdasarkan Gambar 3 dari 100 responden, dapat diketahui bahwa usia masyarakat pengguna BPJS Kesehatan sebesar 0% atau tidak terdapat responden yang berusia dibawah 17 tahun, 16% atau 16 responden berusia 17-25 tahun, 27% atau 27 responden berusia 26-35 tahun, 24% atau 24 responden berusia 36-45 tahun, dan sebesar 33% atau 33 responden berusia lebih dari 45 tahun. hal tersebut dikarenakan sebagian besar masyarakat yang datang ke kantor BPJS Kesehatan Cabang Utama Surabaya adalah ibu hamil dan lansia yang membutuhkan bantuan BPJS Kesehatan karena kebutuhan yang lebih mendesak

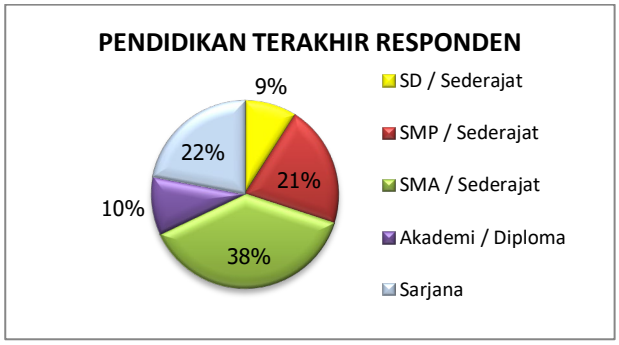
c. Hasil Analisis Karakteristik masyarakat Berdasarkan Domisili



Gambar 4. Persentase Responden Berdasarkan Domisili

Berdasarkan Gambar 4 dari 100 responden, diketahui domisili masyarakat pengguna BPJS Kesehatan sebesar 84% atau 84 responden berdomisili Surabaya dan sebesar 16% atau 16 responden berdomisili luar Surabaya. Dapat dilihat bahwa mayoritas berasal dari kota Surabaya, hal tersebut dikarenakan kantor BPJS Kesehatan memang terdapat pada masing masing kota di Indonesia, oleh karena itu terdapat banyak masyarakat yang mengunjungi kantor BPJS Kesehatan di kota masing masing, termasuk di Surabaya. Namun, ada beberapa responden yang berasal dari daerah lain selain Surabaya namun mengunjungi Kantor BPJS Kesehatan Cabang Utama Surabaya, diantaranya daerah Gresik, Sidoarjo, dan Mojokerto.

d. Hasil Analisis Karakteristik masyarakat Pengguna BPJS Kesehatan Berdasarkan Pendidikan Terakhir

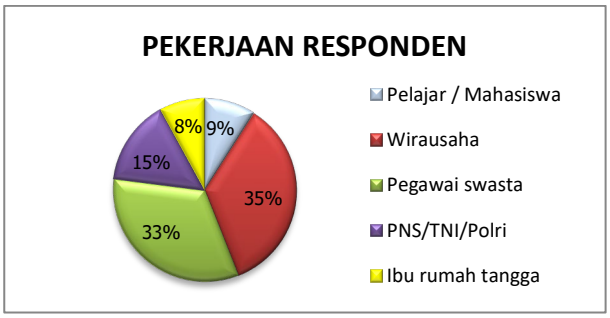


Gambar 5. Persentase Responden Berdasarkan Pendidikan Terakhir

Berdasarkan Gambar 5 dari 100 responden, dapat diketahui bahwa pendidikan terakhir masyarakat pengguna BPJS Kesehatan sebesar 9% atau 9 responden dengan pendidikan terakhir

SD/Sederajat, 21% atau 21 responden dengan pendidikan terakhir SMP/Sederajat, 38% atau 38 responden dengan pendidikan terakhir SMA/Sederajat, 10% atau 10 responden dengan pendidikan terakhir Akademi/Diploma, dan 22% atau 22 responden dengan pendidikan terakhir Sarjana. Menurut data sensus Badan Pusat Statistik Kota Surabaya, mayoritas tingkat pendidikan tertinggi yang ditamatkan adalah SMA atau sederajat [10]. Berdasarkan hal tersebut, maka bisa menjadi salah satu faktor mayoritas pendidikan terakhir masyarakat di Kantor BPJS Kesehatan Cabang Utama Surabaya adalah tamatan SMA atau sederajat.

