EISSN : 2654-7511
PISSN : 2089-0028

J STATISTIKA

JURNAL ILMIAH TEORI DAN APLIKASI STATISTIKA

Volume 16 | Nomor 2 | 2023

EDITORIAL TEAM

INTRODUCTION

We are delighted to announce the current publication of Volume 16, Number 2 of JStatistika, affiliated with the Statistics Department at PGRI Adi Buana University Surabaya, has been released in December 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 Arfima-Garch Menggunakan Variasi Estimasi Parameter Pembeda D Pada Data Long Memory; Pemodelan Regresi Data Panel Harga Beras di Wilayah Indonesia Bagian Barat; Hubungan Faktor Demografis dengan Kejadian Malaria di Kecamatan Wewewa Timur: Pendekatan Analisis Chi-Square; Comparison of K-Means and K-Medoids Clustering for Grouping The Sub-Districts In Bojonegoro Regency Based On Educational Supporting Factors; Fuzzy C-Means for Regional Clustering in East Java Province Based on Human Development Index Indicators; Investigating the Impact of Mobile Legends Gameplay on Students' Academic Performance with Ordinal Logistic Regression; Clustering Villages in the Mountain Areas in West Java Based on Tourism Potential Using K-Prototype Algorithm; Kernel Nonparametric Regression Modeling with the Nadaraya-Watson Estimator (Case Study: Fertility in the Southern Sumatra Region); Model Persamaan Struktural Faktor – Faktor Yang Mempengaruhi Kepuasan Masyarakat Dalam Pemeriksaan Kesehatan di UPTD Puskesmas Pasir Putih Sawangan Depok; Analysis of the Timeliness of Graduation of FMIPA College KIP Students at Bengkulu University Using Binary Logistic Regression; Diversification of Jakarta Islamic Index (JII) Stock Optimal Portfolio for the Period 2018-2023; Forecasting PT Triputra Agro Persada Tbk (TAPG) Share Prices Using Multivariate Time Series Analysis; Identifying Factors that Influence Life Expectancy in Central Java Using Spatial Regression Models"

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, Worldcat, Scilit, ROAD, Onesearch, Journal Stories, Dimensions, Base, Open Alex, Wikidata, Internet Archieve, Root Indexing, Core, Harvard Library, Universiteit Leiden Library, Semantic Scholar, Open Air Explore, ASCI, Cite Factor, University of Saskatchewan Library, The University of Queensland Library, George University Library and Boston University Library.

Surabaya, December 2023

Editor in Chief

LIST OF CONTENTS

Penerapan Model Arfima-Garch Menggunakan Variasi Estimasi Parameter Pembeda D Pada Data *Long Memory*

Isran K. Hasan [(1)], Muhammad Janur [(2)], Nurwan [(3)]

[1]Jurusan Matematika, Fakultas Matematika dan IPA, Universitas Negeri Gorontalo

Jl. Prof. Dr. Ing. B. J. Habibie, Kabupaten Bone Bolango, Gorontalo, Indonesia

e-mail: isran.hasan@ung.ac.id[(1)], muhammadjanur514@gmail.com[(2)], nurwan@ung.ac.id[(3)]

ABSTRAK

Emas menjadi salah satu aset keuangan bagi negara dan menjadi komponen cadangan moneter global untuk perdagangan dan perlindungan ketika menghadapi krisis keuangan secara tiba-tiba. Beberapa data ekonomi sering mengalami ketergantungan atau dependensi jangka panjang (long memory). Salah satu model yang mampu mengatasi masalah tersebut adalah model *Autotegressive Fractionally Integrated Moving Average* (ARFIMA). Ada beberapa metode yang digunakan untuk menentukan estimasi parameter pembeda d yaitu metode *Geweke and Porter Hudak* dan metode *Rescaled Range Statistics* (R/S). Pada beberapa tipe data runtun waktu terkadang mengalami pengelompokan volatilitas (residual tidak konstan). Metode yang dapat digunakan untuk mengatasi masalah tersebut adalah metode *Generalized Autoregressive Conditional Heteroskedasticity* (GARCH). Tujuan dari penelitian ini adalah untuk memodelkan harga emas antam produksi PT. Aneka Tambang menggunakan metode ARFIMA-GARCH serta membandingkan metode estimasi parameter pembeda d terbaik dari model tersebut. Hasil penelitian ini menunjukan model terbaik dilihat dari nilai AIC untuk dgph = 0, 105 adalah ARFIMA(1,d,1)-GARCH(1,1) dan model terbaik untuk dR/S = 0, 288 adalah ARFIMA(1,d,1)-GARCH(1,1). Tingkat akurasi peramalan didasarkan pada nilai MAPE. Nilai error validasi model ARFIMA-GARCH dengan dgph = 0, 105 adalah MAPE=3,474%, sedangkan model ARFIMA-GARCH dengan dR/S = 0, 288 adalah MAPE=3,444%.
Kata kunci: ARFIMA-GARCH; long memory; estimasi parameter pembeda d; peramalan; emas

ABSTRACT

Gold is becoming one of the financial assets for countries and a component of global monetary reserves for trade and protection when facing a sudden financial crisis. Some economic data often experience long memory dependencies. One model that can overcome this problem is the Autotegressive Fractionally Integrated Moving Average (ARFIMA) model. There are several methods used to determine the estimation of the difference parameter d, namely the Geweke Porter Hudak method and the Rescaled Range Statistics (R/S) method. Some types of time series data sometimes experience volatility clustering (residuals are not constant). The method that can be used to overcome this problem is the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) method. The purpose of this study is to model the price of antam gold produced by PT Aneka Tambang using the ARFIMA-GARCH method and compare the best parameter estimation method of the model. The results of this study show that the best model seen from the AIC value for dgph = 0, 105 is ARFIMA(1,d,1)-GARCH(1,1) and the best model for dR/S = 0, 288 is ARFIMA(1,d,1)-GARCH(1,1). The level of forecasting accuracy is based on the MAPE value. The validation error value of the ARFIMA-GARCH model with dgph = 0, 105 is MAPE = 3.474%, while the ARFIMA-GARCH model with dR/S = 0, 288 is MAPE = 3.444%.
Keywords: *ARFIMA-GARCH; long memory; estimation of discriminating parameter d; forecasting; gold*.

PENDAHULUAN

Emas antam salah satu aset yang banyak diminati untuk investasi karena banyak gunakan sebagai komponen cadangan global untuk perdagangan dan Ketika menghadapi krisis keuangan global suatu negara [1]. Meramalkan harga emas menjadi solusi penting bagi para investor, perusahaan pertambangan karena dapat digunakan untuk memeriksa fluktuasi yang nantinya dapat digunakan untuk membuat keputusan di masa mendatang [2].Emas Antam adalah emas berupa batangan yang diproduksi oleh perusahaan milik negara yaitu PT. Aneka Tambang yang bergerak di bidang pertambangan yang salah satu hasil produksinya adalah logam mulia emas. Berdasarkan buku laporan tahunan PT. Aneka Tambang, pada tahun 2022 PT. Aneka Tambang mencatatkan capaian penjualan emas tertinggi sepanjang sejarah perusahaan yakni, 34,97 ton atau tumbuh sekitar 19% *year on year* dibandingkan dengan penjualan emas pada tahun 2021 sebesar 29,38 ton. Hal ini menunjukan bahwa terjadi peningkatan minat masyarakat dalam melakukan investasi emas Antam.

Long memory adalah proses stasioner dimana ada ketergantungan statistik jarak jauh antara nilai saat ini dan nilai seri diwaktu yang berbeda. Hal ini dapat dilihat melalui nilai autokorelasinya yang turun lambat dalam waktu yang lama sehingga membentuk pola data yang hiperbolik [3]. Metode Autoregressive Fractionally Integrated Moving Average (ARFIMA) merupakan salah satu metode peramalan time series yang efektif dalam mengatasi masalah ketergantungan jangka panjang (*long memory*) pada data [4]. Penentuan estimasi parameter pembeda pada model ARFIMA banyak menggunakan metode *Geweke Porter Hudak* (GPH) dan metode *Rescaled Range Statistics* (R/S). Kelebihan dari metode ini adalah sifatnya yang fleksibel dalam menentukan parameter pembeda d meskipun nilai parameter p dan q belum diketahui [5].

Salah satu model yang dapat dikombinasikan dengan metode ARFIMA untuk menangani masalah heteroskedastisitas adalah metode *Generalized Autoregressive Conditional Heteroskedasticity* (GARCH) [6]. Bollerslev pertama kali menggunakan metode ini untuk Melakukan pemodelan inflasi uang di Amerika Serikat [7]

Terdapat beberapa penelitian yang serupa menggunakan metode ARFIMA-GARCH antara lain, penelitian yang dilakukan oleh Aliyu, dkk membandingkan model ARFIMA-GARCH dan ARFIMA-FIGARCH untuk memodelkan volatilitas nilai tukar naira per dollar [8]. Gajda, dkk, memodelkan penggunaan air menggunakan model ARFIMA-GARCH [6]. Dari penelitian-penelitian di atas didapatkan bahwa model ARFIMA-GARCH merupakan model terbaik untuk memodelkan data long memory karena menghasilkan nilai RMSE mendekati 0. Hanifa, dkk, mengaplikasikan metode ARFIMA-EGARCH untuk meramalkan harga beras, yang menghasilkan nilai MAPE sebesar 3,37% [9].

Berdasarkan informasi yang telah disampaikan di atas, peneliti tertarik untuk melakukan penelitian menggunakan metode ARFIMA-GARCH dengan menggunakan estimasi pembeda d metode *Geweke and Porter Hudak* (GPH) dan estimasi pembeda d statistik *Hurst* melalui metode *Rescaled Range Statistics* (R/S) untuk meramalkan harga emas antam produksi PT.Aneka Tambang.

METODE

Penelitian ini merupakan penelitian kuantitatif, data yang dipakai pada penelitian ini merupakan data sekunder berupa harga Emas Antam Produksi PT. Aneka Tambang dari 1 Januari

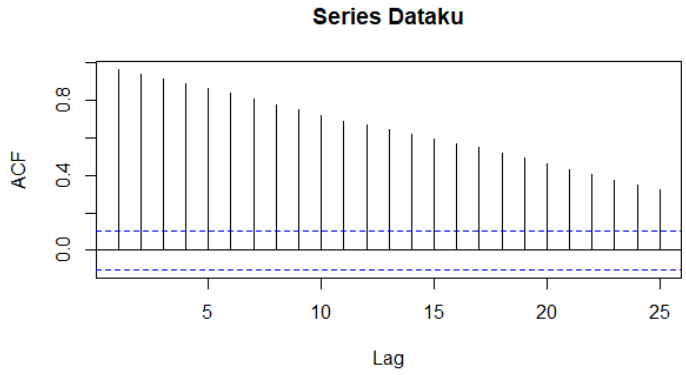
2022 hingga 31 Januari 2023. Data ini diperoleh dari website Katadata (katadata.co.id). Data dibagi menjadi dua bagian, yaitu data *in-sample* yang digunakan untuk membentuk model, dan data *out-sample* yang digunakan untuk menguji validitas hasil peramalan. Pembagian data dilakukan dengan rentang 1 Januari 2022 hingga 31 Desember 2022 sebagai data in sample, dan sisa data merupakan out sample. *Software* pengolahan data pada penelitian ini *menggunakan software RStudio*[10].
Langkah-langkah analisis data yang dilakukan dalam penelitian ini adalah sebagai berikut:

1. Mempersiapkan data penelitian;
2. Melakukan plot data runtun waktu;
3. Mengidentifikasi pola long memory melalui plot ACF atau menghitung nilai statistik *Hurst*;
4. Melakukan pengujian stasioner data menggunakan uji ADF untuk melihat kestasioneran data dalam *mean* dan uji *Box-cox* untuk melihat kestasioneran dalam varian;
5. Melakukan differencing ketika data tidak stasioner dalam mean dan transformasi data ketika tidak stasioner dalam varian;
6. Menentukan parameter pembeda d menggunakan metode *Rescaled Range Statistics* dan metode *Geweke and Porter-Hudak*;
7. Melakukan fractional differencing menggunakan nilai estimasi parameter pembeda d;
8. Membuat plot ACF dan plot PACF untuk masing-masing estimasi parameter pembeda d
9. Penentuan model ARFIMA (p, d, q) serta mengestimasi parameter ϕ dan θ pada setiap model;
10. Melakukan pemilihan model yang signifikan berdasarkan nilai AIC terkecil untuk masing-masing estimasi parameter pembeda d;
11. Melakukan uji kelayakan model menggunakan uji *L-Jung Box* untuk melihat apakah residual data bersifat white noise atau tidak;
12. Melakukan pengujian heteroskedastisitas dengan uji *Langrange Multiplier*;
13. Mengidentifikasi dan melakukan pemodelan metode GARCH;
14. Menggunakan nilai *Akaike's Information Criterion* (AIC) terkecil untuk memilih model terbaik;
15. Melakukan peramalan menggunakan model terbaik metode ARFIMA (p, d, q) dan GARCH (p, q) serta evaluasi akurasi model dengan melihat nilai MAPE.

HASIL DAN DISKUSI

1. Identifikasi Pola Long Memory

Identifikasi long memory di lakukan untuk melihat apakah data memiliki ketergantungan atau persistensi jangka panjang. Berikut ini plot ACF data in sample harga emas produksi PT. Aneka Tambang.

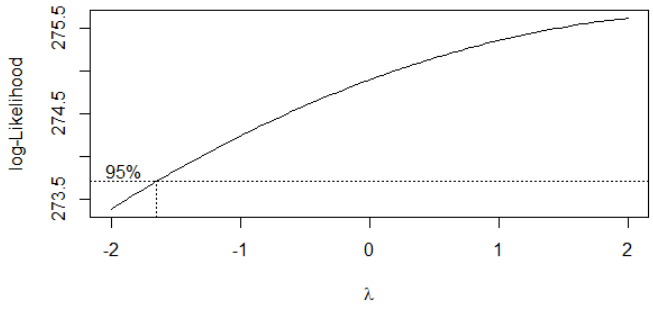


Gambar 1. Plot ACF data in sample harga emas Produksi PT.Aneka Tambang

Berdasarkan gambar 1, data in sample harga emas menunjukan autokorelasi yang *lag*-nya turun lambat menuju angka 0 sehingga membentuk pola hiperbolik, sehingga menujukan adanya ketergantungan jangka panjang (long memory). Selanjutnya dilakukan perhitungan nilai hurst untuk melihat adanya indikasi long memory. Hasil perhitungan nilai statistik hurst untuk data in sample menggunakan *software R Studio* diperoleh nilai Hurst (H) dari data in sample adalah sebesar 0,788 yang berarti berada di rentang 0,5<H<1, hal ini menunjukan bahwa data memiliki ketergantungan jangka panjang.

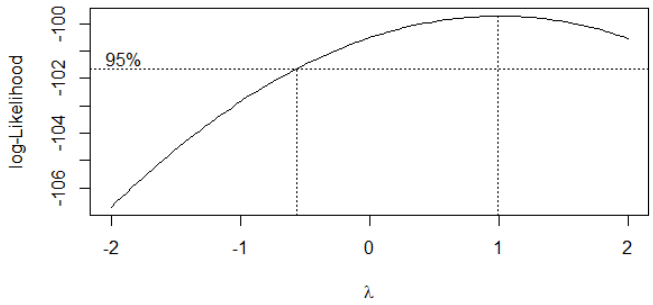
2. Uji Stasioneritas Data

Sebelum melakukan pemodelan, terlebih dahulu dilaukan uji stasioneritas data dalam varian. Berikut Plot *Box-Cox* data in sample uji stasioneritas dalam varian.



Gambar 2. Plot Box-Cox data in sample harga emas Produksi PT.Aneka Tambang

Berdasarkan gambar 2 , dapat dilihat bahwa data tidak stasioner karena nilai lamda (λ) ≠ 1 dan perlu dilakukan transformasi. Berikut plot *Box- Cox* hasil transformasi.



Gambar 3. *Plot Box-Cox* hasil transformasi data in sample

Berdasarkan gambar 3, setelah dilakukan transformasi diperoleh nilai lamda $(\lambda) = 1$, maka dapat disimpulkan data sudah stasioner dalam varian. Analisis selanjutnya adalah melakukan uji kestasioneran dalam mean menggunakan uji *Augmented Dickey-Fuller*. Hasil uji *Augmented Dickey-Fuller* didapatkan nilai p-value uji *Augmented Dickey-Fuller* sebesar 0,5682 lebih besar dari taraf signifikansi $\alpha = (0{,}05)$, sehingga data belum stasioner dalam mean dan harus dilakukan differencing. Setelah dilakukan differencing didapatkan nilai p-value sebesar 0,01 lebih kecil dari taraf signifikansi $\alpha = (0{,}05)$, sehingga dapat disimpulkan data telah stasioner dalam *mean*.

3. Estimasi Parameter Pembeda d

Nilai estimasi parameter pembeda d pada model ARFIMA dapat dihitung dengan menggunakan metode *Geweke and Porter Hudak* (GPH) dan metode *Rescaled Range Statistics* (R/S). Hasil perhitungan nilai estimasi parameter pembeda d data in sample harga emas Antam Produksi PT.Aneka Tambang menggunakan *software R Studio* dapat dilihat pada tabel 1 berikut.

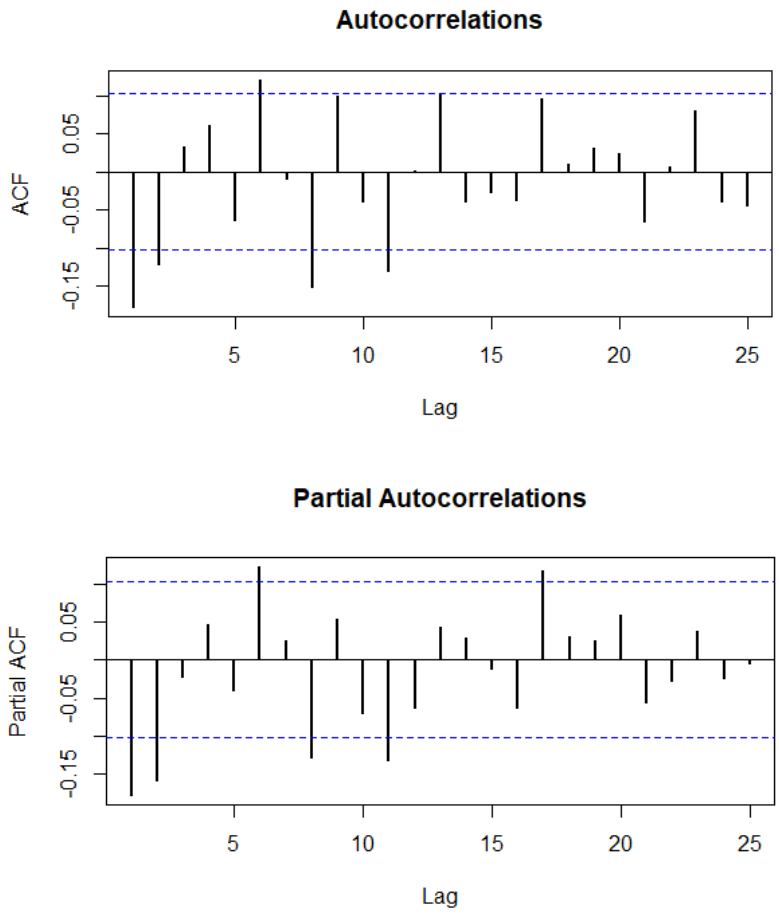
Tabel 1. Estimasi Parameter Pembeda *d*

	GPH	**R/S**
Nilai estimasi parameter pembeda d	0,105	0,288

Berdasarkan tabel 1, dapat dilihat bahwa nilai pembeda *d* berada pada interval 0<d<0,5, yang berarti bahwa data memiliki korelasi positif jangka panjang antar pengamatan yang terpisah jauh.

4. Identifikasi Model ARFIMA(p,d,q) Berdasarkan Plot ACF dan PACF

Identifikasi model ARFIMA(p,d,q) ditentukan berdasarkan plot ACF dan PACF data setelah dilakukan differencing menggunakan masing-masing nilai parameter pembeda d. Penentuan model ARFIMA(p,d,q) dilihat dari *lag* yang keluar garis pada plot ACF dan PACF. Plot ACF dan PACF setelah didifferencing dengan d_{gph} = 0,105 dapat dilihat pada gambar berikut.



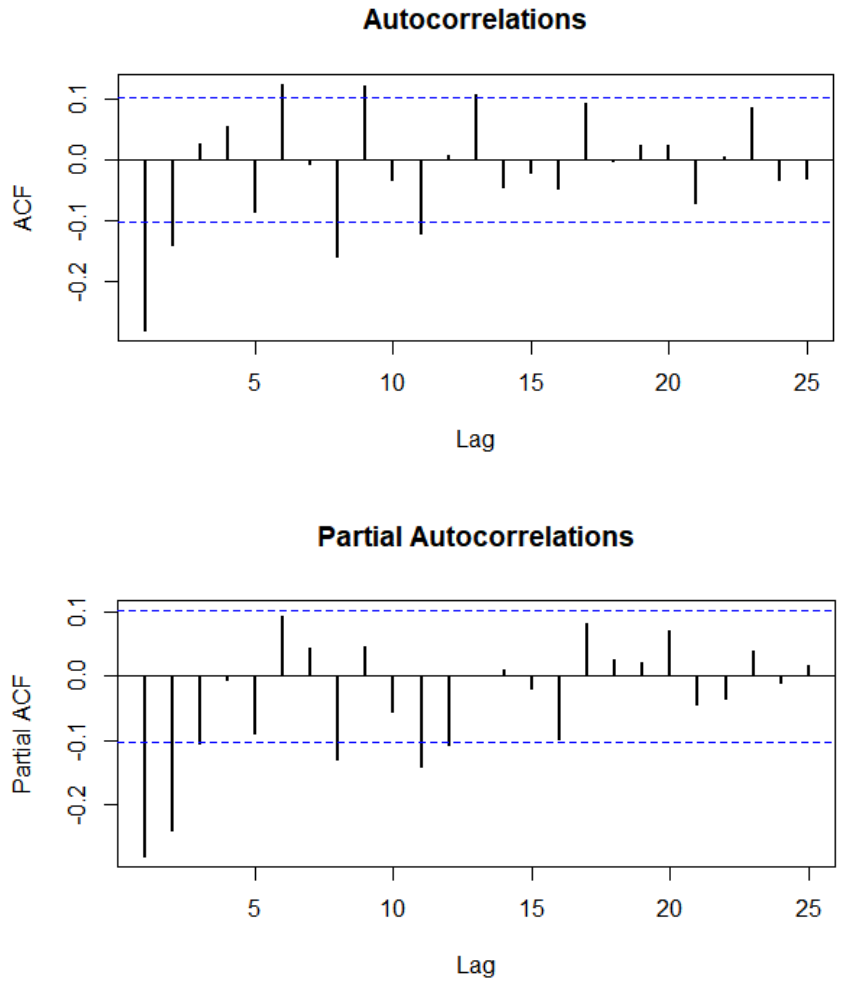
Gambar 4. Plot ACF dan PACF Differencing d_{gph} = 0,105

Berdasarkan Gambar 4, ada beberapa *lag* yang melewati batas sehingga dengan menggunakan prinsip *persimony* model yang mungkin terbentuk dapat dilihat pada tabel 2 berikut.

Tabel 2. Pembentukan Model ARFIMA(p,d,q) d_{gph} = 0,105

Model	MA(0)	MA(1)	MA(2)	MA(6)
AR(0)	-	ARFIMA(0,d,1)	ARFIMA(0,d,2)	ARFIMA(0,d,6)
AR(1)	ARFIMA(1,d,0)	ARFIMA(1,d,1)	ARFIMA(1,d,2)	ARFIMA(1,d,6)
AR(2)	ARFIMA(2,d,0)	ARFIMA(2,d,1)	ARFIMA(2,d,2)	ARFIMA(2,d,6)
AR(6)	ARFIMA(6,d,0)	ARFIMA(6,d,1)	ARFIMA(6,d,2)	ARFIMA(6,d,6)

Plot ACF dan PACF setelah dilakukan differencing dengan $d_{R/S}$ = 0,288 dapat dilihat pada gambar berikut.



Gambar 5. Plot ACF dan PACF Differencing $d_{R/S}$ = 0,288

Berdasarkan Gambar 5, ada beberapa lag yang melewati batas sehingga dengan menggunakan prinsip *parsimony* model yang mungkin terbentuk dapat dilihat pada tabel 3 berikut.

Tabel 3. Pembentukan Model ARFIMA(p,d,q) $d_{R/S}$ = 0,288

Model	MA(0)	MA(1)	MA(2)	MA(6)
AR(0)	-	ARFIMA(0,d,1)	ARFIMA(0,d,2)	ARFIMA(0,d,6)
AR(1)	ARFIMA(1,d,0)	ARFIMA(1,d,1)	ARFIMA(1,d,2)	ARFIMA(1,d,6)
AR(2)	ARFIMA(2,d,0)	ARFIMA(2,d,1)	ARFIMA(2,d,2)	ARFIMA(2,d,6)
AR(3)	ARFIMA(3,d,0)	ARFIMA(3,d,1)	ARFIMA(3,d,2)	ARFIMA(3,d,6)

5. Model Terbaik ARFIMA(p,d,q)

Hasil estimasi parameter dan nilai AIC model ARFIMA(p,d,q) d_{gph} = 0,105 dapat dilihat pada tabel 4 berikut.

Tabel 4. Pemilihan Model ARFIMA(p,d,q) d_{gph} = 0,105

No	Model	Parameter	Nilai AIC
1	ARFIMA(1,d,0)	Signifikan	25138,44
2	ARFIMA(2,d,0)	Signifikan	25315,45
3	ARFIMA(6,d,0)	Signifikan	25290,68
4	ARFIMA(1,d,1)	Signifikan	25127,82
5	ARFIMA(2,d,1)	Signifikan	25137,06
6	ARFIMA(6,d,1)	Signifikan	25387,36
7	ARFIMA(1,d,2)	Tidak Signifikan	15139,11
8	ARFIMA(2,d,2)	Signifikan	25305,79
9	ARFIMA(6,d,2)	Signifikan	25486,64
10	ARFIMA(1,d,6)	Signifikan	25143,03
11	ARFIMA(2,d,6)	Signifikan	25320,06
12	ARFIMA(6,d,6)	Tidak Signifikan	25602,24
13	ARFIMA(0,d,1)	Signifikan	25641,26
14	ARFIMA(0,d,2)	Signifikan	25692,09
15	ARFIMA(0,d,6)	Signifikan	25735,69

Berdasarkan Tabel 4, model ARFIMA(p,d,q) terbaik dengan menggunakan estimasi parameter pembeda d_{gph} = 0,105 terbaik adalah model ARFIMA(1,d,1) dengan nilai AIC 25127,82. Hasil estimasi parameter dan nilai AIC model ARFIMA(p,d,q) $d_{R/S}$ = 0,288 dapat dilihat pada tabel 5 berikut.

Tabel 5. Pemilihan Model ARFIMA(p,d,q) $d_{R/S}$ = 0,288

No	Model	Parameter	Nilai AIC
1	ARFIMA(1,d,0)	Signifikan	25159,31
2	ARFIMA(2,d,0)	Signifikan	25278,83
3	ARFIMA(3,d,0)	Signifikan	25310,98
4	ARFIMA(1,d,1)	Signifikan	25128,79
5	ARFIMA(2,d,1)	Signifikan	25161,34
6	ARFIMA(3,d,1)	Signifikan	25247,36
7	ARFIMA(1,d,2)	Tidak Signifikan	25161,32
8	ARFIMA(2,d,2)	Signifikan	25247,55
9	ARFIMA(3,d,2)	Signifikan	25303,77
10	ARFIMA(1,d,6)	Signifikan	25160,59
11	ARFIMA(2,d,6)	Signifikan	25278,77
12	ARFIMA(3,d,6)	Signifikan	25315,05
13	ARFIMA(0,d,1)	Signifikan	25398,36
14	ARFIMA(0,d,2)	Signifikan	25453,64
15	ARFIMA(0,d,3)	Signifikan	25485,02

Berdasarkan Tabel 5, model ARFIMA(p,d,q) terbaik dengan menggunakan estimasi parameter pembeda $d_{R/S}$ = 0,288 terbaik adalah model ARFIMA(1,d,1) dengan nilai AIC 25128,79.

Tabel 6. Hasil Uji Diagnostik Model ARFIMA Terbaik

Estimasi Parameter d	Model	Uji White Noise
d_{gph} = 0,105	ARFIMA(1,d,1)	0,7386
$d_{R/S}$ = 0,288	ARFIMA(1,d,1)	0,4806

Berdasarkan Tabel 6, hasil pengujian *white noise* menggunakan uji L-Jung Box diperoleh nilai p-value> $\alpha = (0{,}05)$, sehingga dapat disimpulkan bahwa residual model-model tersebut bersifat *white noise*, sehingga asumsi *white noise* terpenuhi dan model layak digunakan untuk peramalan.

Persamaan model ARFIMA(1,d,1) dengan d_{gph} = 0,105 adalah sebagai berikut:

$$\phi_p(B)(1-B)^d G_t = \theta_q(B)a_t \quad (1)$$

$$(1 - 0{,}975685B)(1-B)^{0.105} G_t = (1 - 0{,}221644B)a_t \quad (2)$$

Persamaan model ARFIMA(1,d,1) dengan $d_{R/S}$ = 0,288 adalah sebagai berikut:

$$\phi_p(B)(1-B)^d G_t = \theta_q(B)a_t \quad (3)$$

$$(1 - 0{,}953435B)(1-B)^{0.288} G_t = (1 - 0{,}404556B)a_t \quad (4)$$

Sebelum dilanjutkan dengan model GARCH, residual dari model-model tersebut akan dilakukan uji heteroskedastisitas menggunakan uji Langrange Multiplier. Hasil Uji *Langrange Multiplier* dapat dilihat pada tabel 7 berikut.

Tabel 7. Hasil Uji Heteroskedastisitas Model ARFIMA(p,d,q) Terbaik

Estimasi Parameter d	Model	Uji LM
d_{gph} = 0,105	ARFIMA(1,d,1)	0,001042858
$d_{R/S}$ = 0,288	ARFIMA(1,d,1)	0,003914108

Berdasarkan Tabel 7, hasil uji heteroskedastisitas dengan menggunakan uji *Langrange Multiplier* menunjukan bahwa semua model menghasilkan nilai p-value< $\alpha(0{,}05)$, yang berari semua model memiliki efek heteroskedastisitas pada residual, sehingga residual perlu dilanjutkan dengan model GARCH.

6. Identifikasi Model GARCH(p,q)

Hasil estimasi parameter dan nilai AIC model GARCH(p,q) d_{gph} = 0,105 dan $d_{R/S}$ = 0,288 dapat dilihat pada tabel 8 berikut.

Tabel 8. Pemilihan Model GARCH(p,q) d_{gph} = 0,105 dan $d_{R/S}$ = 0,288

No	Model	Konstanta	Parameter	Nilai AIC
1	GARCH(1,0)	$\omega = 70{,}32293$	Signifikan	7,2336
2	GARCH(1,1)	$\omega = 6{,}784841$	Signifikan	7,2257
3	GARCH(0,1)	$\omega = 0{,}20918$	Signifikan	7,2708

Berdasarkan Tabel 8, model GARCH(p, q) d_{gph} = 0,105 dan $d_{R/S}$ = 0,288 terbaik adalah model GARCH(1, 1) dengan nilai yaitu 7,2257. Persamaan model GARCH(1,1) adalah sebagai berikut:

$$\sigma_t^2 = 6{,}784841 + 0{,}066078\varepsilon_{t-1}^2 + 0{,}851871\sigma_{t-1}^2 \quad (5)$$

Pengujian efek heteroskedastisitas perlu dilakukan kembali untuk memastikan bahwa varian residual model telah kostan. Hasil uji heteroskedastisitas menggunakan uji Langrange Multiplier dapat dilihat pada tabel 9 berikut.

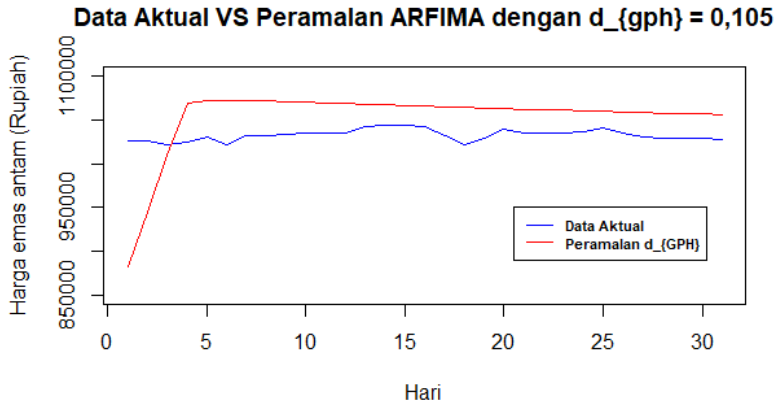
Tabel 9. Hasil Uji Heteroskedastisitas Model GARCH(1,1) Terbaik

Model	Uji LM
GARCH(1,1)	0,4581

Berdasarkan Tabel 9, hasil uji heteroskedastisitas dengan menggunakan uji Langrange Multiplier menunjukan bahwa residual model menghasilkan nilai p-value>$\alpha(0{,}05)$, yang berarti setelah dimodelkan dengan model GARCH efek heteroskedastisitas menjadi hilang.

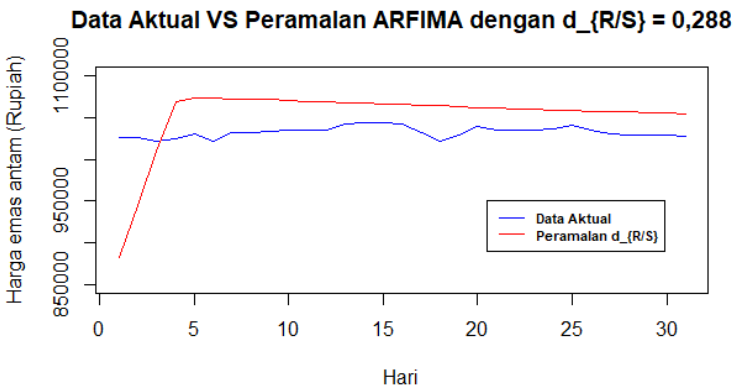
7. Peramalan

Peramalan dilakukan sebanyak 31 periode ke depan. Data yang dimodelkan sebelumya telah dilakukan transformasi dalam bentuk kuadrat, maka setelah diperoleh hasil peramalan hasil dikembalikan ke bentuk asli dari data dengan cara melakukan akar pada data hasil peramalan. Plot data aktual dan data hasil peramalan model ARFIMA(1d,1)-GARCH(1,1) d_{gph} = 0,105 dapat dilihat pada gambar berikut.



Gambar 6. Plot Data Aktual dan Data Peramalan ARFIMA-GARH d_{gph} = 0,105

Plot data aktual dan data hasil peramalan model ARFIMA(1d,1)-GARCH(1,1) $d_{R/S}$ = 0,288 dapat dilihat pada gambar berikut.



Gambar 7. Plot Data Aktual dan Data Peramalan ARFIMA-GARH $d_{R/S}$ = 0,288

Hasil akurasi peramalan harga emas Antam Produksi PT.Aneka Tambang dapat dilihat pada tabel 10 berikut.

Tabel 10. Hasil Akurasi Peramalan Model ARFIMA-GARCH

No	Parameter d	Model	Nilai MAPE
1	d_{gph} = 0,105	ARFIMA(1,d,1)-GARCH(1,1)	3,474223%
2	$d_{R/S}$ = 0,288	ARFIMA(1,d,1)-GARCH(1,1)	3,44475%

Berdasarkan Tabel 10, model terbaik untuk meramalkan harga emas Antam Produksi PT.Aneka Tambang adalah model ARFIMA(1,d,1)-GARCH(1,1) $d_{R/S}$ = 0,288, karena menghasilkan nilai MAPE sebesar 3,44475%.

KESIMPULAN

Berdasarkan hasil analisis dan pembahasan yang telah dilakukan dapat disipulkan bahwa model terbaik untuk meramalkan harga emas Antam Produksi PT.Aneka Tambang adalah model ARFIMA(1,d,1)-GARCH(1,1) $d_{R/S}$ = 0,288 karena menghasilkan nilai MAPE sebesar 3,44475%.

Pada penelitian selanjutnya dapat menggunakan estimasi parameter pembeda d lainnya seperti metode sperio serta apabila model GARCH memiliki efek asimetris disarankan menggunakan model TGACH dengan menghasilkan nilai peramalan yang lebih optimal.

UCAPAN TERIMA KASIH

Penelitian ini dibiayai oleh LPPM-UNG via PNBP-Universitas Negeri Gorontalo berdasarkan SK No. 495/UN47/HK.02/2023, dengan nomor kontrak B/673/UN47.DI/PT.01.03/2023.

DAFTAR PUSTAKA

[1] F. Capie, T. C. Mills, and G. Wood, “Gold As a Hedge Against the Dollar,” *Journal of International Financial Markets, Institutions and Money*, vol. 15, no. 4, pp. 343–352, Oct. 2005, doi: 10.1016/j.intfin.2004.07.002.

[2] S. Ben Jabeur, S. Mefteh-Wali, and J. L. Viviani, "Forecasting Gold Price with the XGBoost Algorithm and SHAP Interaction Values," *Ann Oper Res*, 2021, doi: 10.1007/s10479-021-04187-w.

[3] H. J. Sadaei, R. Enayatifar, F. G. Guimarães, M. Mahmud, and Z. A. Alzamil, "Combining ARFIMA Models and Fuzzy Time Series for the Forecast of Long Memory Time Series," *Neurocomputing*, vol. 175, pp. 782–296, 2016, doi: 10.1016/j.neucom.2015.10.079.

[4] P. Kartikasari, H. Yasin, and D. A. I. Maruddani, "Autoregressive Fractionally Integrated Moving Average (ARFIMA) Model to Predict Covid-19 Pandemic Cases in Indonesia," *MEDIA STATISTIKA*, vol. 14, no. 1, pp. 44–55, 2021, doi: 10.14710/medstat.14.1.44-55.

[5] M. J. I. Akbar and I. Kharisudin, "Model ARFIMA untuk Analisis Data Kecepatan Angin di Bandara Internasional Ahmad Yani," *Unnes Journal of Mathematics*, vol. 8, no. 2, pp. 89–101, 2019.

[6] J. Gajda, G. Bartnicki, and K. Burnecki, "Modeling of Water Usage by Means of ARFIMA–GARCH Processes," *Physica A: Statistical Mechanics and its Applications*, vol. 512, pp. 1–29, 2018, doi: 10.1016/j.physa.2018.08.134.

[7] T. Bollerslev, "Generalized Autoregressive Conditional Heteroskedasticity," *J Econom*, vol. 31, no. 3, pp. 307–327, 1986.

[8] M. A. Aliyu, H. G. Dikko, and U. A. Danbaba, "Statistical Modeling for Forecasting Volatility in Naira per Dollar Exchange Rate Using ARFIMA-GARCH and ARFIMA-FIGARCH Models," *World Sci News*, vol. 176, pp. 27–42, 2023, [Online]. Available: www.worldscientificnews.com

[9] R. D. Hanifa, M. Mustafid, and A. R. Hakim, "Pemodelan Autoregressive Integrated Moving Average dengan Efek Exponential GARCH (ARFIMA-EGARCH) untuk Prediksi Harga Beras di Kota Semarang," *Jurnal Gaussian*, vol. 10, no. 2, pp. 279–292, 2021, doi: 10.14710/j.gauss.v10i2.29933.

[10] G. Anuraga, A. Indrasetianingsih, and M. Athoillah, "Pelatihan Pengujian Hipotesis Statistika Dasar dengan Software R," *BUDIMAS: Jurnal Pengabdian Masyarakat*, vol. 3, no. 2, 2021.

Pemodelan Regresi Data Panel Harga Beras di Wilayah Indonesia Bagian Barat

Yogi Adam Firdaus(1), Ngatini(2), Sekarsari Utami Wijaya(3)

[12]Informatika, [3]Teknik Logistik

Universitas Internasional Semen Indonesia

e-mail: yogi.firdaus17@student.uisi.ac.id(1), ngatini@uisi.ac.id(2), sekarsari.wijaya@uisi.ac.id(3)

ABSTRAK

Beras merupakan kebutuhan pokok atau utama bagi masyarakat di Indonesia. Kenaikan harga beras berpengaruh sangat signifikan dalam berbagai aspek yang dapat mempengaruhi kebijakan ekonomi pemerintah. Sentra beras nasional didominasi oleh wilayah Indonesia bagian barat. Pemenuhan jumlah beras di setiap wilayah dilakukan oleh sentra beras melalui pendistribusian ke wilayah-wilayah lain. Harga pada wilayah yang merupakan sentra beras mempengaruhi harga beras di setiap wilayah-wilayah sekitarnya. Oleh karena itu, peramalan harga beras dibutuhkan. Penelitian ini bertujuan untuk melakukan pemodelan harga beras dengan metode Regresi Data Panel di Wilayah Indonesia Bagian Barat. Model Regresi Data Panel adalah hasil dari penggabungan data cross section dan time series. Dalam penelitian ini, pemodelan dibangun dengan menggunakan data dari semua provinsi di Indonesia bagian barat (cross sectional) pada beberapa tahun sebelumnya dengan tingkat bulanan (time series), sehingga pemilihan metode yang sesuai adalah menggunakan regresi data panel. Model Regresi Data Panel yang terpilih adalah REM (Random Effect Model) dengan rata-rata MAPE sebesar 3.28%. Pemodelan harga beras yang terbentuk dapat digunakan sebagai acuan dalam peramalan harga beras kedepannya, sehingga penentuan kebijakan ekonomi dapat dilakukan secara tepat.