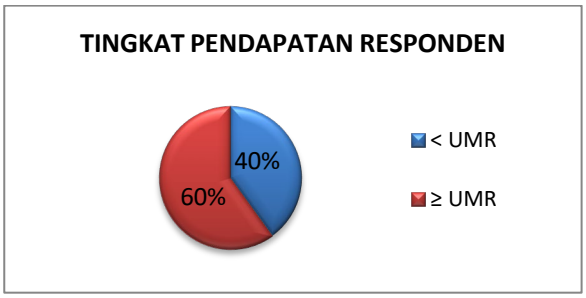
e. Hasil Analisis Karakteristik masyarakat Berdasarkan Pekerjaan



Gambar 6. Persentase Responden Berdasarkan Pekerjaan

Berdasarkan Gambar 6, dapat diketahui bahwa pekerjaan masyarakat pengguna BPJS Kesehatan sebesar 9% atau 9 responden dengan pekerjaan sebagai Pelajar/Mahasiswa, 35% atau 35 responden dengan pekerjaan Wirausaha, 33% atau 33 responden dengan pekerjaan Pegawai Swasta, 15% atau 15 responden dengan pekerjaan PNS/TNI/Polri, dan sebesar 8% atau 8 responden dengan pekerjaan sebagai Ibu Rumah Tangga. Mayoritas memiliki pekerjaan sebagai wirausaha, hal tersebut disebabkan karena masyarakat yang bekerja sebagai wirausaha tidak tergabung di BPJS Kesehatan secara otomatis seperti halnya Pegawai Negeri Sipil (PNS), hal tersebut menyebabkan mayoritas wirausaha mengurus BPJS Kesehatan secara pribadi di Kantor BPJS Kesehatan Cabang Utama Surabaya.

f. Hasil Analisis Karakteristik Berdasarkan Tingkat Pendapatan Perbulan



Gambar 7. Persentase Responden Berdasarkan Tingkat Pendapatan

Berdasarkan Gambar 7 dari 100 responden, dapat diketahui bahwa tingkat pendapatan masyarakat pengguna BPJS Kesehatan di Kantor BPJS Kesehatan Cabang Utama Surabaya sebesar 40% atau 40 responden dibawah UMR Surabaya yaitu 4.375.479,19 dan sebesar 60% atau 60

responden dengan tingkat pendapatan lebih dari sama dengan UMR Surabaya. Hal tersebut dikarenakan mayoritas bekerja sebagai wirausaha, pegawai swasta, dan PNS/TNI/Polri yang kemungkinan besar memiliki penghasilan atau tingkat pendapatan ≥ UMR Surabaya.

4. *Customer Satisfaction Indeks* (CSI)

Customer Satisfaction Indeks diperlukan untuk mengetahui tingkat kepuasan pengguna secara menyeluruh dengan memperhatikan tingkat kepentingan dari atribut atribut produk atau jasa [11]. Hasil perhitungan *Customer Satisfaction Indeks* sebagai berikut:

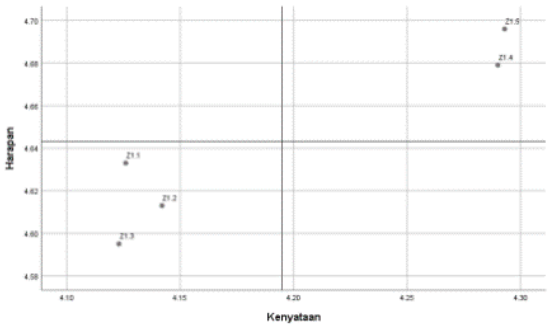
Tabel 3. Hasil *Customer Satisfaction Indeks*

j	Variabel	Rata-rata skor Harapan (MIS_j)	Weight Factor (WF_j)	Rata-rata Skor Kepuasan (MSS_j)	Weight Score (WS_j)
1	$Z_{1.1}$	4.633	4.554	4.126	18.761
2	$Z_{1.2}$	4.613	4.534	4.142	18.771
3	$Z_{1.3}$	4.595	4.516	4.123	18.621
4	$Z_{1.4}$	4.679	4.599	4.29	19.729
5	$Z_{1.5}$	4.696	4.616	4.293	19.815
6	$Z_{2.1}$	4.43	4.354	3.931	17.112
7	$Z_{2.2}$	4.77	4.688	2.962	13.877
8	$Z_{2.3}$	4.72	4.639	4.073	18.895
9	$Z_{2.4}$	4.71	4.629	4.123	19.087
10	$Z_{3.1}$	4.851	4.768	4.549	21.689
11	$Z_{3.2}$	4.851	4.768	4.548	21.685
12	$Z_{3.3}$	4.711	4.630	4.237	19.619
13	$Z_{3.4}$	4.681	4.601	4.251	19.558
14	$Z_{4.1}$	4.554	4.476	4.106	18.379
15	$Z_{4.2}$	4.612	4.533	4.161	18.862
16	$Z_{4.3}$	4.721	4.640	4.379	20.319
17	$Z_{4.4}$	4.6	4.521	4.171	18.858
18	$Z_{5.1}$	4.639	4.560	4.13	18.831
19	$Z_{5.2}$	3.963	3.895	3.183	12.398
20	$Z_{5.3}$	4.608	4.529	4.119	18.655
21	$Z_{5.4}$	4.455	4.379	3.977	17.414
22	$Z_{5.5}$	4.65	4.570	4.213	19.255
Total					**410.191**
Nilai CSI					$\frac{410.191}{5} = 82.038$

Berdasarkan Tabel 3 didapatkan hasil perhitungan CSI sebesar 82,038% yang berarti masyarakat pengguna BPJS Kesehatan di Kantor BPJS Kesehatan Cabang Utama Surabaya sangat puas terhadap pelayanan yang diberikan.

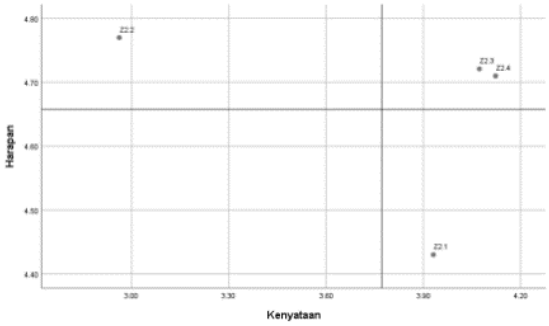
5. *Importance Performance Analysis* (IPA)

Analisis IPA digunakan untuk membandingkan antara penilaian konsumen terhadap kepentingan kualitas layanan *(Importance)* dengan tingkat performansi kualitas layanan *(Performance)*[12,13].



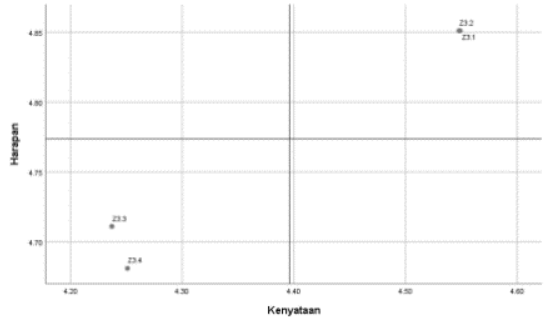
Gambar 8. Hasil IPA Dimensi Kendalan

Berdasarkan gambar 8 yang masuk pada kuadran I yang merupakan pelayanan yang unggul sehingga perlu dipertahankan yaitu $(Z_{1.4})$ dan $(Z_{1.5})$. Pada kuadran III yang merupakan Prioritas rendah dan unsur pelayanan kurang diperhatikan oleh pihak Kantor BPJS Kesehatan Cabang Utama Surabaya serta kualitas pelayanannya yang rendah yaitu $(Z_{1.1})$, $(Z_{1.2})$, dan $(Z_{1.3})$.