Kata kunci: Harga Beras; Regresi Data Panel; MAPE.

ABSTRACT

Rice is a basic or primary need for people in Indonesia. Increasing rice prices significantly affects various aspects that can affect the government's economic policy. The western part of Indonesia dominates the national rice center. Rice centers fulfill the amount of rice in each region through distribution to other areas. Prices in areas that are rice centers affect rice prices in each of the surrounding areas. Therefore, rice price forecasting is needed. This study aims to model the price of rice using the Panel Data Regression method in the western part of Indonesia. The Panel Data Regression Model is the result of combining cross-section and time series data. In this research, the modeling was built using data from all provinces in western Indonesia (cross-sectional) in previous years at a monthly level (time series), so the appropriate method used panel data regression. The selected Panel Data Regression Model is REM (Random Effect Model) with an average MAPE of 3.28%. The formed rice price modeling can serve as a reference in future rice price forecasting, enabling us to determine economic policy appropriately.

Keywords: *Rice Price; Panel Data Regression; MAPE.*

PENDAHULUAN

Indonesia merupakan negara agraris dengan lebih dari 20% penduduk Indonesia bertani padi [1]. Hasil pertanian tersebut menyumbang pemenuhan kebutuhan pokok beras nasional. Pentingnya kebutuhan tersebut menjadi perhatian utama [2], khususnya dalam pemenuhan kebutuhan bahan pokok nasional. Beras merupakan komoditi bahan pokok nasional yang sangat penting karena lebih dari 97% masyarakat Indonesia mengonsumsi beras [3]. Rata-rata konsumsi beras masyarakat Indonesia hingga tahun 2019 yaitu 78.7 kg/kapita/tahun [4]. Tingginya konsumsi beras yang tidak disertai dengan peningkatan luas lahan pertanian berpeluang mengakibatkan semakin menurunnya ketersediaan beras di Indonesia. Permasalahan ketersediaan beras akan berdampak terhadap kestabilan harga beras di Indonesia. Selain jumlah pemenuhan yang harus diperhatikan, pengendalian atau controlling harga bahan pokok tersebut harus selalu dilakukan untuk mengatur kestabilan harga pasar. Rata-rata harga beras nasional hingga Agustus 2022 saat ini adalah Rp 9,069 (untuk kualitas rendah) dan Rp 9,901 untuk kualitas premium [5]. Peningkatan harga beras sangat berpengaruh signifikan terhadap tuntutan kenaikan gaji atau upah pekerja karena adanya peningkatan harga bahan pokok [6]. Hal tersebut menunjukkan bahwa beras memiliki pengaruh kuat bagi roda perekonomian di Indonesia [7]. Kebutuhan pokok akan konsumsi beras bagi masyarakat sangat penting diperhatikan mulai dari proses penanaman hingga tahap pemanenan. Di tahap pemanenan, sangat penting dilakukan adanya upaya prediksi jumlah produksi beras [8] yang sangat berguna bagi petani, pemerintah, ilmuwan maupun lembaga pertanian dalam pengambilan kebijakan [9]. Salah satu upaya prediksi dapat menggunakan Machine Learning[10] maupun dengan pembentukan model prediksi [11]. Sentra beras nasional didominasi oleh wilayah Indonesia bagian barat [12]. Pemenuhan jumlah beras di setiap wilayah dilakukan oleh sentra beras melalui pendistribusian ke wilayah-wilayah lain. Harga pada wilayah yang merupakan sentra beras mempengaruhi harga beras di setiap wilayah-wilayah sekitarnya. Gambaran atau prediksi permintaan, cadangan (pasokan) serta harga beras sangat penting untuk dilakukan dalam rangka penentuan kebijakan ekonomi yang sangat dipengaruhi oleh bahan pokok tersebut. Berdasarkan hal tersebut, diperlukan peramalan harga beras khususnya di sentra beras nasional wilayah Indonesia bagian barat. Beberapa algoritma yang dapat diimplementasikan untuk peramalan tersebut antara lain yaitu metode regresi data panel, metode simple moving average dan metode ARIMA.

Penelitian terdahulu yang digunakan dalam acuan penelitian ini, diantaranya adalah penelitian pemodelan harga beras di daerah Sumatera yang dilakukan oleh Dwi Yulianti dkk dengan Generalized Space Time ARIMA [13]. Penelitian serupa juga telah dilakukan oleh Sekarsari Utami Wijaya dan Ngatini (2020) untuk pemodelan harga beras di wilayah Indonesia bagian barat dengan Clustering Time Series [14]. Penelitian yang mengimplementasikan regresi data panel dilakukan oleh Yusuf Prawira Putra (2010) pada analisis pengaruh harga beras dan PDRB terhadap inflasi di Indonesia tahun 2015 [15]. Penelitian lain mengenai regresi data panel antara lain yaitu regresi data panel dalam memodelkan impor beras [16], penelitian tentang analisis jumlah penduduk[17] maupun penentuan penerima bantuan[18]. Berdasarkan penelitian-penelitian yang telah dilakukan sebelumnya pemodelan harga beras dengan menggunakan pendekatan regresi data panel dapat dilakukan. Data panel adalah gabungan data deret waktu (time series) dan data silang (cross section). Data cross section dideskripsikan sebagai suatu observasi yang terdiri dari banyak objek dan variabel yang berkorespondesi dengan objek tersebut terjadi pada titik waktu tertentu. Data time series mengobservasi satu objek dari waktu ke waktu. Data panel terdiri dari

dua data tersebut yang digabungkan menjadi sebuah model dengan mengumpulkan data dari banyak objek dari waktu ke waktu. Pada penelitian ini pemodelan dibentuk dengan menggunakan data seluruh provinsi yang ada di wilayah Indonesia bagian barat (cross sectional) pada beberapa tahun sebelumnya dalam satuan bulanan (time series), sehingga metode yang tepat adalah pemodelan dengan menggunakan regresi data panel. Pemodelan tersebut disimulasikan dengan menggunakan program R dan perhitungan nilai akurasi dengan menggunakan nilai MAPE (Mean Absolute Persentage Error).

METODE

Data panel adalah data yang menggabungkan informasi dari data cross section dan data time series. Dengan kata lain, data panel mencakup pengamatan berulang kali terhadap individu yang sama dalam konteks data cross section. Sebagai tambahan, model regresi data panel umumnya dijelaskan seperti berikut [19].

$$y_{it} = \alpha + X_{it}\beta + \varepsilon_{it} \quad (1)$$

dengan $i = 1,2,\dots,n$ adalah individu ke-i (data cross section), $t = 1,\dots,T$ adalah periode waktu ke-t, y_{it} adalah individu ke-i untuk periode ke-t pada variabel dependen, α adalah intercept (konstanta) yang didapat dari analisis regresi data panel, X_{it} merupakan vektor variabel independen berukuran $(1xk)$, variabel-variabel independen dari individu ke-i pada periode waktu ke-t (yakni terdapat k variabel independen, dimana setiap variabel merupakan data panel), β parameter regresi (slope koefisien) berupa vektor berukuran $(kx1)$, dan ε_{it} adalah error pada individu ke-i untuk periode waktu ke-t. Dalam analisis regresi data panel, terdapat tiga jenis model estimasi yang umum digunakan: Model CEM (Common Effect Model), Model FEM (Fixed Effect Model), dan Model REM (Random Effect Model). Setelah mengestimasi ketiga model ini, langkah selanjutnya adalah membandingkannya untuk menentukan model yang paling cocok. Untuk melakukan pemilihan model regresi data panel yang sesuai, beberapa uji statistik digunakan yaitu sebagai berikut.

a. Uji Chow

Uji Chow digunakan untuk menentukan pilihan antara Model Common Effect atau Model Fixed Effect yang lebih cocok digunakan dalam analisis. Hipotesis yang diajukan dalam uji Chow dapat dirumuskan yaitu [20]:

H_0: Common Effect Model; H_1: Fixed Effect Model

b. Uji Hausman

Uji ini bertujuan untuk mengetahui model yang sebaiknya dipakai yaitu Fixed Effect Model atau Random Effect Model. Dalam FEM setiap objek memiliki intersep yang berbeda-beda, akan tetapi intersep masing-masing objek tidak berubah seiring waktu. Hal ini disebut dengan time-invariant. Hipotesis yang digunakan dalam uji Hausman adalah sebagai berikut [21]:

H_0: Random Effect Model; H_1: Fixed Effect Model

c. Uji Lagrange Multiplier (LM)

Uji LM digunakan untuk mengetahui apakah model random effect lebih baik dari pada model common effect. Hipotesis yang digunakan dalam uji Hausman adalah sebagai berikut [22]:

H_0: Common Effect Model; H_1: Random Effect Model

Tahapan Implementasi Regresi Data Panel

Pemodelan harga beras dibentuk dengan metode regresi data panel. Tahapan dalam implementasi metode pada penelitian tersebut adalah sebagai berikut:

1. Penentuan variabel
   Variabel dependen (Y) : Harga beras di Indonesia wilayah bagian barat. Variabel independen (X): Produksi padi (X1), tingkat inflasi (X2) dan jumlah pengeluaran konsumsi beras (X3) di Indonesia wilayah bagian barat.
2. Eksplorasi data
   Objek penelitian ini adalah delapan belas provinsi yang ada di Indonesia wilayah bagian bara yang terdiri dari Provinsi Aceh, Sumatera Barat, Sumatera Utara, Sumatera Selatan, Riau, Kepulauan Riau, Jambi, Lampung, Bengkulu, Banten Jawa Barat, Jawa Timur, Jawa Tengah, DKI Jakarta, DI Yogyakarta.
3. Pembentukan model
   Pada tahap ini dilakukan pembentukan model dengan estimasi model Common Effect, model Fixed Effect dan model Random Effect.
4. Pemilihan model terbaik dengan menggunakan uji Uji Chow, uji Langrange Multiplier (LM) dan uji Hausman.
5. Pengujian
   Pada tahap ini dilakukan uji multikolineritas dengan nilai Variance Inflation Factors (VIFs) dan uji residual independen.
6. Peramalan variabel independen
   Dari model yang terbentuk, dilakukan peramalan variabel bebas dengan menggunakan ARIMA. Hasil peramalan tersebut kemudian disubtitusikan pada model tersebut.
7. Perhitungan tingkat akurasi model dengan MAPE
   Pada tahap ini dilakukan perhitungan nilai MAPE (Mean Percentage Absolute Error) antara harga aktual dengan harga dari model yang telah terbentuk untuk mengetahui tingkat akurasi dari model tersebut.
8. Analisis hasil dan pengambilan kesimpulan

Jenis dan Sumber Data

Data yang digunakan dalam penelitian ini di antaranya harga beras, produksi padi, tingkat inflasi dan jumlah pengeluaran konsumsi beras di wilayah Indonesia bagian barat pada Januari 2010 hingga Desember 2020. Adapun provinsi yang termasuk dalam wilayah Indonesia bagian barat di antaranya Aceh, Sumatera Utara, Sumatera Barat, Riau, Jambi, Sumatera Selatan, Bengkulu, Lampung, Bangka Belitung, Kepulauan Riau, DKI Jakarta, Jawa Barat, Jawa Tengah, DI Yogyakarta, Jawa Timur, Banten, Kalimantan Barat dan Kalimantan Tengah. Data tersebut diperoleh dari Badan Pusat Statistika (BPS).

Variabel Penelitian

Variabel yang digunakan dalam penelitian ini terdiri dari:

1. Variabel terikat (*dependent variable*)
   Variabel yang dipengaruhi oleh variabel lain dikenal sebagai variabel dependen (*dependent variable*). Variabel dependen adalah variabel yang menerima pengaruh dari data karena adanya variabel independen. Dalam penelitian ini, variabel dependen yang digunakan adalah harga beras, yang diwakili oleh simbol "Y".
2. Variabel bebas (*independent variable*)
   Variabel yang tidak bergantung pada variabel lainnya disebut variabel independen (*independent variable*). Variabel independen adalah variabel yang menyebabkan perubahan pada variabel tergantung. Faktor-faktor penentu yang telah diidentifikasi sebagai pengaruh terhadap harga beras mencakup produksi padi, tingkat inflasi, konsumsi beras, impor beras tahun sebelumnya, dan nilai tukar riil. Variabel independen yang digunakan dalam penelitian ini mencakup produksi padi (X1), tingkat inflasi (X2), dan jumlah pengeluaran konsumsi beras (X3).

Variance Inflation Factors (VIF)

Variance Inflation Factors (VIFs) digunakan untuk mendeteksi kolinearitas antar prediktor [23]. Formula untuk perhitungan VIF menggunakan model sebagai berikut [24].

$$VIF_j = \frac{1}{1-\tilde{R}_j^2} \quad (2)$$

Dengan, $\tilde{R}_j^2 = 1 - \frac{x_j' M_j x_j}{x_j' C x_j}$, $M_j = \boldsymbol{I}_n - \boldsymbol{X}_j\left(\boldsymbol{X}_j'\boldsymbol{X}_j\right)^{-1}\boldsymbol{X}_j'$ yang merupakan koefisien determinasi terpusat dimana variabel x_j diregresi pada remaining variabel independen (termasuk intersep).

Residual Independen

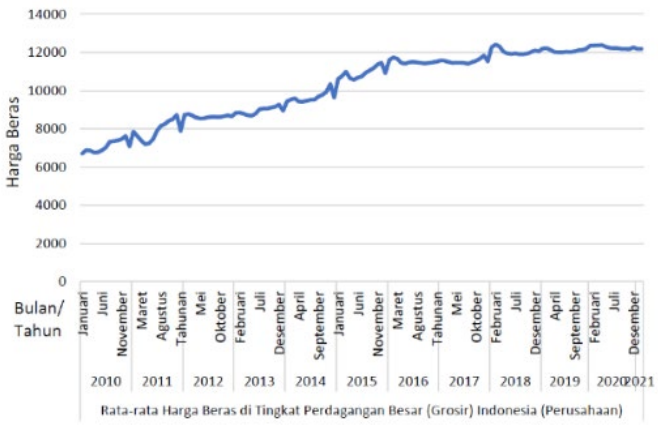
Data panel memiliki keunggulan utama yaitu bersifat robust dalam mengatasi beberapa jenis pelanggaran asumsi Gauss Markov seperti heterokedastisitas dan normalitas [25]. Berdasarkan hal tersebut, pemeriksaan asumsi residual dilakukan untuk residual independen yaitu tidak adanya korelasi antar variabel bebas. Uji residual independen dilakukan untuk melihat apakah residual memenuhi asumsi independent. Asumsi independen berarti tidak adanya autokorelasi pada residual atau residual bersifat saling independen yang ditunjukkan oleh nilai kovarian antara εi dan εj adalah sama dengan nol. Uji Durbin Watson dapat digunakan untuk mendeteksi adanya kasus autokorelasi [26].

HASIL DAN PEMBAHASAN

Eksplorasi Data

Berdasarkan Gambar 1 pola harga beras di Indonesia cenderung meningkat setiap tahunnya, hal ini secara tidak langsung dapat menyebabkan meningkatnya inflasi. Kenaikan harga beras yang merupakan kebutuhan pokok masyarakat mendorong pekerja untuk mendapatkan pemasukan

(income) yang lebih, salah satunya melalui tuntutan kenaikan upah/gaji. Hal tersebut yang menjadikan beras berpengaruh dalam segala aspek perekonomian di Indonesia.



Gambar 1. Pola Harga Beras Indonesia Januari 2010 – Desember 2021[27]

Pembentukan Model Prediksi Harga Beras

Pembentukan model dilakukan secara komputasi numerik dengan menggunakan R Studio dengan library *plm*. Hasil estimasi modelnya adalah sebagai berikut.

1. Model *Common Effect*

   Model yang terbentuk dengan *Common Effect Model* (CEM) adalah sebagai berikut dengan hasil estimasi koefisien dapat dilihat pada Tabel 1.

$$Y = 8515.4 - 0.0247X_1 + 147.85X_2 + 0.0442X_3 \quad (3)$$

Tabel 1. Hasil Estimasi Model CEM

No.	Coefficients	Estimate	*p-value*	Ket.
1.	Intercept	8515.4	$2.2\text{x}10^{-16}$	Signifikan
2.	Produksi	-0.0247	$2.2\text{x}10^{-16}$	Signifikan
3.	Inflasi	147.85	0,00565	Signifikan
4.	Konsumsi	0.0442	$2.2\text{x}10^{-16}$	Signifikan

2. Model *Fixed Effect*

   Model kedua dengan menggunakan *Fixed Effect Model* (FEM) adalah sebagai berikut.

$$Y = \alpha_i - 0.0278X_1 + 100.27X_2 + 0.0446X_3 \quad (4)$$

   Koefisien intersep (*intercept*) dari masing-masing provinsi dapat dilihat pada Tabel *2*.

3. Model *Random Effect*

   Model yang ketiga adalah Model Random Effect yang dapat dilihat pada persamaan (5) dengan nilai error untuk setiap provinsinya dapat dilihat dapat Tabel 3.

$$Y = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + \quad (5)$$

Tabel 2. Koefisien intersep FEM

Indeks(i)	Provinsi	α_i
1	Aceh	8553.8
2	Sumatera Utara	8571.2
3	Sumatera Barat	9662.4
4	Riau	9957.8
5	Jambi	8937.3
6	Sumatera Selatan	7430.2
7	Bengkulu	7967.8
8	Lampung	7866.3
9	Bangka Belitung	9301.6
10	Kepulauan Riau	10063.0
11	DKI Jakarta	9806.4
12	Jawa Barat	8230.0
13	Jawa Tengah	8224.0
14	DI Yogyakarta	8336.2
15	Jawa Timur	8053.8
16	Banten	8105.2
17	Kalimantan Barat	8784.7
18	Kalimantan Tengah	9231.4

Tabel 3. Estimasi nilai error model REM

Indeks(i)	Provinsi	ε
1	Aceh	164.99
2	Sumatera Utara	148.62
3	Sumatera Barat	895.12
4	Riau	1178.43
5	Jambi	201.29
6	Sumatera Selatan	1241.14
7	Bengkulu	726.58
8	Lampung	823.70
9	Bangka Belitung	549.31
10	Kepulauan Riau	1278.91
11	DKI Jakarta	1033.44
12	Jawa Barat	475.25
13	Jawa Tengah	481.22
14	DI Yogyakarta	1033.44
15	Jawa Timur	644.44
16	Banten	595.43
17	Kalimantan Barat	55.37
18	Kalimantan Tengah	483.72

Penentuan Model Terbaik

1. Uji Chow untuk pemilihan model CEM atau FEM

   Perhitungan uji Chow dilakukan dengan bantuan program R Studio menggunakan fungsi *pooltest*. *Pooltest* digunakan untuk membandingkan model *pooling* dan *within*. Hipotesis uji Chow sebagai berikut:

   H_0: *Common Effect Model;* H_1: *Fixed Effect Model*

```
> #uji Chow
> pooltest(cem,fem)

        F statistic

data:  Harga ~ Produksi + Inflasi + Konsumsi
F = 24.923, df1 = 17, df2 = 2355, p-value < 2.2e-16
alternative hypothesis: unstability
```

Gambar 2. Hasil Uji Chow

```
> #uji hausman
> phtest(rem,fem)

        Hausman Test

data:  Harga ~ Produksi + Inflasi + Konsumsi
chisq = 7.1869, df = 3, p-value = 0.06617
alternative hypothesis: one model is inconsistent
```

Gambar 3. Hasil Uji Hausman

Berdasarkan Gambar *2* diperoleh *nilai p-value* sebesar 2.2×10^{-16}. Nilai tersebut lebih kecil dari 0.05, sehingga H_0 ditolak. Dengan demikian, dapat disimpulkan bahwa terdapat efek individu pada model persamaan harga beras di Indonesia wilayah bagian barat, sehingga model yang sesuai adalah model *Fixed Effect* (FEM). Selanjutnya adalah melakukan pengujian untuk memilih model FEM atau model REM dengan uji Hausman.

2. Uji Hausman untuk pemilihan model FEM atau REM
   Perhitungan uji Hausman dilakukan dengan program R Studio dengan *phtest* untuk membandingkan model *within* dan *random*. Untuk menjalankan fungsi *phtest* dapat dilakukan dengan sintaks *phtest(rem,fem)*. Hipotesis uji Hausman sebagai berikut:
   H_0: *Random Effect Model;* H_1: *Fixed Effect Model*
   Berdasarkan Gambar 3 diperoleh nilai *p-value* sebesar 0.066. Nilai tersebut lebih besar dari 0.05, sehingga H_0 tidak ditolak. Dengan demikian dapat disimpulkan bahwa model REM yang dipilih. Selanjutnya pemilihan antara model REM atau model CEM dengan uji Lagrange Multiplier.
3. Uji Lagrange Multiplier untuk pemilihan Model REM dan CEM
   Sama halnya dengan uji model sebelumnya, uji Lagrange Multiplier juga dilakukan dengan program R Studio. Fungsi *plmtest* digunakan untuk membandingkan model random dan pooling. Untuk menjalankan fungsi *plmtest* dapat dilakukan dengan sintaks *plmtest(rem)*.
   H_0: *Common Effect Model*; H_1: *Random Effect Model*

```
> #uji lm
> plmtest(rem)

        Lagrange Multiplier Test - (Honda) for balanced panels

data:  Harga ~ Produksi + Inflasi + Konsumsi
normal = 57.238, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Gambar 4. Hasil Uji Lagrange Multiplier

```
Durbin-Watson statistic
(original):    NA , p-value: NA
(transformed): 2.93571 , p-value: 1e+00

 coefficients:
 (Intercept)     Produksi      Inflasi     Konsumsi
61207.130982    -0.004507    16.996132     0.003681
```

Gambar 5. Hasil Uji Durbin Watson

Berdasarkan Gambar 4 diperoleh nilai *p-value* sebesar $2.2\text{x}10^{-16}$. Nilai tersebut lebih kecil dari 0.05, sehingga H_0 ditolak. Dengan demikian dapat disimpulkan bahwa model REM yang dipilih.

$$Y = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + \varepsilon$$

Berdasarkan uji pemilihan model yang telah dilakukan, model regresi yang tepat untuk harga beras di Indonesia wilayah bagian barat adalah menggunakan *Random Effect Model*. Koefisien variabel X_1 (produksi padi) adalah -0.00276. Tanda negatif di depan koefisien menunjukkan bahwa produksi padi (X_1) memiliki hubungan negatif dengan harga beras (Y). Interpretasi dari koefisien tersebut adalah jika produksi padi meningkat sebesar 1 ton maka harga beras akan turun dengan penurunan sebesar Rp 0.00276 dan sebaliknya jika produksi padi turun sebesar 1 kg maka harga beras akan mengalami peningkatan sebesar Rp 0.00276. Koefisien variabel X_2 (tingkat inflasi) sebesar 102.37. Hal tersebut menunjukkan tingkat inflasi memiliki pola hubungan positif terhadap harga beras yang artinya jika tingkat inflasi meningkat sebesar 1% maka harga beras akan mengalami peningkatan sebesar Rp 102.37, dan sebaliknya jika tingkat inflasi turun maka harga beras akan mengalami penurunan sebesar Rp 102.37. Koefisien variabel X_3 (jumlah pengeluaran konsumsi beras) sebesar 0.0446. Pola hubungan antara konsumsi beras dan harga beras adalah positif yang artinya jika konsumsi beras meningkat sebesar 1 kg maka harga beras akan meningkat sebesar Rp 0.0446 dan sebaliknya jika konsumsi beras turun maka harga beras akan mengalami penurunan sebesar Rp 0.00446.

Pengujian Residual Independen

Uji autokorelasi seringkali digunakan untuk melihat ada atau tidaknya hubungan antara residual satu observasi dengan residual observasi lainnya. Dalam penelitian ini, untuk melakukan uji autokorelasi digunakan untuk melihat ada atau tidaknya hubungan antara residual satu observasi dengan residual observasi lainnya. Dalam penelitian ini untuk melakukan uji autokorelasi menggunakan uji durbin watson. Hipotesis uji Durbin Watson sebagai berikut.

H_0: Tidak ada korelasi antar residual; H_1: Ada korelasi residual

Kriteria pengujian Durbin Watson jika *p-value* kurang dari 0.05 maka H_0 ditolak yang artinya terjadi autokorelasi. Sebaliknya, jika *p-value* lebih dari 0.05 maka H_0 tidak ditolak artinya residual tidak terjadi autokorelasi. Berdasarkan pengujian yang telah dilakukan (Gambar 5) didapatkan nilai *p-value* sebesar 1.00. Nilai ini lebih besar dari 0.05 maka keputusan yang diperoleh adalah H_0 tidak ditolak yang artinya tidak terdapat autokorelasi pada penelitian ini, artinya model regresi data panel tidak memiliki korelasi antar residual.

Pengujian Bebas Multikolinieritas

Multikolinieritas terlihat ketika dua atau lebih variabel independen dalam regresi saling berkorelasi [28]. Asumsi ini dapat didiagnostik dengan menggunakan *Variance Inflation Factors* (VIF). Jika ada korelasi antar variabel independen, maka standar eror dari koefisien regresi akan meningkat sehingga varians dari koefisien regresi mengalami inflasi. Nilai VIF lebih dari 5 menunjukkan bahwa ada korelasi antar variabel independen atau ada multikolinieritas. VIF yang diperoleh dari hasil pengolahan data dapat dilihat pada Tabel *4*. Berdasarkan Tabel 4, variabel independen tidak saling berkorelasi atau bebas multikolinieritas karena nilai VIF yang diperoleh kurang dari 5.

Tabel 4. Nilai VIF

Variabel Independen	VIF
Produksi	1,02
Inflasi	1,00
Konsumsi	1,02

Peramalan Variabel Independen

Model prediksi harga beras ditunjukkan pada Persamaan 5 di atas. Sebelum dilakukan prediksi harga beras (*Y*), terlebih dahulu dilakukan peramalan terhadap variabel independen yaitu produksi padi (X1), tingkat inflasi (X2) dan konsumsi (X3). Prediksi variabel independen tersebut dilakukan dengan menggunakan metode ARIMA. Pemilihan model terbaik diambil berdasarkan nilai MAPE yang terkecil. Perhitungan model ARIMA akan dilakukan dengan bantuan program R Studio. Variabel produksi padi dan inflasi menggunakan metode ARIMA dengan model SARIMA (Seasonal Autoregressive Moving Average) karena plot data produksi padi dan inflasi menunjukan pola musiman. Periode musiman untuk produksi padi yaitu periode 4 bulan, sedangkan tingkat inflasi menunjukkan pola musiman dengan periode 12 bulan. Periode musiman tersebut akan tercantum di model SARIMA yang terbentuk. Variabel konsumsi menggunakan metode ARIMA karena plot data tidak menunjukan pola musiman. Berdasarkan pengolahan data, didapatkan model terbaik untuk peramalan produksi padi (X1), tingkat inflasi (X2) dan konsumsi (X3) di Indonesia wilayah bagian barat yang dapat dilihat pada Tabel 5. Model ARIMA(p,d,q) tersebut menjelaskan

tentang p: ordo AR, d:banyaknya differencing dan q:ordo MA, untuk data musiman ditambahkan dengan ARIMA(p,d,q)(P,D,Q)s yaitu P: ordo musiman AR, D: banyaknya musiman yang di differencing, dan Q: ordo musiman MA dengan s merupakan periode musiman.

Tabel 5. Ordo ARIMA dan SARIMA dari Variabel Independen

No	Provinsi	Produksi Padi (X1)	Tingkat Inflasi (X2)	Konsumsi (X3)
1	Aceh	(0,1,1)(1,1,2)4	(0,1,1)(0,1,1)12	(3,1,0)
2	Sumatera Utara	(2,1,2)(0,1,1)4	(0,1,1)(0,1,1)12	(1,0,1)
3	Sumatera Barat	(1,1,1)(0,1,1)4	(2,1,1)(0,1,1)12	(2,0,2)
4	Riau	(2,1,0)(2,1,0)4	(2,1,2)(0,1,1)12	(1,0,3)
5	Jambi	(2,1,3)(1,1,0)4	(0,1,1)(0,1,1)12	(2,1,3)
6	Sumatera Selatan	(1,1,1)(1,1,0)4	(1,1,1)(1,1,0)12	(3,1,4)
7	Bengkulu	(1,1,1)(3,1,0)4	(2,1,0)(2,1,0)12	(3,0,1)
8	Lampung	(2,1,3)(1,1,0)4	(1,1,1)(0,1,1)12	(2,1,1)
9	Bangka Belitung	(2,1,1)(0,1,1)4	(2,1,0)(0,1,1)12	(1,1,2)
10	Kepulauan Riau	(1,1,3)(1,1,0)4	(2,1,0)(3,1,0)12	(1,1,1)
11	DKI Jakarta	(0,1,1)(0,1,1)4	(3,1,0)(1,1,0)12	(2,1,2)
12	Jawa Barat	(2,1,3)(0,1,1)4	(0,1,1)(0,1,1)12	(2,1,2)
13	Jawa Tengah	(1,1,1)(0,1,1)4	(0,1,2)(0,1,1)12	(1,1,1)
14	DI Yogyakarta	(1,1,1)(0,1,1)4	(1,1,2)(1,1,0)12	(1,1,1)
15	Jawa Timur	(2,1,3)(1,1,0)4	(2,1,3)(0,1,1)12	(2,1,1)
16	Banten	(0,1,1)(0,1,1)4	(2,1,0)(1,1,0)12	(3,1,2)
17	Kalimantan Barat	(0,1,1)(1,1,1)4	(2,1,2)(1,1,0)12	(3,1,4)
18	Kalimantan Tengah	(0,1,1)(0,1,1)4	(2,1,1)(1,1,0)12	(1,1,1)

Model Peramalan Harga Beras dan Perhitungan tingkat akurasi model dengan MAPE

Model regresi data panel yang diperoleh adalah REM (Random Effcet Model) yang dapat digunakan untuk memrediksi harga beras di Indonesia Wilayah Bagian Barat yaitu $Y = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + \varepsilon$. Koefisien intercept: 8717.7, koefisien produksi: -0.00276, koefisien inflasi: 102.37, koefisien konsumsi: 0.0446 dan ε merupakan estimasi error model REM. Keseluruhan model tiap provinsi dapat dilihat pada Tabel 6. Peramalan harga beras dilakukan dengan mensubtitusikan nilai variabel independen X_1, X_2 dan X_3 di Tabel 5 pada model Tabel 6. Tingkat akurasi dari model dihitung dengan menggunakan MAPE (Mean Absolute Percentage Error). Berdasarkan Tabel 6, rata-rata nilai MAPE model regresi data panel sebesar 3.28%. Rata-rata MAPE yang diperoleh kurang dari 10%. Hal tersebut menunjukkan model harga beras dengan regresi data panel sangat baik dan layak digunakan sebagai peramalan.

Tabel 6. Model Regresi Data Panel dan MAPE Setiap Provinsi di Indonesia Wilayah Bagian Barat

No	Provinsi	Model Regresi Data Panel	MAPE
1.	Aceh	$Y_{Aceh} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 164.99$	1.96%
2.	Sumatera Utara	$Y_{Sumut} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 148.62$	7.45%

3.	Sumatera Barat	$Y_{Sumbar} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 895.12$	2.43%
4.	Riau	$Y_{Riau} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 1178.43$	3.85%
5.	Jambi	$Y_{Jambi} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 201.29$	3.59%
6.	Sumatera Selatan	$Y_{Sumsel} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 1241.14$	1.16%
7.	Bengkulu	$Y_{Bengkulu} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 726.58$	2.48%
8.	Lampung	$Y_{Lampung} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 823.70$	2.51%
9.	Bangka Belitung	$Y_{Bangka} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 549.31$	0.67%
10.	Kepulauan Riau	$Y_{Kepri} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 1278.91$	7.25%
11.	DKI Jakarta	$Y_{Jakarta} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 1033.44$	1.35%
12.	Jawa Barat	$Y_{Jabar} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 475.25$	3.14%
13.	Jawa Tengah	$Y_{Jateng} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 481.22$	6.58%
14.	DI Yogyakarta	$Y_{Jogja} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 1033.44$	5.92%
15.	Jawa Timur	$Y_{Jatim} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 644.44$	0.6%
16.	Banten	$Y_{Banten} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 595.43$	1.42%
17.	Kalimantan Barat	$Y_{Kalbar} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 55.36$	3.85%
18.	Kalimantan Tengah	$Y_{Kalteng} = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + 483.72$	2.84%
Rata-rata			3.28%

KESIMPULAN

Model Regresi Data Panel harga beras di wilayah Indonesia bagian barat telah terbentuk. Dari hasil yang telah dilakukan didapatkan kesimpulan sebagai berikut:

1. Model regresi data panel yang sesuai dalam peramalan harga beras di wilayah Indonesia Bagian Barat adalah REM (Random Effect Model)

$$Y = 8717.7 - 0.00276X_1 + 102.37X_2 + 0.0446X_3 + \varepsilon$$

   Dengan Y: harga beras (variabel dependen), serta variabel independen yaitu produksi padi (X_1), tingkat inflasi (X_2), dan jumlah pengeluaran konsumsi beras (X_3).
2. Prediksi variabel independen dilakukan dengan ARIMA untuk perhitungan rata-rata nilai MAPE (Mean Absolute Persentage Error) dari model prediksi setiap provinsi. Rata-rata MAPE nya adalah 3.28%. Error model di bawah 5% yang menunjukan bahwa tingkat akurasi sangat baik dan model matematika harga beras tersebut dapat digunakan untuk peramalan harga beras di Wilayah Indonesia Bagian Barat.

Saran untuk penelitian kedepannya adalah dapat menambah variabel independen. Selain itu juga dapat dilakukan pembentukan model matematika dengan memperhatikan hubungan antar provinsi. Serta pembuatan sistem informasi untuk peramalan harga beras di Indonesia.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada LPPM UISI yang telah memberikan hibah dana penelitian melalui skema hibah Riset Bersaing (HRB) Universitas Internasional Semen Indonesia.

DAFTAR PUSTAKA

[1] E. Frimawaty, A. Basukriadi, J. A. Syamsu, and T. E. B. Soesilo, "Sustainability of Rice Farming based on Eco-Farming to Face Food Security and Climate Change: Case Study in Jambi Province, Indonesia," Procedia Environ Sci, vol. 17, pp. 53–59, 2013, doi: 10.1016/j.proenv.2013.02.011.

[2] W. Anggraeni, F. Mahananto, A. Q. Sari, Z. Zaini, K. B. Andri, and Sumaryanto, "Forecasting the price of Indonesia's rice using hybrid artificial neural network and autoregressive integrated moving average (hybrid NNS-ARIMAX) with exogenous variables," in Procedia Computer Science, Elsevier B.V., 2019, pp. 677–686. doi: 10.1016/j.procs.2019.11.171.

[3] J. Louhenapessy, Sagu Harapan dan Tantangan. PT Bumi Aksara, Jakarta, 2010.

[4] Kementerian Pertanian RI, "Rencana Strategis Kementerian Pertanian Tahun 2020-2024," 2021.

[5] Badan Pusat Statistik, "Rata-Rata Harga Beras Bulanan di Tingkat Penggilingan Menurut Kualitas (Rupiah/Kg), 2022," bps.go.id.

[6] E. Marsudi, N. Fathia, and T. Makmur, "PENGARUH PENINGKATAN HARGA BERAS TERHADAP LAJU INFLASI DI INDONESIA (INCREASING PRICE OF RICE INFLUENCE TO INFLATION RATE IN INDONESIA)," 2018.

[7] H. Noviar, "Analisis Kecukupan dan Ketersediaan Beras di Indonesia (1975-2009)," Jurnal Penelitian Ilmu-Ilmu Sosial Universitas Malikussaleh, Lhokseumawe, NAD, vol. 1, no. 1, pp. 93–104, 2013.

[8] M. Noorunnahar, A. H. Chowdhury, and F. A. Mila, "A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh," PLoS One, vol. 18, no. 3, p. e0283452, Mar. 2023, doi: 10.1371/journal.pone.0283452.

[9] K. Choudhary, W. Shi, Y. Dong, and R. Paringer, "Random Forest for rice yield mapping and prediction using Sentinel-2 data with Google Earth Engine," Advances in Space Research, vol. 70, no. 8, pp. 2443–2457, 2022.

[10] M. Singh Boori, K. Choudhary, R. Paringer, and A. Kupriyanov, "Machine learning for yield prediction in Fergana valley, Central Asia," Journal of the Saudi Society of Agricultural Sciences, vol. 22, no. 2, pp. 107–120, Feb. 2023, doi: 10.1016/j.jssas.2022.07.006.

[11] Li Tian, Chun Wang, Hailiang Li, and Haitian Sun, "Yield prediction model of rice and wheat crops based on ecological distance algorithm," Environ Technol Innov, vol. 20, pp. 2352–1864, 2020.

[12] S. U. Wijaya and Ngatini, "Pengembangan Pemodelan Harga Beras di Wilayah Indonesia Bagian Barat dengan Pendekatan Clustering Time Series," Limits: Journal of Mathematics and Its Applications, vol. 17, no. 1, pp. 51–66, 2020.

[13] D. Yulianti, I. Made Sumertajaya, and I. D. Sulvianti, "Pemodelan Harga Beras di Pulau Sumatera dengan Menggunakan Model Generalized Space Time ARIMA," 2018.

[14] S. U. Wijaya and N. N. Ngatini, "Pengembangan Pemodelan Harga Beras di Wilayah Indonesia Bagian Barat dengan Pendekatan Clustering Time Series," Limits: Journal of Mathematics and Its Applications, vol. 17, no. 1, p. 51, Jul. 2020, doi: 10.12962/limits.v17i1.5994.

[15] Y. P. Putra, "ANALISIS PENGARUH HARGA BERAS, PRODUKSI BERAS DAN PDRB TERHADAP INFLASI DI INDONESIA TAHUN 2010– 2015," Malang, 2010.

[16] E. N. Kencana, D. Arnawa, and K. Jayanegara, "Memodelkan Impor Beras Menggunakan Regresi Data Panel," Jurnal Matematika, vol. 10, no. 2, p. 135, Jan. 2021, doi: 10.24843/jmat.2020.v10.i02.p130.

[17] P. R. Arum and M. Al Haris, "Analisis Faktor-Faktor yang Mempengaruhi Jumlah Penduduk di Kota Semarang Menggunakan Metode Regresi Data Panel," J Statistika, vol. 12, no. 2, pp. 36–41, 2019, [Online]. Available: www.unipasby.ac.id
[18] D. Venosia, Suliyanto, Sediono, and N. Chamidah, "Pemodelan Persentase Kepesertaan Bpjs Non Penerima Bantuan Iuran Dengan Pendekatan Regresi Data Panel," J Statistika, vol. 15, no. 1, pp. 116–126, 2022, [Online]. Available: www.unipasby.ac.id
[19] B. H. Baltagi, Econometric Analysis of Panel Data. John Wiley & Sons, 2005.
[20] W. Agus, Ekonometrika Pengantar dan Aplikasinya. Yogyakarta: Ekonesia, 2009.
[21] D. Gujarati, Dasar-dasar Ekonometrika (diterjemahkan oleh: Mangunsong, R.C.), Edisi Lima. Jakarta: Salemba Empat, 2012.
[22] A. T. Basuki and Imamudin. Yuliadi, "ELECTRONIC DATA PROCESSING (SPSS 15 dan EVIEWS 7)," Sleman, 2014.
[23] D. Liao and R. Valliant, "Variance inflation factors in the analysis of complex survey data," Dan Liao, RTI International, 2012. [Online]. Available: http://psidonline.isr.umich.
[24] J. Groß, "Variance Inflation Factors," R News, vol. 3, no. 1, pp. 13–14, 2003.
[25] J. M. Wooldridge, Introductory Econometrics: A Modern Approach. Boston: Addison Wesley, 2013.
[26] D. N. Gujarati and D. C. Porter, Basic Econometrics. New York: Mc Graw-Hill Education, 2004.
[27] BPS (Badan Pusat Statistik), "Rata-rata Harga Beras di Tingkat Perdagangan Besar/Grosir Indonesia (Rupiah/Kg), 2010-2020." Accessed: Sep. 07, 2023. [Online]. Available: https://www.bps.go.id/linkTableDinamis/view /id/963
[28] J. I. Daoud, "Multicollinearity and Regression Analysis," Journal of Physics: Conf. Series, pp. 1–6, 2017.

Hubungan Faktor Demografis dengan Kejadian Malaria di Kecamatan Wewewa Timur: Pendekatan Analisis Chi-Square

Junaldo Umbu Moto [(1)], Maria Agustina Kleden[(2)*], Robertus Dole Guntur [(3)]

[123]Department of Mathematics, Nusa Cendana University, Jalan Adisucipto, Penfui Kupang, Nusa Tenggara Timur Indonesia, Kupang, Indonesa

e-mail: maria_kleden@staf.undana.ac.id.

ABSTRAK

Pendekatan chi-square dalam analisis faktor risiko kesehatan telah menjadi topik yang menarik dan relevan dalam penelitian terkini. Malaria, sebagai penyakit infeksi menular yang ditularkan melalui gigitan nyamuk Anopheles betina, masih menjadi ancaman bagi masyarakat di beberapa daerah di Indonesia, termasuk Kecamatan Wewewa Timur. Meskipun kasus malaria di kecamatan ini mengalami penurunan dari tahun 2018 hingga 2021, namun pada tahun 2022, kasusnya kembali mengalami peningkatan. Dalam upaya untuk memahami lebih lanjut tentang faktor-faktor yang mempengaruhi kejadian malaria di Kecamatan Wewewa Timur, penelitian ini menggunakan pendekatan chi-square(χ^2) yang termasuk dalam uji non-parametrik digunakan untuk menentukan apakah terdapat perbedaan yang signifikan antara distribusi frekuensi yang diamati (observed frequencies) dengan distribusi frekuensi yang diharapkan (expected frequencies) dalam satu atau lebih kategori melibatkan 400 data primer yang diambil di wilayah tersebut. Hasil penelitian menunjukkan bahwa terdapat lima faktor yang secara signifikan berhubungan dengan kejadian malaria di Kecamatan Wewewa Timur, yaitu jenis kelamin, usia, pekerjaan, penggunaan kelambu, dan tingkat pendidikan. Temuan ini memberikan wawasan yang berharga dalam perumusan kebijakan dan strategi penanggulangan malaria yang lebih efektif dan tepat sasaran

Kata kunci: Malaria; Demografi; Chi-Square

ABSTRACT

Chi-square approach in health risk factor analysis has become an interesting and relevant topic in recent research. Malaria, as an infectious disease transmitted through the bite of female Anopheles mosquitoes, remains a threat to communities in several regions of Indonesia, including Wewewa Timur District. Despite a decline in malaria cases in the district from 2018 to 2021, there was an increase in cases in 2022. In an effort to further understand the factors influencing malaria incidence in Wewewa Timur District, this study utilized the chi-square approach and involved 400 primary data collected from the area. The research findings reveal five factors significantly associated with malaria incidence in Wewewa Timur District: gender, age, occupation, bed net usage, and education level. These findings provide valuable insights for formulating effective and targeted policies and strategies to combat malaria

Keywords: *Malaria; Demography; Chi-Square*

PENDAHULUAN

Malaria adalah penyakit infeksi menular yang menyebar melalui gigitan nyamuk Anopheles betina. Malaria umumnya ditandai dengan demam dan menggigil selama beberapa hari (Kementrian Kesehatan RI, 2022). Morbiditas malaria dapat diketahui dengan menilai indikator Annual Parasite Incidence (API) per 1000 penduduk berisiko. Indikator ini diperoleh dengan menghitung proporsi antara pasien positif malaria terhadap penduduk berisiko di wilayah tertentu. Menurut rilis Kementrian Kesehatan pada tahun 2022, Indonesia berhasil menekan API menjadi kurang dari 1,0 per 1000 penduduk sejak tahun 2015 sampai tahun 2020. Namun, demikian pada tahun 2021, API meningkat hingga 1,1 per 1000 penduduk. Kasus positif malaria menunjukkan konsentrasi kabupaten/kota endemis tinggi malaria di wilayah Indonesia Timur. Wilayah Indonesia Timur yang merupakan kabupaten/kota endemis tinggi malaria adalah Papua dengan tinggi API sebesar 80,05 per 1.000 penduduk, Papua Barat dengan tinggi API sebesar 7,56 per 1.000 penduduk, dan Nusa Tenggara Timur dengan tinggi API sebesar 1,69 per 1.000 penduduk Kementerian Kesehatan RI. 2022. Profil Kesehatan Indonesia Tahun 2021. Jakarta: Kementerian Kesehatan RI, [1]

Kecamatan Wewewa Timur adalah daerah dengan dataran rendah dan merupakan kawasan pertanian dan kehutanan. Hal ini berkontribusi pada timbulnya beragam masalah kesehatan masyarakat seperti penyakit malaria. Kasus malaria yang terjadi di Kecamatan Wewewa Timur menurun sejak tahun 2018–2021 yaitu dari 5,13% atau 314 kasus malaria pada tahun 2018 (Badan Pusat Statistik SBD, 2019) turun menjadi 1,36% atau 38 kasus malaria pada tahun 2021. Namun, pada tahun 2022, kasus malaria mulai meningkat menjadi 2,14% atau 123 kasus Badan Pusat Statistik SBD. 2023. Kabupaten SBD dalam Angka 2023. Weetabula: BPS Sumba Barat Daya, [2] Kasus malaria yang terjadi di Kecamatan Wewewa Timur penting untuk diketahui hubungan antara kejadian malaria dengan faktor yang mempengaruhinya agar mempermudah penanggulangan kasus kejadian malaria di Kecamatan Wewewa Timur. Maka dari itu, dilakukan penelitian untuk menentukan hubungan variabel demografis dengan kejadian malaria di Kecamatan Wewewa Timur dengan menggunakan pengujian chi-square.

Pendekatan chi-square adalah salah satu alat statistik yang paling umum digunakan dalam berbagai bidang ilmu, termasuk kesehatan. Uji chi-square didasarkan atas kesesuaian antara frekuensi amatan dengan frekuensi harapan yang diperoleh [3]. Metode ini memiliki kemampuan untuk menganalisis data yang terkategori atau diskrit [4]–[12], di mana variabel-variabel yang diamati berupa kategori atau kelompok. Dalam bidang kesehatan, pendekatan chi-square telah banyak digunakan dalam epidemiologi, analisis data klinis, penelitian kesehatan masyarakat, dan berbagai studi lainnya yang bertujuan untuk memahami hubungan antara faktor-faktor tertentu dengan kesehatan manusia. Beberapa penelitian terkait penggunaan metode Chi-square telah dilakukan sebelumnya.. [13]penelitian tentang hubungan antara pengetahuan masyarakat dengan kepatuhan penggunaan masker sebagai upaya pencegahan penyakit Covid-19 di Ngronggah menggunakan pendekatan analisis Chi-Square. [14] memfokuskan penelitiannya pada hubungan antara aspek kesehatan lingkungan dalam Perilaku Hidup Bersih dan Sehat (PHBS) rumah tangga dengan kejadian penyakit diare di Kecamatan Karangreja pada tahun 2012 menggunakan pendekatan analisis Chi-Square. [15] menyelidiki hubungan antara sistem pembelajaran daring dengan kesehatan mental mahasiswa di era COVID-19 menggunakan pendekatan analisis Chi-Square Test dan Dependency Degree sebagai model analisis. [16], melakukan penelitian terkait Chi-Square Tests dengan satu derajat kebebasan dan memperluas prosedur Mantel-Haenszel. [17] fokus pada aspek statistik analisis data dari studi retrospektif penyakit. [18] melakukan penelitian untuk menggali asosiasi antara profil pembuang sampah dan perilaku pembuangan sampah menggunakan pendekatan analisis Chi-Square. [19] melakukan penelitian yang berfokus pada penggunaan model analisis Chi-Square Automatic Interaction Detector Decision Tree untuk memprediksi respons cefmetazole pada infeksi intra-abdominal.

Berdasarkan penelitian yang telah disebutkan, belum ada yang secara khusus mengeksplorasi hubungan antara faktor-faktor demografi dengan kejadian malaria. Oleh karena itu, peneliti tertarik untuk mengisi kekosongan pengetahuan ini dengan melakukan penelitian yang berfokus pada hubungan antara faktor-faktor demografi (seperti usia, jenis kelamin, pendidikan, pekerjaan, dan status perkawinan) dengan kejadian malaria. Peneliti menggunakan pendekatan analisis Chi-Square karena penelitian ini akan melibatkan variabel-variabel kategorikal yaitu kelompok usia, jenis kelamin dalam kategori pria/wanita, pendidikan dalam beberapa tingkat pendidikan, pekerjaan dalam beberapa kategori pekerjaan, dan status perkawinan dalam beberapa kategori status perkawinan). Pendekatan Chi-Square sangat tepat digunakan untuk menguji hubungan antara variabel kategorikal seperti ini.

Metode penelitian ini akan melibatkan pengumpulan data dari populasi atau sampel yang relevan, dan kemudian data akan dianalisis menggunakan uji Chi-Square untuk mengevaluasi apakah ada hubungan yang signifikan antara faktor-faktor demografi dan kejadian malaria di Kecamatan tersebut. Hasil dari penelitian ini diharapkan dapat memberikan pemahaman yang lebih mendalam tentang peran faktor-faktor demografi dalam risiko terkena malaria, dan dapat membantu dalam perumusan kebijakan atau intervensi kesehatan yang lebih tepat dan efektif dalam mengurangi kejadian malaria. Kebaharuan terkait pendekatan chi-square dalam analisis faktor risiko kesehatan telah menjadi hal yang menarik dan relevan. Chi-square, sebagai metode statistik yang digunakan untuk menguji hubungan antara dua variabel kategori, telah menemukan penerapan yang semakin luas dalam berbagai aspek analisis faktor risiko kesehatan. Penelitian terkini telah menunjukkan bagaimana chi-square dapat digunakan untuk mengidentifikasi hubungan antara variabel-variabel kategorikal yang berperan sebagai faktor risiko penyakit. Chi-square juga telah diterapkan dalam analisis multivariat untuk mengidentifikasi kombinasi faktor risiko yang berkontribusi pada risiko penyakit secara bersama-sama. Berdasarkan uraian masalah di atas yang menjadi fokus dalam penelitian ini yaitu mengetahui bagaimana hubungan faktor -faktor demografi dengan kejadian malaria di Kecamatan Wewewa Timur di analisis menggunakan pendekatan Chi-square.

METODE

Penelitian telah dilakukan di Kecamatan Wewewa Timur, Kabupaten Sumba Barat Daya, Provinsi NTT selama 3 minggu terhitung pada tanggal 13 April 2023 sampai dengan 05 Mei 2023. Dalam penelitian ini menggunakan data primer dan data sekunder. Data primer menggunakan instrumen penelitian berupa kuesioner yang diambil dari penelitian terdahulu yaitu penelitian Guntur RD, Kingsley J, dan Islam FMA pada tahun 2021 dalam penelitian yang berjudul Epidemiology of Malaria in East Nusa Tenggara Province in Indonesia sedangkan data sekunder yaitu data dari Badan Pusat Statistik dan buku rekam medis penyakit malaria tahun 2022 dari Puskesmas Elopada Kecamatan Wewewa Timur. Variabel penelitian ditampilkan pada Tabel 1.

Table 1. Variabel penelitian

Variabel dependen			
Variabel	Simbol	Kategori	Variabel dummy
Kejadian Malaria	Y	Malaria	1
		Tidak malaria	0
Variabel independen			
Jenis kelamin	X_1	Laki – laki	1

		Perempuan	0
Umur	X_2	<25 tahun	1
		≥25 tahun	0
Pekerjaan	X_3	Di dalam ruangan	1
		Di luar ruangan	0
Penggunaan kelambu	X_4	Ya	1
		Tidak	0
Jenis Kelambu	X_5	Kelambu pembagian	1
		Kelambu bukan pembagian	0
Tipe dinding rumah	X_6	Bambu/papan	1
		Tembok	0
Tingkat Pendidikan	X_7	Tidak sekolah, SD, SMP	1
		SMA, D3, Sarjana	0

Uji Chi – square

Pengujian *Chi – square* dilakukan untuk mengetahui hubungan antara setiap variabel independen dan variabel dependen [19]–[24].

Proses kerja metode Chi-square dalam menguji hipotesis tentang hubungan antara variabel kategorikal melibatkan beberapa langkah penting. Berikut ini adalah penjelasan langkah demi langkah:

1. Mengumpulkan dan Mengatur Data :
   Awalnya, data dikumpulkan dan diatur dalam bentuk tabel kontingensi. Tabel ini menggambarkan frekuensi pengamatan pada kombinasi berbagai kategori dari dua variabel atau lebih.
2. Menghitung Frekuensi yang Diharapkanpenulis :
   Untuk setiap sel dalam tabel kontingensi, frekuensi yang diharapkan dihitung. Ini didasarkan pada asumsi bahwa tidak ada hubungan antara variabel, sehingga frekuensi yang diharapkan dihitung dengan mengalikan frekuensi marginal (total baris dan kolom) dan membaginya dengan total jumlah pengamatan.

$$E_i = \frac{(total\ baris \times total\ kolom)}{total\ pengamatan}.$$

3. Menghitung Statistik Uji Chi-squarepenulis :
   Selanjutnya, statistik uji Chi-square dihitung dengan formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

   di manaOiadalah frekuensi yang diamati, dan Ei adalah frekuensi yang diharapkan. Perhitungan ini dilakukan untuk setiap sel di tabel kontingensi.

4. Penulis Menentukan Derajat Kebebasan (Degree of Freedom, df)penulis :
   Derajat kebebasan dihitung berdasarkan jumlah baris dan kolom dalam tabel kontingensi. Biasanya,

$$df = (jumlah\ baris - 1) \times (jumlah\ kolom - 1).$$

5. Penulis Membandingkan dengan Nilai Kritispenulis :
   Nilai Chi-square yang dihitung kemudian dibandingkan dengan nilai kritis dari tabel distribusi Chi-square yang sesuai dengan derajat kebebasan dan tingkat signifikansi yang dipilih (misalnya, 0.05 atau 0.01).
6. Pengambilan Keputusanpenulis :
   Jika nilai Chi-square dihitung lebih besar dari nilai kritis, maka hipotesis nol (tidak ada perbedaan signifikan) ditolak. Ini menandakan adanya perbedaan yang signifikan antara frekuensi yang diamati dan yang diharapkan, yang mengindikasikan adanya hubungan antara variabel kategorikal yang diuji.
7. Interpretasi Hasilpenulis :
   Hasilnya kemudian diinterpretasikan dalam konteks penelitian. jika hipotesis nol ditolak, dapat disimpulkan bahwa terdapat hubungan yang signifikan antara variabel-variabel yang diuji.

Proses ini dilakukan menggunakan perangkat lunak statistik, Dalam melakukan pengujian ini menggunakan bantuan aplikasi *SPSS V.22.*

Hipotesis dalam uji *chi – square* yaitu:

H_0: $\chi^2 = 0$

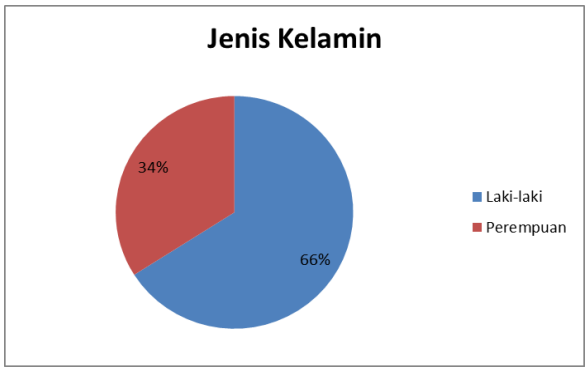
H_1: $\chi^2 \neq 0$

Dasar dalam pengambilan keputusan uji *chi – square* yaitu tolak H_0 jika nilai signifikan < 0,05 atau terima H_1 yang artinya ada hubungan signifikan antara variabel independen dan dependen, jika nilai signifikan > 0,05 maka terima H_0 atau tolak H_1 yang artinya tidak ada hubungan signifikan antara variabel independen dan variabel dependen.

HASIL DAN PEMBAHASAN

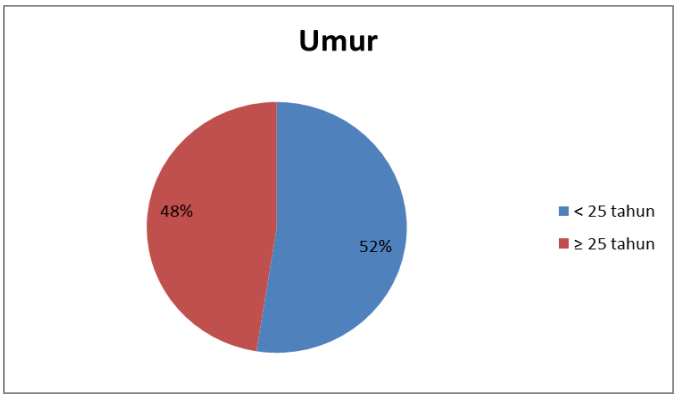
Analisis Deskriptif

Data yang terkumpul yaitu sebanyak 400 data yang berasal dari tiga desa yaitu dari Desa Dangga Mango sebanyak 149 data, Desa Tema Tana sebanyak 129 data dan Desa Kalembu Ndara Mane sebanyak 122 data yang menjadi sampel untuk menganalisis kejadian malaria di Kecamatan Wewewa Timur. Berdasarkan jenis kelamin terdapat 223 responden yang berjenis kelamin laki – laki dan 117 yang berjenis kelamin perempuan.



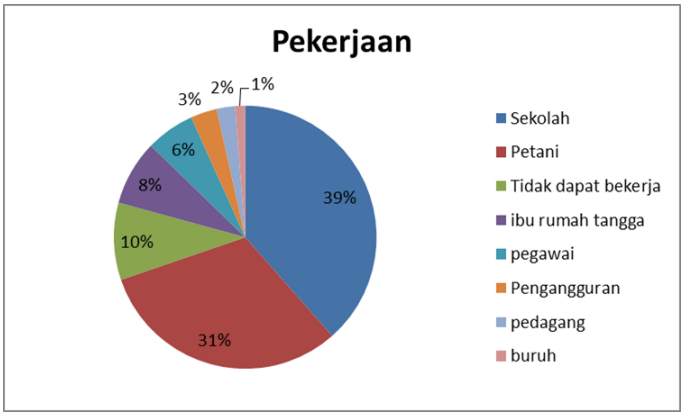
Gambar 1. Jenis Kelamin

Berdasarkan umur terdapat 210 responden yang umurnya kurang dari 25 tahun dan 190 responden yang berumur lebih dari sama dengan 25 tahun.



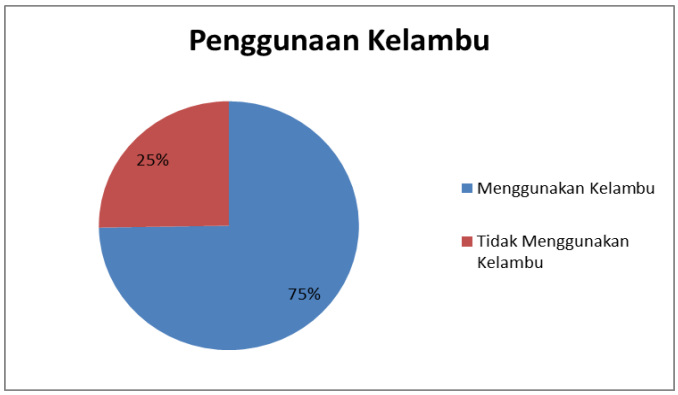
Gambar 2. Umur

Berdasarkan pekerjaan terdapat 154 responden yang masih sekolah, 125 responden yang bekerja sebagai petani, 38 responden yang tidak dapat bekerja, 32 responden yang bekerja sebagai ibu rumah tangga, 24 responden yang bekerja sebagai pegawai, 13 responden pengangguran, 9 responden yang bekerja sebagai pengusaha, dan 5 responden yang bekerja sebagai buruh. Jadi berdasarkan pekerjaan responden terdapat 261 atau 66% responden yang bekerja di dalam ruangan dan 139 atau 34% responden yang bekerja di luar ruangan.



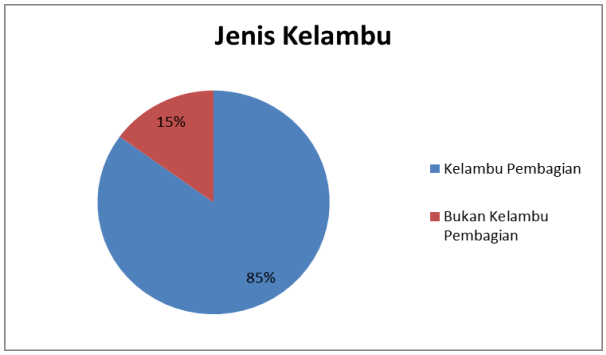
Gambar 3. Pekerjaan

Berdasarkan penggunaan kelambu oleh responden terdapat 101 responden yang tidur tidak menggunakan kelambu pada malam hari dan 299 responden yang tidur menggunakan kelambu pada malam hari.



Gambar 4. Penggunaan Kelambu

Berdasarkan jenis kelambu yang digunakan responden terdapat 340 responden yang menggunakan kelambu pembagian dari puskesmas yaitu kelambu yang berinsektisida dan 60 responden yang menggunakan bukan kelambu pembagian yaitu kelambu yang dibeli sendiri oleh responden dan tidak berinsektisida.



Gambar 5. Jenis Kelambu

Berdasarkan tipe dinding rumah yang menjadi tempat tinggal responden terdapat 179 responden yang memiliki rumah dengan dinding yang terbuat dari bambu dan 221 responden yang memiliki rumah dengan dinding yang terbuat dari semen parmanen.

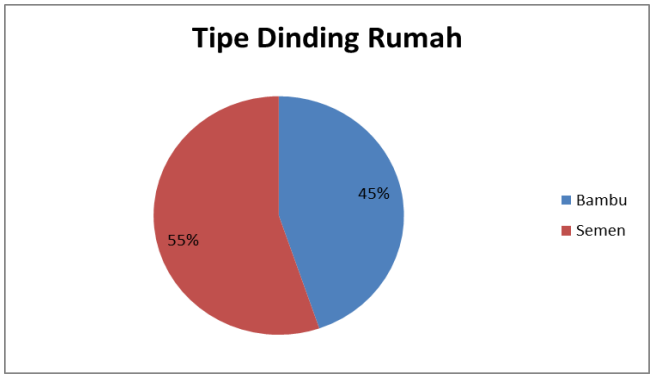


Figure 6. Tipe Dinding Rumah

Berdasarkan tingkat pendidikan dari responden terdapat 52 responden yang tidak sekolah, 118 responden yang berpendidikan SD, 97 responden yang berpendidikan SMP, 107 responden yang berpendidikan SMA dan 26 responden yang berpendidikan sarjana. Jadi berdasarkan tingkat pendidikan responden dapat dikelompokkan terdapat 267 atau 66% responden yang berpendidikan rendah dan 133 atau 34% responden yang berpendidikan tinggi.

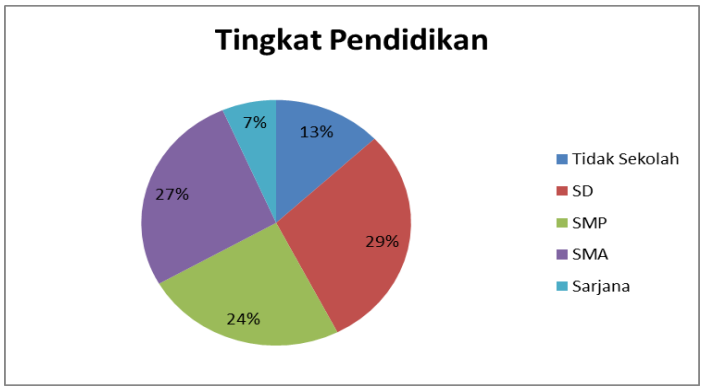


Figure 7. Tingkat Pendidikan

Berdasarkan hasil wawancara yang dilakukan di Kecamatan Wewewa Timur mengenai kejadian malaria, dari tiga desa yang menjadi sampel terdapat 227 responden yang pernah mengalami sakit malaria dan 173 responden yang tidak atau belum pernah mengalami sakit malaria.

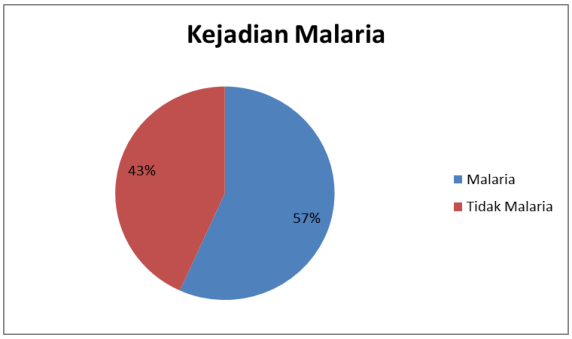


Figure 8. Kejadian Malaria

Uji Chi – square

Dalam analisis statistik yang dilakukan, uji Chi-square digunakan untuk mengevaluasi hubungan antara jenis kelamin dan kejadian malaria. Hipotesis nol (H0) diajukan dengan asumsi bahwa tidak ada hubungan signifikan antara kedua variabel tersebut, sementara hipotesis alternatif (H1) menyatakan adanya hubungan yang signifikan. Hasil uji menunjukkan nilai signifikansi sebesar 0,011, yang lebih kecil dari tingkat kepercayaan 5% atau 0,05 yang umumnya digunakan. Oleh karena itu, hipotesis nol dapat ditolak. Keputusan ini mengindikasikan bahwa terdapat hubungan yang signifikan antara jenis kelamin dan kejadian malaria dalam sampel yang diteliti. Meskipun temuan ini dapat diandalkan pada tingkat kepercayaan yang dipilih, penting untuk mempertimbangkan berbagai faktor validitas seperti ukuran sampel, representativitas, dan kontrol variabel. Dengan tingkat signifikansi yang rendah, yaitu 0,011, kesimpulan ini memberikan dasar

untuk menyimpulkan bahwa perbedaan dalam kejadian malaria antara jenis kelamin tidak terjadi secara kebetulan.

Namun, perlu diingat bahwa hasil statistik hanya memberikan informasi tentang hubungan antara variabel tersebut dalam konteks sampel yang diuji. Interpretasi yang cermat diperlukan, dan hasil ini sebaiknya diterapkan dengan hati-hati pada populasi lebih luas. Kesimpulan ini dapat membuka peluang untuk penelitian lanjutan atau intervensi yang lebih mendalam terkait faktor-faktor yang memengaruhi kejadian malaria berdasarkan jenis kelamin. Pengujian *Chi – square* dilakukan untuk mengetahui hubungan antara setiap variabel independen dan variabel dependen. Hipotesis dalam uji *chi – square* yaitu:

H_0: tidak ada hubungan signifikan antara variabel independen dan kejadian malaria.

H_1: Ada hubungan signifikan antara variabel independen dan kejadian malaria.

Dasar dalam pengambilan keputusan uji *chi – square* yaitu tolak H_0 jika nilai signifikan < 0,05 artinya ada hubungan signifikan antara variabel independen dan kejadian malaria.

Table 2. Hubungan Jenis Kelamin Dengan Kejadian Malaria

Jenis Kelamin	Kejadian Malaria				Total		Asymp. Sig. (2-sided)
	Tidak Malaria		Malaria				
	n	%	n	%	n	%	
Perempuan	89	22,2	88	22,0	177	44,2	
Laki – laki	84	21,0	139	37,8	223	55,8	0,011
Total	173	43,2	227	56,8	400	100	

Berdasarkan Tabel 2 dikethui nilai signifikan dari jenis kelamin yaitu 0,011 < 0,05 maka artinya menolak H_0 sehingga dapat disimpulkan bahwa terdapat hubungan yang signifikan antara jenis kelamin dan kejadian malaria.

Table 3. Hubungan Umur Dengan Kejadian Malaria

Umur	Kejadian Malaria				Total		Asymp. Sig. (2-sided)
	Tidak Malaria		Malaria				
	n	%	n	%	n	%	
≥25	56	14,0	134	33,5	190	47,5	
<25	117	29,2	93	23,3	210	52,5	0,000
Total	173	43,2	227	56,8	400	100	

Berdasarkan Tabel 3 nilai signifikan dari umur yaitu 0,000 < 0,05 artinya menolak H_0 dan menerima H_1 maka umur berhubungan signifikan dengan kejadian malaria.

Table 4. Hubungan Pekerjaan Dengan Kejadian Malaria

Pekerjaan	Kejadian Malaria				Total		Asymp. Sig. (2-sided)
	Tidak Malaria		Malaria				
	n	%	n	%	n	%	
Luar ruangan	28	7,0	111	27,8	139	34,8	
Dalam ruangan	145	36,2	116	29,0	261	65,2	0,000
Total	173	43,2	227	56,8	400	100	

Berdasarkan Tabel 4 nilai signifikan dari pekerjaan yaitu 0.000 < 0,05 yang artinya variabel independen pekerjaan memiliki hubungan yang signifikan dengan kejadian malaria.

Tabel 5. Hubungan Penggunaan Kelambu Dengan Kejadian Malaria

Penggunaan kelambu	Kejadian Malaria				Total		Asymp. Sig. (2-sided)
	Tidak Malaria		Malaria				
	n	%	n	%	n	%	
Ya	148	37,0	151	37,8	299	74,8	
Tidak	25	6,2	76	19,0	101	25,2	0,000
Total	173	43,2	227	56,8	400	100	

Berdasarkan Tabel 5 bahwa ada hubungan yang signifikan antara penggunaan kelambu dan kejadian malaria karena nilai signifikan penggunaan kelambu yaitu 0,000 < 0,05.

Tabel 6. Hubungan Jenis Kelambu Dengan Kejadian Malaria

Jenis Kelambu	Kejadian Malaria				Total		Asymp. Sig. (2-sided)
	Tidak Malaria		Malaria				
	n	%	n	%	n	%	
Kelambu Pembagian	147	36,7	193	48,3	340	85,0	
Bukan Kelambu Pembagian	26	6,5	34	8,5	60	15,0	0,989
Total	173	43,2	227	56,8	400	100	

Berdasarkan Tabel 6 tidak terdapat hubungan yang signifikan antara jenis kelambu dan kejadian malaria karena nilai signifikan jenis kelambu yaitu 0,989 > 0,05.

Tabel 7. Hubungan Tipe Dinding Rumah Dengan Kejadian Malaria

Tipe Dinding Rumah	Kejadian Malaria				Total		Asymp. Sig. (2-sided)
	Tidak Malaria		Malaria				
	n	%	n	%	n	%	
Semen	103	25,7	118	29,5	221	55,2	
Bambu	70	17,5	109	27,3	179	44,8	0,132
Total	173	43,2	227	56,8	400	100	

Berdasarkan Tabel 7 diketahui bahwa nilai signifikan dari tipe dinding rumah yaitu 0,132 > 0,05. Artinya menerima H_0 sehingga dapat disimpulkan bahwa tidak ada hubungan yang signifikan antara tipe dinding rumah dan kejadian malaria.

Tabel 8. Hubungan Tingkat Pendidikan Dengan Kejadian Malaria

Tingkat Pendidikan	Kejadian Malaria				Total		Asymp. Sig. (2-sided)
	Tidak Malaria		Malaria				
	n	%	N	%	n	%	

Tinggi	68	17,0	65	16,3	133	33,3	
Rendah	105	26,2	162	40,5	267	66,7	0,025
Total	173	43,2	227	56,8	400	100	

Berdasarkan Tabel 8 nilai signifikan dari tingkat pendidikan yaitu 0,025 < 0,05 artinya ada hubungan yang signifikan dari tingkat pendidikan dan kejadian malaria. Dari hasil uji *chi - square* diperoleh bahwa terdapat 2 variabel independen yaitu jenis kelambu (X_5) dan tipe dinding rumah (X_6) yang tidak memiliki hubungan yang singnifikan dengan kejadian malaria di Kecamatan Wewewa Timur.

Pembahasan

Hasil penelitian menunjukkan variabel jenis kelamin mempengaruhi kejadian malaria di Kecamatan Wewewa Timur. Jenis kelamin mempengaruhi dikarenakan memiliki hubungan dengan kebiasaan keluar rumah dan bekerja [25]. Dari hasil wawancara yang telah dilakukan diketahui bahwa hal semacam itu terjadi juga di Kecamatan Wewewa Timur karena kebanyakan responden memiliki pekerjaan sebagai petani, baik itu sebagai pekerjaan utama atau pekerjaan sampingan karena didukung dengan kondisi geografis yang merupakan kawasan pertanian, hutan dan persawahan. Pekerjaan sebagai petani tidak saja dilakukan oleh laki – laki tetapi dilakukan juga oleh perempuan.

Berdasarkan hasil penelitian proporsi jenis kelamin laki – laki dengan kejadian malaria di Kecamatan Wewewa Timur sebanyak 37,8%. Persentase ini lebih tinggi dibandingkan perempuan dengan kejadian malaria di Kecamatan Wewewa Timur sebanyak 22%. Artinya laki – laki lebih beresiko terkena malaria. Hal terjadi karena dari hasil wawancara diketahui bahwa laki – laki di kecamatan ini lebih banyak melakukan aktivitas di luar rumah pada malam hari seperti mengairi sawah, memancing, ronda malam dan nongkrong sehingga lebih banyak mengalami kontak langsung dengan nyamuk malaria dan lebih mudah terkena malaria. Hasil penelitian ini sejalan dengan [26] yaitu seorang laki – laki lebih berisiko terkena malaria dibandingkan perempuan karena laki – laki sering malakukan kegiatan pada malam hari.

Secara umum penyakit malaria ini dapat menyerang semua masyarakat tanpa mengenal umur seseorang. Namun demikian hasil penelitian menunjukan proposi kelompok umur ≥ 25 tahun dengan kejadian malaria di Kecamatan Wewewa Timur sebanyak 33,5%. Persentase ini lebih tinggi dibandingkan kelompok umur < 25 tahun dengan kejadian malaria sebanyak 23,3%. Artikan bahwa masyakat yang umurnya ≥ 25 lebih berisiko terkena malaria. Dari hasil wawancara dengan responden banyak orang dewasa yaitu kelompok umur ≥ 25, lebih banyak melakukan aktivitas diluar rumah seperti bertani, berkebun dan mengairi sawah, karena hal tersebut orang dewasa lebih banyak melakukan kontak langsung dengan nyamuk malaria yang menyebabkan peluang terkena penyakit malaria lebih besar. Dengan demikian umur memiliki hubungan bermakna dengan kejadian malaria. Hasil penelitian ini sejalan dengan penelitian [27], yang menjelaskan bahwa umur memiliki hubungan bermakna dengan kejadian malaria yaitu pada kelompok umur 16 – 35 tahun.

Bedasarkan hasil wawancara diketahui bahwa kondisi geografis Kecamatan Wewewa Timur yang banyak kawasan hutan, perairan seperti sawah dan rawa berpotensi membantu perkembangbiakan nyamuk sehingga pekerjaan mempengaruhi kejadian malaria di kecamatan

Wewewa Timur. Berdasarkan data yang sudah terkumpul lebih banyak responden memiliki pekerjaan utama yang tempat bekerjanya di dalam ruangan. Proporsi responden yang bekerja di dalam ruangan dengan tidak terjadi malaria di kecamatan ini sebanyak 36,2%, persentase ini lebih tinggi dibandingkan responden yang bekerja di luar ruangan dengan tidak terjadi malaria di kecamatan ini sebanyak 7%. Banyaknya masyarakat yang bekerja dalam ruangan akan mengurangi kontak langsung dengan nyamuk *Anopheles* yang menyebabkan sakit malaria. Hasil penelitian ini relevan dengan [28] mengatakan bahwa jenis pekerjaan di luar ruangan memberikan kontribusi untuk menyebabkan malaria sebesar 3 kali dibandingkan yang memiliki pekerjaan di dalam ruangan. Pekerjaan berisiko (bekerja di luar ruangan) yang dilakukan oleh masyarakat di Kecamatan Wewewa Timur antara lain bertani, berkebun, menebang pohon, buruh dan membajak sawah. [2] menjelaskan "bahwa keadaan lingkungan seperti danau, area pesawahan, tambak ikan, hutan, dan pertambangan meningkatkan kemungkinan timbulnya malaria karena tempat – tempat tersebut tempat penyebaran nyamuk malaria yang biasa dikenal dengan nyamuk *Anopheles* betina,". Selain pekerjaan tersebut sebagian besar masyarakat di Kecamatan Wewewa Timur memiliki rutinitas memberikan makan ternak setiap hari. Hasil penelitian [29], menjelaskan keberadaan kandang ternak mempengaruhi kejadian malaria karena kandang ternak merupakan sala satu tempat perindukan nyamuk malaria.