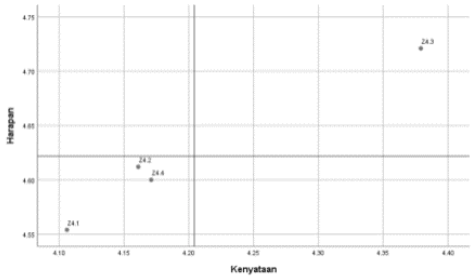
Gambar 9. Hasil IPA Dimensi Daya Tanggap

Berdasarkan gambar 9 variabel yang masuk pada kuadran I yang merupakan pelayanan yang unggul sehingga perlu dipertahankan yaitu $(Z_{2.3})$, dan $(Z_{2.4})$, Pada kuadran II yang merupakan pelayanan yang prioritas tinggi dan butuh perbaikan yaitu $(Z_{2.2})$, dan pada kuadran IV yang merupakan pelayanan yang kurang penting namun pelayanannya memuaskan yaitu $(Z_{2.1})$.



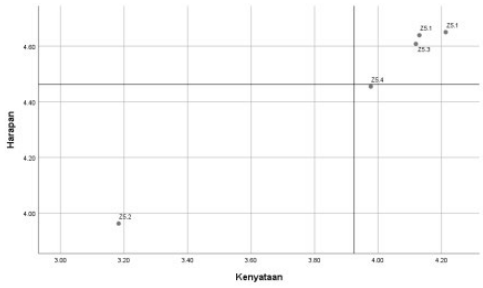
Gambar 10. Hasil IPA Dimensi Jaminan

Berdasarkan gambar 10 variabel yang masuk pada kuadran I yang merupakan pelayanan yang unggul sehingga perlu dipertahankan yaitu $(Z_{3.1})$, dan $(Z_{3.2})$, Pada kuadran III yang merupakan merupakan Prioritas rendah dan unsur pelayanan kurang diperhatikan oleh pihak Kantor BPJS Kesehatan Cabang Utama Surabaya serta kualitas pelayanannya yang rendah yaitu $(Z_{3.3})$, dan $(Z_{3.4})$.



Gambar 11. Hasil IPA Dimensi Empati

Berdasarkan gambar 11 variabel yang masuk pada kuadran I yang merupakan pelayanan yang unggul sehingga perlu dipertahankan yaitu $(Z_{4.3})$. Pada Pada kuadran III yang merupakan Prioritas rendah dan unsur pelayanan kurang diperhatikan oleh pihak Kantor BPJS Kesehatan Cabang Utama Surabaya serta kualitas pelayanannya yang rendah yaitu $(Z_{4.1})$, $(Z_{4.2})$, dan $(Z_{4.4})$.



Gambar 12. Hasil IPA Dimensi Tampilan Fisik

Berdasarkan gambar 12 variabel yang masuk pada kuadran I pelayanan yang unggul sehingga perlu dipertahankan yaitu $(Z_{5.1})$, $(Z_{5.3})$, dan $(Z_{5.5})$, Pada kuadran III Prioritas rendah dan unsur pelayanan kurang diperhatikan oleh pihak Kantor BPJS Kesehatan Surabaya serta kualitas pelayanannya rendah yaitu $(Z_{5.2})$, serta pada kuadran IV terdapat unsur pelayanan yang dianggap kurang penting namun memuaskan adalah $(Z_{5.4})$.

Berdasarkan gambar diatas, dapat dijelaskan bahwa variabel yang perlu dipertahankan berada di Kuadran I, yaitu ketuntasan pegawai dalam memberikan pelayanan kepada masyarakat, pengetahuan pegawai terhadap persyaratan pelayanan di kantor BPJS Kesehatan Surabaya, Ketanggapan pegawai BPJS Kesehatan Surabaya dalam menjawab/menanggapi pertanyaan dari masyarakat, ketepatan waktu dalam menyelesaikan pembuatan dan pendaftaran BPJS Kesehatan, Ketanggapan pegawai dalam memberikan informasi, Keamanan dokumen yang diserahkan, menjamin kerahasiaan data diri, pelayanan sesuai urutan kedatangan, kebersihan dan kenyamanan ruang tunggu, kerapian penampilan pegawai, dan ketersediaan toilet dan musholla dengan bersih.

Lalu, variabel yang perlu ditingkatkan yaitu berada pada Kuadran II, yaitu meningkatkan kinerja pada kecepatan dalam menanggapi keluhan secara online melalui saluran telepon dan email. Variabel pada kuadran II perlu ditingkatkan agar masyarakat pengguna BPJS Kesehatan di Kantor BPJS Kesehatan Cabang Utama Surabaya senantiasa mendapat kepuasan dari pelayanan Kantor BPJS Kesehatan.