Penggunaan kelambu telah diketahui sebagai salah satu upaya untuk mencegah terjadinya malaria [30]. Hasil penelitian ini menunjukkan proporsi responden yang menggunakan kelambu dengan kejadian malaria sebanyak 37,8%, persentase ini lebih tinggi dibandingkan responden menggunakan kelambu dengan kejadian malaria sebanyak 37%. Dengan demikian di Kecamatan Wewewa Timur faktor penggunaan kelambu tidak sendiri mempengaruhi kejadian malaria. Dari hasil wawancara hal ini bisa terjadi diduga karena responden yang menggunakan kelambu tidak memperhatikan beberapa hal diantaranya frekuensi penggunaan kelambu, perawatan kelambu, kondisi kelambu yang diigunakan tidak baik seperti berlubang atau robek dan penggunaan kelambu tidak dimasukan dibawah kasur. Sehingga kontak dengan nyamuk malaria pada malam hari masih terjadi meskipun sudah menggunakan kelambu.

Hasil analisis deskriptif terdapat 400 responden bahwa sebagian besar memiliki pendidikan yang masih rendah. Hal ini sejalan dengan penelitian yang dilakukan oleh [31], yang mengatakan masyarakat pedesaan di Nusa Tenggara Timur umumnya berpendidikan rendah. Tingkat pendidikan responden yang rendah diduga oleh karena keterbatasan ekonomi untuk bersekolah ke jenjang lebih tinggi, pola pikir responden cenderung bepikir untuk masalah jangka pendek berkaitan kebutuhan sandang dan pangan serta kesadaran diri dari setiap responden untuk bersekolah ke jenjang lebih tinggi tidak ada [31].

Tingkat pendidikan yang rendah erat kaitannya dengan pola pikir responden dalam berperilaku kurang baik terhadap kejadian malaria. Dengan pendidikan yang rendah, pengetahuan dan pemahaman responden mengenai malaria seperti cara pencegahan malaria kurang. Sehingga di kecamatan ini banyak responden yang tidak menerapkan perilaku pencegahan nyamuk malaria seperti menggunakan pakaian pelindung badan (baju lengan panjang dan celana lengan panjang) saat bekerja di sawah atau di kebun, mengurangi aktivitas malam hari, menggunakan kelambu tidak sesuai fungsi dan jarang menggunakan obat nyamuk bakar/semprot/elektrik padahal kondisi lingkungan sekitar rumah merupakan sawah, hutan dan kandang ternak yang merupakan tempat perindukan nyamuk. Hasil penelitian ini juga menunjukkan proporsi responden berpendidikan

rendah dengan kejadian malaria di Kecamatan Wewewa Timur sebanyak 40,8%, persentase ini lebih tinggi dibandingkan responden berpendidikan tinggi dengan kejadian malaria sebanyak 16,3%. Hasil penelitian ini sejalan [31] yaitu masyarakat yang mempunyai pendidikan rendah mempunyai perilaku kurang baik dibandingkan dengan masyarakat yang mempunyai pendidikan tinggi atau semakin tinggi pendidikan semakin baik perilaku terhadap pemberantasan malaria.

KESIMPULAN

Berdasarkan hasil penelitian yang menyatakan adanya hubungan bermakna antara faktor demografis (jenis kelamin, umur, pekerjaan, penggunaan kelambu, dan tingkat pendidikan) dengan kejadian malaria di Kecamatan Wewewa Timur, serta adanya kawasan hutan, pertanian, dan persawahan yang dapat membantu perkembangbiakan nyamuk malaria, serta perilaku masyarakat dalam mencegah penyakit malaria. Berdasarkan temuan penelitian, implikasi yang dapat diambil adalah bahwa penggunaan Chi-Square sebagai alat analisis menegaskan bahwa pemilihan metode yang tepat untuk mengevaluasi hubungan kategorikal antara faktor demografis dan kejadian malaria di Kecamatan Wewewa Timur. Selain itu Validitas Temuan terkait keterhubungan faktor demografis dengan kejadian malaria menunjukkan pentingnya faktor-faktor tersebut dalam mempengaruhi tingkat infeksi malaria di wilayah tersebut. Implikasi ini memberikan dukungan terhadap validitas temuan penelitian dan mengindikasikan bahwa variabel-variabel yang dipilih memiliki dampak yang signifikan terhadap kejadian penyakit malaria. Penelitian ini menyoroti pentingnya kondisi geografis Kecamatan Wewewa Timur (seperti kawasan hutan, pertanian, dan persawahan) dalam mendukung perkembangbiakan nyamuk malaria. Implikasi ini menggarisbawahi perlunya pendekatan pencegahan dan pengendalian yang spesifik terhadap vektor malaria dalam konteks lingkungan geografis tertentu. Penelitian ini memberikan informasi penting bagi pengambil kebijakan tentang faktor-faktor demografis yang perlu dipertimbangkan dalam merancang program pencegahan dan pengendalian malaria di Kecamatan Wewewa Timur. Implikasi ini menyoroti perlunya pendekatan yang komprehensif dan berbasis bukti dalam mengatasi masalah kesehatan masyarakat yang berkaitan dengan malaria.

UCAPAN TERIMA KASIH

Kami ingin mengucapkan terima kasih sebesar-besarnya kepada Dinas Kesehatan Kecamatan Wewewa Timur atas dukungan dan izin yang diberikan dalam penelitian ini. Terima kasih kepada pembimbing yang memberikan saran dan masukan berharga dalam penulisan artikel ini. Dukungan dan telah memberikan kontribusi besar.

DAFTAR PUSTAKA

[1] S. SUWITO, U. K. HADI, S. H. SIGIT, and S. SUKOWATI, “Hubungan Iklim, Kepadatan Nyamuk Anopheles dan Kejadian Penyakit Malaria,” *J Entomol Indones*, vol. 7, no. 1, 2015, doi: 10.5994/jei.7.1.42.

[2] T. P. Utami *et al.*, “Faktor Risiko Penyebab Terjadinya Malaria di Indonesia : Literature Review,” *Jurnal Surya Medika*, vol. 7, no. 2, 2022, doi: 10.33084/jsm.v7i2.3211.

[3] W. Pramesti, “Tabel Kontingensi Untuk Mengetahui Hubungan Antara Jenis Penyakit, Jenis Kelamin, Usia, Lama Rawat dan Keadaan Keluar Pasien,” *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, vol. 4, no. 1, 2012, doi: 10.36456/jstat.vol4.no1.a1165.

[4] E. B. Wilson and M. M. Hilferty, "The Distribution of Chi-Square," *Proceedings of the National Academy of Sciences*, vol. 17, no. 12, 1931, doi: 10.1073/pnas.17.12.684.
[5] D. Sharpe, "Your chi-square test is statistically significant: Now what?," *Practical Assessment, Research and Evaluation*, vol. 20, no. 8, 2015.
[6] M. Masih and A. Grant, "Chi square feature extraction based SVMS arabic language text categorization system," *Talent Development and Excellence*, vol. 9, no. 2, 2017, doi: 10.3844/jcssp.2007.430.435.
[7] Ari. Wibowo, "Uji Chi-Square pada Statistika dan SPSS," *Jurnal Ilmiah SINUS*, vol. 4, no. 2, 2017.
[8] F. B. Bryant and A. Satorra, "Principles and Practice of Scaled Difference Chi-Square Testing," *Structural Equation Modeling*, vol. 19, no. 3, 2012, doi: 10.1080/10705511.2012.687671.
[9] A. Satorra and P. M. Bentler, "A scaled difference chi-square test statistic for moment structure analysis," *Psychometrika*, vol. 66, no. 4, 2001, doi: 10.1007/BF02296192.
[10] S. D. Bolboacă, L. Jäntschi, A. F. Sestraş, R. E. Sestraş, and D. C. Pamfil, "Pearson-fisher chi-square statistic revisited," *Information (Switzerland)*, vol. 2, no. 3, 2011, doi: 10.3390/info2030528.
[11] N. Pandis, "The chi-square test," *American Journal of Orthodontics and Dentofacial Orthopedics*, vol. 150, no. 5. 2016. doi: 10.1016/j.ajodo.2016.08.009.
[12] X. Ji, W. Gu, X. Qian, H. Wei, and C. Zhang, "Combined Neyman–Pearson chi-square: An improved approximation to the Poisson-likelihood chi-square," *Nucl Instrum Methods Phys Res A*, vol. 961, 2020, doi: 10.1016/j.nima.2020.163677.
[13] Devi Pramita Sari and Nabila Sholihah 'Atiqoh, "HUBUNGAN ANTARA PENGETAHUAN MASYARAKAT DENGAN KEPATUHAN PENGGUNAAN MASKER SEBAGAI UPAYA PENCEGAHAN PENYAKIT COVID-19 DI NGRONGGAH," *Infokes: Jurnal Ilmiah Rekam Medis dan Informatika Kesehatan*, vol. 10, no. 1, 2020, doi: 10.47701/infokes.v10i1.850.
[14] A. Y. Irawan, "Hubungan antara Aspek Kesehatan Lingkungan dalam PHBS Rumah Tangga dengan Kejadian Penyakit Diare di Kecamatan Karangreja Tahun 2012," *Unnes Journal of Public Health*, vol. 2, no. 4, 2014.
[15] N. G. Annisa, R. Efendi, and L. Chairani, "Hubungan Sistem Pembelajaran Daring dengan Kesehatan Mental Mahasiswa Di Era COVID-19 Menggunakan Chi-Square Test dan Dependency Degree," *SNTIKI (Seminar Nasional Teknologi Informasi Komunikasi dan Industri*, no. SNTIKI, 2020.
[16] N. Mantel, "Chi-Square Tests with One Degree of Freedom; Extensions of the Mantel-Haenszel Procedure," *J Am Stat Assoc*, vol. 58, no. 303, 1963, doi: 10.2307/2282717.
[17] N. Mantel and W. Haenszel, "Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease | JNCI: Journal of the National Cancer Institute | Oxford Academic," *J Natl Cancer Inst*, 1959.
[18] M. Asmui, S. M. Zaki, S. N. S. Wahid, N. M. Mokhtar, and S. S. Harith, "Association between litterers' profile and littering behavior: A chi-square approach," in *AIP Conference Proceedings*, 2017. doi: 10.1063/1.4982841.
[19] M. Hiranuma, D. Kobayashi, K. Yokota, and K. Yamamoto, "Chi-square automatic interaction detector decision tree analysis model: Predicting cefmetazole response in intra-abdominal infection," *Journal of Infection and Chemotherapy*, vol. 29, no. 1, 2023, doi: 10.1016/j.jiac.2022.09.002.
[20] I. C. Negara and A. Prabowo, "Penggunaan Uji Chi–Square untuk Mengetahui Pengaruh Tingkat Pendidikan dan Umur terhadap Pengetahuan Penasun Mengenai HIV–AIDS di

Provinsi DKI Jakarta," *Prosiding Seminar Nasional Matematika dan Terapannya 2018*, vol. 1, no. 1, 2018.
[21] M. L. Mchugh, "The Chi-square test of independence Lessons in biostatistics," *Biochem Med (Zagreb)*, vol. 23, no. 2, 2013.
[22] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 2, 2020, doi: 10.1016/j.jksuci.2018.05.010.
[23] S. T. Nihan, "Karl Pearsons chi-square tests," *Educational Research and Reviews*, vol. 15, no. 9, 2020, doi: 10.5897/err2019.3817.
[24] M. L. McHugh, "The Chi-square test of independence," *Biochem Med (Zagreb)*, vol. 23, no. 2, 2012, doi: 10.11613/BM.2013.018.
[25] A. Ruliansyah and F. Y. Pradani, "Perilaku-Perilaku Sosial Penyebab Peningkatan Risiko Penularan Malaria di Pangandaran," *Buletin Penelitian Sistem Kesehatan*, vol. 23, no. 2, 2020, doi: 10.22435/hsr.v23i2.2797.
[26] A. Farihatun and Z. Mamdy, "FAKTOR-FAKTOR YANG BERHUBUNGAN DENGAN PERILAKU PENCEGAHAN PENYAKIT MALARIA PADA MASYARAKAT DI DESA KARYAMUKTI KECAMATAN CIBALONG KABUPATEN GARUT PROVINSI JAWA BARAT," *Jurnal Kesehatan Bakti Tunas Husada: Jurnal Ilmu-ilmu Keperawatan, Analis Kesehatan dan Farmasi*, vol. 15, no. 1, 2016, doi: 10.36465/jkbth.v15i1.157.
[27] Istiana, M. D. Prenggono, J. E. S. Parhusip, and M. F. A. Rahman, "Angka Kejadian Malaria Berdasarkan Pemeriksaan Raoid Diagnostik Test di Kalimantan Selatan," *Prosiding Seminar Nasional Lingkungan Lahan Basah*, vol. 6, no. 3, 2021.
[28] W. Wibowo, "RISIKO KEJADIAN MALARIA DI WILAYAH KERJA PUSKESMAS KECAMATAN CIKEUSIK," *Media Kesehatan Masyarakat Indonesia*, vol. 13, no. 2, 2017, doi: 10.30597/mkmi.v13i2.1985.
[29] M. P. Moni, K. Br Ginting, and Aryanto, "Analisis regresi logistik terhadap faktor-faktor yang mempengaruhi kejadian malaria pada balita di Kecamatan Lamboya Kabupaten Sumba Barat," *Jurnal MIPA FST UNDANA*, vol. 20, no. 1, 2016.
[30] R. D. Guntur, J. Kingsley, and F. M. A. Islam, "Epidemiology of malaria in east nusa tenggara province in indonesia: Protocol for a cross-sectional study," *JMIR Res Protoc*, vol. 10, no. 4, 2021, doi: 10.2196/23545.
[31] R. D. Guntur, J. Kingsley, and F. M. Amirul Islam, "Malaria awareness of adults in high, moderate and low transmission settings: A cross-sectional study in rural East Nusa Tenggara Province, Indonesia," *PLoS One*, vol. 16, no. 11 November 2021, 2021, doi: 10.1371/journal.pone.0259950.

Comparison of K-Means and K-Medoids Clustering for Grouping The Sub-Districts In Bojonegoro Regency Based On Educational Supporting Factors

Alif Yuanita Kartini[(1)], Syarif Husen[(2)]

[1,2]Universitas Nahdlatul Ulama Sunan Giri

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

ABSTRAK

Pendidikan di Bojonegoro saat ini masih belum merata. Hal ini dikarenakan upaya pemerataan pendidikan yang telah dilakukan mempunyai banyak kendala. Kendala yang sering terjadi adalah masyarakat yang berada di daerah terpencil dan jauh dari perkotaan kesulitan dalam mengakses layanan pendidikan. Oleh karena itu pengelompokan wilayah perlu dilakukan agar pemerintah kabupaten Bojonegoro dapat memperhatikan cluster wilayah yang memperlukan perbaikan pendidikan. Pada penelitian ini menggunakan metode K-Means dan K-Medoids untuk pengelompokan kecamatan di kabupaten Bojonegoro berdasarkan faktor pendukung pendidikan. K-Means merupakan salah satu metode unsupervised learning yang digunakan untuk menganalisis data dengan melakukan pengelompokan. Sementara itu K-Medoids merupakan metode pengelompokan partisi yang mengelompokkan sekumpulan n objek menjadi sejumlah k cluster. Data yang digunakan dalam penelitian ini merupakan data sekunder yang didapatkan dari Dinas Pendidikan kabupaten Bojonegoro berupa data faktor pendukung pendidikan yang meliputi jumlah sekolahan, jumlah tenaga pendidik, dan jumlah rombongan belajar (ROMBEL) tahun 2022 pada masing-masing kecamatan di kabupaten Bojonegoro. Dari hasil penelitian didapatkan metode K-Means lebih baik dibandingkan dengan metode K-Medoids. Hasil pengelompokan menggunakan K-Means didapatkan sebanyak 5 cluster, dimana cluster 1 beranggotakan 1 kecamatan, cluster 2 beranggotakan 7 kecamatan, cluster 3 beranggotakan 1 kecamatan, cluster 4 beranggotakan 12 kecamatan dan cluster 5 beranggotakan 7 kecamatan. Berdasarkan karakteristik dari masing-masing cluster yang didapatkan, diharapkan dapat digunakan sebagai masukan bagi dinas Pendidikan sebagai Upaya pemerataan Pendidikan di kabupaten Bojonegoro.
Kata kunci: K-Means; K-Medoids; Pemerataan Pendidikan; Pengelompokan

ABSTRACT

Education in Bojonegoro is currently still uneven. This is because efforts to equalize education that have been carried out have many obstacles. The obstacle that often occurs is that people who are in remote areas and far from urban areas have difficulty accessing education services. Therefore, regional grouping needs to be done so that the Bojonegoro district government can pay attention to regional clusters that need education improvement. This study used the K-Means and K-Medoids methods to group sub-districts in Bojonegoro district based on educational supporting factors. K-Means is one of the unsupervised learning methods used to analyze data by grouping. Meanwhile, K-Medoids is a partition grouping method that groups a set of n objects into a number of k clusters. The data used in this study is secondary data obtained from the Bojonegoro district Education Office in the form of data on education supporting factors which include the number of schools, the number of educators, and the number of learning groups (ROMBEL) in 2022 in each sub-district in Bojonegoro district. From the research results, it was found that the K-Means method was better than the K-Medoids method. The results of grouping using K-Means obtained as many as 5 clusters, cluster 1 consists of 1 sub-district, cluster 2 consists of 7 sub-districts, cluster 3 consists of 1 sub-district, cluster 4 consists of 12 sub-districts and cluster 5 consists of 7 sub-districts. Based on the characteristics of each cluster obtained, it is expected to be used as input for the Education office for equal distribution of education in Bojonegoro district.
Keywords: *K-Means; K-Medoids; Education Equity; Grouping*

INTRODUCTION

Education is one of the main factors in a nation's progress [1]. A developed nation has a good education. The benchmark of good education is equal distribution of opportunities to get proper education or commonly referred to as equal distribution of education [2]. Education equity has long been an issue that has received a lot of attention, especially in developing countries. Equality of education includes two important aspects, namely equal opportunities to obtain education and equal justice in society [3].

Education in Bojonegoro is currently still uneven. There are still many people who have not received the education they should have received since of 6 years old [4]. This is because efforts to equalize education have many obstacles. The obstacle that often occurs is that people who are in remote areas and far from urban areas have difficulty accessing education services [5]. In the implementation of education, there are several supporting factors in the implementation of effective and efficient teaching and learning activities, including the provision of schools, the provision of learning groups (ROMBEL) and the provision of educators [6]. Supporting factors in providing education must be felt by all regions, especially those in Bojonegoro district. Therefore, regional grouping needs to be done so that the Bojonegoro district government can pay attention to which regional clusters need improved education.

One method that can be used for clustering is cluster analysis [7]. Cluster analysis is the merging of objects or data based on similar characteristics [8]. There are many methods in cluster analysis, including K-Means and K-Medoids. In a study entitled "Mapping Excellent Class Students Using the K-Means Clustering Algorithm" stated that K-Means is an algorithm that is easy to implement [9]. In another study entitled "K-Means and K-Medoids Algorithm for Grouping Subdistricts of Social Assistance Recipients in Bojonegoro Regency" stated that the K-Means method has a relatively small complexity of space and time [10]. Meanwhile, in a study entitled "Hybrid K Means-Multivariate Adaptive Regression Splines For Distribution Of Dengue Fever Risk Mapping In Bojonegoro District" stated that the K-Means method is able to group big data very quickly [11]. Another study showed that K-Means is an effective method for grouping datasets [12], [13],[14]. Meanwhile, the K-Medoids method also has many advantages in several studies. In a study entitled "Implementation of K-Medoids Clustering Method for Grouping Data on Potential Forest/Land Fires Based on the Distribution of Hot Spots (Hotspot)" stated that K-Medoids is a method that is sensitive to outliers [15]. In another study entitled "Analysis of K-Medoids in Grouping the Ratio of Students to Teachers, Students with Rombel, and the Ratio of Rombel to Elementary and Junior High School Education Grades by Province" stated that K-Medoids are more reliable when there is data noise [16]. Based on research entitled "Comparison of Distance Measure on K-Medoids Clustering for ARI Disease Grouping" states that the K-Medoids algorithm can get the closest grouping results to an object [17]. Another study states that K-Medoids are not affected by other extreme data [18], [19].

Based on this background, research will be conducted for the grouping of sub-districts in Bojonegoro sub-district based on educational supporting factors using the K-Means and K-Medoids algorithms. From this research, it is expected to provide advice and input for the government to pay attention to areas that lack schools, educators, and learning groups (ROMBEL) in accordance with the cluster results obtained. This is an effort to equalize education in Bojonegoro district.

METHOD

The data used in this study is secondary data obtained from the Bojonegoro District Education Office which is published on the Bojonegoro one data web. The data is in the form of data on education supporting factors which include the number of schools, the number of educators, and the number of learning groups (ROMBEL) in 2022 in each sub-district in Bojonegoro district. The variables used in this study include number of elementary schools (x_1), number of junior high schools (x_2), number of high schools (x_3), number of elementary school educators (x_4), number of junior high school educators (x_5), number of high school educators (x_6), number of elementary school learning groups (x_7), number of junior high school learning groups (x_8), and number of high school learning groups (x_9).

The methods used in this study are K-Means and K-Medoids. K-Means is a distance-based clustering method that divides data into clusters. K-Means algorithm only works on numeric attributes. K-Means algorithm includes partitioning clustering that separates data into k separate regions. K-Means algorithm is well known for its ease and ability to segment big data and data outliers very quickly. In K-Means algorithm, each data must belong to a specific cluster. In K-Means algorithm it is possible for any data belonging to a particular cluster at one stage of the process, at the next stage moving to another cluster.

Meanwhile, K-Medoids is a classic partitioning technique of clustering that clusters datasets of n objects into k clusters known as a priori. K-Medoids algorithm operates on the principle of minimizing the number of similarities between each object and its corresponding reference points. The K-Medoids algorithm uses data objects as representatives (medoids) of the cluster center. K-Medoids algorithm is used to overcome the weaknesses of the k-means algorithm which is very sensitive to outliers. This is because these objects are very far from the majority of other data, so that if entered into a cluster, this kind of data can distort the average value (mean) of the cluster.
In this study using the help of Rapidminer software. The analysis steps used are as follows.

1. Group sub-districts in Bojonegoro based on education supporting factors using the k-means algorithm with the following steps [20],[21],[22],[23] :
   a. Determine the number of clusters to use (k)
   b. Split data into k clusters
   c. Calculate the centroid value or average of data in each cluster formed
   d. Calculates Euclidean distances to determine the distance to all mean values in each cluster
   e. Allocate data for clusters that have the average closest to the data
   f. Loop starting in step (b) if there is still data moving to the cluster
   g. Get clustering results along with members from each cluster
   h. Validate cluster quality by calculating Silhoutte values Coefficient
2. Grouping sub-districts in Bojonegoro based on education supporting factors using the K-Medoids algorithm with the following steps [15],[16],[17],[24]:
   a. Determine the number of clusters to be used (k)
   b. Determine medoids by randomly taking objects from objects to be grouped
   c. Calculate Euclidean distance to obtain the distance of non-medoids objects
   d. Determine the members of each cluster based on the closest distance to the grouping object
   e. Calculates the total distance of non-medoids objects closest to the grouping object

f. Randomly select one non-medoids object for each cluster that is used as a candidate for new medoids. If the selected object has been a medoid, it is not allowed to be re-selected.
g. Calculates the Euclidean distance at the distance of non-medoids objects to new medoids candidates for each cluster.
h. Determine the members of each cluster based on the closest distance to prospective new medoids
i. Calculates the total distance of nonmedoids objects closest to potential new medoids
j. Calculates the difference between the total distance and the new prospective medoids with the total distance and the medoids
k. If the value of the total distance difference is less than zero, then the candidate medoids become new medoids in the next iteration
l. Repeating from step (f) until there is no change in medoids
m. If the value of the total distance difference is more than zero, the iteration will stop and get the cluster and the members of each cluster

3. Comparing the results of sub-district grouping in Bojonegoro district based on supporting factors of education using K-Means and K-Medoids algorithms by looking at cluster distance performance values [25],[26].
4. Create a clustered map with the best method
5. Getting the characteristics of each cluster from the best method

RESULT AND DISCUSSION

Clusterization Results Using K-Means Algorithm

The first step to grouping using the K-Means algorithm is to preprocess the data. Data preprocessing is preparing data and standardizing it. The next step after preprocessing data is to process data. The number of clusters (k) used in this study was 2, 3, 4 and 5. Furthermore, to evaluate the optimal grouping results, namely by looking at the smallest average value within centroid distance. From the results of the analysis, the smallest average within centroid distance value in the number of clusters (k) was 5, which was 1.833. So that the number of clusters used to group sub-districts in Bojonegoro district based on supporting factors for education with the K-Means algorithm is 5. After obtaining the number of clusters to be used, the next step is to determine the initial centroid value randomly, calculate the Euclidian distance, grouping objects with the shortest distance between objects, until a fixed centroid value is obtained and cluster members do not move to another cluster. Based on these steps, clustering results using K-Means are obtained as shown in Table 1 below.

Table 1. Number and Members of Clusters Formed Using K-Means

Cluster	Number of cluster members	Cluster Members
Cluster 1	1	Kedungadem
Cluster 2	7	Baureno, Dander, Kalitidu, Kepohbaru, Ngasem, Padangan, Sumberejo
Cluster 3	1	Bojonegoro

Cluster	Number of cluster members	Cluster Members
Cluster 4	12	Balen, Bubulan, Gayam, Gondang, Kapas, Kasiman, Kedewan, Margomulyo, Ngambon, Ngraho, Sekar, Trucuk
Cluster 5	7	Kanor, Malo, Purwosari, Sugihwaras, Sukosewu, Tambakrejo, Temayang

Clusterization Results Using K-Medoids Algorithm

Just like grouping using the K-Means algorithm, grouping using the K-Medoids algorithm also begins with determining the number of clusters. The number of clusters (k) used in this study is the same as that used in K-Means, which are 2, 3, 4 and 5. To evaluate the best grouping results, namely by using the smallest average value within centroid distance. From the results of the analysis, the smallest average within centroid distance value in the number of clusters (k) was 5, which was 2.786. So that the number of clusters used to group sub-districts in Bojonegoro based on supporting factors for education with the K-Means algorithm is 5.

The next step after determining the number of clusters is to determine the initial centroid value as medoids. Furthermore, after going through several stages until the clustering process is stopped, the clusterization results are obtained using the K-Medoids algorithm as shown in Table 2 below.

Table 2. Number and Members of Clusters Formed Using K-Medoids

Cluster	Number of cluster members	Cluster Members
Cluster 1	12	Kalitidu , Kanor, Kasiman, Malo, Ngasem, Ngraho, Padangan, Purwosari, Sugihwaras, Sukosewu, Tambakrejo, Temayang
Cluster 2	5	Baureno, Dander, Kedungadem, Kepohbaru, Sumberejo
Cluster 3	1	Bojonegoro
Cluster 4	8	Balen, Bubulan, Gayam, Gondang, Kedewan, Margomulyo, Ngambon, Sekar
Cluster 5	2	Kapas, Trucuk

Comparison of Grouping Results Using K-Means and K-Medoids Algorithms

Based on the results of clusterization using the K-Means and K-Medoids algorithms as shown in Table 1 and Table 2, a comparison of clusterization results is obtained as shown in Table 3 below.

Table 3. Clusterization Results Using K-Means and K-Medoids Algorithms

Cluster	K-Means	K-Medoids
1	Kedungadem	Kalitidu , Kanor, Kasiman, Malo, Ngasem, Ngraho, Padangan, Purwosari, Sugihwaras, Sukosewu, Tambakrejo, Temayang
2	Baureno, Dander, Kalitidu, Kepohbaru, Ngasem, Padangan, Sumberejo	Baureno, Dander, Kedungadem, Kepohbaru, Sumberejo
3	Bojonegoro	Bojonegoro
4	Balen, Bubulan, Gayam, Gondang, Kapas, Kasiman, Kedewan, Margomulyo, Ngambon, Ngraho, Sekar, Trucuk	Balen, Bubulan, Gayam, Gondang, Kedewan, Margomulyo, Ngambon, Sekar
5	Kanor, Malo, Purwosari, Sugihwaras, Sukosewu, Tambakrejo, Temayang	Kapas, Trucuk

Furthermore, to get the best clustering method, namely by using the average value within centroid distance. The best clustering method is the one with the smallest average value within centroid distance. The results of the comparison of K-Means and K-Medoids algorithms by looking at the average value within centroid distance as shown in Table 4 below.

Table 4. Comparison Results of K-Means and K-Medoids Algorithms

Grouping Method	Average Within Centroid Distance
K-Means	**1,833**
K-Medoids	2,786

Based on Table 4, it can be seen that the average value within centroid distance for the K-Means method is 1.833. Meanwhile, using the K-Medoids method, the average value within centroid distance was 2.786. This means that the value of average within centroid distance using the K-Means method is smaller than the value of average within centroid distance using the K-Medoids method. So it can be concluded that the clusterization method using the K-Means method is better for grouping sub-districts in Bojonegoro district based on educational supporting factors.

Mapping Clusterization Results with the Best Method

Based on Table 4, the best clustering method for grouping sub-districts in Bojonegoro based on supporting factors for education is K-Means method. To create a map of clusterization results with the best method based on clusterization results using the K-Means method as shown in Table 1. The results of clusterization are in the form of maps as shown in Figure 1 below.

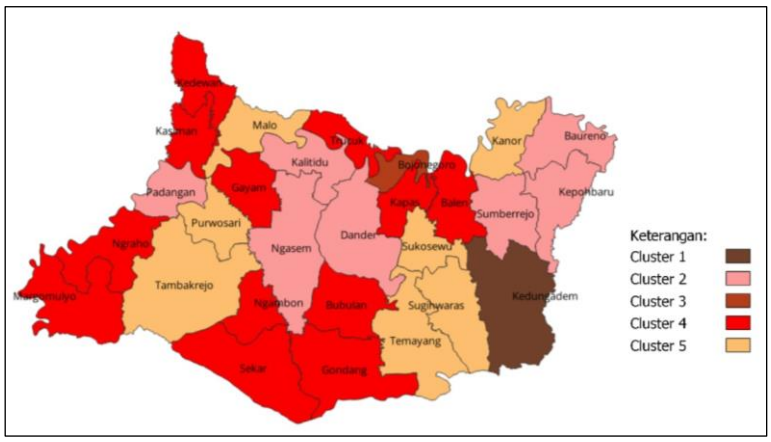


Figure 1. Map of Clusterization Results using the Best Method

Characteristics of Each Cluster Formed Based on the Best Method

To distinguish the results of clusterization formed, profiling is carried out by finding the average value of each variable. The clusterization results used are based on the results of clusterization using the K-Means method as shown in Table 1. The results of the profiling formed as shown in Table 5 below.

Table 5. Characteristics of Each Cluster Using the Best Method

Cluster	District	Characteristics
1	Kedungadem	This cluster has the highest number of elementary schools among other clusters
2	Baureno, Dander, Kalitidu, Kepohbaru, Ngasem, Padangan, Sumberejo	This cluster has the highest number of elementary school educators and the least number of junior high school educators. In addition, this cluster has the least number of junior high school learning groups
3	Bojonegoro	This cluster has the largest number of junior high schools, the largest number of high schools, the largest number of junior high school educators, the largest number of high school educators, the largest number of junior high school learning groups and the largest number of high school learning groups among other clusters
4	Balen, Bubulan, Gayam, Gondang, Kapas, Kasiman, Kedewan, Margomulyo, Ngambon, Ngraho, Sekar, Trucuk	This cluster has the least number of elementary schools, the least number of junior high schools, the least number of high schools and the least number of elementary educators compared to other clusters
5	Kanor, Malo, Purwosari, Sugihwaras, Sukosewu, Tambakrejo, Temayang	This cluster has the highest number of elementary school study groups among other clusters. In addition, this cluster also has the least number of high school educators

and the least number of high school learning groups compared to other clusters.

Based on the results, the characteristics of each cluster as shown in Table 5 can be used as input for the Bojonegoro district government for equal distribution of education in Bojonegoro. For example, cluster 2 has the highest number of elementary educators among other clusters, while cluster 4 has the least number of elementary educators among others. From these conditions, the number of elementary school educators in cluster 2 can be delegated and distributed to cluster 4. In addition, as another example, cluster 1 has the highest number of elementary schools among other clusters, while cluster 4 has the least number of elementary schools among other clusters. From these conditions, the government can reduce the number of elementary schools in cluster 1 and increase the number of elementary schools in cluster 4.

CONCLUSION

Based on the results of the study, it was found that the K-Means method is the best method to group sub-districts in Bojonegoro district based on educational supporting factors. The results of grouping using the K-Means method obtained 5 clusters. Cluster 1 consists of 1 sub-district, namely Kedungadem. Meanwhile, cluster 2 consists of 7 sub-districts, namely Baureno, Dander, Kalitidu, Kepohbaru, Ngasem, Padangan, and Sumberejo. For cluster 3, there is 1 sub-district, namely Bojonegoro. And for cluster 4 consisting of 12 sub-districts, namely Balen, Bubulan, Gayam, Gondang, Kapas, Kasiman, Kedewan, Margomulyo, Ngambon, Ngraho, Sekar and Trucuk. For cluster 5 there are 7 sub-districts, namely Kanor, Malo, Purwosari, Sugihwaras, Sukosewu, Tambakrejo, and Temayang. From the results of clusterization, profilization and characteristics of each cluster are obtained. It is hoped that these results can be used as input for the Education office as an effort to equalize education in Bojonegoro district.

REFERENCE

[1] I. Purwaningsih, O. Oktariani, L. Hernawati, R. Wardarita, and P. I. Utami, "Pendidikan sebagai Suatu Sistem," *J. Vision. Penelit. dan Pengemb. dibidang Adm. Pendidik.*, vol. 10, no. 1, pp. 21–26, 2022.

[2] P. S. Rosmana, S. Iskandar, N. Fadilah, N. Azhar, D. Oktavini, and A. C. Munte, "Upaya Pemerataan Pendidikan Berkelanjutan di Daerah 3T," *Attadib J. Elem. Educ.*, vol. 6, no. 2, pp. 405–418, 2023.

[3] K. Anwar, "Implementasi dan Relevansi Kebijakan Dalam Pemerataan Pendidikan: Studi Literatur Pelayanan Publik," *Coopetition J. Ilm. Manaj.*, vol. 13, no. 3, pp. 419–428, 2022.

[4] M. V. Roesminingsih, T. R. Hariastuti, and F. Agustina, "Perencanaan Peningkatan Mutu Sekolah di SMKN Purwosari Bojonegoro," *J. Pendidik. Tambusai*, vol. 6, no. 1, pp. 1892–1906, 2022.

[5] S. Satiti, "Gerakan Ayo Sekolah Di Kabupaten Bojonegoro: Peningkatan Sumber Daya Manusia Melalui Pendidikan Untuk Menyongsong Bonus Demografi," *J. Kependud. Indones.*, vol. 14, no. 1, pp. 77–92, 2019.

[6] H. Nurhuda, "Masalah-Masalah Pendidikan Nasional; Faktor-Faktor Dan Solusi Yang Ditawarkan," *Dirasah J. Pemikir. Dan Pendidik. Dasar Islam*, vol. 5, no. 2, pp. 127–137, 2022.

[7] A. A. Maulana, A. W. Rafii, Y. A. Anjelina, and E. Widodo, "Pengelompokkan Kecamatan di

Kabupaten Bima Berdasarkan Jumlah Produksi dan Luas Panen Bawang Merah Tahun 2021 Menggunakan K-Means Clustering," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 16, no. 1, pp. 442–451, 2023.

[8] F. Alfiah, A. Almadayani, D. Al Farizi, and E. Widodo, "Analisis Clustering K-Medoids Berdasarkan Indikator Kemiskinan di Jawa Timur Tahun 2020," *J. Ilm. Sains*, vol. 22, no. 1, pp. 1–7, 2022.

[9] J. Hutagalung, "Pemetaan Siswa Kelas Unggulan Menggunakan Algoritma K-Means Clustering," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 1, pp. 606–620, 2022.

[10] 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.

[11] A. Y. Kartini and N. Cahyani, "HYBRID K MEANS-MULTIVARIATE ADAPTIVE REGRESSION SPLINES FOR DISTRIBUTION OF DENGUE FEVER RISK MAPPING IN BOJONEGORO DISTRICT," *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 1, pp. 313–322, 2023.

[12] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci. (Ny).*, 2022.

[13] Y. Mayona, R. Buaton, and M. Simanjutak, "Data Mining Clustering Tingkat Kejahatan Dengan Metode Algoritma K-Means (Studi Kasus: Kejaksaan Negeri Binjai)," *J. Inform. Kaputama*, vol. 6, no. 3, pp. 2548–9739, 2022.

[14] L. Qadrini, "Metode K-Means dan DBSCAN pada Pengelompokan Data Dasar Kompetensi Laboratorium ITS Tahun 2017," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 13, no. 2, pp. 5–11, 2020.

[15] D. F. Pramesti, M. T. Furqon, and C. Dewi, "Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data Potensi Kebakaran Hutan/Lahan Berdasarkan Persebaran Titik Panas (Hotspot)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. e-ISSN*, vol. 2548, p. 964X, 2017.

[16] D. M. Sinaga, A. P. Windarto, and D. Hartama, "Analisis K-Medoids Dalam Pengelompokkan Rasio Murid dengan Guru , Murid dengan Rombel , dan Rasio Rombel dengan Kelas Jenjang Pendidikan SD dan SMP Menurut Provinsi," *J. Ris. Tek. Inform. dan Data Sci.*, vol. 1, no. 1, pp. 1–6, 2022.

[17] M. N. P. Pamulang, M. N. Aini, and U. Enri3, "Komparasi Distance Measure Pada K-Medoids Clustering untuk Pengelompokkan Penyakit ISPA," *Edumatic J. Pendidik. Inform.*, vol. 5, no. 1, pp. 99–107, 2021, doi: 10.29408/edumatic.v5i1.3359.

[18] R. N. H. Hutasuhut, H. Okprana, and B. E. Damanik, "Penerapan Data Mining Untuk Menentukan Penerima Program Bidikmisi Menggunakan Algoritma K-Medoids," *TIN Terap. Inform. Nusant.*, vol. 2, no. 11, pp. 667–672, 2022.

[19] R. Anjariansyah and A. Triayudi, "Clustering Kebutuhan Makanan untuk Meminimasi Standar Deviasi Angka Kebutuhan Gizi Menggunakan Algoritma K-Means dan K-Medoids," *J. MEDIA Inform. BUDIDARMA*, vol. 6, no. 1, pp. 597–607, 2022.

[20] D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat, "The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data," *Qual. Quant.*, vol. 56, no. 3, pp. 1283–1291, 2022.

[21] M. R. Nugroho, I. E. Hendrawan, and P. P. Purwantoro, "Penerapan Algoritma K-Means Untuk Klasterisasi Data Obat Pada Rumah Sakit ASRI," *NUANSA Inform.*, vol. 16, no. 1, pp. 125–133, 2022.

[22] Sanusi and J. Husna, "Utilization of Rapidminer using the K-Means Clustering Algorithm for

Classification of Dengue Hemorrhagic Fever (DHF) Spread in Banda Aceh City," *J. Inotera*, vol. 5, no. 2 SE-Articles, pp. 146–151, Oct. 2020, doi: 10.31572/inotera.Vol5.Iss2.2020.ID119.

[23] C. A. Sugianto, A. H. Rahayu, and A. Gusman, "Algoritma k-means untuk pengelompokkan penyakit pasien pada puskesmas cigugur tengah," *J. Inf. Technol.*, vol. 2, no. 2, pp. 39–44, 2020.

[24] S. Sindi, W. R. O. Ningse, I. A. Sihombing, F. Ilmi R.H.Zer, and D. Hartama, "Analisis algoritma K-Medoids clustering dalam pengelompokan penyebaran Covid-19 di Indonesia," *Jti (Jurnal Teknol. Informasi)*, vol. 4, no. 1, pp. 166–173, 2020.

[25] M. Herviany, S. P. Delima, T. Nurhidayah, and K. Kasini, "Perbandingan Algoritma K-Means dan K-Medoids untuk Pengelompokkan Daerah Rawan Tanah Longsor Pada Provinsi Jawa Barat: Comparison of K-Means and K-Medoids Algorithms for Grouping Landslide Prone Areas in West Java Province," *Malcom Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. 1, pp. 34–40, 2021.

[26] S. Ramadhani, D. Azzahra, and Z. Tomi, "Comparison of K-Means and K-Medoids Algorithms in Text Mining based on Davies Bouldin Index Testing for Classification of Student's Thesis," *Digit. Zo. J. Teknol. Inf. dan Komun.*, vol. 13, no. 1, pp. 24–33, 2022.

Fuzzy C-Means for Regional Clustering in East Java Province Based on Human Development Index Indicators

Marita Qori'atunnadyah [(1)]

[1]Program Studi Informatika, Institut Teknologi dan Bisnis Widya Gama Lumajang

Jalan Gatot Subroto No. 4 Lumajang

e-mail: maritaqori@gmail.com[(1)]

ABSTRAK

Indeks Pembangunan Manusia (IPM) adalah tolok ukur utama PBB untuk mengukur kemajuan manusia di suatu negara. IPM menggabungkan aspek penting dalam kehidupan manusia, termasuk pendapatan per kapita, harapan hidup, dan pendidikan. Di Indonesia, IPM digunakan untuk menilai kesejahteraan masyarakat, dan meskipun beberapa upaya telah dilakukan untuk meningkatkannya, IPM di Jawa Timur masih berada di bawah nilai rata-rata nasional dan target pemerintah. Untuk mengatasi masalah ini, penelitian ini menggunakan metode Fuzzy C-Means untuk mengelompokkan wilayah di Jawa Timur berdasarkan indikator IPM. Hasil penelitian menunjukkan bahwa ada lima kelompok yang optimal berdasarkan uji pseudo F-statistic. Hasil analisis One-Way MANOVA menunjukkan adanya variasi dalam karakteristik antara berbagai kelompok, sementara uji One-Way ANOVA mengonfirmasi bahwa keempat variabel indikator IPM berperan dalam pengelompokkan ini. Hasil pengelompokkan berdasarkan indikator IPM menunjukkan bahwa kelompok 3 berstatus tinggi, sementara kelompok 1 berstatus rendah. Kelompok 2 memiliki status cukup tinggi, kelompok 4 memiliki status sedang, dan kelompok 5 memiliki status cukup rendah. Oleh karena itu, dianjurkan kepada pemerintah untuk lebih berfokus pada upaya perbaikan kelompok dengan indikator IPM yang rendah, dengan tujuan meningkatkan kesejahteraan masyarakat di Jawa Timur. Penelitian ini dapat menjadi dasar bagi pemerintah dan pemangku kepentingan lainnya dalam merancang kebijakan untuk meningkatkan IPM di wilayah ini.

Kata kunci: *Fuzzy C-Means*; Indikator; IPM, Klaster

ABSTRACT

The Human Development Index (HDI) is the UN's key metric for gauging human advancement within a country, blending vital elements like per capita income, life expectancy, and education. In Indonesia, the HDI assesses societal well-being, with East Java's HDI lagging behind national and governmental targets despite mitigation efforts. To address this, the study utilizes Fuzzy C-Means clustering to classify East Java's regions based on HDI indicators, revealing five optimal groups via pseudo-F-statistic analysis. One-way MANOVA confirms variations among these groups, while One-Way ANOVA validates the significance of the four HDI indicators in categorization. The HDI-based categorization denotes Group 3 as high-status, Group 1 as low-status, Group 2 as moderately high-status, Group 4 as moderate, and Group 5 as moderately low-status. Consequently, it's advised that the government concentrates on improving low-HDI groups to uplift East Java's populace. This research can serve as a cornerstone for policymakers and stakeholders in their efforts to enhance the HDI in this region.

Keywords: *Cluster, Fuzzy C-Means, HDI, Indicator*

INTRODUCTION

The Human Development Index (HDI) is a parameter used by the United Nations to assess the social progress of humans in a country. The HDI combines several vital dimensions of human life, such as life expectancy, education, and per capita income [1]. HDI is used to measure the level of well-being of the population in a country, and one of the countries that uses HDI as a tool to gauge its social and economic development is Indonesia. In 1990, the Badan Pusat Statistik (BPS) introduced the Human Development Index (HDI) for the first time in Indonesia. HDI measures three primary dimensions: life expectancy, education, and per capita income. Life expectancy reflects the health and quality of life of the population and is calculated based on the average age of life expectancy at birth. Education comprises two components: mean years of schooling and expected years of schooling. Per capita income reflects the average income level of the population in a country. Improving HDI is usually a crucial indicator for governments in their efforts to enhance the well-being of the population.

The development of HDI in Indonesia has shown improvement over time, although challenges remain in enhancing access to quality education, achieving balanced development, and reducing social and economic inequalities. Efforts continue to be made to increase the HDI, which is a key factor in assessing human progress in Indonesia. In 2022, Indonesia's Human Development Index (HDI) reached a figure of 72.91, which is still below the target set by the government since 2018, ranging between 73.41 and 73.46 [2]. Out of the 38 provinces in Indonesia, East Java Province's HDI is still below the national average and has not yet reached the 2022 HDI target of 72.75 [3]. Therefore, it is necessary to establish groups of regencies/cities based on the HDI indicators in East Java to assist the government in designing effective policies to enhance the HDI in this region.

Cluster analysis is a multivariate analysis method used to group objects based on similar characteristics they possess. Within a specific cluster, there is a high level of similarity among objects, whereas the similarity between different clusters is low [4]. There are two types of clustering methods: hierarchy and non-hierarchy [5]. One example of a non-hierarchical clustering method is the fuzzy c-means and c-means method. C-means is an example of a non-hierarchical clustering method that partitions data into one or more clusters. This method partitions data by grouping similar data into one cluster, while data with different characteristics are placed in different clusters. The goal is to optimize a specified objective function in the clustering process, which aims to minimize the variation within a cluster and maximize the variation between clusters [6]. Fuzzy c-means is an extension of c-means that utilizes fuzzy weighting. Previous research has compared hierarchical and non-hierarchical clustering methods by simulating a dataset. The research results indicate that fuzzy c-means provide the best outcomes, especially in cases involving data with outliers and overlaps, compared to hierarchical clustering methods (such as single linkage, complete linkage, and average linkage), Self-Organizing Maps (SOM), and c-means [7].

Several previous studies related to clustering, such as regional clustering based on road conditions using K-Means [8], and regional clustering based on the teacher-student ratio at various educational levels using K-Means [9], [10]. The results of the previous research open up opportunities to develop methods for clustering teacher data, clustering based on undergraduate or non-graduate qualifications does not reflect the distribution of teachers well. There is a gap between provinces in eastern Indonesia and some outside Java in the ratio of undergraduate and between provinces in eastern Indonesia and some outside Java in the ratio of undergraduate to non-graduate teachers; and non-graduate teachers, provinces with student-teacher ratios and teacher-school ratios below the teacher-school ratios below the set standards. In addition, it shows a gap in the distribution of teachers between provinces in Java and outside Java. Thus, the use of the K-Means Clustering method in this study contributes to providing a better perspective of contributes to

providing a better perspective on the distribution and quality of teachers in Indonesia as well as providing a basis for recommendations for improvement.

Furthermore, research on regional classification based on HDI indicators has been conducted using the c-means method [11]. The commonality in all three studies is the adoption of the c-means method. Therefore, in this research, a different approach is applied, which is the fuzzy c-means method. The use of the K-Means method in this study resulted in the determination of the optimum number of groups of 4, with statistical test results showing significant differences between groups. The four variables used influence the average differences between groups, and the analysis shows the characteristics and indicators of the Human Development Index (HDI) that need improvement. The findings guide policies and interventions appropriate to the characteristics of each group to improve the quality of human development in these regions. Some previous studies that have used the fuzzy c-means method include the clustering of Junior High Schools in Indonesia based on the National Education Standards [12], clustering of Indonesian provinces based on indicators of the well-being of the population [13], clustering of regions based on health indicators [14], clustering of tax revenue types in Makassar City [15], and clustering of COVID-19 cases in Indragiri Hilir Regency [16]. From several studies presented, it can be concluded that the Fuzzy C-Means (FCM) method provides advantages in overcoming various complex data clustering problems. The advantage of FCM lies in its ability to handle data uncertainty and complexity by providing a finer degree of membership in each cluster. FCM is also flexible in handling variations and complex patterns, providing more accurate and informative clustering results to support decision-making.

This study extends the existing body of research on regional classification by employing the Fuzzy C-Means method to cluster East Java Province based on the Human Development Index (HDI) in 2022. In contrast to previous studies that utilized methods like K-means and c-means, the fuzzy c-means approach provides enhanced flexibility in addressing uncertainties and overlaps within the data, allowing for more nuanced cluster assignments. The researcher aims to determine optimal clusters and rigorously test the obtained results, contributing novel insights into the development characteristics of East Java's regencies/cities.

METHOD

This research utilized secondary data obtained from the source of the BPS-Statistics of Jawa Timur Province [17]. The data utilized encompassed 38 regencies and cities within that region and included various indicators within the Human Development Index (HDI). The HDI indicators used in this study involved Life Expectancy (LE), Expected Years of Schooling (EYS), Mean Years of Schooling (MYS), and adjusted per capita income.

A. Fuzzy C-Means

The fuzzy c-means clustering method reallocates data into groups using the concept of non-binary membership. In this method, membership function variables are employed, indicating the extent to which data can belong to a specific group. There is also a variable '*m*', referred to as the weighting exponent, which governs the extent to which data can be a member of a group. The value of '*m*' is typically greater than 1, with the standard value often being set at 2 [6].

In the fuzzy c-means method, the initial cluster centers are determined as the mean locations of each cluster. In this concept, each data point can belong to multiple clusters simultaneously, making the boundaries between clusters fuzzy. By iteratively refining the cluster centers and membership degrees through the fuzzy c-means algorithm, we can obtain cluster centers that approximate the appropriate locations. The iteration process is based on minimizing the objective

function in the given equation. The value of 'm' used in this algorithm influences the extent to which the membership degrees of data in each cluster can change during the iteration. The iteration in this algorithm is based on minimizing the objective function in the equation (1).

$$J\left(U, c_1, \ldots, c_g\right) = \sum_{c=1}^{g} J_c = \sum_{c=1}^{g} \sum_{j}^{n} u_{ci}^{m} d_{ci}^{2} \quad (1)$$

Notation :

u_{ci} : Membership degree of object *i* to cluster *c*

c_g : The matrix of centroids for all clusters

n : Number of data points

c : Number of clusters

m : Weighting exponent

$d_{ci} = \| c_g - x_i \|$: Euclidean distance between cluster *c* and cluster center *i*

This objective function reflects the distance between the given data point and the cluster center, weighted by the membership degree of that data point. The membership degree of a data point in a particular cluster can be calculated using the following equation (2).

$$u_{ci} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ci}}{d_{ki}} \right)^{2/(m-1)}} \quad (2)$$

Notation :

u_{ci} : Membership degree of data point *i* to cluster *c*

d_{ci} : Centroid value of data point *i* in cluster *c*

d_{ki} : Centroid value of data point *i* in cluster *k*

c : Number of clusters

m : Weighting exponent

The membership function has a range of values between $0 \leq u_{ci} \leq 1$. To calculate the cluster centers, the following equation (3) can be used.

$$c_i = \frac{\sum_{i=1}^{n} u_{ci}^{m} x_i}{\sum_{i=1}^{n} u_{ci}^{m}} \quad (3)$$

where x_i represents the object or data point *i*.

The algorithm applied in the fuzzy c-means method to identify cluster centers c_i *with the membership matrix* **U** is as follows:

1. Determine the number of clusters or groups to be formed (c).
2. Choose the value of the weighting exponent ($m > 1$), with the common value being2.
3. Specify the tolerance threshold or stopping criteria for iterations.
4. Create the initial partition matrix **U** (membership degree matrix). Matrix U is filled with random numbers between 0 and 1 based on the following equation (4):

$$U = \begin{bmatrix} \mu_{11}(x_1) & \mu_{12}(x_2) & \cdots & \mu_{1i}(x_i) \\ \mu_{21}(x_1) & \mu_{22}(x_2) & \cdots & \mu_{2i}(x_i) \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{c1}(x_1) & \mu_{c2}(x_2) & \cdots & \mu_{ci}(x_j) \end{bmatrix} \quad (4)$$

5. Calculate the fuzzy centroids c_i , $i = 1$,..., c using Equation (3) and create a new matrix for centroid values to calculate the objective function.
6. Compute the fuzzy c-means objective function value based on Equation (1). The objective function value is used to determine whether the iteration continues or stops. Iteration stops if the objective function value falls below the tolerance threshold.
7. If the objective function value is still above the threshold, update the calculation of the membership matrix or partition matrix **U** based on Equation (2), and repeat these steps starting from step 5. This matrix is used to determine the groups of each observation after the iteration stops.

B. Pseudo F-Statistic

The pseudo-F-statistic was first introduced by Calinski and Harabasz. This statistic is used to measure the validity of clustering to determine the optimum or best number of clusters [18]. The highest pseudo-F-statistic reflects an optimal clustering outcome, where the similarity within a cluster is high, while the differences between clusters are significant [19]. Here is the equation (5) used to calculate the pseudo F-statistic.

$$Pseudo\ F-statistic = \frac{SSB/(k-1)}{SSW/(n-k)} \quad (5)$$

Notation: :

SSB = variation between clusters, also known as sum of squares between.
SSW = variation within clusters, also known as sum of squares within.
k = the number of clusters generated by the clustering algorithm.
n = the number of data points.

C. One Way MANOVA and One Way ANOVA

The use of One-Way MANOVA is to compare means of two or more populations when there are multiple dependent variables or to assess the effects of a treatment on responses [5]. MANOVA is used to evaluate the similarity among the formed groups. Before conducting the One-Way MANOVA test, it is important to undergo multivariate normality testing and homogeneity testing. Multivariate normality testing is an extension of univariate normality testing that involves at least two observed variables. In multivariate analysis, multivariate normality testing is necessary to ensure that the observed data follows a multivariate normal distribution [20]. Furthermore, to check the homogeneity of covariance matrices, Box's M test is used. One-way ANOVA, on the other hand, is used to test differences between groups when only one dependent variable is used or to test differences between variables among group members [5].

RESULT AND DISCUSSION

In this study, a regional cluster analysis in East Java Province based on the 2022 Human Development Index indicators was conducted using the fuzzy c-means method. There are 38

regencies and cities in East Java that will be grouped into several clusters. Clustering was performed by trying different numbers of clusters, ranging from 2 to 5 clusters, and the most optimal cluster configuration was chosen using R software. The results of the clustering for each cluster are displayed in Table 1.

Table 1. Clustering using the Fuzzy C-Means Method

Cluster	Number of Clusters			
	2	**3**	**4**	**5**
1	26	18	12	7
2	12	16	9	7
3		3	3	3
4			14	11
5				10

Table 1 presents data on the number of members (regencies/cities) in each cluster resulting from the clustering of 38 regencies/cities in East Java Province. Clustering was performed using the fuzzy c-means method with varying numbers of clusters ranging from 2 to 5. The selection of the optimal number of clusters can be determined based on the highest pseudo F-statistic value among these variations. Here are the pseudo F-statistic values associated with each cluster.

Table 2. Pseudo F-Statistic Values of the Fuzzy C-Means Method

Number of Clusters	*Psudo F-Statistic*
2	64,80039
3	95,53214
4	144,1018
5	**170,1298**

Table 2 shows the calculation of pseudo F-statistics using the fuzzy c-means method for the number of clusters ranging from 2 to 5. The optimal number of clusters for grouping regencies/cities in East Java based on the Human Development Index indicators is 5 clusters. This can be observed from the highest pseudo F-statistic value, which is 170.1298, found in the clustering with 5 clusters. The fuzzy c-means clustering process involves several initial steps, including determining the desired number of clusters (in this example, 5 clusters), setting the initial weighting exponent (m) to 2, and setting the tolerance threshold or iteration stopping criterion to approximately 10^{-6}. Next, the initial partition matrix U is created using random numbers between 0 and 1, with the number of rows corresponding to the number of regencies/cities in East Java Province (38 rows) and the number of columns corresponding to the number of clusters (5 columns) to be formed.

The next step is to determine the centroid (center point) for each cluster and calculate the objective function value. The objective function value is used to determine whether the iteration continues or stops based on a comparison with the stopping criterion value. If the objective function value is greater than the tolerance threshold of approximately 10^{-6}, then the calculation of the new partition matrix U (membership function) is performed. After that, the centroid values and objective function are recalculated until the objective function value reaches less than 10^{-6}. When the

objective function value reaches this threshold, the iteration is stopped, and the membership for each of the five clusters is determined based on the membership degree matrix in the last iteration. For example, if the first regency has the highest membership degree in cluster three, it will be a member of cluster three, and the same applies to the other regencies up to the 38th regency. The results of clustering using 5 clusters can be seen in Table 3.

Table 3. List of Regencies/Cities in 5 Clusters

Cluster	**Regencies/Cities**			
1	Pacitan Bangkalan	Lumajang Sampang	Jember Pamekasan	Sumenep
2	Sidoarjo Blitar City	Mojokerto Kota Batu	Gresik Mojokerto City	Pasuruan City
3	Malang City	Madiun City	Surabaya City	
4	Kediri Jombang Magetan	Banyuwangi Nganjuk Ngawi	Probolinggo Madiun Lamongan	Probolinggo City Kediri City
5	Ponorogo Malang Situbondo	Trenggalek Bondowoso Blitar	Tulungagung Pasuruan	Tuban Bojonegoro

Table 3 contains the members of each cluster. In clustering the regencies/cities in East Java Province based on the Human Development Index indicators using the fuzzy c-means method, it is expected that there are differences in characteristics within each group related to all Human Development Index indicators. To assess whether there are characteristic differences in the formed groups, this can be done through one-way MANOVA and one-way ANOVA methods. Before conducting these tests, the initial steps are to test whether the data is normally distributed in multivariate form and to test the homogeneity of variances among groups.

The multivariate normality distribution test is used to evaluate whether the data follows a multivariate normal distribution or not. The result of the multivariate normality distribution test shows a correlation value of 0.986 as indicated above.

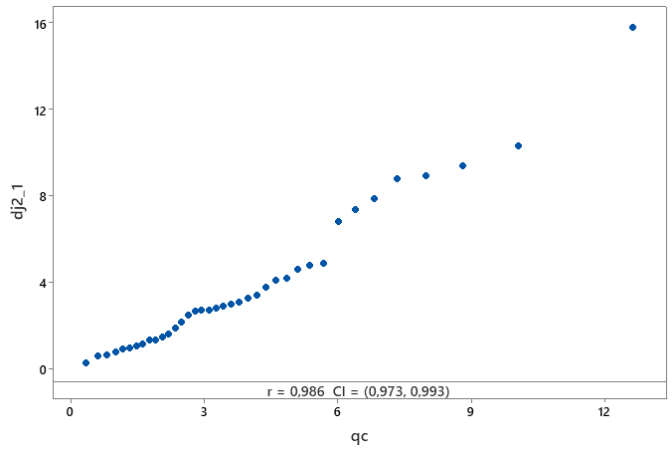


Figure 1. Multivariate Normal Probability Plot

These values will be compared with the critical point from the table of the probability plot of the correlation coefficient for normality (PPCC). The critical point obtained at a 5% significance level is 0.9700. The result indicates that there is not enough evidence to reject H0, which means that the data follows a multivariate normal distribution. To test the homogeneity of the covariance matrix, Box's M test is used at a 5% significance level. The test result shows that the value of Box's M is 46.847.

Table 4. Box's M Test Result

	Value
Box's M	46,847
F	1,155
df1	30
df2	1884,898
Sig.	0,258

At a significance level of 5% (alpha 0.05) and with 40 degrees of freedom, the value is obtained as 55,758. The found Box's M value is smaller than the expected critical value, which is $\chi^2_{\frac{1}{2}(g-1)p(p+1)}$.

Furthermore, considering the significance value of 0.258, which exceeds the significance level alpha (0.05), the conclusion is that H_0 fails to be rejected. Therefore, it can be concluded that the covariance matrix is homogeneous.

After undergoing multivariate normality testing and checking for data homogeneity, the results indicate that the data follows a multivariate normal distribution and has a homogeneous covariance matrix. Therefore, the test for differences in characteristics using One-Way MANOVA adopts Pillai's Trace test statistic. In this analysis, we want to assess the factors or treatments that are suspected to have a significant impact on the response variable, which in this case is the formed clusters. The response variable in the One-Way MANOVA test is the Human Development Index indicators.

Table 5. One-Way MANOVA Test Results

Pillai's Trace Value	*F*	Degrees of Freedom for Hypothesis	Degrees of Freedom for Error	Sig.	*Partial Eta Squared*
1,243	3,718	16	132	0,000	0,311

Based on TABLE 5, it was found that the One-Way MANOVA test results showed a Pillai's Trace test statistic value of $F = 3.718$. The critical value $F_{64;528;0.05}$, which is the significance threshold, is 1.334. When comparing these two values, the found F value is greater than the critical value $F_{64;528;0.05}$, resulting in the rejection of the null hypothesis (H_0). This means there is a significant difference among the formed clusters.

One-way ANOVA testing is used to evaluate differences among variables among group members. The results of the One-Way ANOVA test are as follows.

Table 6. One-Way ANOVA Test Results

Variable	F	Sig
Life Expectancy	3,552	0,016
Expected Years of Schooling	8,413	0,000
Mean Years of Schooling	24,843	0,000
Adjusted Gross National Income per Capita	170,132	0,000

From Table 6, the F-values for each variable were found. These values will be compared to the $F_{4;33;0,05}$ value, which is equivalent to 2.659. When comparing the F-values of the variables with the $F_{4;33;0,05}$ value, it is evident that all four variables have higher F-values. Therefore, the null hypothesis (H_0) is rejected, indicating that there are significant differences in the characteristics of the four variables among the formed clusters. This suggests that these four variables have a significant impact on the cluster formation.

By applying the fuzzy c-means method, the districts/cities in East Java Province were successfully grouped into 5 clusters based on the IPM indicators. The results of the one-way MANOVA test indicate significant differences among the five formed clusters, and the four variables have varying effects on these cluster differences, as revealed through the one-way ANOVA results. Here is a description of each of the formed clusters.

Table 7. List of Regencies/Cities in 5 Clusters

Cluster	**1**	**2**	**3**	**4**	**5**
n	7	7	3	11	10
Life Expectancy	70,24	73,32	73,78	72,02	72,05
Expected Years of Schooling	12,80	14,07	15,01	13,56	12,99
Mean Years of Schooling	6,43	10,03	10,96	8,29	7,49
Adjusted Gross National Income per Capita	9251,43	13750,29	17248,33	11953,64	10559,60

Based on Table 7, you can observe the average characteristics of each cluster formed from the 2022 IPM indicator data in East Java Province. It's noted that Cluster 3 has the highest average values among the groups for each indicator. This indicates that Cluster 3 has a high IPM indicator, reflecting good quality in those indicators. Cluster 2 has the second-highest average values for each indicator after Cluster 3, but it's important to notice that there's a significant difference in the per capita expenditure indicator when compared to Cluster 3. Therefore, improvement is needed in this indicator. Following Cluster 2, the next-highest averages are found in Cluster 4, although the life expectancy indicator has a slightly lower average than Cluster 5. Hence, improvement should focus on the life expectancy indicator. Cluster 1 has the lowest average values for all indicators, thus requiring comprehensive improvements in all indicators. Meanwhile, Cluster 5 has the second-lowest averages, particularly in the indicators of expected years of schooling, mean years of schooling, and per capita expenditure. Although their averages are not significantly different from Cluster 1, improvements should be focused on these three indicators. Additionally, when examining the number of districts/cities in each group, Cluster 4 and 5 have a large number of members, even though their average indicators fall into the "fair" category. Nonetheless, more targeted improvements are still needed in both of these groups. Based on the average values of each indicator, the ranking status can be assigned to the groups of districts/cities as follows.

Table 8. Status of Each Cluster

Cluster	Status
1	Low HDI Indicators
2	Moderately High HDI Indicators
3	High HDI Indicators
4	Moderate HDI Indicators
5	Moderately Low HDI Indicators

CONCLUSION

Based on the results and discussion, the regional clustering in East Java based on the Human Development Index (HDI) indicators using the fuzzy c-means method yielded an optimal clustering result of 5 clusters. This result was obtained based on the largest pseudo-F-statistic value. In the one-way MANOVA test using Pillai's Trace statistic, significant differences were observed among the formed clusters. Additionally, the one-way ANOVA test indicated that all four variables had a significant impact on the differences in characteristics among these clusters. Cluster 3 showed high HDI indicators. Cluster 2, while having moderately high HDI indicators, requires adjustment, particularly in the per capita expenditure indicator. Cluster 4 exhibited moderate HDI indicators, thus necessitating improvements in life expectancy indicators. Cluster 5 displayed moderately low HDI indicators, requiring more focused efforts to enhance indicators related to expected years of schooling, mean years of schooling, and per capita expenditure. Cluster 1 represented the category with low HDI indicators, necessitating comprehensive improvements across all indicators. The Human Development Index (HDI) is a crucial parameter in assessing the success of efforts to improve the quality of human life. Therefore, the recommendation to the government is to concentrate improvement efforts on indicators with low values within each cluster. Other clustering methods like Revised Fuzzy C-Means (RFCM) can be applied to data with uneven cluster sizes and contamination from noise and outliers, as demonstrated in the study [21].

REFERENCE

[1] Badan Pusat Statistik, "Statistik Indonesia," Jakarta, 2019.
[2] Kementerian PPN/Bappenas, "Pemuktahiran Rencana Kerja Pemerintah (RKP) Tahun 2022," 2021. Accessed: Mar. 29, 2023. [Online]. Available: https://www.bappenas.go.id/show-result-satudata?name=publikasi&key=rkp&tahun=
[3] Badan Pusat Statistik, " Indeks Pembangunan Manusia menurut Provinsi 2020-2022," 2023.
[4] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, "MULTIVARIATE DATA ANALYSIS EIGHTH EDITION," 2019. [Online]. Available: www.cengage.com/highered
[5] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis 6th Edition*, 6th ed. United States of America: Pearson Prentice Hall, 2007.
[6] Y. Agusta, "K-Means-Penerapan, Permasalahan dan Metode Terkait," 2007.
[7] S. A. Mingoti and J. O. Lima, "Comparing SOM neural network with Fuzzy c-means, K-means, and traditional hierarchical clustering algorithms," *Eur J Oper Res*, vol. 174, no. 3, pp. 1742–1759, Nov. 2006, doi: 10.1016/j.ejor.2005.03.039.

[8] M. Qori'atunnadyah and F. D. Rahmawati, "Pengelompokkan Kabupaten dan Kota Berdasarkan Kondisi Infrastruktur Jalan Menggunakan Hierarchical Clustering," *Journal of Informatics Development*, vol. 1, no. 1, pp. 1–5, 2022.
[9] M. Qori'atunnadyah, "Pengelompokkan Wilayah Berdasarkan Rasio Guru-Murid Pada Jenjang Pendidikan Menggunakan Algoritma K-Means," *Journal of Informatics Development*, vol. 1, no. 2, pp. 33–38, 2022.
[10] F. Idris, F. Azmi, and P. S. Daru Kusuma, "PENGELOMPOKAN DATA GURU DI INDONESIA MENGGUNAKAN K-MEANS CLUSTERING TEACHER DATA GROUPING IN INDONESIA USING K-MEANS CLUSTERING," *eProceedings of Engineering*, vol. 6, no. 2, pp. 5648–5653, 2019.
[11] M. Qori'atunnadyah, "Metode C-Means untuk Pengelompokkan Kabupaten/Kota Provinsi Jawa Timur berdasarkan Indikator Indeks Pembangunan Manusia (IPM)," *Journal of Informatics Development*, vol. 1, no. 2, pp. 51–58, 2023, doi: 10.30741/jid.v2i2.1013.
[12] H. A. Chusna and A. T. Rumiati, "Penerapan Metode K-Means dan Fuzzy C-Means untuk Pengelompoan Sekolah Menengah Pertama (SMP) di Indonesia Berdasarkan Standar Nasional Pendidikan (SNP)," *Jurnal Sains dan Seni ITS*, vol. 9, no. 2, Feb. 2021, doi: 10.12962/j23373520.v9i2.58349.
[13] N. Dwitiyanti, N. Selvia, and F. R. Andrari, "Penerapan Fuzzy C-Means Cluster dalam Pengelompokkan Provinsi Indonesia Menurut Indikator Kesejahteraan Rakyat," *Faktor Exacta*, vol. 12, no. 3, p. 201, Nov. 2019, doi: 10.30998/faktorexacta.v12i3.4526.
[14] G. S. Nugraha and B. A. Riyandari, "IMPLEMENTASI FUZZY C-MEANS UNTUK PENGELOMPOKAN DAERAH BERDASARKAN INDIKATOR KESEHATAN," *Jurnal Teknologi Informasi*, vol. 4, no. 1, pp. 52–62, Jun. 2020, doi: 10.36294/jurti.v4i1.1222.
[15] I. Irwan, S. Sidjara, and A. P. Aryati, "Pengelompokan Jenis Penerimaan Pajak di Kota Makassar Menggunakan Fuzzy Clustering," *Euler : Jurnal Ilmiah Matematika, Sains dan Teknologi*, vol. 10, no. 1, pp. 98–102, May 2022, doi: 10.34312/euler.v10i1.14225.
[16] S. F. Octavia and M. Mustakim, "Penerapan K-Means dan Fuzzy C-Means untuk Pengelompokan Data Kasus Covid-19 di Kabupaten Indragiri Hilir," *Building of Informatics, Technology and Science (BITS)*, vol. 3, no. 2, pp. 88–94, Sep. 2021, doi: 10.47065/bits.v3i2.1005.
[17] BPS-Statistics of Jawa Timur Province, "Provinsi Jawa Timur Dalam Angka 2023," Surabaya, 2023.
[18] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Commun Stat Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.
[19] T. Hochin *et al.*, "Quasi-optimality under pseudo F statistic in clustering data Quasi-optimality under pseudo F statistic in clustering data Quasi-optimality under pseudo f statistic in clustering data," 2018. [Online]. Available: www.sciencepubco.com/index.php/IJET
[20] A. C. Rencher and W. F. Christensen, *Methods of Multivariate Analysis*, 3rd ed. Wiley, 2012.
[21] S. Askari, "Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development," *Expert Syst Appl*, vol. 165, p. 113856, Mar. 2021, doi: 10.1016/j.eswa.2020.113856.

Investigating the Impact of Mobile Legends Gameplay on Students' Academic Performance with Ordinal Logistic Regression

Muhamad Irawan [(1)], Nurul Fitriyani [(2,*)], I Gede Adhitya Wisnu Wardhana [(1)], Irwansyah [(1)], Zulhan Widya Baskara [(2)]

[1] Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Mataram, Jl. Majapahit No.62, Mataram, Nusa Tenggara Barat. 82115
[2] Department of Statistics, Faculty of Mathematics and Natural Sciences, University of Mataram, Jl. Majapahit No.62, Mataram, Nusa Tenggara Barat. 82115
e-mail: [(*)]nurul.fitriyani@unram.ac.id

ABSTRAK

Perkembangan teknologi informasi dan game online seperti Mobile Legends Bang Bang telah merambah ke berbagai lapisan masyarakat di Indonesia, mulai dari anak-anak, pelajar, hingga mahasiswa. Meskipun pemerintah telah mendukung industri e-sports, penelitian mengenai pengaruh minat bermain Mobile Legends terhadap prestasi akademik siswa masih terbatas. Oleh karena itu, penelitian ini bertujuan untuk mengidentifikasi dampak minat bermain Mobile Legends dan faktor signifikan yang mempengaruhi prestasi akademik siswa. Kami menggunakan metode regresi logistik ordinal dalam analisis kami, suatu teknik statistik untuk mengukur hubungan antara variabel independen dan variabel dependen ordinal, seperti tingkat kinerja akademik yang dikategorikan dalam IPK rendah, sedang, atau tinggi. Analisis kami menghasilkan dua model: Indeks Prestasi Kumulatif (IPK) rendah dan IPK sedang. Faktor yang signifikan adalah tingkat minat yang meliputi kategori 'sangat berminat' dan 'tertarik', serta jenis kelamin dengan kategori 'laki-laki'. Analisis kami juga menunjukkan bahwa model yang diperoleh memberikan hasil yang baik dan dapat diterima karena semua variabel penjelas signifikan secara statistik.
Kata Kunci : *Mobile Legends, Game Online, Regresi Logistik Ordinal*

ABSTRACT

The development of information technology and online games, such as Mobile Legends Bang Bang, has spread to various segments of society in Indonesia, including children, students, and university students. Although the government has supported the e-sports industry, research on the influence of interest in playing Mobile Legends on students' academic performance is still limited. Therefore, this research aims to identify the impact of interest in playing Mobile Legends and the significant factors affecting students' academic performance. We used the ordinal logistic regression method in our analysis, a statistical technique to measure the relationship between independent variables and ordinal dependent variables, such as academic performance levels categorized as low, moderate, or high GPA. Our analysis results in two models: low Cumulative Grade Point Average (GPA) and moderate GPA. The significant factors are the level of interest, including the 'very interested' and 'interested' categories, and Gender with the category 'male.' Our analysis also indicates that the obtained model provides good results and is acceptable since all the explanatory variables are statistically significant.
Keywords: *Mobile Legends, Online Games, Ordinal Logistic Regression*

INTRODUCTION

Technology has undeniably become indispensable to daily life, permeating work, households, education, and entertainment. Devices like smartphones, tablets, and personal computers are now essential tools, particularly for students engaging in online gaming. These games, stemming from the development of information technology, have evolved beyond personal computers, with smartphones and other advanced gadgets offering access to a diverse range of online games. Among these games, genres like role-playing games (RPG), puzzles, and multiplayer online battle arenas (MOBA) have gained immense popularity. MOBA games involve multiple players divided into teams striving to secure victory by dismantling their opponents' unique structures. This game has also transcended mere entertainment and has become the subject of various competitions both at the national and international levels. Indonesian players also have a good reputation in this game, as evidenced by the many Indonesian players on the Leaderboard for Ranked games. Moonton, the developer of Mobile Legends Bang Bang, has honored Indonesian players by creating a hero named "*Gatot Kaca*" as a special tribute (Playstore, 2022).

Government support has played a pivotal role in driving the growth of Indonesia's esports industry, particularly within the community of Mobile Legends enthusiasts. What was once merely a form of entertainment has now transformed into a substantial source of income for numerous individuals. However, many problems have arisen, such as students becoming more focused on playing games than on their studies and the emergence of addictive behaviors. These issues underscore the urgency of examining how playing Mobile Legends impacts students' academic performance. Therefore, it is essential to know how much interest in playing mobile legends games influences students' academic achievement (Iskandar et al., 2019).

There are many statistical methods used to see the influence of an object on other objects, one of which is regression analysis. Various statistical methods, such as ordinal logistic Regression, can help understand the extent of this influence. Using ordinal logistic Regression allows the researchers to understand how the level of interest in Mobile Legends correlates with different academic performance levels, and it provides insights into the strength and direction of this relationship. This method is appropriate for analyzing such relationships when independent and dependent variables are categorical or ordinal, making it a suitable choice for this research. Logistic Regression is a statistical analysis method describing the relationship between a dependent variable with two or more categories and one or more independent variables on a categorical or continuous scale. Logistic Regression can be divided into binary logistic Regression, multinomial logistic Regression and ordinal logistic Regression (Hosmer and Lemeshow, 2000; Supratno, 2000; Gujarati, 2007; Imaslihkah, 2013; Fitriyani et al., 2016; Sugiyono, 2018).

Ordinal logistic Regression is an analytical method used to determine the relationship between dependent and independent variables, where the dependent variable is on an ordinal scale consisting of three or more categories and the measurement scale is level. The advantage of ordinal logistic Regression over logistic Regression is that the dependent variable is multilevel categorical scale data. Moreover, this Regression type can be considered a generalization of either multiple linear Regression or binomial logistic Regression. Like other forms of Regression, ordinal Regression can predict dependent variables through interactions between independent variables. This requires that the dependent variable data scale be ordinal and the independent variable data scale may be categorical or quantitative. Ordinals have different degrees in each category, with better or worse and high or low (Hosmer and Lemeshow, 2000; Akbar et al., 2010).

In this research, we will see how influential a person's interest in playing mobile legends games is on academic achievement, which means that the dependent variable used is categorical data and a multilevel category scale, namely, no influence, influence and very influence. This analytical approach is chosen for its suitability for assessing the relationship between the level of interest in Mobile Legends and the different academic performance levels, categorized as low, moderate, or high GPA. Therefore, this study aims to determine the model of the influence of interest in playing mobile legends games on academic achievement and also to determine what factors significantly influence academic achievement using the ordinal logistic regression method.

RESEARCH METHODOLOGY

The data used in this research are primary data obtained through online questionnaire surveys. The main objective of this study is to develop a model that can depict the influence of interest in playing Mobile Legends on the academic performance of students enrolled in the Mathematics Program at the Faculty of Mathematics and Natural Sciences, University of Mataram. Additionally, this research aims to identify the factors that significantly impact this relationship.

To achieve these objectives, this research involves several stages of analysis as follows:

1. Literature Review
   The literature review is the first step in conducting research, to expand references on the subject to be examined, thereby providing a clear foundation and guidance for the study.
2. Collecting Temporary Data
   Temporary data is collected to test whether the questionnaire created aligns with the researcher's intentions.
3. Validity and Reliability Testing
   Validity and reliability testing is used to verify whether the questions in the questionnaire distributed are valid and reliable. Validity and reliability testing can be conducted using the following equations (Simamora, 2004):
   a. Validity testing
   $$r_k = \frac{n(\sum xy) - (\sum x \sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \quad (1)$$
   b. Reliability testing
   $$r = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum \sigma^2{}_b}{\sigma^2{}_t}\right) \quad (2)$$
4. Collecting Sample Data
   The collected data matches the predetermined sample size through a survey using online questionnaires (Hosmer & Lemeshow, 2000).
5. Ordinal Logistic Regression
   This research will use the ordinal logistic regression method to analyze the available data. The ordinal logistic regression model to be employed is as follows (Hosmer & Lemeshow, 2000):
   $$\pi_r(x_i) = \frac{\exp(\beta_{0r} + x_i{}^T\beta)}{1 + \exp(\beta_{0r} + x_i{}^T\beta)};\ r = 1,2,\dots,s \quad (3)$$
6. Parameter Estimation

Parameter estimation in the ordinal logistic regression model uses the Maximum Likelihood Estimation (MLE) method. The MLE method is commonly used for estimating parameters in the ordinal logistic regression model (Hosmer & Lemeshow, 2000; Gujarati, 2007).

7. Parameter significance testing

   Parameter significance testing is conducted collectively using the G-test statistic and individually using the Wald test statistic (Imaslihkah et al., 2013):

   a. Simultaneous test

      For simultaneous testing, it can be conducted using the following formula:

$$G^2 = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}\left(\frac{n_3}{n}\right)^{n_3}}{\prod_{i=1}^{n}[\pi_1(x_i)^{y_{1i}}\pi_2(x_i)^{y_{2i}}\pi_3(x_i)^{y_{3i}}]}\right] \quad (4)$$

   b. Partial Test

      For partial testing, it can be conducted using the following formula:

$$W = \frac{\widehat{\beta}_k}{SE(\widehat{\beta}_k)} \quad (5)$$

8. Model Suitability Test

   The suitability of the ordinal logistic regression model is evaluated through the coefficient of determination (R^2) test. This test aims to measure the extent of the influence of variables on the ordinal logistic regression model. To conduct this test, the following formula can be used (Ghozali, 2018):

$$R^2 = r^2 \times 100\ \% \quad ; r^2 = 1 - e^{2\left(ln(L_0) - ln(L_M)\right)/n} \quad (6)$$

RESULTS AND DISCUSSION

1. Validation and Reliability Testing

The accuracy of the data used determines the quality of research because data represents the variables under study and serves as evidence in testing hypotheses. A good instrument must meet essential requirements, namely validity and reliability.

a. Validity Test

Validity indicates the extent to which an instrument is valid or reliable. The product-moment correlation formula is used to test the validity level. If the value of r_k is greater than or equal to the value of r_{table}, the instrument is considered valid. However, if the value of r_k is less than the value of r_{table}, the instrument is considered invalid. The value of r_k obtained from 34 respondents is presented in the following Table 1.

Table 1. Results of the Validity Test