KESIMPULAN

Berdasarkan penelitian terhadap 100 masyarakat yang datang ke Kantor BPJS Kesehatan Cabang Utama Surabaya tanggal 12 Desember 2023 sampai 23 Desember 2023 mayoritas berjenis kelamin perempuan, berusia lebih dari 45 tahun, berasal dari Surabaya, pendidikan terakhir SMA/Sederajat, memiliki pekerjaan sebagai Wirausaha, dan memiliki pendapatan ≥ UMR Surabaya. Berdasarkan nilai *Customer Satisfaction Indeks* 82, 038 % yang berarti masyarakat merasa sangat puas terhadap pelayanan di Kantor BPJS Kesehatan Cabang Utama Surabaya. Atribut yang perlu ditingkatkan oleh Kantor BPJS Kesehatan Cabang Utama Surabaya berdasarkan analisis *Importance Performance Analysis* yaitu Kecepatan dalam menanggapi keluhan secara online melalui saluran telepon dan email.

UCAPAN TERIMA KASIH

Saya ucapkan terimakasih kepada Kantor BPJS Kesehatan Cabang Utama Surabaya dikarenakan memperbolehkan kami untuk melakukan penelitian terkait kepuasan masyarakat terhadap pelayanan Kantor BPJS Kesehatan Cabang Utama Surabaya.

DAFTAR PUSTAKA

[1] W. Krisdayanti, "Pemanfaatan Aplikasi Mobile Jaminan Kesehatan Nasional (JKN) Untuk Meningkatkan Efektivitas Pelayanan BPJS Kesehatan Di Kota Medan," 2021.
[2] D. P. Dian Anggraini, "Analisis Persepsi Konsumen Menggunakan Metode Importance Performance Analysis dan Customer Satisfaction Index," Jurnal Industri, vol. 4, no. 2, pp. 74-81.
[3] D. M. Mandasari, "Analisis Kepuasan Konsumen Terhadap Kualitas Produk dan Pelayanan Mangsi Grill and Coff Denpasar," Jurnal Rekayasa dan Manajemen Agroindustri, vol. 7, no. 3, pp. 336-346, 2019.
[4] D. Ruhimat, "Kepuasan Pelanggan," Jakarta: PT. Gramedia Pustaka Utama, 2015.
[5] A. Deharja, F. Putri, dan L. Oktaviotika, "Analisis Kepuasan Pasien Bpjs Rawat Jalan Dengan Metode Servqual, CSI dan IPA di Klinik Dr. M. Suherman," Jurnal Kesehatan, vol. 5, no. 2, pp. 106-115, April 2017.
[6] "Besar Sampel Dalam Penelitian Kesehatan," Yogyakarta: U. G. Mada, 1997.
[7] Sugiyono, "Metode Penelitian Pendidikan Pendekatan Kuantitatif, Kualitatif, dan R & D," Bandung: Alfabeta, 2019.
[8] L. Amanda, F. Yanuar, dan D. Devianto, "Uji Validitas Dan Reliabilitas Tingkat Partisipasi Politik Masyarakat Kota Padang," Jurnal Matematika UNAND, vol. VIII, pp. 179-188, 2019.
[9] "Badan Pusat Statistik Jawa Timur," [Tanggal diakses], from https://jatim.bps.go.id/subject/28/pendidikan.html
[10] "Data Sensus BPS Kota Surabaya," [Tanggal diakses], from: https://surabayakota.bps.go.id/subject/6/tenaga-kerja.html

[11] Y. Siyamto, "Kualitas Pelayanan Bank Dengan Menggunakan Metode Importance Performance Analysis (IPA) Dan Customer Satisfaction Index (CSI) Terhadap Kepuasan Nasabah," Jurnal Ilmiah Ekonomi Islam, vol. 03, no. 01, 2017.
[12] C. Martinez, "Evaluation Report: Tools Cluster Networking Meeting #1," Arizona: CenterPoint Institute, Inc., 2003.
[13] G. Anuraga, A. Indrasetianingsih, and M. Athoillah, "Pelatihan Pengujian Hipotesis Statistika Dasar dengan Software R," BUDIMAS J. Pengabdi. Masy., vol. 3, no. 2, 2021.