No.	r_k	$r_{table(\frac{\alpha}{2},n)}$	Description
1	0,843	0,338	Valid
2	0,412	0,338	Valid
3	0,553	0,338	Valid
4	0,913	0,338	Valid
5	0,830	0,338	Valid

Based on Table 1, all items or questions are valid, so all items/questions can be used to obtain data.

b. Reliability Test

Reliability refers to the questionnaire's consistency and dependability in measuring something over time. The Cronbach's Alpha formula can be used to test the instrument's reliability. Data is considered reliable if the Cronbach's Alpha value exceeds 0.6. Based on the testing results from 34 respondents, a Cronbach's Alpha value of 0.781 was obtained, indicating that it is greater than 0.6. Therefore, it can be concluded that the questionnaire is reliable and can be used for decision-making in this research.

2. Population and Sample

In this research, the data used is in the form of primary data obtained through a survey. Data is collected through an online questionnaire and supports the research being conducted. The population in this study consists of students in the Mathematics Program at the Faculty of Mathematics and Natural Sciences, University of Mataram. Purposive sampling was chosen as the sampling method to obtain a sample representing the population. The criteria used in this case are students in the Mathematics Program at the Faculty of Mathematics and Natural Sciences, University of Mataram, who are actively enrolled and have a GPA. The proportion used is the number of students who play Mobile Legends. Seventy-nine students play Mobile Legends, and the total number of active students in the Mathematics Program at the Faculty of Mathematics and Natural Sciences, University of Mataram, is 380 students. To determine the sample size for the research, you can use the following formula:

$$n = \frac{\left(Z_{1-\frac{\alpha}{2}}\right)^2 p\,(1-p)N}{d^2(N-1) + \left(Z_{1-\frac{\alpha}{2}}\right)^2 p\,(1-p)} = \frac{(1{,}96)^2\,(0{,}208)\,(1-0{,}208)(380)}{(0{,}05)^2(380-1) + (1{,}96)^2\,(0{,}208)\,(1-0{,}208)}$$

$$= 152{,}1707 \approx 153$$

Based on the sample size calculation above, the minimum required sample size is 153 samples.

3. Categorization of GPA Data

The GPA variable (Y) was obtained through a questionnaire, with the lowest GPA being 1.89 and the highest GPA being 3.54. Then, steps were taken to categorize the research data, such as calculating the ideal mean and standard deviation, and determining the raw scores for the interval class levels of the research variable. The result is the GPA tendency distribution table as follows:

Table 2. Categorization of GPA Distribution

No.	Score	Frequency		Descriptions
		Frequency	%	
1	$X \geq 3{,}00$	56	35	High GPA
2	$2{,}45 \leq X < 3{,}00$	55	34,375	Moderate GPA

3	$X < 2{,}45$	49	30,625	Low GPA
	Total	160	100	

Based on Table 2, the percentage of GPA values for students in the Mathematics Program, out of a total sample of 160 students, shows that 56 students (35%) fall into the high GPA category, 55 students (34.375%) fall into the moderate GPA category, and 49 students (30.625%) fall into the low GPA category.

4. Categorization of Interest Data

The interest variable (X_1) is measured through a questionnaire consisting of 5 items/questions. From the available questions, the highest score is 12, and the lowest is 5. Subsequently, steps were taken to categorize the research data, such as calculating the ideal mean and standard deviation and determining the raw scores for the interval class levels of the research variable. The result is the interest tendency distribution table as follows:

Table 3. Categorization of Interest Distribution

No.	Score	Frequency		Descriptions
		Frequency	%	
1	$X \geq 9{,}67$	12	7,5	Highly Interested
2	$7{,}33 \leq X < 9{,}67$	28	17,5	Interested
3	$X < 7{,}33$	120	75	Not Interested
	Total	160	100	

Based on Table 3, the percentage of interest among students in the Mathematics Program in playing Mobile Legends, out of a total sample of 160 students, shows that 12 students (7.5%) fall into the highly interested category, 28 students (17.5%) fall into the interested category, and 120 students (75%) fall into the not interested category.

5. Ordinal Logistic Regression

a. Parameter Estimation

The parameter β is estimated by optimizing the likelihood function and obtaining its derivative concerning β. Since the first derivative of the likelihood function is nonlinear, the Newton-Raphson iteration method is used to obtain parameter estimates. After the iteration process, the estimated parameter values are presented in Table 4 below:

Table 4. Estimated Parameter Values for Ordinal Logistic Regression

Parameter	Estimation
β_{01}	-0,641
β_{02}	0,176
$\beta_{1(1)}$	-0,193

$\beta_{1(2)}$	0,287
$\beta_{1(3)}$	0
$\beta_{2(1)}$	0,092
$\beta_{2(2)}$	0

Therefore, considering the coefficients obtained, the ordinal logistic regression model for the influence of interest in playing Mobile Legends on the academic performance of mathematics students in the Faculty of Mathematics and Natural Sciences is as follows:

$$\pi_1(x) = \frac{\exp\left(\beta_{01} + \beta_{1(1)}x_{1(1)} + \beta_{1(2)}x_{1(2)} + \beta_{2(1)}x_{2(1)}\right)}{1 + \exp\left(\beta_{01} + \beta_{1(1)}x_{1(1)} + \beta_{1(2)}x_{1(2)} + \beta_{2(1)}x_{2(1)}\right)}$$
$$= \frac{\exp\left(-0{,}641 - 0{,}193x_{1(1)} + 0{,}287x_{1(2)} - 0{,}092x_{2(1)}\right)}{1 + \exp\left(-0{,}641 - 0{,}193x_{1(1)} + 0{,}287x_{1(2)} - 0{,}092x_{2(1)}\right)}$$

$$\pi_2(x) = \frac{\exp\left(\beta_{02} + \beta_{1(1)}x_{1(1)} + \beta_{1(2)}x_{1(2)} + \beta_{2(1)}x_{2(1)}\right)}{1 + \exp\left(\beta_{02} + \beta_{1(1)}x_{1(1)} + \beta_{1(2)}x_{1(2)} + \beta_{2(1)}x_{2(1)}\right)}$$
$$= \frac{\exp\left(0{,}176 - 0{,}193x_{1(1)} + 0{,}287x_{1(2)} - 0{,}092x_{2(1)}\right)}{1 + \exp\left(0{,}176 - 0{,}193x_{1(1)} + 0{,}287x_{1(2)} - 0{,}092x_{2(1)}\right)}$$

b. Parameter Significance Test

1. Simultaneous Test

Simultaneous parameter significance testing is carried out using the likelihood ratio test. Hypothesis:

H_0: $\beta_{1(1)} = \beta_{1(2)} = \cdots = \beta_{p(r)} = 0$

H_1: there is at least one $\beta_{j(k)} \neq 0, j = 1,2,\ldots,p, k = 1,2,\ldots,r$

After performing calculations for the simultaneous test statistic, the results are obtained as shown in Table 5 below:

Table 5. Simultaneous Test Result

L_M	L_0	G^2
−125,02782	−175,98415	101,91266

Based on Table 5, it can be seen that the value of G^2 is 101,91266, and the critical χ^2_{tabel} value is 5,591. The obtained value indicates that G $>$ χ^2_{tabel}, so H_0 is rejected, meaning that the independent variables simultaneously affect the dependent variable.

2. Partial Test

Separate parameter significance testing (partial) is performed using the Wald test.

Hypothesis:

$H_0: \beta_{j(k)} = 0$

$H_1: \beta_{j(k)} \neq 0, j = 0,1,2, k = 1,2,3$

After conducting calculations for the partial test statistic, the results are obtained as shown in Table 6 below:

Table 6. Partial Test Value

Parameter	Estimation	$SE(\hat{\beta}_{j(k)})$	$\|W\|$	Z_{table}	Conclusions
$\beta_{0(1)}$	-0,641	0,023	27,869	2.35	Significant
$\beta_{0(2)}$	0,176	0,023	7,652	2.35	Significant
$\beta_{1(1)}$	-0,193	0,021	9,190	2.35	Significant
$\beta_{1(2)}$	0,287	0,025	11,480	2.35	Significant
$\beta_{2(1)}$	-0,092	0,022	4,182	2.35	Significant

Based on the partial test conducted using the Wald test statistic, the conclusion is that the variables significantly influencing are the "interest" variable with the categories "very interested" and "interested" and the "gender" variable with the category "male". The ordinal logistic regression model for the influence of interest in playing Mobile Legends on the academic performance of mathematics students in the Faculty of Mathematics and Natural Sciences is as follows:

a. Low GPA Model

$$\pi_1(x) = \frac{\exp(-0{,}641 - 0{,}193x_{1(1)} + 0{,}287x_{1(2)} - 0{,}092x_{2(1)})}{1 + \exp(-0{,}641 - 0{,}193x_{1(1)} + 0{,}287x_{1(2)} - 0{,}092x_{2(1)})}$$

b. Moderate GPA Model.

$$\pi_1(x) = \frac{\exp(-0{,}641 - 0{,}193x_{1(1)} + 0{,}287x_{1(2)} - 0{,}092x_{2(1)})}{1 + \exp(-0{,}641 - 0{,}193x_{1(1)} + 0{,}287x_{1(2)} - 0{,}092x_{2(1)})}$$

6. Coefficient of Determination for the Model

The result of the R^2 value is 0.471123, which means that the independent variables in the ordinal logistic model can explain 47.1123% of the variation in the dependent variable. In comparison, the remaining 52.8877% is explained by other variables not included in the model. Based on this result, it can be concluded that the obtained model provides good results and is acceptable since all the explanatory variables are statistically significant. The R^2 value is non-negative and corresponds to the square of the multiple correlation coefficient, and the utility of goodness of fit measures depends on whether the analysis focuses on explaining the outcome or explaining the effects of some regressors on the result. This might happen because the independent variables used in the model do not contribute to raising the value of R^2 very much. Therefore, to

obtain a higher value of R^2, researchers need to put more attention in the choice of variables in the model, be aware of the presence of outliers, and provide more samples (Hagquist and Stenbeck, 1998; Chicco et al., 2021; Chen & Qi 2023; Ozili, 2023).

CONCLUSION

Taking into careful consideration the results of the data analysis and the methodologies employed, it can be confidently concluded that:

1. The model of the relationship between interest in playing Mobile Legends and the academic performance of students in the Mathematics Program of the Faculty of Mathematics and Natural Sciences, Universitas Mataram, is as follows:

$$\pi_1(x) = \frac{\exp(-0{,}641 - 0{,}193x_{1(1)} + 0{,}287x_{1(2)} - 0{,}092x_{2(1)})}{1 + \exp(-0{,}641 - 0{,}193x_{1(1)} + 0{,}287x_{1(2)} - 0{,}092x_{2(1)})}$$

$$\pi_2(x) = \frac{\exp(0{,}176 - 0{,}193x_{1(1)} + 0{,}287x_{1(2)} - 0{,}092x_{2(1)})}{1 + \exp(0{,}176 - 0{,}193x_{1(1)} + 0{,}287x_{1(2)} - 0{,}092x_{2(1)})}$$

2. The factors that significantly influence the academic performance of students in the Mathematics Program of the Faculty of Mathematics and Natural Sciences are Interest (x_1) with the categories "very interested" and "interested" and Gender (x_2) with the category "male."

REFERENCES

[1] S.J. Akbar, et al., "Bagging Regresi Logistik Ordinal pada Status Gizi Balita," in *Media Statistika*, vol. 3, no. 2, 2010, pp. 103-116.

[2] Q. Chen & J. Qi, "How Much Should We Trust R2 and Adjusted R^2: Evidence from Regressions in Top Economics Journals and Monte Carlo Simulations," in *Journal of Applied Economics*, vol. 26, no. 1, 2207326, 2023.

[3] D. Chicco, M.J. Warrens, & G. Jurman, "The Coefficient of Determination R-squared is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation," in *PeerJ Computer Science*, 7, e623, 2021.

[4] N. Fitriyani, L. Awalushaumi, & A. Kurnia, "Polytomous Logistic Regression in Analyzing the Presence of National Pilot Mosque in Karang Baru Mataram", in the *Proceeding, 1st ICST* Mataram University, 2016

[5] I. Ghozali, "*Aplikasi Analisis Multivariate dengan Program SPSS 25*. Semarang, Universitas Diponegoro, 2018.

[6] D. Gujarati, *Dasar-dasar Ekonometri Edisi Ketiga, Jilid I dan II*. Erlangga, Jakarta, 2007.

[7] C. Hagquist & M. Stenbeck, "Goodness of Fit in Regression Analysis–R^2 and G^2 Reconsidered", in *Quality and Quantity*, vol. 32, no. 3, pp. 229-245, 1998.

[8] D.W. Hosmer and S. Lemeshow, "Applied Logistic Regression," John Wiley & Sons, Inc., New York, 2000.

[9] S. Imaslihkah, et al., "Analisis Regresi Logistik Ordinal Terhadap Predikat Kelulusan Mahasiswa S1 di ITS Surabaya," in *Jurnal Sains dan Seni POMITS*, vol. 2, no. 2, pp. 177-182, 2013.

[10] F.R. Iskandar, et al., "Dampak Permainan Mobile Legends Terhadap Motivasi Belajar Siswa Sekolah Dasar," in *EduBasic Journal: Jurnal Pendidikan Dasar*, vol. 1, no. 2, pp. 116-122, 2019.

[11] P.K. Ozili, “The Acceptable R-Square in Empirical Modelling for Social Science Research,” in *Social Research Methodology and Publishing Results*, 5 June, 4128165, 2022.
[12] Playstore, "Games Kategori," 2022. [*Online, Accessed 26 July 2022*]. Available: https://play.google.com/store/games.
[13] B. Simamora, *Panduan Riset Prilaku Konsumen*. Gramedia, Jakarta, 2004.
[14] Sugiyono, *Metode Penelitian Kuantitatif*. Bandung, Alfabeta, 2018.
[15] J. Supratno, *Statistika: Teori dan Aplikasi*. Erlangga, Jakarta, 2000.

Clustering Villages in the Mountain Areas in West Java Based on Tourism Potential Using K-Prototype Algorithm

Ainun Salsabila [(1)], L. M. Risman Dwi Jumansyah [(2)], Anwar Fitrianto [(3)], Erfiani [(4)], Alfa Nugraha P. [(5)]

[1,2,3,4,5]Department of Statistics and Data Science, FMIPA, IPB *University*

Meranti Street, IPB University Dramaga, Bogor 16680, West Javat
e-mail: ainunsalsabila@apps.ipb.ac.id[(1)], rismandwijumansyah@apps.ipb.ac.id[(2)], anwarstat@gmail.com[(3)], erfiani@apps.ipb.ac.id[(4)], alfanugraha@apps.ipb.ac.id[(5)]

ABSTRAK

Analisis klaster adalah salah satu metode analisis multivariat yang digunakan untuk mengelompokkan suatu objek berdasarkan kesamaan karakteristiknya. Secara umum, data yang digunakan pada proses klasterisasi adalah data numerik atau kategorik saja. Tetapi, terkadang juga banyak ditemui kasus yang menggunakan data campuran numerik dan kategorik. Oleh karena itu, algoritma yang dapat digunakan adalah *K-Prototype*. *K-Prototype* adalah pengembangan dari K-Means yang dapat digunakan pada data berjumlah besar bertipe numerik dan kategorik. Dasar pengembangan *K-Prototype* adalah mengukur jarak antara objek dan pusat *prototype*. Banyaknya *prototype* tergantung dengan jumlah *cluster* yang dibentuk. Pada penelitian ini, peneliti memanfaatkan algoritma tersebut untuk mengklasterkan desa-desa wilayah pegunungan di Jawa Barat berdasarkan potensi wisatanya, agar mengetahui potensi-potensi yang perlu dikembangkan berdasarkan lima komponen, yaitu Daya Tarik, Keterjangkauan, Akomodasi, Fasilitas Pendukung, dan Kesadaran. Berdasarkan *Silhouette* dan *McClain Index*, diperoleh bahwa jumlah klaster optimal adalah dua. Klaster 1 terdiri dari 103 desa dan klaster 2 terdiri dari 703 desa. Klaster 1 adalah desa-desa yang secara umum sudah lebih unggul pada komponen Keterjangkauan, Kesadaran, dan Fasilitas Pendukung, tetapi masih kurang pada komponen Daya Tarik dan Akomodasi dibandingkan desa-desa di klaster 2.
Kata kunci: Analisis Klaster; *K-Prototype*; Potensi Wisata

ABSTRACT

Cluster analysis is a multivariate analysis method used to group objects based on their similar characteristics. In general, in the clustering process only use numerical or categorical data. But, sometimes we also encounter cases that use both numerical and categorical data. Therefore, the algorithm that can be used is K-Prototype. K-Prototype is a development of K-Means that can be used on large data with numerical and categorical types. The basis of K-Prototype development is to measure the distance between the object and its centroid prototype. The number of prototypes depends on the number of clusters formed. In this study, researchers use this algorithm to group mountainous villages in West Java based on their tourism potential, in order to find out the potential that needs to be developed based on five components, namely Attractions, Access, Accommodation, Amenities and Awareness. Based on the Silhouette and McClain Index, the optimal number of clusters is two. Cluster 1 *consists of* 103 *villages and cluster* 2 *consists of* 703 *villages. Cluster* 1 *are villages that are generally better in the Access, Awareness and Amenities, but are still lacking in the Attraction and Accommodation components compared to villages in cluster* 2.
Keywords: *Cluster Analysis; K-Prototype; Tourism Potential*

INTRODUCTION

Cluster analysis is grouping objects who have the same characteristics in one cluster [1, pp. 342–353]. Cluster analysis requires a distance measure that is defined for each pair of objects to be grouped. Smaller the distance, the more similar one object is to another object. The distances commonly used are distance for numerical and categorical data [2]. So, if the data is a combination of numerical and categorical, the method that is used is a method that combines distances for numerical and categorical data, such as K-Prototype. K-Prototype is a development of K-Means, introduced by Huang that can be used on large data with numerical and categorical types. The basis of K-Prototype development is to measure the distance between the object and its centroid prototype [3]. Cluster analysis and K-Prototype can be used to cluster areas like villages. So, in this research, researcher interest to cluster villages in the mountain areas in West Java based on their tourism potential which consists of five components, Attraction, Access, Accommodation, Amenities, and Awareness, in order to help the parties concerned to find out the potential, because research regarding exploring village tourism potential in mountain areas is still very difficult to find. Researcher also interest to proposed the use of K-Prototype because the data that is used in this research are mix numerical and categorical.

Research using cluster analysis was carried out by [4] with the title "Pengembagan Desa Wisata dengan Pendekatan Klaster" which applies components for developing the potential of tourist villages consisting of *Attraction*, *Accessibility*, *Amenity*, *Ancillary Services*, *Activity*, and *Accommodation*, which results in 6 clusters to arrange strategies to develop tourism villages. Meanwhile, the application of K-Prototype algorithm for grouping villages was carried out by [5] in his research entitled "Penerapan Algoritma *K-Prototypes* untuk Pengelompokan Desa-Desa di Provinsi Jawa Barat Berdasarkan Indikator Indeks Desa Membagun Tahun 2020". This research produced 5 clusters with the lowest development achievement criteria for cluster 1 to the best development achievement for cluster 5.

One of the ways in which cluster analysis can be used to cluster villages is to help explore the village's potential. One of the interesting potentials to be explored is tourism potential, especially tourism potential mountain villages. Tourism potential needs to be explored, because it is one of the sectors driving growth and improvement of the national economy. Tourism is a combination of activities, services, and industries that provide travel experiences including culinary, accommodation, transportation, entertainment, activity facilities, and hospitality services for individuals or groups traveling far from their home [6]. According to Badan Pusat Statistik, the tourism sector in 2021 gave a contribution to Gross Domestic Product (GDP) of 4.97% and increase every year, with state foreign exchange earnings of US$ 15 billion per year [7]. Currently, the trend in tourism development in Indonesia is a tourism development model based on tourist villages. This happened because the interest of tourists began to shift who wanted entertainment with beautiful views, far from the city, and cheaper costs. Apart from that, the phenomenon of the increasing number of development of tourist villages is in line with the enactment of the Law on Villages which gives villages development authority to each village government and provides multiple villages funds [8]. A tourist villages itself is a rural area that offers an overall atmosphere that reflects the authenticity of the village, both in terms of social-economic life, social-culture, daily life, unique of village structures, interesting

economic activities, and has the tourism potential to develop, such as attractions, accommodation, food and drinks, and other tourist facilities [9].

Exploring the tourism potential of mountain villages is interesting because mountain villages have very different characteristics from valley and mainland villages. Mountain villages have limited basic facilities and employment opportunities. The geographical location of mountain villages is one of the reasons why the social and economic conditions of mountain village residents are lower. Thus, additional efforts are needed to solve existing limitations, so the economic disparities can be solved [10]. One of the efforts that can be made is to develop tourist villages in mountain areas, because not only beautiful mountain views can be offered, but the use of existing resources around the mountain such as forests, rivers or lakes as well as the sale of superior products can also be used as an additional attraction for tourists. Tourism potential needs to be developed considering that the West Java area consists of many mountains with interesting resources.

METHOD

This research uses secondary data from Badan Pusat Statistik, namely PODES (Potensi Desa) data for 2021. The data used only data for villages in mountain areas (top and valley) where there are settlements in West Java Province. The amount of data is used consist of 806 observations with 16 variables which are divided into five components as in the Table 1 below:

Table 1. Description of Research Variables

Component	Variable	Description	Explanation	Scale
Attraction	X_1	Location of village in forest areas	1: In the forest area 2: Around the forest area 3: Outside the forest area	Categorical
	X_2	Use of rivers and lakes/reservoirs/dams for tourism	1: River 2: Lake/reservoir/dam 3: River and lake 4: Neither	Categorical
	X_3	Main product: a. Main products	1: Food 2: Non Food 3: Food and Non Food 4: Neither	Categorical
	X_4	b. Main products exported to other countries	1: Yes, mostly 2: Yes, a small part 3: None	Categorical
	X_5	Tourist attractions/public baths	1: Exist 2: None	Categorical
Amenities	X_6	Total number of places to eat: a. Restaurant b. Stall/food stall	-	Numeric
	X_7	Number of shopping places (mini market/ supermarket)	-	Numeric

	X_8	Number of shops/ grocery stalls	-	Numeric
	X_9	Total number of health facilities: a. Hospital b. Community health center with inpatient care c. Polyclinic/treatment center d. Doctor's practice place e. Village Health Post f. Pharmacy	-	Numeric
	X_{10}	Total number of worship place: a. Mosque b. Musholla	-	Numeric
	X_{11}	The number of police posts (including police stations) used	-	Numeric
	X_{12}	Total number of operating financial institutions: a. Government Commercial Banks (BRI, BNI, Mandiri, BPD, BTN) b. Private Commercial Banks (BCA, Permata, Sinarmas, CIMB, etc) c. Rural Bank (BPR)	-	Numeric
Access	X_{13}	Land roads between villages can be passed by motorized vehicles with 4 or more wheels	1: Throughout the year 2: Throughout the year, except certain moment (when rain, etc) 3: During dry season 4: Can't passed throughout the year	Categorical
	X_{14}	The existence of public transportation	1: Exist, with fixed route 2: Exist, with unfixed route 3: None	Categorical
Accommodation	X_{15}	Total number of hotels and homestays	-	Numeric
Awareness	X_{16}	Build, maintenance or normalization of rivers, canals, embankments, ditches, drainage, reservoirs, etc	1: Yes 2: No	Categorical

The stages of data analysis in this research are as follows:

1. Describe the research data in general such as the number of villages according to the topography of region in percentage.
2. Determine the optimal number of clusters using the Silhouette and McClain Index using equations: Silhouette Index is a combination of cohesion (average distance from object *i* with all objects that are in one cluster, denoted a_i) and separation (average distance from object *i* with all objects that are in other cluster, denoted b_i) [11]:

   $$S_i = \frac{b_i - a_i}{max(a_i, b_i)} \quad (1)$$

   McClain Index is ratio of variety of within cluster distance (S_W) average with variety of between cluster distance (S_B) average [12]:

   $$McClain = \frac{\bar{S}_w}{\bar{S}_b} \quad (2)$$

3. Initialization of the prototype
   Random selection of *k* prototypes from the data according to the specified number of clusters. The limit of the number of clusters *k* to be formed is a minimum of 2 and a maximum of *n* or *n*/2, where *n* is the number of samples [13].
4. Measure the distance of all objects to all prototypes

   $$d(X,Y) = \sum_{j=1}^{p}(x_{jn} - y_{jn})^2 + \gamma \sum_{j=p+1}^{m} d(x_{jc}, y_{jc}) \quad (3)$$

   $$d(x_{jc}, y_{jc}) = \begin{cases} 0, x_{jc} = y_{jc} \\ 1, x_{jc} \neq y_{jc} \end{cases} \quad (4)$$

5. Allocate objects to the cluster with the closest prototype.
6. Calculating the new prototype.
7. Reallocate all objects to the new prototype
   After all objects have been allocated, the distance between all objects and all existing prototypes will be re-measured. If there is an object that turns out to be closer to another prototype, membership will be transferred and then updates will be made to the old and new cluster prototypes.
8. If the prototype still changes significantly or the stopping criteria have not been met, then return to step 4 until no object movement or the maximum iteration is reached.
   The analysis stage can be described with a flowchart as in Figure 1 below:

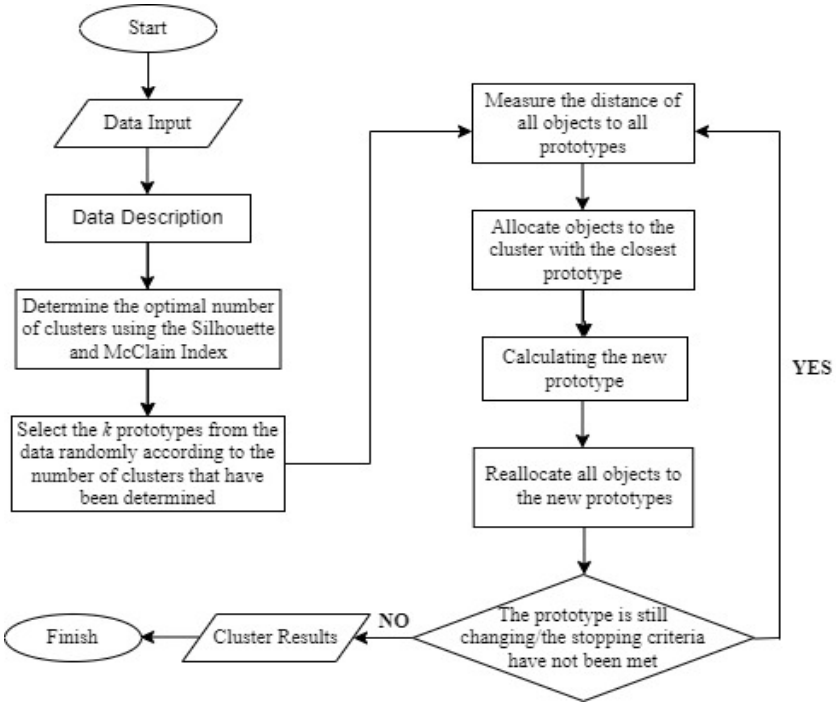
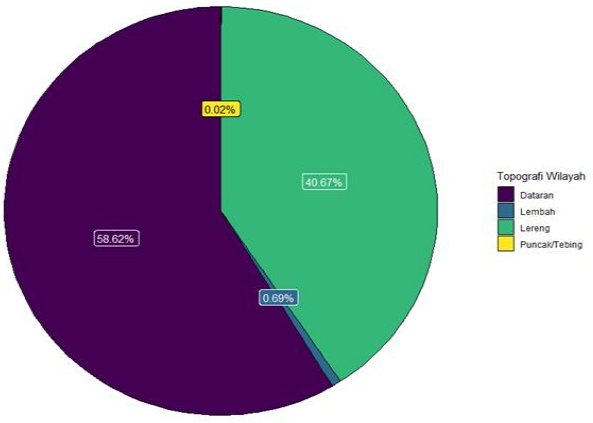


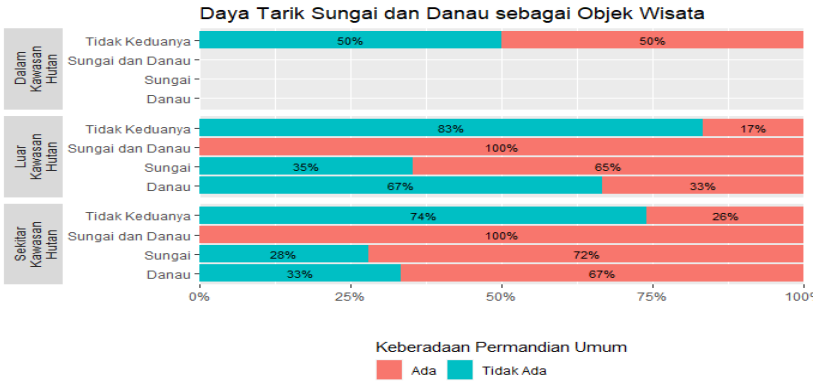
Figure 1. Flowchart of Data Analysis

RESULT AND DISCUSSION

West Java is a province that offers a very interesting topography. More than half, approximately 58.62% of the province's territory is land that is often the center of settlement and agricultural activity. Furthermore, about 40.67% of the villages are located downhill. The villages in the region have natural beauty and attractive resources that can be used as tourist destinations to support the economy of the villagers. But, to realize that, there is a lot of tourism potential that needs to be explored and developed.



(a)



(b)

Figure 2. (a) Topographic Description of Village in West Java; (b) Description of Rivers and Lakes as Tourist Attractions

Tourist potential components such as attractions, forest areas, and the presence of rivers and lakes are the main magnet for tourists both from the local area and from outside the area, as when the river and lake are used as public bathing places. As in the study conducted by [14], the Lau Seruwai River in North Sumatra is used as a place for bathing or swimming by tourists. For west Java (see Figure 2(b)), about 65% of the rivers outside the forest area are used as public bathing facilities. Interestingly, some villages around the Forest Area have rivers and lakes that are entirely or 100% allocated as public bathing facilities. Based on this, it can be known that the people or tourists in the village are already aware of the potential of rivers and lakes if used as tourist attractions.

Another potential that can be a major attraction for tourists is the products produced by local villagers, whether food or non-food products, which have been successfully exported to various countries. A description of this in the village around downhill in West Java can be seen in Figure 3. The forested territory, about 50% of the products produced, both food and non-food, have expanded the international market. In the villages outside and around the forest, the food and non-food products exported are still very small. Based on this, the villages in the outer territory and around the forest area should be more keen in marketing their superior products, to be a tourist attraction and attract tourists to come.

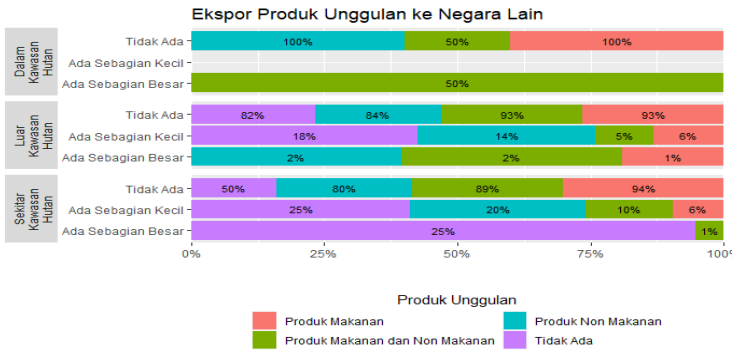


Figure 3. Description of Exported Villages Main Product

The clustering of villages begins by searching the optimal number of clusters because the K-Prototype algorithm is one of the non-hierarchical grouping methods, in which the cluster number must be determined at an early stage. The selection of the optimal number of clusters is done using the method of Silhouette and McClain coefficients.

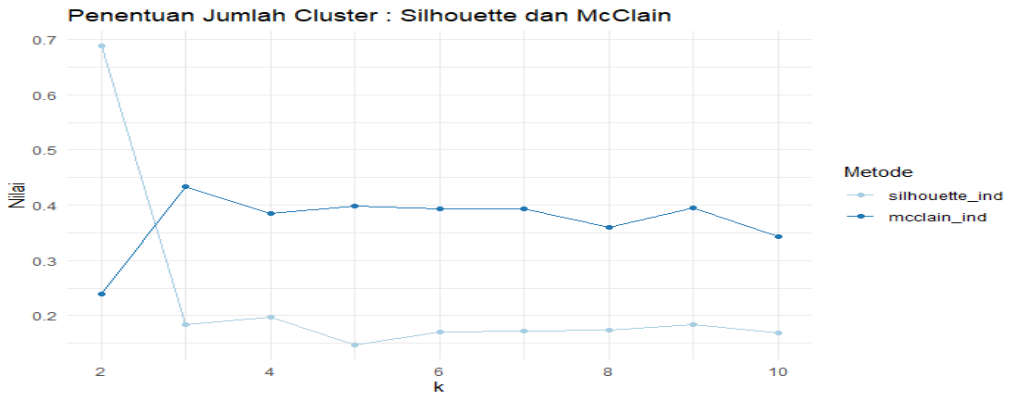


Figure 4. Silhouette and McClain Coefficient Graph

Figure 4 shows a graph of the values of the Silhouette and McClain coefficients for each number of clusters. From the graph, it can be concluded that the optimum number of Clusters is 2 because it has the highest value of Silhouettes and the lowest of McClain. The range of selection of the number of clusters is limited from 2 to 10 so that the interpretation of the characteristics of the cluster is easier.

The optimum number of clusters obtained was then used to cluster with the K-Prototype algorithm. The result of a cluster using a K-prototype with a value of k = 2 mapped as many as 103 villages into cluster 1 and 703 villages in cluster 2. The district of Bandung has the most members in cluster 1 of 25 villages and the city of Tasikmalaya has no villages gathered in cluster 1. Whereas in the second cluster, Garut district has the highest number of members, 106 villages, and the town of Tasikmalaya has the lowest membership of 3 villages. But in general, mountain villages in West Java are much more in Cluster 2 than Cluster 1. The cluster results obtained are as follows:

Table 2. Cluster Results

Region	Number of Cluster Members		Region	Number of Cluster Members	
	1	2		1	2
Bandung	25	40	Kota Tasikmalaya	0	3
Bandung Barat	7	20	Kuningan	1	58
Bogor	11	30	Majalengka	4	31
Ciamis	4	86	Pangandaran	0	27
Cianjur	12	74	Purwakarta	2	29
Cirebon	1	21	Subang	1	24
Garut	7	106	Sumedang	12	75
Kota Bandung	7	5	Tasikmalaya	5	67
Kota Bogor	4	7			

The characteristics of each cluster based on categorical variables are as follows:

Table 3. Characteristics of Cluster Results for Categorical Variable

Component	Variable	Category	Percentage (%)	
			1	2
Attraction	X_1	In the Forest Area	0	0.6
		Around the Forest Area	71.8	68.8
		Outside the Forest Area	28.2	30.6
	X_2	Lake	1.0	1.6
		River	4.9	7.7
		River and Lake	1.9	0.6
		Neither	92.2	90.2
	X_3	Food Product	47.6	55.3
		Non Food Product	20.4	30.6
		Both	30.1	12.2
	X_4	Neither	1.9	1.8
		Yes, Mostly	1.9	1.4
		Yes, a Small Part	11.7	6.7
		None	86.4	91.9
Access	X_5	Exist	20.4	24.8
		None	79.6	75.2
	X_{13}	During the Dry Season	0	0.1
		Throughout Year	98.1	97.3
		Throughout Year (Depending on Conditions)	1.9	2.6
	X_{14}	Exist (Fixed Route)	60.2	51.2
		Exist (Unfixed Route)	37.9	46.4
		None	1.9	2.4
Awareness	X_{16}	Neither	92.2	90.2
		Yes	55.3	43.5
		No	44.7	56.5

Meanwhile, the characteristics of each cluster based on numerical variables are as follows:

Table 4. Characteristics of Cluster Results for Numerical Variable

Component	Variable	Minimum		Average		Maximum	
		1	2	1	2	1	2
Amenities	X_6	0	0	37.05	14.43	211	177
	X_7	0	0	1.62	0.47	12	10
	X_8	105	0	181.26	41.82	411	117
	X_9	1	0	6.46	3.44	24	24
	X_{10}	11	1	46.66	32.85	127	121

	X_{11}	0	0	0.12	0.07	3	2
	X_{12}	0	0	0.73	0.21	12	17
Accommodation	X_{15}	0	0	1.09	1.31	30	105

So, the characteristics based on Tables 3 and 4 can be summarized as follows:

Table 5. Summary of Cluster Characteristics

Cluster	Characteristics
1	The location of the village tends to be far away/outside the forest, have easy access and have lots of public transportation. Has more adequate dining facilities, shopping places, health facilities, places of worship and police posts. However, the use of rivers and lakes as tourist attractions is still lacking and the number of accommodations is a little, even though the population is more aware of cleanliness and environmental maintenance, as well as exporting main products, even though the numbers are still little.
2	The location of the village is around a forest area, some even inside the forest, so many rivers and lakes are used as tourist attractions such as public baths, and have many main products, both food and non-food, but not many are exported. Access to roads and public transportation, awareness of the environment, and facilities such as places to eat, shopping, health, places of worship, and police posts are still lacking, but there are more accommodations.

CONCLUSION

Mountain villages in West Java are divided into two clusters, namely cluster 1 comprises 103 villages and Cluster 2 comprises 703 villages. Cluster 1 are villages that are generally superior in the components of Accessibility, Awareness, and Supporting Facilities, but still less in the elements of Attraction and Accommodation compared to the villages of Cluster 2. Therefore, villages within Cluster 1 are given more attention and dig into the existing tourist attractions, such as using rivers and lakes as their main tourist attraction. In addition, the village in Cluster 1, must also create more excellent products to attract tourists to come. Whereas for villages inside cluster 2, it is preferable to be the focus of government attention in improving and improving the accessibility and infrastructure of villages for tourist convenience. So, when these potential tourist components can be improved, it is expected to make the village an attractive tourist village that can bring many visitors, to help and boost the economy of the village people.

REFERENCE

[1] A. Y. Kartini and A. M. Jamiluddin, "Application of Agglomerative Hierarchical Clustering Method for Grouping Non-Cash Food Assistance Recipients in Ngambon Bojonegoro," 2023. Accessed: Dec. 10, 2023. [Online]. Available: www.unipasby.ac.id

[2] R. Wijayati and D. R. S. Saputro, "Clustering Data Numerik dan Kategorik Menggunakan Algoritme K-Prototype. PRISMA, Prosiding Seminar Nasional Matematika 6," pp. 702–706, 2023, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[3] S. Annas, I. Irwan, R. H. Safei, and Z. Rais, "K-Prototypes Algorithm for Clustering The Tectonic Earthquake in Sulawesi Island," *Jurnal Varian*, vol. 5, no. 2, pp. 191–198, May 2022, doi: 10.30812/varian.v5i2.1908.

[4] I. Kurniawan and L. Fitriani, "PENGEMBANGAN DESA WISATA DENGAN PENDEKATAN KLASTER," *Indonesian Journal of Economics, Entrepreneurship and Innovation*, vol. 2, no. 2, pp. 52–59, 2021, doi: 10.31960/ijoeei.v2i2.1477.

[5] M. Ganmanah, A. Kudus, P. Statistika, F. Matematika, D. Ilmu, and P. Alam, "Penerapan Algoritme K-Prototypes untuk Pengelompokkan Desa-Desa di Provinsi Jawa Barat Berdasarkan Indikator Indeks Desa Membangun Tahun 2020", doi: 10.29313/.v0i0.28974.

[6] L. A. Birahmatika and E. Ahyudanari, "Jurnal Aplikasi Teknik Sipil Analisis Clustering Objek Daya Tarik Wisata Kabupaten Banyuwangi."

[7] K. Putri and F. Sari, "PENGARUH SEKTOR PARIWISATA TERHADAP PERTUMBUHAN EKONOMI DI KABUPATEN LOMBOK TENGAH DAN KABUPATEN MALANG Oleh," *Jurnal Ekonomi dan Bisnis*, vol. 11, no. 2, 2022.

[8] A. Y. Asmoro, T. B. Bachri, and A. Detmuliati, "ANALISIS POTENSI WISATA DESA DENGAN KERANGKA 6A STUDI KASUS DESA NGAJUM, MALANG," vol. 18, no. 2, 2020, doi: 10.36275/mws.

[9] D. Olivia, W. Tisno Atmojo, A. Guna, P. Studi, and P. Wilayah Dan Kota, "Analisis Potensi Desa Wisata Sebagai Upaya Peningkatan Aksesibilitas Dan Konektivitas Di Desa Wisata Cikolelet." [Online]. Available: https://journals.upi-yai.ac.id/index.php/ikraith-teknologi/issue/archive

[10] A. Y. Saputra, E. Rustiadi, and W. Rindayati, "Perkembangan dan Karakterisasi Desa-desa Pegunungan Jawa Tengah," *Journal of Regional and Rural Development Planning*, vol. 6, no. 1, pp. 1–13, Feb. 2022, doi: 10.29244/jp2wd.2022.6.1.1-13.

[11] L. Qadrini *et al.*, "Metode K-Means dan DBSCAN pada Pengelompokan Data Dasar Kompetensi Laboratorium ITS Tahun 2017," 2020. [Online]. Available: www.unipasby.ac.id

[12] R. Aschenbruck and G. Szepannek, "Cluster Validation for Mixed-Type Data", doi: 10.5445/KSP/1000098011/02.

[13] R. Nooraeni, J. Suprijadi, and P. Statistika STIS, "K-PROTOTYPE UNTUK PENGELOMPOKAN DATA CAMPURAN," *Jurnal Statistika Teori dan Aplikasi: Biomedics, Industry & Business And Social Statistics*, vol. 13, no. 1.

[14] J. S. P. Pakpahan, O. H. Syahputra, and B. Slamet, "Jurnal Persepsi Masyarakat," *Jurnal Penelitian Kehutanan Bonita*, vol. 4, no. 1, pp. 20–29, 2022.

Kernel Nonparametric Regression Modeling with the Nadaraya-Watson Estimator (Case Study: Fertility in the Southern Sumatra Region)

Indah Wahyuliani [(1)], Muhammad Arib Alwansyah [(2)], Dyah Setyo Rini [(3)], Winalia Agwil [(4)]

[1,2,3,4] Statistics Study Program, Bengkulu University, Bengkulu

Road WR. Supratman, Kandang Limun Village, Muara Bangkahulu District, Bengkulu City
e-mail: indahwahyuliani.9e@gmail.com[(1)], muhammadaribalwanyah232@gmail.com[(2)], dyah.setyorini@unib.ac.id[(3)], winaliaagwil@unib.ac.id[(4)]

ABSTRAK

Fertilitas adalah kelahiran hidup (*live birth*) yaitu terlepasnya bayi dari rahim seorang perempuan dengan adanya tanda-tanda kehidupan seperti berteriak, bernafas, jantung berdenyut, dan sebagainya. Sumber data penelitian ini berasal dari publikasi website resmi Badan Pusat Statistika (BPS). Penelitian ini bertujuan untuk memodelkan dan memprediksi data fertilitas pada tahun 2020 dengan regresi nonparametrik kernel dengan estimator *Nadaraya-Watson*. Model nonparametrik kernel menunjukkan hubungan fertilitas (Y) dengan persentase usia perkawinan pertama dibawah umur (X_1), persentase wanita 15-49 tahun yang tidak menggunakan KB/alat tradisional (X_2) jumlah peserta KB aktif (X_3), jumlah pasangan usia subur (X_4), persentase rata-rata lama sekolah (X_5), dan jumlah pengeluaran perkapita (X_6) berdasarkan nilai *bandwidth* yang hasilkan dan fungsi kernel *Gaussian*. Berdasarkan hasil analisis diperoleh variabel bebas yang berpengaruh signifikan yaitu X_1, X_3, X_4, X_5terhadap variabel terikat (Y) dengan nilai *bandwidth* optimum sebesar 0,490 dan nilai R^2 sebesar 99,6% serta nilai MSE sebesar 0,332. Pemodelan fertilitas penting karena dapat membantu memahami dan memprediksi tren populasi. Hal ini memberikan wawasan tentang potensi tingkat kelahiran dalam suatu populasi di masa depan. Informasi ini dapat digunakan untuk perencanaan kebijakan, termasuk kesehatan, pendidikan, dan kebijakan sosial.
Kata kunci: Fertilitas, Regresi Nonparametrik Kernel, MSE, *Nadaraya-Watson*, R^2.

ABSTRACT

Fertility is a live birth, namely the release of a baby from a woman's womb with signs of life such as screaming, breathing, a throbbing heart, and so on. The source of this research data comes from the publication of the official website of the Central Statistics Agency (BPS). This study aims to model and predict fertility data in 2020 with kernel nonparametric regression using the Nadaraya-Watson estimator. The nonparametric kernel model shows the relationship between fertility (Y) and the percentage of underage women at first marriage (X_1), the percentage of women 15-49 years who do not use traditional KB or conventional methods (X_2), the number of active family planning participants (X_3), the number of couples of childbearing age (X_4), the percentage of the average length of schooling (X_5), and the total expenditure per capita (X_6) based on Gaussian kernel function and bandwidth values. Based on the results of the analysis, the independent variables that have a significant effect are X_1, X_3, X_4, X_5 on the dependent variable with the optimum bandwidth value of 0.490 and the value of R^2 of 99.6%, and the MSE value of 0.332. Modeling fertility is important as it helps understand and predict population trends. It provides insights into the potential number of births in a population in the future. This information can be used for policy planning, including health, educations, and social policies.
Keywords: *Fertility, Kernel Nonparametric Regression, MSE, Nadaraya-Watson, R^2*

INTRODUCTION

A nonparametric approach is an analysis that does not focus on assuming a particular curve shape, there by providing greater flexibility, the data is expected to find its own estimated shape. In carrying out nonparametric analysis, it is necessary to estimate the nonparametric regression function, which is carried out based on smoothing techniques [1]. One of the estimators used to estimate an unknown regression function is the Nadaraya-Watson estimator. The Nadaraya-Watson estimator has clear generalizations for explanatory variables and kernel functions, and Nadaraya and Watson in 1964 defined a kernel regression estimator called the Nadaraya-Watson estimator, or Nadaraya-Watson kernel estimator [2]. Choosing nonparametric kernel regression for fertility data is advantageous due to the inherent complexity of fertility patterns, which often deviate from linear or parametric assumptions. Nonparametric models, facilitated by kernels, offer increased flexibility in capturing intricate variations within fertility data. This approach is particularly suitable when the relationships influencing fertility exhibit nonlinear characteristics, allowing the model to adapt without being confined to a predefined functional form. Additionally, kernel regression tends to be more robust in the presence of outliers or minor fluctuations commonly found in fertility datasets. Another notable advantage is that nonparametric methods alleviate the need to assume a specific function form, making them well-suited for handling fertility data with unpredictable patterns. However, it is essential to consider challenges such as the proper tuning of kernel parameters and the requirement for an adequate amount of data to support a nuanced model. The selection of nonparametric kernel regression should align with the specific characteristics and complexities inherent in the fertility data under analysis.

The population in the southern Sumatra region continues to increase. This can be seen from the results of the 2000 population census, the population at that time was 18.52 million people. Then, in the 2010 population census, it increased to 21.08 million people [3]. As of the 2020 population census, the population in the Southern Sumatra region reached 24.49 million people [4]. If the population increase is not balanced with an increase in economic capacity, it will have an impact on the welfare of the population in a country, such as high rates of poverty, unemployment, crime, and worsening other social conditions. This is in line with the research conducted by Arum and Haris Which says that population growth is caused by tree components, one of which is fertility [5]. This problem can be limited by controlling the birth rate (fertility). Fertility is a live birth, namely the release of a baby from a woman's womb with signs of life, for example, screaming, breathing, heartbeat, and so on [6]. The relationship between fertility and several factors that influence it has an unknown regression curve shape, so a nonparametric approach is used to estimate the unknown regression function [7].

In the context of fertility data, nonparametric kernel regression offers significant advantages. Fertility, influenced by a myriad of complex factors, often exhibits patterns that are challenging to capture linearly or parametrically. Nonparametric kernel methods provide the flexibility needed to address the intricate relationships among variables in fertility data. This allows the model to adapt dynamically to changes in fertility patterns over time without relying on specific assumptions about the functional form. Moreover, nonparametric kernel regression can identify local patterns in the data, understanding regional or group variations in fertility characteristics. By leveraging kernel techniques, research on fertility data can more effectively manage uncertainty that may arise and

gain a deeper understanding of factors influencing fertility rates without sacrificing model complexity [8].

METHOD

Nonparametric regression is one of the methods used to estimate the relationship pattern between the free variable and the bound variable, where the form of its regress curve is unknown [9]. Nonparametric regression does not require assumptions to be met, and the data is expected to find its own form of estimation so that it has high flexibility. Estimation of regression functions is performed based on observational data using certain recovery techniques. Therefore, the systematic model of nonparametric regression can be written as follows:

$$y_i = m(x_i) + \varepsilon_i \qquad i = 1,2,\dots,n.$$

With ε_I is a random variable that is assumed to be independent with an average of zero and a variance σ^2. The function $m(x)$ It is an unknown function called the regression function. In nonparametric regression, there is no assumption of the form of the regression function m(x), thus providing flexibility in possible forms of tre regression function [7]. There are several techniques for reducing the value of a response in response to nonparametric, namely, kernel, spline, local polynomial, fourier dye, and wavelet [10]. Although nonparametric regression is used to overcome data modeling that does not form a particular pattern of relationships, the nonparametric regression model can still be used to model data that has a linear or nonlinear pattern because of the absence of assumptions to be met [9]. One way to estimate the form of relationship between the free variables and the bound variable is by looking at the form of relationship patterns on the scatter diagram or *scatterplot*. By knowing the pattern of relationships that formed, the appropriate approach can be determined to estimate the regression function [11].

The kennel density function is one of the nonparametric methods to determine the probability density function of a random variable. An estimator of kernel density is a development of a histogram estimator. This method is often used because it has a more flexible form. A good density function has smooth functionality variances in sampling are not great and important information from the data is not lost [9]. So it is required to be a constitutive estimator of the function. Kennel density for density functions f(x) and f(x,y) is defined as follows [12]:

$$\hat{f}_x(x) = \frac{1}{nh_x}\sum_{i=1}^{n} K\left(\frac{x - X_i}{nh_x}\right) \tag{1}$$

$$\hat{f}(x,y) = \frac{1}{nh_x h_y}\sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right) K_y\left(\frac{y - Y_i}{h_y}\right) \tag{2}$$

The nonparametric regression model of the kernel can be expressed as follows:

$$Y_i = m(x_i) + \varepsilon_i \tag{3}$$

Regarding kernel, is one of the nonparametric techniques to estimate conditional expectations of random variables with the aim of finding relationships between free variables and variables bound by using kernel function weight [13]. Conditional expectations of the bound variable relative to the free variable can be written as follows:

$$m(x) = E(Y|X = x)$$

$$m(x) = \int_{-\infty}^{\infty} y f(y|x) dy$$

$$m(x) = \int_{-\infty}^{\infty} y \frac{f(x,y)}{f_x(x)} dy \quad (4)$$

Estimation $m(x)$ obtained from substitution of equation (1) and (2) to equation (4)

$$\hat{m}(x) = \int_{-\infty}^{\infty} y \frac{f(x,y)}{f_x(x)} dy$$

$$\hat{m}(x) = \frac{1}{f_x(x)} \int_{-\infty}^{\infty} y f(x,y) dy$$

$$\hat{m}(x) = \frac{1}{f_x(x)} \int_{-\infty}^{\infty} y \frac{1}{nh_x h_y} \sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right) K_y\left(\frac{y - Y_i}{h_y}\right) dy \quad (5)$$

$$\hat{m}(x) = \frac{1}{f_x(x)} \frac{1}{nh_x} \sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right) \int_{-\infty}^{\infty} \frac{y}{h_y} K_y\left(\frac{y - Y_i}{h_y}\right) dy \quad (6)$$

Suppose $u = \frac{y - Y_i}{h_y}$, then:

$$\hat{m}(x) = \frac{1}{f_x(x)} \frac{1}{nh_x} \sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right) \int_{-\infty}^{\infty} (u + Y_i) K_y(u) dy \quad (7)$$

$$\hat{m}(x) = \frac{1}{f_x(x)} \frac{1}{nh_x} \sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right) \int_{-\infty}^{\infty} u K_y(u)\, du + Y_i \int_{-\infty}^{\infty} K_y(u)\, du \quad (8)$$

By using the kernel properties $\int_{-\infty}^{\infty} K(u) du = 1$ and $\int_{-\infty}^{\infty} u K(u) du = 0$, so obtained estimator for $m(x)$ as follows:

$$\hat{m}(x) = \frac{\frac{1}{nh_x} \sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right) Y_i}{f_x(x)} \quad (9)$$

On kernel regression, there is a known estimator that can be used to estimate the regression function estimator, *Nadaraya-Watson*. The density function $f_x(x)$ at equation (1) is subtitized to equation (9) it will be obtained by the estimator *Nadaraya-Watson* with $h_x = h$ and $K_x = K$ as follows:

$$\hat{m}(x) = \frac{\frac{1}{nh_x} \sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right) Y_i}{\frac{1}{nh_x} \sum_{i=1}^{n} K_x\left(\frac{x - X_i}{h_x}\right)}$$

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)} \quad (10)$$

Estimator *Nadaraya-Watson* is an estimator used to estimate*ad* a local weighted average by using the kernel as a weighting function. Estimator *Nadaraya-Watson* multivariate forms can be written in the same equation [14]:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} \prod_{j=1}^{d} K(\frac{x_{ij} - x_j}{h}) Y_i}{\sum_{i=1}^{n} \prod_{j=1}^{d} K(\frac{x_{ij} - x_j}{h})} \quad (11)$$

The main philosophy of nonparametric regression is to estimate the regression function m(x) using the weighted average of raw data, where the weight is a distance function in space-x. In particular, the weight is a distance decline function. This type of weighting scheme is proposed by *Nadaraya-Watson* (1964), where the weighting is associated with observation Y_I, for predictions x_I obtained from:

$$W_i(x) = \frac{K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)} = \frac{K(u)}{\sum_{j=1}^{n} K(u)} \quad (12)$$

where the K(u) is a declining function of u and $h > 0$ called *bandwidth* or the preparation parameters. K(u) is the kernel function that can be considered as a function of density, like opportunities such as *Gaussian*. $W_I(x)$ is a position of positive weight and has characteristics $\sum_{I=1}^{n} W_I(x) = 1$, then obtained:

$$\hat{m}(x) = \sum_{i=1}^{n} W_i(x) Y_i \quad , i = 1,2,\ldots,n \quad (13)$$

with each one n data are given different weights [14]. So that *Nadaraya-Watson* is the average weight of Y_I.

RESULT AND DISCUSSION

The presentation of data with descriptive statistics aims to give a general description of the percentage of the birth fertility of people living in the Sumbagsel region (Y) In 2020, the percentage of the age of marriage was first age (X_1), the percentage of women 15-49 years who do not use KB / traditional tools (X_2), the number of KB participants (X_3), the number of feet of fertile age (X_4), the average percentage of the school's long (X_5), the amount of perpaitable expenditure (X_6). The amount of data used in this study was as many as 60, consisting of districts and cities in the Sumbagsel region.

Tabel 1. Statistics Descriptive

Variable	Minimum	Median	*Mean*	Maximum	Standard Deviation
Y	33,490	66,100	64,500	87,540	9,782
X_1	6,180	15,230	15,240	28,090	4,430
X_2	14,830	29,160	31,310	58,160	10,058
X_3	12.308,000	42.727,000	60.964,000	246.345,000	51.565,970
X_4	9,000	55.263,000	74.881,000	345.802,000	73.043,590
X_5	6,350	7,865	8,209	11,790	1,158
X_6	7.892,000	10.220,000	10.499,000	15.663,000	1.699,560

Based on Table 4.1, average birthday living (Y) by 2020 will be 64,500%, with the lowest percentage of 33,490% in the city of Pangkal Pinang and the highest percentage of the baby's birth of the highest life of 87.540% in the city of Tanggamus, with the diversity of the baby's diet living

in every district or city in the region of Sumbagsel by 9.782%. Then the average percentage of the age of marriage is first under age (X_1) By 2020 that is 15.240%. The percentage of the first major marriage age was 28.090% in Kerinci District, while the percentage of the first marriage age was below the lowest of 6,180% in the city of Jambi, with a diversity rate of 4,430%.

At the percentage of women 15-49 years old who do not use KB or Traditional tools (X_2) Having a diversity rate of 10.058% and an average of 31.310%, with the lowest percentage value of 14.830% in the rotary of the rotary day and the highest percentage value of 58,160% found in the Regency of the Belitung East. Then the diversity rate on the number of participant KB participants (X_3) 51.565.970, with the average user of KB active of 60.964.000 It has a minimum value of 12.308.000 people in the full river district and a maximum population of 246.345.000 people in Central Lampung Regency.

As for the average PUS (X_4) in the Sumbagsel region of 74.881.000 with the level of diversity of PUS 73.043.590 and It has a minimum value of 9,000 pairs contained in West Bangka regency while the maximum value of 345.802.000 pairs contained in Central Lampung Regency. Then the diversity rate of the RSL percentage (X_5) of 1.156% with a minimum value of 6.350% there is in Tanjung Jabung Timur Regency and the maximum value of 11,790% found in Bengkulu City and the average percentage of RLS in the Sumbagsel region of 8.209%. Average perpendicular expenditure (X_6) In the Sembagsel region of 10.499.000 with minimum perkapita expenditure of 7.892.000 in Pesargan District while the perpendapa expenditure Maximum of 15.663.000 in the Pangkal Pinang city and the diversity of perpenditable expenditure amount 1.699.563.

Tabel 2. Test Linearity and Identification Data Pattern.

Variable	P-*Value*	Information
X_1 terhadap Y	0,086	Nonlinier
X_2 terhadap Y	0,562	Nonlinier
X_3 terhadap Y	0,365	Nonlinier
X_4 terhadap Y	0,233	Nonlinier
X_5 terhadap Y	0,000	Linier
X_6 terhadap Y	0,000	Linier

Based on the result of the linearity test and the identification of data patterns with a *scatterplot,* the variable $X_1, X_2, X_3, X_4$and X_5 it is a nonparametric component because these four variables do not contain previous function-form information that can be shown from data patterns that do not form certain patterns. Where variables X_6 are a parametric component for forming a linear data pattern. Election Optimum Bandwidth

Tabel 3. Test Linearity and Identification Data Pattern.

No	*Bandwidth* (h)	GCV
1	0,4900	0,0000
2	0,5000	0,0001
3	0,0000	0,5100

4	0,0100	0,5200
5	0,0200	0,5300
6	0,0300	0,5400
7	0,0400	0,5500
8	0,0500	0,5600
9	0,0600	0,5700
10	0,0700	0,5800

Based on table 4.5, it can be seen that the minimum GCV value is equal 0,0000 there is on *bandwidth* 0.490 so$h$ =0.490 is the optimal *bandwidth*. The amount of value *bandwidth* Optimum is then used in the kernel method by means of subtituting the value *bandwidth* on the estimator *Nadaraya-Watson*. Based on the election *bandwidth* Optimum using GCV approach then obtained value *bandwidth* Optimum of 0.490 and by using variables Nonparameteric is the fertility of living babies of life, the percentage of age of marriage is first age (X_1) ,Percentage of women 15-49 years who do not use KB / Traditional tools (X_2), the number of participant KB participants (X_3), the number of feet of fertile age (X_4), the estimated result is obtained by the mesignitual kennel function of estimator as *Nadaraya-Watson*as follows:

$$\hat{m}(x_i) = \frac{\sum_{i=1}^{n}\prod_{j=1}^{d}(\frac{1}{\sqrt{2\pi}}exp\,(-\frac{1}{2}(\frac{x_{ij}-x_j}{h_j})^2)Y_i}{\sum_{i=1}^{n}\prod_{j=1}^{d}(\frac{1}{\sqrt{2\pi}}exp\,(-\frac{1}{2}(\frac{x_{ij}-x_j}{h_j})^2)}$$

So it is obtained the following equation:

$$\hat{m}(x_i) = \frac{\sum_{i=1}^{60}\prod_{j=1}^{5}(\frac{1}{\sqrt{2\pi}}exp\,(-\frac{1}{2}(\frac{x_{ij}-x_j}{0{,}490})^2)Y_i}{\sum_{i=1}^{60}\prod_{j=1}^{5}(\frac{1}{\sqrt{2\pi}}exp\,(-\frac{1}{2}(\frac{x_{ij}-x_j}{0{,}490})^2)}$$

After the kernel estimation, *Nadaraya-Watson* With the Gaussian function, the form of the nonparametric regression model as follows:

$$Y_i = m(x_i) + \varepsilon_i$$

$$Y_i = \frac{\sum_{i=1}^{60}\prod_{j=1}^{5}(\frac{1}{\sqrt{2\pi}}exp\,(-\frac{1}{2}(\frac{x_{ij}-x_j}{0{,}490})^2)Y_i}{\sum_{i=1}^{60}\prod_{j=1}^{5}(\frac{1}{\sqrt{2\pi}}exp\,(-\frac{1}{2}(\frac{x_{ij}-x_j}{0{,}490})^2)} + \varepsilon_i$$

$$Y_i = \frac{\sum_{i=1}^{60}\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_{i1}-x_1}{0{,}490}\right)^2\right)Y_i\right)}{\sum_{i=1}^{60}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_{i1}-x_1}{0{,}490}\right)^2\right)} + \frac{\sum_{i=1}^{60}\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_{i2}-x_2}{0{,}490}\right)^2\right)Y_i\right)}{\sum_{i=1}^{60}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_{i2}-x_2}{0{,}490}\right)^2\right)} + \frac{\sum_{i=1}^{60}\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_{i3}-x_3}{0{,}490}\right)^2\right)Y_i\right)}{\sum_{i=1}^{60}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_{i3}-x_3}{0{,}490}\right)^2\right)} + \frac{\sum_{i=1}^{60}\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_{i4}-x_4}{0{,}490}\right)^2\right)Y_i\right)}{\sum_{i=1}^{60}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_{i4}-x_4}{0{,}490}\right)^2\right)} + \frac{\sum_{i=1}^{60}\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_{i5}-x_5}{0{,}490}\right)^2\right)Y_i\right)}{\sum_{i=1}^{60}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x_{i5}-x_5}{0{,}490}\right)^2\right)} + \varepsilon_i$$

The kernel nonparametric regression model shows the form of the weight of the Gaussian kernel function with a scaling size of the weighting function of 0.490, which means that the estimated curve fits the fertility data pattern in the optimum weight measure of 0.490. If we do not use the optimum weight, the following conditions will occur: if the weight size used is less than 0.490 then the estimated curve will undersmooth, if the weight size used is greater than 0.490 then the estimated curve will oversmooth.

CONCLUSION

Based on the non-parametric regression analysis of the kernel estimated with the estimator *Nadaraya-Watson* In the case of fertility in the Sumbagsel region by 2020, we will obtain a value *bandwidth* optimum of 0.490 and the value of the determination coefficient of 99.65%, and the independent variable that significantly influences on the bound variable is variable X_1,X_3,X_4, and X_5that is the percentage of women who had married first age, the number of participant KB participants, the number of feet of fertile age, and the average percentage of the school's sstudent's.

REFERENCE

[1] N. E. Chandra, S. Haryatmi, and Z. Zulaela, "Regresi Nonparametrik Kernel Adjusted," *J. Ilm. Mat. dan Pendidik. Mat.*, vol. 7, no. 1, p. 1, 2015, doi: 10.20884/1.jmp.2015.7.1.2894.

[2] W. Hardle, "Applied Nonparametric Methods," *Handb. Econom.*, vol. IV, no. 26, pp. 2295–2339, 1992.

[3] B. 2010, *Penduduk Indonesia Penduduk Indonesia Hasil SP2010*. Jakarta: Badan Pusat Statistik, 2023.

[4] BPS Indonesia, *Statistik Indonesia Statistical Yearbook Of Indonesia 2020*, vol. 1101001. Jakarta: Badan Pusat Statistik, 2020.

[5] P. rismawati Arum, "Analisis Faktor-Faktor yang Mempengaruhi Jumlah Penduduk di Kota Semarang Menggunakan Metode Regresi Data Panel," *J Stat. J. Ilm. Teor. dan Apl. Stat.*, vol. 12, no. 2, pp. 36–41, 2019, doi: 10.36456/jstat.vol12.no2.a2227.

[6] B. 2010, *Fertilitas Penduduk Indonesia*. Jakarta: Badan Pusat Statistik.

[7] M. Abdy, "Tinjauan Singkat Tentang Regresi Parametrik dan Regresi non Parametrik,"

Saintifik, vol. 5, no. 1, pp. 58–62, 2019, doi: 10.31605/saintifik.v5i1.199.

[8] S. Alfiani and P. R. Arum, "Pemodelan Pertumbuhan Ekonomi di Jawa Barat Menggunakan Metode Geographically Weighted Panel Regression," *J Stat.*, vol. 15, no. 2, pp. 219–227, 2022.

[9] N. Salam, Y. Sukmawaty, and A. Halida, "Estimasi Model Regresi Nonparametrik Dengan Metode B-Spline," vol. 3, no. 7, pp. 6843–6848, 2022.

[10] I. Puspitasari, Suparti, and Y. Wilandari, "Analisis Indeks Harga Saham Gabungan (IHSG) dengan Menggunakan Model Regresi Kernel," *J. Gaussian*, vol. 1, no. 1, pp. 93–102, 2012.

[11] A. H. Maharani, H. Yozza, and Y. Asdi, "Pemodelan Berat Badan Balita dengan Menggunakan Regresi Kernel," *J. Mat. UNAND*, vol. 4, no. 3, p. 31, 2019, doi: 10.25077/jmu.4.3.31-40.2015.

[12] R. Blundell, A. Duncan, and R. Blundell, "Board of Regents of the University of Wisconsin System Kernel Regression in Empirical Microeconomics Kernel Regression in Empirical Microeconomics," vol. 33, no. 1, pp. 62–87, 2015.

[13] Istiqomatul Fajriyah Yuliati and P. Sihombing, "Pemodelan Fertilitas Di Indonesia Tahun 2017 Menggunakan Pendekatan Regresi Nonparametrik Kernel dan Spline," *J. Stat. dan Apl.*, vol. 4, no. 1, pp. 48–60, 2020, doi: 10.21009/jsa.04105.

[14] W. Hardle, "Applied Nonparametric Regression.," *Biometrics*, vol. 50, no. 2, p. 592, 1994, doi: 10.2307/2533418.

Model Persamaan Struktural Faktor – Faktor Yang Mempengaruhi Kepuasan Masyarakat Dalam Pemeriksaan Kesehatan di UPTD Puskesmas Pasir Putih Sawangan Depok

Nisa Utari [(1)], Besse Arnawisuda Ningsi [(2),] Irvana Arofah [(3)]

[1,2,3]Program Studi Matematika, FMIPA, Universitas Pamulang

Jl. Raya Puspitek, Buaran, Kec. Pamulang, Kota Tangerang Selatan, Banten 15310

e-mail: nisautarii@gmail.com [(1)], dosen00205@unpam.ac.id[(2)], dosen00351@unpam.ac.id[(3)]

ABSTRAK

Salah satu faktor penting kepuasan masyarakat pada layanan kesehatan adalah mutu layanan. Mutu pelayanan yang berhubungan dengan medis, mewakili tingkat kesempurnaan pelayanan medis melalui terciptanya rasa kepuasan pada setiap masyarakat. Pada penelitian ini kepuasan masyarakat diukur dengan adanya mutu pelayanan yang mempunyai kaitan dengan kualitas pelayanan dan kompetensi pegawai di Puskesmas Pasir Putih kota Depok. Penelitian ini bertujuan untuk mengetahui pengaruh langsung kualitas pelayanan terhadap kepuasaan masyarakat, untuk mengetahui pengaruh langsung kompetensi pegawai terhadap kepuasan masyarakat, untuk mengetahui pengaruh tidak langsung antara kualitas pelayanan terhadap kepuasan masyarakat melalui kompetensi pegawai, dan pengaruh total antara kualitas pelayanan, kompetensi pegawai terhadap kepuasan masyarakat. Jumlah sampel pada penelitian ini adalah 379 responden yang ditentukan dengan menggunakan teknik *accidental sampling*, dimana responden yang dipilih adalah masyarakat yang mendapatkan pelayanan kesehatan di UPTD Puskesmas Pasir Putih. Teknik analisis yang digunakan adalah *Structural Equation Modelling (SEM).* Berdasarkan hasil analisis diperoleh bahwa terdapat pengaruh langsung kualitas pelayanan terhadap kepuasan masyarakat, terdapat pengaruh langsung kompetensi pegawai terhadap kepuasan masyarakat, terdapat pengaruh tidak langsung antara kualitas pelayanan masyarakat melalui kompetensi pegawai, dan terdapat pengaruh total antara kualitas pelayanan dan kompetensi pegawai terhadap masyarakat di UPTD Puskesmas Pasir Putih.

Kata kunci: Kepuasan Masyarakat, Kompentensi Pegawai, Kualitas Pelayanan,

ABSTRACT

One important factor in public satisfaction with health services is service quality. The quality of medical-related services represents the level of perfection of medical services by creating a sense of satisfaction in every community. In this research, community satisfaction is measured by the quality of service, which is related to the quality of service and employee competency at the UPTD Puskesmas Pasir Putih, Depok City. This research aims to determine the direct influence of service quality on community satisfaction, the direct influence of employee competence on community satisfaction, the indirect influence of service quality on community satisfaction through employee competency, and the total influence of service quality and employee competency on community satisfaction. The number of samples in this study was 379 respondents who were determined using the accidental sampling technique. The respondents selected were people who received health services at the UPTD Puskesmas Pasir Putih. The analysis technique used is structural equation modeling (SEM). Based on the results of the analysis, it was found that there was a direct influence of service quality on community satisfaction, there was a direct influence of employee competence on community satisfaction, there was an indirect influence between the quality of community service through employee competency, and there was a total influence between service quality and employee competency on the UPTD Puskesmas Pasir Putih.

Keywords: ***Community Satisfaction, Employee Competency, Service Quality***

PENDAHULUAN

Kepuasan merupakan perasaaan seseorang mengenai kesenangan atau kepuasan atau kekecewaan yang diperoleh dari hasil membandingkan penampilan produk yang telah disediakan dengan harapan pelanggan. Jadi dapat disimpulkan bahwa kepuasan adalah perasaan senang atau kecewa yang dimiliki seseorang berdasarkan perbandingan antara kenyataan yang diperoleh dengan harapan yang dinginkan [16]. Karena kepuasaan masyarakat tergantung pada kualitas pelayanan. Kepuasan masyarakat adalah salah satu tujuan dari peningkatan kualitas layanan kesehatan. Ada bukti bahwa masyarakat atau orang yang puas dengan layanan kesehatan yang terorganisir cenderung mematuhi saran, setia atau berpegang teguh pada rencana perawatan yang telah disepakati [10]. Konsep kualitas pelayanan yang dihubungkan dengan kepuasan pasien ditentukan oleh lima faktor yang disebut dengan kualitas pelayanan "SERVQUAL". Kelima faktor tersebut yaitu *responsiveness* (daya tanggap), *assurance* (jaminan), *tangible* (bukti fisik), *empathy* (empati), dan *reliability* (keandalan). Mutu pelayanan yang berhubungan dengan medis, mewakili tingkat kesempurnaan pelayanan medis melalui terciptanya rasa kepuasan pada setiap pasien. Semakin besar kepuasaan maka semakin baik bantuan medis. Pada kenyataannya masih banyak pengaduan (keluhan) atau pernyataan ketidakpuasan masyarakat terhadap kualitas pelayanan. Pelayanan kesehatan menjadi titik strategis yang mana kepercayaan masyarakat secara luas kepada pihak puskesmas dipertaruhkan [5]. Berbagai kebijakan nasional dalam rangka meningkatkan kualitas pelayanan tersebut telah memberikan fondasi bagi puskesmas untuk melakukan upaya nyata dalam mereformasi pelayanan untuk meningkatkan kepuasan masyarakat [7]. Beberapa penelitian terdahulu mengenai faktor faktor yang mempengaruhi kepuasan masyarakat, diantaranya Sefri Imanuel Fallo menganalisis tentang pengaruh kepuasan pasien pada Puskesmas Halmahera Kota Semarang. Hasil penelitian yang diperoleh yaitu indikator yang mempengaruhi kepuasan pasien yaitu kualitas pelayanan dan kompetensi pegawai. Pada penelitian ini dimensi yang paling berpengaruh signifikan dalam membentuk variabel kualitas pelayanan [6]. Selanjutnya B. A. Ningsi, menurut menunjukkan bahwa kualitas produk berpengaruh terhadap kepuasan pelanggan yang artinya tidak semua pelanggan akan merasa puas jika hanya dengan kualitas produk, melainkan harus ada faktor lain. Pada penelitian ini indikator yang paling berpengaruh positif terhadap kualitas produksi sehingga kepuasan pelanggan dapat dicapai adalah indikator *performance*. Sedangkan indikator yang kurang berpengaruh terhadap kualitas produksi sehingga menyebabkan kepuasan pelanggan kurang tercapai adalah indikator *design* [8].

Dalam penelitian ini akan dilakukan analisis faktor-faktor yang mempengaruhi kepuasan masyarakat Puskesmas Pasir Putih. Analisis akan dilakukan dengan menggunakan model persamaan struktural atau biasa disebut dengan *Structural Equations Model* **(**SEM). SEM merupakan sekumpulan cara atau teknik statistik yang pengujiannya dapat memungkinkan menganalisis serangkaian hubungan secara simultan. SEM merupakan gabungan dari metode analisis faktor dan regresi [7]. Penelitian ini bertujuan untuk mengetahui pengaruh langsung kualitas pelayanan terhadap kepuasaan masyarakat, untuk mengetahui pengaruh langsung kompetensi pegawai terhadap kepuasan masyarakat, untuk mengetahui pengaruh tidak langsung antara kualitas pelayanan terhadap kepuasan masyarakat melalui kompetensi pegawai, dan pengaruh total antara kualitas pelayanan,kompetensi pegawai terhadap kepuasan masyarakat.

METODE

Variabel Penelitian

Variabel adalah suatu konsep tentang atribut ataupun sifat yang terdapat pada subjek penelitian yang beraneka ragam secara kuantitatif maupun kualitatif [1][17]. Variabel dalam penelitian ini yaitu kualitas pelayanan, kompetensi pegawai, dan kepuasan masyarakat.

Tabel 1. Variabel dan Indikator Penelitian

Variabel Laten		Indikator	Kode
Variabel Eksogen	Kualitas Pelayanan	*Tangible* (Bukti Fisik)	KP1
		Responsiveness (Daya Tanggap)	KP2
		Realiability (Keandalan)	KP3
		Assurance (Jaminan)	KP4
		Empathy (Empati)	KP5
	Kompetensi Pegawai	Pengetahuan (*Knowlegde)*	KOP1
		Pemahaman *(Understanding)*	KOP2
		Kemampuan (*Skill)*	KOP3
		Nilai (*Value)*	KOP4
		Sikap (*Attitude)*	KOP5
		Minat (*Interest)*	KOP6
Variabel Endogen	Kepuasan Masyarakat	Keterampilan Dan Komunikasi	KM1
		Disiplin Kerja Dan Fasilitas	KM2
		Biaya Perawatan Dan Efisien Kerja	KM3

Teknik Pengambilan sampel

Populasi penelitian yang dipakai dalam penelitian ini adalah seluruh masyarakat kelurahan Pasir Putih. Dikarenakan peneliti tidak memungkinkan mengobservasi semua anggota populasi, maka dari itu dibuat suatu perwakilan populasi yang disebut sampel. Untuk mendapatkan sampel yang dapat mewakili populasi maka metode pengambilan sampel dalam penelitian ini menggunakan metode *non probability sampling*, dan teknik yang digunakan untuk mengambil sampel dalam penelitian ini adalah *accidental sampling.* Karena populasi pada penelitian ini bersifat homogen yaitu masyarakat yang berada di kelurahan Pasir Putih dengan jumlah masyarakat sebanyak 27.219 jiwa [3]. *Accidental sampling* adalah mengambil responden sebagai sampel berdasarkan kebetulan, yaitu siapa saja yang secara kebetulan bertemu dengan peneliti dapat digunakan sebagai sampel, bila orang yang kebetulan ditemui cocok sebagai sumber data dengan kriteria utamanya adalah masyarakat di kelurahan Pasir Putih yang sudah mendapatkan pelayanan kesehatan di UPTD Puskesmas Pasir Putih. Dalam menentukan jumlah sampel pada penelitian ini dengan mengembangkan rumus yang dihasilkan *National Educational Association* (NEA), dalam tabel Krejcie dan Morgan (1970). Jadi, jika jumlah N sebesar 30.000 orang, maka sampel yang dibutuhkan adalah 379 [1]. Dengan cara tersebut tidak perlu dilakukan perhitungan yang rumit. Krejcie dalam melakukan perhitungan sampel didasarkan atas kesalahan 5 %. Jadi sampel yang diperoleh itu mempunyai kepercayaan 95% terhadap populasi.

Teknik Pengumpulan Data

Teknik pengumpulan data dalam penelitian ini menggunakan instrumen sampel. kuesioner merupakan metode pengumpulan data yang dilakukan dengan memberi pertanyaan atau pernyataan tertulis kepada responden untuk dijawabnya. Jawaban kuesioner yang digunakan yaitu dengan skala pengukuran likert yang terdiri dari: Sangat Sesuai (SS), Sesuai (S), Ragu-Ragu (RR), Tidak Sesuai (TS), dan Sangat Tidak Sesuai (STS) [14].

Teknik Analisis Data

Analisis data yang digunakan dalam penelitian ini menggunakan metode *structural equation modelling* (SEM) adalah menggunakan teknik statistic multivariat yang merupakan kombinasi antara analisis faktor dan analisis regresi (korelasi) dengan bantuan *software* AMOS untuk menganalisis pengaruh antar variabel. Model penelitian yang digunakan dalam penelitian ini bersifat first order [6]. Dimana kualitas pelayanan, kompetensi pegawai, dan kepuasan masyarakat diukur secara reflektif. Analisis *Structural Equations Model* (SEM) berawal dari penggabungan antara sistem persamaan simultan, atau analisis jalur, atau analisis regresi dengan analisis faktor. Data variabel laten didapatkan dari penggunaan metode analisis faktor. Proses pendugaan parameterdan pengujian berdasarkan pada konsep matriks varians-kovarians, sehingga sering juga disebut dengan SEM berbasis kovarians (*covariance).* SEM berbasis kovarians hanya dapat digunakan pada model indikator reflektif dikarenakan metode tersebut hanya menyediakan analisis faktor untuk mendapatkan data variabel laten . Variabel dalam SEM tidak berbeda jauh dengan variabel dari metode analisis lainnya [2]. Istilah variabel dalam SEM berbeda dengan istilah variabel dalam analisis regresi. Dalam SEM, variabel X dan Y merupakan lambing untuk indikator dari variabel laten. Variabel laten adalah variabel yang tidak bisa diukur secara langsung oleh karena itu diperlukan indikator atau variabel manifest untuk menguur laten. Variabel laten sendiri dibedakan menjadi 2, yaitu variabel laten endogen dan variabel eksogen. Selain dibedakan berdasarkan sifat pengaruhnya, variabel dalam SEM juga dikelompokkan berdasarkan bentuk datanya atau skala. Skala Nominal, Skala Ordinal, dan Skala Interval [4].

Tahapan analisis

Berikut langkah-langkah teknik analisis data yaitu:

1. mengembangan model berdasarkan teori;
2. mengkontruksi diagram jalur untuk hubungan kausal;
3. mengkonversi diagram jalur ke dalam Persamaan seperti berikut ini [12]:
   a. Persamaan Struktural
      Dalam menyusun persamaan structural, persamaan yang dibangun terdiri dari persamaan struktural (*structural equation)* dan persamaan model pengukuran (*measurement model)*. Persamaan struktural dirumuskan untuk menyatakan kausalitas antara berbagai konstruk. Pada dasarnya persamaan structural tersusun atas formula sebagai berikut [12]:
      Variabel Endogen1 = Variabel Eksogen + Variabel Endogen2 + Error
   b. Persamaan Model Pengukuran
      Dalam membuat persamaan model pengukuran hanya melibatkan indikator pengukur kontruk [12].
      Model pengukuran untuk variabel eksogen:
      Xn = $\lambda_n \xi_n + \delta_n$
      Xm = $\lambda_m \xi_m + \delta_m$
      Model pengukuran untuk variabel endogen:
      Yn = $\lambda_n \eta_n + e_n$
      Ym = $\lambda_m \eta_m + e_m$
4. kemudian langkah selanjutnya memilih input matriks dan estimasi model;
5. menilai identifikasi model struktural;
6. evaluasi kecocokan model;
7. interpretasi dan modifikasi model.

HASIL DAN PEMBAHASAN

Pengujian Validitas dan Reliabilitas

Dalam penelitian ini terdapat beberapa pengujian diantaranya uji valiitas dan uji reliabilitas. Berikut ini hasil pengujian validitas dan reliabilitas yang bisa dilihat dari tabel 2 dan 3 [6].

Tabel 2. Hasil Uji Validitas

Variabel	Indikator	r tabel	Estimate	Keterangan
Kepuasan Masyarakat	KM1	0,098	0,936	Valid
	KM2	0,098	0,829	Valid
	KM3	0,098	0,916	Valid
Kualitas Pelayanan	KP1	0,098	0,961	Valid
	KP2	0,098	0,954	Valid
	KP3	0,098	0,915	Valid
	KP4	0,098	0,837	Valid
	KP5	0,098	0,826	Valid
Kompetensi Pegawai	KOP1	0,098	0,945	Valid
	KOP2	0,098	0,964	Valid
	KOP3	0,098	0,965	Valid
	KOP4	0,098	0,991	Valid
	KOP5	0,098	0,886	Valid
	KOP6	0,098	0,885	Valid

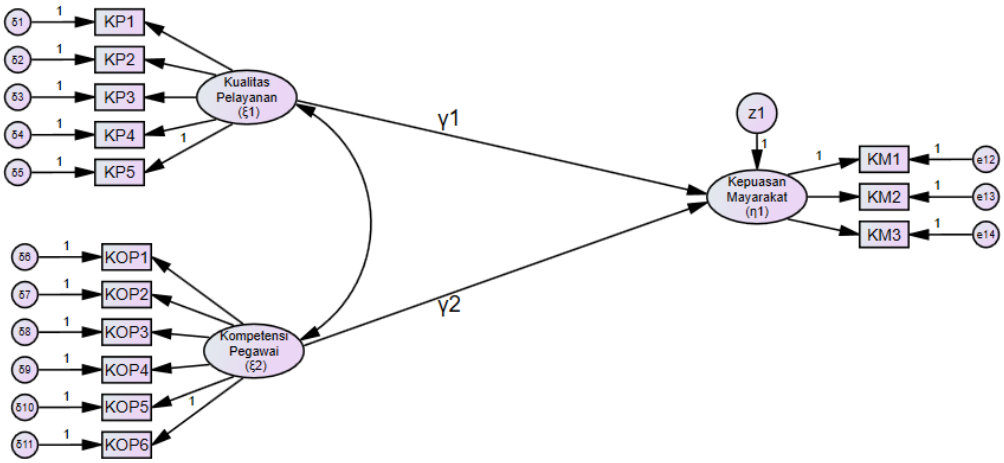
Berdasarkan hasil perhitungan uji validitas variabel pertama yaitu kepuasan masyarakat dimana indikator KM1, KM2, dan KM3 menunjukkan hasil yang valid karena r hitung > r tabel. Selanjutnya begitupun pada variabel kualitas pelayanan dan kompetensi pegawai dimana indikator KP1,KP2,KP3,KP4,KP5,KOP1,KOP2, KOP3,KOP4,KOP5, dan KOP6 menunjukkan hasil yang valid karena r hitung > r tabel [6]. Sehingga dapat disimpulkan bahwa semua indikator dinyatakan valid. Selanjutnya Pengujian Reliabilitas berikut ini :

Tabel 3. Hasil Uji Reliabilitas

Variabel	Indikator	CR	AVE	Keterangan
Kualitas Pelayanan	KP1	0,95522	0,81074	Reliabel
	KP2			
	KP3			
	KP4			
	KP5			
Kompetensi Pegawai	KOP1	0,97855	0,88397	Reliabel
	KOP2			
	KOP3			
	KOP4			
	KOP5			
	KOP6			
Kepuasan Masyarakat	KM1	0,92324	0,80080	Reliabel
	KM2			
	KM3			

Berdasarkan hasil perhitungan data reliabilitas, dinyatakan bahwa data yang diuji memiliki nilai *Consturct Reliability* (CR) menunjukkan hasil yang lebih besar dari batas kritis yang telah ditentukan yaitu 0,70 [6]. Dari hasil tabel di atas dapat dilihat nilai *Consturct Reliability* (CR) variabel kualitas pelayanan sebesar 0,95522 nilai *Consturct Reliability* (CR) [6] dari variabel kompetensi pegawai 0,97855, serta nilai *Consturct Reliability* (CR) dari variabel kepuasan masyarakat sebesar 0,92324. Dan nilai *Average Variance Extract* (AVE) menunjukkan hasil yang lebih besar dari batas kritis yang telah ditentukan yaitu 0,50 [6], dimana nilai *Average Variance Extract* (AVE) dari variabel kualitas pelayanan sebesar 0,81074. nilai *Average Variance Extract* (AVE) dari variabel kompetensi pegawai sebesar 0,88397 dan nilai *Average Variance Extract* (AVE) dari variabel kepuasan masyarakat sebesar 0,80080 [6]. Hal ini berarti bahwa secara umum instrumen yang telah dibuat sudah valid dan reliabel sehingga dapat digunakan dalam mengukur faktor-faktor yang mempengaruhi kepuasan masyarakat dalam pemeriksaan kesehatan.

Diagram Jalur



Gambar 1. Diagram Jalur

Dari diagram jalur di atas kemudian dilakukan konversi ke dalam persamaan, baik persamaan struktural maupun persamaan model pengukuran [12]. Berikut model persamaan struktural dan persamaan pengukuran:

a. Persamaan Model Struktural

Pada dasarnya persamaan struktural tersusun dalam formula sebagai berikut:

$$KM = \gamma_1 KP + \gamma_2 KOP + z_1$$

Dari persamaan di atas menunjukkan bahwa variabel kepuasan masyarakat dipengaruhi oleh variabel kualitas pelayanan dan kompetensi pegawai.

b. Model Pengukuran

Model pengukuran variabel eksogen:

$KP1 = \lambda_1 \xi_1 + \delta_1$
$KP2 = \lambda_2 \xi_1 + \delta_2$
$KP3 = \lambda_3 \xi_1 + \delta_3$
$KP4 = \lambda_4 \xi_1 + \delta_4$
$KP5 = \lambda_5 \xi_1 + \delta_5$

Dari persamaan di atas menunjukkan bahwa variabel kualitas pelayanan diukur melalui 5 indikator yaitu KP1, KP2, KP3, KP4 dan KP5.

KOP1 = $\lambda_6\xi_2+\delta_6$
KOP2 = $\lambda_7\xi_2+\delta_7$
KOP3 = $\lambda_8\xi_2+\delta_8$
KOP4 = $\lambda_9\xi_2+\delta_9$
KOP5 = $\lambda_{10}\xi_2+\delta_{10}$
KOP6 = $\lambda_{11}\xi_2+\delta_{11}$

Dari persamaan di atas menunjukkan bahwa variabel kompetensi pegawai diukur melalui 6 indikator yaitu KOP1, KOP2, KOP3, KOP4, KOP5 dan KOP6.

Model pengukuran endogen:

KM1 = $\lambda_{12}\eta_1 + e_{12}$
KM2 = $\lambda_{13}\eta_1 + e_{13}$
KM3 = $\lambda_{14}\eta_1 + e_{14}$

Sedangkan variabel kepuasan masyarakat diukur melalui 3 indikator yaitu KM1, KM2 dan KM3.

Pengujian Hipotesis

Tabel 4. Uji Hipotesis secara Langsung

	Path	**C.R.**	**P**	**Hasil**
H1	Kualitas Pelayanan → Kepuasan Masyarakat	19,323	0,001	Diterima dan Signifikan
H2	Kompetensi Pegawai → Kepuasan Masyarakat	-8,464	0,001	Diterima dan signifikan

Pengaruh langsung yang terjadi antara kualitas pelayanan terhadap kepuasan masyarakat di UPTD Puskesmas Pasir Putih ialah semakin tinggi tingkat kualitas pelayanan semakin tinggi pula tingkat kepuasan masyarakat dalam melakukan pemeriksaan kesehatan di UPTD Puskesmas Pasir Putih. Dimana kualitas pelayanan di puskesmas tersebut memberikan pelayanan yang sangat positif, kenyamanan, keamanan, keramahan, dan kecepatan yang dilakukan sangatlah baik dan membuat masyarakat yang melakukan pemeriksaan di puskesmas tersebut puas dan tidak kecewa atas pelayanan yang diberikan oleh pihak puskesmas. Hasil pengujian menunjukkan bahwa nilai *P-value* kualitas pelayanan terhadap kepuasan masyarakat menghasilkan nilai yang positif, artinya kualitas pelayanan memiliki pengaruh positif secara langsung terhadap kepuasan masyarakat sebesar 0,001. Dari hasil ini dinyatakan bahwa *P-value < 0,05* maka dapat disimpulkan bahwa terdapat pengaruh langsung kualitas pelayanan terhadap kepuasan masyarakat sehingga hipotesis diterima. Pengaruh langsung yang terjadi antara kompetensi pegawai terhadap kepuasan masyarakat di UPTD Puskesmas Pasir Putih yaitu dengan melihat pengetahuan, pemahaman, kemampuan, dan sikap pegawai. Kompetensi pegawai memang cukup baik dan berpengaruh yang positif, walaupun masih ada msayarakat yang mengeluh dengan kompetensi pegawai terutama dengan sikap pegawai baik dalam melayani ataupun sikap pegawai dalam memeriksa masyarakat. Hasil pengujian menunjukkan bahwa nilai *P-value* kompetensi pegawai terhadap kepuasan masyarakat menghasilkan nilai yang positif, artinya kompetensi pegawai memiliki pengaruh positif secara langsung terhadap kepuasan masyarakat sebesar 0,001. Dari hasil ini dinyatakan bahwa *P-value < 0,05* artinya dapat disimpulkan bahwa terdapat pengaruh langsung kompetensi pegawai terhadap kepuasan masyarakat akan tetapi signifikan.

Tabel 5. Uji Hipotesis secara Tidak Langsung

	Pengaruh tidak langung	
	Kompetensi Pegawai	**Kualitas Pelayanan**
Kepuasan Masyarakat	0,000	0,000

Pengaruh tidak langsung antara kualitas pelayanan terhadap kepuasan masyarakat melalui kompetensi pegawai di UTPD Puskesmas Pasir Putih diperoleh bahwa kualitas pelayanan akan memiliki pengaruh akan tetapi secara tidak langsung melalui kompetensi pegawai. Dilihat dari kompentensi pegawai memiliki pengaruh yang cukup positif juga walaupun ada beberapa masyarakat masih ada yang merasa kecewa dengan beberapa sikap pegawai yang kurang efisien akan tetapi kualitas pelayanan memiliki pengaruh sangat positif dan masyarakat merasa puas dengan pelayanan yang ada di UPTD Puskesmas Dengan melihat hal tersebut menjadikan terjadinya pengaruh tidak langsung antara kualitas pelayanan terhadap kepuasan masyarakat melalui kompetensi pegawai.

Tabel 6. Uji Hipotesis secara Total

	Pengaruh total	
	Kompetensi Pegawai	**Kualitas Pelayanan**
Kepuasan Masyarakat	-0,673	1,681

Untuk menghitung pengaruh total secara langsung dan tidak langsung yaitu dengan melihat output AMOS *options total effect.* Dari tabel 4.11 menunjukkan bahwa hasil pengaruh total yang pertama yaitu kualitas pelayanan terhadap kepuasan masyarakat dengan nilai sebesar 1,681 yang dimana lebih besar dari probabilitas 0,05. Dan hasil pengaruh total yang kedua yaitu kompetensi pegawai terhadap kepuasan masyarakat dengan nilai sebesar -0,673 yang dimana lebih besar dari probabilitas 0,05 hanya saja bersifat negatif. Hanya saja masih ada masyarakat yang merasa kecewa akan sikap pegawai yang belum efisien. Pengaruh total antara kualitas pelayanan dan kompetensi pegawai terhadap kepuasan masyarakat yang terjadi di UPTD Puskesmas Pasir Putih yaitu dengan melihat kualitas pelayanan lebih memiliki pengaruh yang cukup dominan dan sangat positif dimana masyarakat begitu merasa puas akan pelayanan yang dilakukan di UPTD Pasir Putih. Baik keamanan, kenyamanan, keramahan, kecepatan bahkan fasilitas yang ada juga canggih, membuat masyarakat menjadi terjamin melakukan pemeriksaan kesehatan. Begitupun kompetensi pegawai juga berpengaruh terhadap kepuasan masyarakat hanya saja ada sedikit kekecewaan yang terjadi pada masyarakat akibat sikap pegawai yang masih ada beberapa tidak begitu memuaskan masyarakat baik sikap dalam melayani atau sikap dalam memeriksa sebagai tim dokter atau tenaga medis

KESIMPULAN DAN SARAN

Berdasarkan hasil analisis dapat ditarik kesimpulan bahwa terdapat pengaruh langsung kualitas pelayanan terhadap kepuasan masyarakat di UPTD Puskesmas Pasir Putih, terdapat pengaruh langsung kompetensi pegawai terhadap kepuasan masyarakat di UPTD Puskemas Pasir Putih, terdapat pengaruh tidak langsung antara kualitas pelayanan masyarakat melalui kompetensi pegawai di UPTD Puskesmas Pasir Putih dan terdapat pengaruh total antara kualtas pelayanan dan kompetensi pegawai terhadap masyarakat. Dan faktor yang paling mempengaruhi kepuasan masyarakat adalah kualitas pelayanan. Kualitas pelayanan disini begitu berpengaruh sangat positif dan masyarakat merasa puas dengan pelayanan yang ada di UPTD Puskesmas Pasir Putih.

UCAPAN TERIMA KASIH

Terima kasih kepada seluruh jajaran Program Studi Matematika Fakultas MIPA Universitas Pamulang dalam menunjang pelaksanaan penelitian ini. Dan terima kasih juga kepada pihak- pihak yang membantu dalam penelitian ini.

DAFTAR PUSTAKA

[1] H. Ahyar *et al.*, *Buku Metode Penelitian Kualitatif & Kuantitatif*, no. March. 2020.
[2] A. Amos, Junaidi, S.E., Ak., M.Ak., CA., Ph.D, "Aplikasi Amos dan Structural Equation Modeling (SEM)." Unhaspress, 2021.
[3] S. Amrina, "Cover Depan," *J. Pendidik. Teknol. dan Kejuru.*, vol. 23, no. 1, 2016, doi: 10.21831/jptk.v23i1.10441.
[4] I. Arofah, "METODE STATISTIKA." Unpampress, 2023.
[5] A. J. Djohan, "Faktor-Faktor yang Mempengaruhi Kepuasan dan Kepercayaan untuk Mencapai Loyalitas Pasien Rawat Inap Faktor-Faktor yang Mempengaruhi Kepuasan dan Kepercayaan untuk Mencapai Loyalitas Pasien Rawat Inap pada Rumah Sakit Swasta di Kota Banjarmasin," *J. Apl. Manaj.*, vol. 13, no. 66, hal. 257–271, 2015.
[6] S. Equation *et al.*, "Sefri Imanuel Fallo adanya variabel laten yang diukur oleh satu atau lebih variabel yang," vol. 16, no. 1, hal. 52–67, 2022.
[7] K. Masyarakat dan D. Pemeriksaan, "Analisis faktor-faktor yang mempengaruhi kepuasan masyarakat dalam pemeriksaan kesehatan dengan menggunakan structural equations model (sem)," 2023.
[8] B. A. Ningsi, L. Agustina, S. Matematika, F. Matematika, P. Alam, dan U. Pamulang, "Analisis Kepuasan Pelanggan Atas Kualitas Produk dan Pelayanan Dengan Metode SEM-PLS," vol. 2, no. 2, hal. 8–16, 2018.
[9] C. P. Raymond, S. Hatane, dan J. Hutabarat, "Analisis Kualitas Sumber Daya Manusia, Kualitas Pelayanan, Kinerja Organisasi, Kepercayaan Masyarakat dan Kepuasan Masyarakat (Studi Kasus : Dinas Kependudukan Dan Catatan Sipil Kabupaten Nabire)," *J. Teknol. dan Manaj. Ind.*, vol. 1, no. 1, hal. 1–8, 2018.
[10] S. Riyadi, A. Hermawan, dan U. Sumarwan, "Kepuasan Masyarakat terhadap Kualitas Pelayanan Kantor Pertanahan Kabupaten Indramayu," *J. Ilmu Kel. dan Konsum.*, vol. 8, no. 1, hal. 49–58, 2015, doi: 10.24156/jikk.2015.8.1.49.
[11] I. Sholihati, D. Handayani, dan M. S. Nusyirwan, "Magister Manajemen Rumah Sakit, Universitas Muhammadiyah Yogyakarta," *Persepsi Kepuasan Pasien Pada Kualitas Pelayanan Di Rumah Sakit Gigi Dan Mulut Univ. Muhammadiyah Yogyakarta*, 2013.
[12] S. Si, M. Si, A. Sauddin, S. Pd, dan M. Si, *Irwan, S.Si., M.Si. Adnan Sauddin, S.Pd., M.Si.*
[13] Y. Susanti, Y. Azis, dan D. Kusnadi, "Pengaruh *Appointment Registration System* terhadap Waktu Tunggu dan Kepuasan Pasien," *Global Medical & Health Communication (GMHC)*, vol. 3, no. 1. hal. 40, 2015. doi: 10.29313/gmhc.v3i1.1545.
[14] R. Veronica, "Analisis Faktor yang Mempengaruhi Loyalitas Perawat di Rumah Sakit Tugu Ibu Depok Analysis on the Influencing Factors of Nurse Loyalty at Tugu Ibu Hospital Depok [1] Rina Veronica [1] Institut Kesehatan Indonesia Email : riena.veronica3@gmail.com Alamat Kores," *J. Manaj. Kesehat.*, vol. 6, no. 2, hal. 192–208, 2020.
[15] T. Wijaya dan S. Budiman, *PENELITIAN MANAJEMEN*.
[16] F. K and Maxima Ari Saktiono, "Tingkat Kepuasan Mahasiswa Terhadap Pelayanan Akademik Pada Program Diploma Pelayaran Universitas Hang Tuah Surabaya", J Statistika, vol. 13, no. 2, pp. 20–24, Dec. 2020.
[17] G. Anuraga, A. Indrasetianingsih, and M. Athoillah, "Pelatihan Pengujian Hipotesis Statistika Dasar dengan Software R," BUDIMAS: Jurnal Pengabdian Masyarakat, vol. 3, no. 2, 2021.

Analysis of the Timeliness of Graduation of FMIPA College KIP Students at Bengkulu University Using Binary Logistic Regression

Riki Crisdianto(1), Muhammad Fathi Abdillah Rakafalih(2), Alya Saputri(3), Laga Sopiansyah(4), Athaya Fairuzindah(5), Dyah Setyo Rini(6)

[1,2,3,4,5,6] S1-Statistics Study Program, Bengkulu University

Jl. WR. Supratman, Kandang Limun, Kec. Muara Bangka Hulu, Sumatera, Bengkulu 38371
e-mail: rikic628@gmail.com*(1), FathiFalih017@gmail.com(2), alyasaputri9a@gmail.com(3), laga9621@gmail.com(4), athayazindah@gmail.com(5), dyah.setyorini@unib.ac.id(6)

ABSTRAK

Penelitian-penelitian terdahulu yang membahas mengenai ketepatan waktu lulus mahasiswa pada umumnya lebih difokuskan pada populasi mahasiswa secara keseluruhan, tanpa mempertimbangkan pengaruh dari program beasiswa secara spesifik seperti KIP Kuliah. Penelitian ini bertujuan untuk mendapatkan model ketepatan waktu lulus mahasiswa dan mendapatkan faktor-faktor yang mempengaruhi ketepatan waktu lulus mahasiswa KIP kuliah FMIPA Universitas Bengkulu menggunakan metode Regresi Logistik Biner. Hasil dari penelitian ini diharapkan dapat memberikan informasi yang berharga bagi pengelola KIP kuliah Universitas Bengkulu dalam meningkatkan efektivitas program beasiswa KIP Kuliah dan membantu mahasiswa KIP kuliah FMIPA Universitas Bengkulu agar bisa lulus tepat waktu. Dalam penelitian ini digunakan metode Regresi Logistik Biner karena variable respon memiliki skala biner nominal. Berdasarkan hasil pembahasan penelitian yang dilakukan diperoleh model Regresi Logistik Biner dari ketepatan waktu lulus mahasiswa KIP kuliah FMIPA Universitas Bengkulu dengan faktor-faktor yang memiliki pengaruh signifikan terhadap ketepatan waktu lulus mahasiswa KIP kuliah FMIPA Universitas Bengkulu adalah Asal prodi S1-Fisika, S1-Biologi dan S1-Statistika dan IPK. Kemudian Hasil klasifikasi menggunakan Regresi Logistik Biner memiliki tingkat ketepatan klasifikasi sebesar 77.97%. Sehingga dapat disimpulkan bahwa dalam klasifikasi ketepatan waktu lulus mahasiswa FMIPA Universitas Bengkulu pada model Regresi Logistik Biner cukup baik.
Kata kunci: KIP Kuliah; Regresi Logistik Biner; *Odds Ratio*, Klasifikasi

ABSTRACT

Previous studies that discuss the timeliness of graduating students are generally more focused on the student population as a whole, without considering the influence of specific scholarship programs such as KIP Kuliah. This study aims to obtain a model of the timeliness of student graduation and obtain factors that affect the timeliness of graduating KIP Kuliah students of FMIPA Bengkulu University using the Binary Logistic Regression method. The results of this study are expected to provide valuable information for KIP Kuliah managers of Bengkulu University in improving the effectiveness of the KIP Kuliah scholarship program and helping KIP Kuliah students of FMIPA Bengkulu University to graduate on time. In this study, Binary Logistic Regression method is used because the response variable has a nominal binary scale. Based on the results of the discussion of the research conducted, a Binary Logistic Regression model of the timeliness of graduating KIP Kuliah students of FMIPA Bengkulu University is obtained with factors that have a significant influence on the timeliness of graduating KIP Kuliah students of FMIPA Bengkulu University is the origin of the S1-Physics, S1-Biology and S1-Statistics study programs and GPA. Then the classification results using Binary Logistic Regression have a classification accuracy rate of 77.97%. So it can be concluded that the classification of the timeliness of graduating students of FMIPA Bengkulu University in the Binary Logistic Regression model is good enough.
Keywords: *College KIP; Binary Logistic Regression; Odds Ratio, Classification*

INTRODUCTION

Higher education has an important role in developing quality and highly competitive human resources. The percentage of students graduating on time is one of the indicators of assessing the success and feasibility of a study program in the implementation of higher education [1]. The timeliness of student graduation has different criteria for each program available at the tertiary level. D3 (Diploma) program students are said to graduate on time if they can complete their studies in less than or equal to three years. Students of the S1 (Bachelor) program are said to graduate on time if they can complete their studies in less than or equal to four years. Students of the S2 (Master) program are said to graduate on time if they can complete a study of less or equal to two years and the S3 (Doctoral) program if they can complete a study of less or equal to three years [2]. The longer students complete their studies in college, the more they need a lot of money. One of the government programs that helps students in terms of costs to continue their education in higher education is the education fee for poor students with achievements (Bidikmisi) or what is now known as the Indonesia Smart College Card (KIP-K).

KIP-K is higher education assistance in the form of cash, expanding access, and learning opportunities from the government given to students who come from poor/vulnerable families to finance education [3]. The period of KIP-K provision for Diploma 3 programs is a maximum of 6 semesters, Bachelor programs are a maximum of 8 semesters and Doctoral programs are a maximum of 4 semesters. It is feared that if KIP-K students do not graduate on time or study beyond the KIP-K granting period, they will not be able to continue their studies due to financial constraints. So the timeliness of graduating KIP-K students is important and needs to be considered by study programs and universities in the implementation of higher education. This research is important to understand the factors that specifically affect the timeliness of graduating KIP-K students and can provide valuable information to related parties and also the students themselves, to improve the effectiveness of the KIP-K scholarship program.

Previous research conducted by Agwil, Fransiska, and Hidayati in 2020 discussed the timeliness of student graduation [2]. The research focused on the student population as a whole, without considering the influence of specific scholarship programs such as KIP-K. One of the public universities in Bengkulu province that has the trust to manage KIP-K is Universitas Bengkulu (UNIB). There are still many KIP-K UNIB students who do not graduate on time, especially in FMIPA UNIB. The method used to analyze the timeliness of graduating KIP-K students at FMIPA UNIB in this study is Binary Logistic Regression.

Binary Logistic Regression is one of the effective analysis methods in predicting binary variables, such as student graduation on time. So in this study, the Binary Logistic Regression method is applied in modeling the timeliness of student graduation where category 1 represents students who graduate on time (success) and category 0 represents students who do not graduate on time (failure). In addition, based on previous research related to the Binary Logistic Regression method, namely research by Nikie Ramsi Tamnge on the effect of service quality on student satisfaction at the Faculty of Teacher Training and Education, Muhammadiyah University of Surabaya which has a classification accuracy of 86% included in the Good Classification category [4].

LITERATURE REVIEW

Binary Logistic Regression

Logistic regression is a regression analysis technique used to explore the relationship between dichotomous response variables (have two categories) or polycotomous (have more than two categories) with a group of predictor variables that are continuous or categorical [5]. According to Hosmer & Lemeshow (2000) [6], Binary Logistic Regression is a statistical method used to find the relationship between the response variable (y) which has a nominal data scale (two categories or binary) with predictor variables (x) which are categorical or continuous.

The logistic regression model with p predictor variables is as follows:

$$\pi(x) = \frac{e^{(\beta_O+\beta_1 X_1+\cdots+\beta_p X_p)}}{1+e^{(\beta_O+\beta_1 X_1+\cdots+\beta_p X_p)}} \quad (1)$$

Description:

$\pi(x)$: probability of "success" or chance of a successful event

$X_1, \ldots, X_p$: Independent variable

β_O: Constant of the model

$\beta_O, \ldots, \beta_p$: Regression coefficient parameter.

The logit transformation model of $\pi(x)$ from the equation above can be written as follows:

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_O + \beta_1 X_1 + \cdots + \beta_p X_p \quad (2)$$

with g(X): the logit transformation of π(x).

Biner Parameter Estimation of Binary Logistic Regression Model

The Maximum Likelihood Estimation (MLE) method is used to estimate parameters in logistic regression. The MLE method was chosen because it has several advantages compared to other methods, including that it can be used to form non-linear models such as logistic regression, and the estimation results are unbiased [6]. Mathematically, the likelihood function $x_i, y_i)$ can be expressed [7]:

$$f(x_i) = \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \quad (3)$$

If each observation has been assumed to be an independent variable, the likelihood function is also a multiplication of each likelihood function, which is as follows:

$$L(\beta) = \Pi_{i=1}^{n} f(x_i) \quad (4)$$

By using the Maximum Likelihood method to estimate the Binary Logistic Regression parameters, the estimator is obtained $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$ by solving the system of equations $\Sigma_i y_i x_{ia} - \Sigma_i y_i x_{ia}\pi(x_i) = 0, a = 1, 2, \ldots, k$ with $\pi(x_i) = \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}$ using numerical methods.

Parameter Estimation Testing

The parameter estimation test aims to determine whether there is a significant relationship between the independent variables (parameters) and the dependent variable.

Simultaneous Test

The simultaneous test is used to test parameter estimates simultaneously (together). The test hypothesis is as follows:

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1 : There is at least one $\beta_j \neq 0; j = 1,2, \ldots, p$

Test statistics:

$$G = -2ln \frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_o}{n}\right)^{n_o}}{\sum_{i=1}^{n} \hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{1-y_i}} \quad (5)$$

Reject H_0 if $G > x^2_{(a,p)}$ with p number of predictor variables in the model [6].

Partial Test

Partial tests are used to test parameter estimates partially (separately). The test hypothesis is as follows:

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0; j = 1,2, \ldots, p$

Test statistics:

$$W^2 = \left[\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}\right]^2 \quad (6)$$

Reject H_0 if $W^2 > \chi^2\ (a, p)$ with p number of predictor variables in the model [6].

Fit Test

The Goodness of fit test is used to evaluate whether the model obtained is suitable or not with the observed data The test hypothesis is as follows [8]:

H_0 : Model is fit

H_1 : Model does not fit

Test statistics:

$$\hat{C} = \sum_{j=1}^{k} \frac{(o_j - n_j{}'\overline{\pi}_j)^2}{n_j{}'\overline{\pi}_j(1-\overline{\pi}_j)} \quad (7)$$

Reject H_0 if $\hat{C} > \chi^2\ (p-2)$ or $P - value < \alpha$.

Parameter Coefficient Interpretation

The parameter coefficient can be interpreted using the *odds ratio* (ψ) value. The definition of odds ratio for $x = 1$ dan $x = 0$ is as follows:

$$OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} = \frac{\exp\ (\beta_O + \beta_1)}{\exp\ (\beta_O)} = \exp\ (\beta) \quad (8)$$

The odds ratio is a measure of association that can be interpreted broadly, especially in epidemiology. From the equation above, the odds ratio is the average of the tendency of the response variable to have a certain value if $x = 1$ compared to $x = 0$.[6].

Classification Accuracy

Classification accuracy of the model is used to determine whether the data is classified correctly or not [9]. The evaluation metrics used are APER and Accuracy. The APER value expresses the proportion of samples incorrectly classified by the classification function [10]. The Accuracy value expresses the proportion of samples that are correctly classified by the classification function. Calculating the APER value can use the confusion matrix as follows [11]:

Table 1. Contingency Table of Classification Accuracy

Actual Class	Prediction Class		Total
	Yes	No	
Yes	TP	FN	P
No	FP	TN	N
Total	P'	N'	P+N

$$APER = \frac{FP+FN}{P+N} \quad (9)$$

$$Accuracy = \frac{TP+TN}{P+N} = 100\% - APER \quad (10)$$

Smart Indonesia Card College (KIP-K)

According to [3], KIP-K is higher education assistance in the form of cash, expansion of access, and learning opportunities from the government provided to students from poor/vulnerable families to finance education. KIP-K aims to increase the economic potential and social mobility for students from poor/vulnerable families to attend college. KIP-K is an expansion of the Bidikmisi scholarship.

METHOD

Research Type and Data Source

The type of research used is quantitative. This method must also use quantitative tools in the form of the R program in processing the data. The data used is data on undergraduate students receiving Bidikmisi scholarships 2017-2019 FMIPA Bengkulu University. The type of data used is secondary data. Secondary data is data obtained from a previously available source. The data source comes from the Academic section of FMIPA Bengkulu University.

Research Variables

The variables used in this study are as follows:

Table 2. Research Variables

Label	Variable	Category	Scale
Y	Timeliness of graduation	0 = Not on time (> 8 semester) 1 = on time (≤ 8 semester)	Binary
X_1	Study Program	1 = S1 Mathematics 2 = S1 Chemistry 3 = S1 Physics 4 = S1 Biology 5 = S1 Statistics 6 = S1 Geophysics 7 = S1 Pharmacy	Nominal
X_2	Gender	0 = Female 1 = male	Nominal

X_3	District of Origin	0 = Outside Bengkulu City 1 = Bengkulu City	Nominal
X_4	School of Origin	1 = SMA 2 = SMK 3 = MAN 4 = Lainnya	Nominal
X_5	Entrance Path	0 = SBMPTN 1 = SNMPTN	Nominal
X_6	GPA Score	-	Numerical

Research Stages

Research activities outline starting from designing, collecting references and data, making analysis, making models, model testing, and interpretation. In detail, the stages of this research include activities:

1. Literature review and methodology exploration
2. Data collection.
3. Data exploration and categorization
4. Creating an initial model
5. Perform parameter significance testing
6. Create a new model with significant variables
7. Performing model fit test
8. Calculating and interpreting the odds ratio
9. Assessment of classification accuracy.

RESULT AND DISCUSSION

Initial Model of Binary Logistic Regression

The initial Binary Logistic Regression model obtained from the timeliness of graduating KIP students of FMIPA Bengkulu University is as follows:

$$\hat{g}(x) = -11.60552 + 0.24629x_{12} - 2.05500x_{13} - 1.65856x_{14} - 2.56339x_{15} - 1.09088x_{16} - 0.58078x_{17} - 0.83477x_{21} + 0.23606x_{31} - 0.41358x_{42} + 1.04611x_{43} + 0.04034x_{44} - 0.77481x_{51} + 3.75888x_6 \quad (11)$$

Parameter Estimation Testing

The parameter estimation test aims to determine whether there is a significant relationship between the independent variables (parameters) and the dependent variable.

Test Partial

The simultaneous test is used to test parameter estimates simultaneously (together). The test is as follows:

1) Hypothesis

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_8 = 0$

H_1 : There are at least $\beta_j \neq 0; j = 1,2, \ldots ,6$

2) Required quantities
   n = 168
   α = 5%
3) Test statistics

Table 3. G Test(simultaneously)

G	df	$x^2_{(a,p)}$
66.0853579	6	12.59159

4) Rejection criteria
   Reject H_0 if $p - value <$ α or $G > x^2_{(a,p)}$
   Accept H_0 if $p - value >$ α or $G < x^2_{(a,p)}$
5) Conclusion
   Based on the results of the G test in Table 3, it can be seen that the value of the G test on the model obtained is 66.0853579 with a free degree of 6. Because the value of $G = 66.0853579 > x^2_{(0,05,6)} = 12.59159$ then H_0 is rejected. This means that there is at least one predictor variable that is simultaneously significant to the graduation of KIP College students at FMIPA Bengkulu University.

Partial Test

Partial tests are used to test parameter estimates partially (separately). The test is as follows:

1) Hypothesis
   $H_0 : \beta_j = 0$
   $H_1 : \beta_j \neq 0; j = 1,2, \dots, p$
2) Required quantities
   n = 168
   α = 5%
3) Test statistics

Table 4. Wald Test (Partial)

Parameters	***Coef***	***S.E.***	***Wald Z***	***Pr(>|Z|)***
Intercept	-11.606	3.7045	-3.13	0.0017
X_{12}	0.2463	0.5952	0.41	0.679
X_{13}	-2.055	0.638	-3.22	0.0013
X_{14}	-1.6586	0.6551	-2.53	0.0114
X_{15}	-2.5634	0.8669	-2.96	0.0031
X_{16}	-1.0909	0.9151	-1.19	0.2332
X_{17}	-0.5808	1.338	-0.43	0.6642
X_{21}	-0.8348	0.5852	-1.43	0.1537
X_{31}	0.2361	0.4115	0.57	0.5662
X_{42}	-0.4136	1.278	-0.32	0.7462
X_{43}	1.0461	0.7402	1.41	0.1575

X_{44}	0.0403	0.6131	0.07	0.9475
X_{51}	-0.7748	0.43	-1.8	0.0715
X_6	3.7589	1.1124	3.38	0.0007

4) Rejection criteria

   Reject H_0 if $p - value <$ α or $W^2 > \chi^2\ (a, p)$

   Accept H_0 if $p - value >$ α or $W^2 < \chi^2\ (a, p)$

5) Conclusion

   Based on the Wald test results in Table 4, it can be seen that in some variables the $p - value < a = 0{,}05$, then H_0 is rejected. student graduation. The variables that have a significant effect on the timeliness of graduating KIP students at FMIPA Bengkulu University are the variable origin of study programs in categories 3, 4, and 5 and the GPA variable.

Binary Logistic Regression Model

Based on testing the significance of parameters both simultaneously and partially, it shows that the variables of Study Program Origin and GPA have a significant effect so in the new model the variables of Study Program Origin and GPA are included. The Binary Logistic Regression model obtained is as follows:

$$\hat{g}(x) = -12.76537 + 0.09213x_{12} - 1.78109x_{13} - 1.26683x_{14} - 2.16772x_{15} - 1.23262x_{16} - 0.75135x_{17} + 3.97552x_6 \quad (12)$$

Model Fit Test

The goodness of fit test is used to evaluate whether the model obtained is suitable or not with the observed data. The test is as follows:

1) Hypothesis

   H_0 : Model is fit

   H_1 : Model does not fit

2) Required quantities

   n = 168

   α = 5%

3) Test statistics

Table 5. Model fit test

$\widehat{C}$	Df	$x^2_{(a,p)}$	$P - value$
6.9512	2	15.50731	0.5419

4) Rejection criteria

   Reject H_0 if $p - value <$ α or $\hat{C} > x^2_{(a,p)}$

   Accept H_0 if $p - value >$ α or $\hat{C} < x^2_{(a,p)}$

5) Conclusion

Based on the model fit test on the model, it can be seen that the $\hat{C}$ or chi-square value is 120 with a free degree of 4. Because the value of $\hat{C} = 6.9512 < x^2_{(a,p)} = 15.50731$ and $p - value = 0.5419 > a = 0{,}05$, then H_0 is accepted. This means that the model used is appropriate.

Parameter Coefficient Interpretation

The interpretation of the parameter coefficient is to determine the functional relationship between the response variable and the predictor variable and define any changes in the response variable caused by the predictor variable. The odds ratio results are as follows:

Table 6. *Odds ratio value*

Variable	***Coef***
Intercept	$2.858049e-06$
X_{12}	1.096509
X_{13}	0.1684552
X_{14}	0.2817225
X_{15}	0.1144383
X_{16}	0.2915274
X_{17}	0.4717298
X_6	53.27807

Based on Table 6 above, it can be seen that the odds ratio value is obtained from the exponential value (β) interpretation that the odds ratio value of variable X_{12} as a student who comes from the S1-Chemistry Study Program tends of 1.096509 times compared to students who come from other study programs at FMIPA Bengkulu University to graduate on time. The odds ratio value of variable X_{13} as a student who comes from the S1-Physics Study Program tends of 0.1684552 times compared to students who come from other study programs at FMIPA Bengkulu University to graduate on time. The odds ratio value of variable X_{14} as a student who comes from the S1-Biology Study Program tends of 0.2817225 times compared to students who come from other study programs at FMIPA Bengkulu University to graduate on time. The odds ratio value of variable X_{15} as a student who comes from the S1-Statistics Study Program tends of 0.1144383 times compared to students who come from other study programs at FMIPA Bengkulu University to graduate on time. The odds ratio value of variable X_{16} as a student who comes from the S1-Geophysics Study Program tends of 0.2915274 times compared to students who come from other study programs at FMIPA Bengkulu University to graduate on time. The odds ratio value of variable X_{17} as students who come from the S1-Pharmacy Study Program tends of 0.4717298 times compared to students who come from other study programs at FMIPA Bengkulu University to graduate on time. The odds ratio value of variable X_6 is GPA, for a one-unit change in the GPA value, it can be said that if the GPA value increases by one unit, the odds ratio or the student's chance to graduate on time will increase by 53.27807 times.

So it can be concluded that the higher the GPA value, the greater the chance for students to graduate on time.

Classification Accuracy

The classification accuracy results of the model are as follows:

Table 7. Confusion Matrix

Actual Class	Prediction Class		Total
	1	0	
1	55	17	72
0	20	76	96
Total	75	93	168

$$APER = \frac{20 + 17}{168} = \frac{37}{168} = 0.220238$$

$$Accuracy = \frac{55 + 76}{168} = \frac{131}{168} = 0.7797619$$

The results of the APER calculation, it can be seen that the classification error value is 22.03%. From the accuracy calculation, it can be seen that the classification accuracy value is 77.97%. So it can be concluded that the Binary Logistic Regression model obtained in classifying the graduation time of college KIP FMIPA Bengkulu University students on time and not on time is good enough.

CONCLUSION

Based on the results of the discussion of the research conducted, a Binary Logistic Regression model of the timeliness of graduating KIP College students FMIPA Bengkulu University is obtained with variables that have a significant effect as follows:

$$\hat{g}(x) = -12.76537 + 0.09213x_{12} - 1.78109x_{13} - 1.26683x_{14} - 2.16772x_{15} - 1.23262x_{16} - 0.75135x_{17} + 3.97552x_{6}$$

Factors that have a significant influence on the timeliness of graduating Bidikmisi students of FMIPA Bengkulu University are the Origin of S1-Physics, S1-Biology, and S1-Statistics study programs and GPA. Then the classification results using Binary Logistic Regression have a classification accuracy rate of 77.97%. So it can be concluded that the Binary Logistic Regression model obtained in classifying the graduation time of college KIP FMIPA students at Bengkulu University on time and not on time is good enough.

ACKNOWLEDGMENTS

Thank you to the supervisors and the friends as well as relatives who have helped.

REFERENCE

[1] Badan Akreditasi Nasional Perguruan Tinggi, "Naskah Akademik IAPT 3.0.," BANPT, Jakarta, 2019.

[2] W. Agwil, H. Fransiska and N. Hidayati, "Analisis Ketepatan Waktu Lulus Mahasiswa Dengan Menggunakan Bagging CART," *Jurnal Pendidikan Matematika dan Matematika*, vol. 6, no. 2, pp. 155-166, 2020.

[3] Pusat Layanan Pembiayaan Pendidikan, "Pedoman Pendaftaran Kartu Indonesia Pintar Kuliah KIP Kuliah Merdeka," Puslapdik, Jakarta, 2023.

[4] N. R. Tamnge, "Regresi Logistik Biner Dalam Menentukan Pengaruh Kualitas Pelayanan Terhadap Kepuasan Mahasiswa Fakultas Keguruan dan Ilmu Pendidikan Universitas Muhammadiyah Surabaya," *Journal of Mathematics Education, Science and Technology*, vol. 1, no. 2, pp. 222-233, 2016.

[5] A. Agresti, Categorical Data Analysis, New York: John Wiley & Sons, Inc, 1990.

[6] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, Canada: John Wiley & Sons, Inc, 2000.

[7] W. Alwi, Ermawati and S. Husain, "Analisis Regresi Logistik Biner Untuk Memprediksi Kepuasan Pengunjung Pada Rumah Sakit Umum Daerah Majene," *Jurnal MSA*, vol. 6, no. 1, pp. 20-26, 2018.

[8] B. Peeters, R. Dewil and I. Y. Smets, "Improved Process Control of an Industrial Sludge Cemtrifuge-dryer Installation Through Binary Logistic Modelling of The Fouling Issues," *Journal of Process Control*, vol. 22, no. 7, pp. 1387-1396, 2012.

[9] A. Agresti, Categorical Data Analysis Second edition, New Jersey: John Wiley & Sons, 20002.

[10] R. Johnson and D. Wichern, Applied Multivariate Statistical Analysis, New Jersey: Prentice Hall, 1992.

[11] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, USA: Elsevier, 2012.

Diversification of Jakarta Islamic Index (JII) Stock Optimal Portfolio for the Period 2018-2023

Khairul Alim(1)*, Bayun Matsaany(2), Anisa Rahmawati(3)

[1]Program Studi Matematika, Jurusan MIPA, Fakultas Sains dan Teknologi, Universitas Jambi

Jl. Jambi – Muara Bulian No.KM. 15, Mendalo Darat, Jambi

[2]Program Studi S1 Terapan Teknologi Rekayasa Manufaktur, Jurusan Teknik Mesin, Politeknik Negeri Jakarta

Jl. Prof. DR. G.A. Siwabessy, Kampus Universitas Indonesia Depok, Jawa Barat

[3]Program Studi Manajemen Logistik Industri Elektronika, Politeknik APP Jakarta

Jl. Timbul No.34, RT.6/RW.5, Cipedak, Kec. Jagakarsa, Kota Jakarta Selatan

e-mail: khairulalim@unja.ac.id(1)*, bayun.matsaany@mesin.pnj.ac.id(2), anisarara07@gmail.com(3)

ABSTRAK

Investasi adalah tindakan menempatkan dana pada suatu aset dengan tujuan meraih keuntungan di masa depan melalui perubahan harga aset atau capital gain. Di Indonesia, investasi saham, terutama melalui Bursa Efek Indonesia (BEI), menjadi pilihan masyarakat. BEI memiliki berbagai indeks, termasuk Jakarta Islamic Index (JII), yang menarik perhatian. JII terdiri dari saham-saham perusahaan yang menerapkan prinsip syariah. Dalam kegiatan investasi, analisis yang cermat sangat penting untuk menghindari kesalahan dalam pemilihan saham. Diversifikasi adalah strategi dengan menyebarkan investasi pada beberapa saham, yang bertujuan untuk memaksimalkan keuntungan dan meminimalkan risiko. Meskipun cara termudah bagi investor untuk berinvestasi di pasar saham adalah dengan membeli satu jenis saham terbaik, namun memiliki hanya satu jenis saham dapat menjadi berisiko jika harga saham tersebut turun. Penelitian ini membandingkan hasil investasi antara portofolio tunggal dan diversifikasi dengan dua atau tiga saham, dengan pemilihan saham berdasarkan indeks Sharpe, Treynor, dan Jensen. Penelitian ini menganalisis harga penutupan saham dalam indeks JII dari Maret 2018 hingga Februari 2023. Hasil penelitian menunjukkan bahwa melakukan diversifikasi beberapa saham dapat mengurangi risiko investasi, meskipun mengurangi potensi keuntungan.
Kata kunci: Investasi, Saham, Syariah, Diversifikasi

ABSTRACT

Investment is placing funds into an asset to gain profits in the future through changes in asset prices or capital gains. In Indonesia, stock investment, primarily through the Indonesia Stock Exchange (BEI), is popular among the public. BEI offers various indices, including the Jakarta Islamic Index (JII), which has garnered significant attention. JII comprises stocks of companies that adhere to Sharia principles. In investment, careful analysis is crucial to avoid errors in stock selection. Diversification is a strategy that involves spreading investments across several stocks, aiming to maximize profits while minimizing risks. While the easiest way for investors to enter the stock market is by purchasing the best-performing single stock, relying solely on one stock can be risky if its price drops. This research compares investment outcomes between single-stock and diversified portfolios of two or three stocks, selected based on the Sharpe, Treynor, and Jensen indices. The research analyzes the closing stock prices within the JII index from March 2018 to February 2023. The study results show that diversifying investments across multiple stocks can reduce investment risks, even though it may reduce profit potential.
Keywords: *Investment; Stocks; Sharia, Diversification*

INTRODUCTION

Investment is an important activity to gain future profits by placing current funds. This profit is known as capital gain or cash income [1]. A subcategory that is now increasingly popular is sharia-based investment [2], which involves trading shares on the Sharia capital market. OJK data shows an increase in sharia shares from 375 in 2017 to 552 in 2022 [3]. Fighting inflation and achieving profits are the two main reasons investing is essential. When investing, it is important to consider market conditions, the economy, and the company's financial performance [4]. Analysis of company performance on the Indonesian Stock Exchange (BEI) is crucial in investing in the capital market. BEI creates an index known as the Composite Stock Price Index (IHSG). IHSG is used to see the performance of all shares listed on the IDX [5]. The IDX also divides shares into various indices, such as the Jakarta Islamic Index (JII), which follows sharia principles [6].

Benchmarking different investment methods is essential to identify the most effective strategies in different market conditions [7]. This study uses the Sharpe, Treynor, and Jensen indices to create portfolios with one, two, and three stocks [8]. This approach is interesting because it explains how diversification and stock selection based on different metrics can affect investment results. The risks arising in shares come from the volatility of price movements. Volatility can appear in various types of assets and prices of goods [9]. One way to control this emerging risk is to diversify. Diversification is a way of investing in various assets, an essential strategy for controlling profits and risks [10]. Diversification aims to reduce dependence on one particular asset or type of risk, thereby achieving a balance of risk and return. The advantage of diversification is achieving higher long-term returns and lower risk than relying on just one type of asset or stock [11].

This research aims to look at differences in Sharia stock portfolio management, focusing on the impact of diversification on profits and risks. Through comparisons between portfolios consisting of various numbers of shares, this study will reveal the effectiveness of diversification in optimizing returns and reducing risk. This is important to see better Islamic investment strategies and make more informed investment decisions.

Our previous study focused on forming an optimal portfolio for JII shares from March 2018 to February 2023, involving only three shares [11]. This study will compare portfolios of one, two, and three stocks. Thus, this research will provide new insights into the impact of diversification in Islamic investments and reveal differences in the profits and risks associated with the various investment strategies implemented.

METHOD

Portfolio Performance Measurement

The performance of a stock portfolio can be determined through the Sharpe, Treynor, and Jensen Index approaches. These three methods use historical data to project future profits and risks, and the results generated from each method can vary [12]. Therefore, in evaluating the performance of a portfolio, it is necessary to consider the investment objectives and the nature of the portfolio owned.

Sharpe Index

The Sharpe Index compares the investment return minus the return of a risk-free asset, and then the return is normalized based on the level of risk in the investment. Better stock performance is reflected in a higher index value. The Sharpe Index not only considers the rate of return on investment but also considers the risks involved. By measuring a stock's success not just based on profit alone, this index provides a more comprehensive view of investment performance by considering the level of risk that may occur. Therefore, the Sharpe Index provides a solid foundation for evaluating and selecting stocks that produce an optimal balance between return and risk, a crucial factor in investment decision-making. The calculation of the Sharpe Index is based on a specific mathematical formula [13].

$$S_i = \frac{\bar{R}_i - \bar{R}}{\sigma_i} \tag{1}$$

S_i : Sharpe index
$\bar{R}_i$: average return of stock-i
σ_i : standard deviation of the return of stock-i
$\bar{R}$: average return of the risk-free interest rate

Treynor Index

The Treynor Index is a tool for evaluating the performance of an investment by considering risk. This index is measured by calculating the ratio of profits generated for each risk that arises. Good stock performance is reflected in a high index value. The Treynor Index provides a more detailed perspective on investment performance by incorporating risk factors into its analysis. By assessing the return ratio obtained per level of risk, this index helps investors understand the extent to which a stock provides relative profit compared to the risk taken [14].

$$T = \frac{\bar{R}_i - \bar{R}}{\beta_i} \tag{2}$$

T_i : Treynor index
$\bar{R}_i$: average return of stock-i
$\bar{R}$: average return of the risk-free interest rate
β_i : beta coefficient for stock-i

Jensen Index

The third approach uses the Jensen Index, often referred to as Jensen's Alpha. The Jensen ratio is used to assess the extent to which an investment manager can generate above-average results. The Jensen's Alpha index measures how efficiently a manager can manage investments and achieve results that exceed market expectations. Jensen's Alpha index offers a unique perspective in evaluating investment performance, as it considers the final results and assesses the manager's ability to optimize returns relative to the level of risk taken. In this way, the index provides a more comprehensive picture of the manager's contribution to investment performance. Using the Jensen's Alpha index, investors can identify whether an investment manager can create added value or fall below market expectations. Therefore, the Jensen index becomes a valuable tool in evaluating and

selecting investments that yield optimal results. The calculation of the Jensen Index follows a specific mathematical formula [15].

$$\alpha_i = (\bar{R}_i - \bar{R}) - [\beta_i(\bar{R}_M - \bar{R})] \quad (3)$$

α_i : Indeks Jensen
$\bar{R}_i$: average return of stock-i
$\bar{R}$: average return of the risk-free interest rate
β_i : beta coefficient for stock-i
$\bar{R}_M$: average market profit

If the Jensen's alpha value shows a positive number, it indicates that the investment performance has an advantage in terms of return. Conversely, the investment performance is considered neutral if the value is zero. However, if Jensen's alpha value is negative, the investment is deemed to have suboptimal performance. In evaluating investment performance, the Treynor and Jensen index methods utilize a beta coefficient, which is calculated using a mathematical formula:

$$\beta = \frac{\sigma_{i,M}}{\sigma_M^2} = \frac{[n\sum(R_M.R_i] - [\sum R_M \sum R_i]}{n(\sum {R_M}^2) - (\sum R_M)^2} \quad (4)$$

β : beta coefficient
$\sigma_{i,M}$: covariance between the return of stock-i and the return of the market portfolio
σ_M^2 : variance of market
n : number of data samples
R_M : market return
R_i : return of stock-i

The beta coefficient is vital in assessing how much an investment moves in line with market movements. Using beta helps understand the level of risk involved in an investment and the extent to which the investment is defensive or aggressive towards market fluctuations. Market conditions can be inferred from the beta value, which follows.

Table 1. Beta value

Indicator	Explanation
$\beta < 1$	Stocks have a lower risk compared to the market
$\beta = 1$	Stocks have the same risk as the market
$\beta > 1$	Stocks are considered to have a higher risk compared to the market

Selection of the Best Stocks

By evaluating the previous indices, each of the 30 stocks included in the JII index will be given a score ranging from 1 to 30. The highest score, 30, is awarded to the top-ranked stock from each index analysis, while the lowest score, 1, is given to the last-ranked stock in each analysis. The highest cumulative score from all indices determines the best stock. This assessment process allows for identifying stocks that excel not just in one aspect but also in evaluating their overall

performance through different criteria. The scores reflect the extent to which a stock meets specific parameters considered necessary in the index analysis.

Formation of the Best Portfolio

An optimal portfolio is formed by minimizing variance, which is a measure of risk. As a result, the formed portfolio will have the smallest risk. The profit of a portfolio is the weighted average of the expected profit of each stock, and is expressed as follows:

$$E(r_p) = \sum_{i=1}^{n} w_i \cdot E(r_i) \tag{5}$$

$E(r_p)$: expected return of a portfolio
w_i : proportion of stock-i
$E(r_i)$: expected return of the stock

The importance of low variance in the search for an optimal portfolio emphasizes efforts to reduce risk as far as possible. Thus, minimizing variance through intelligent weighting composition of different assets becomes vital in forming a portfolio that performs well overall and has an acceptable level of risk. Since the risk of a portfolio refers to the variance of its constituent stocks, the variance of two stocks is expressed as follows:

$$\sigma_p^2 = w_x^2\sigma_x^2 + w_y^2\sigma_y^2 + 2w_x w_y \rho_{xy}\sigma_x\sigma_y \tag{6}$$

and the variance of three stocks is expressed as

$$\sigma_p^2 = w_x^2\sigma_x^2 + w_y^2\sigma_y^2 + w_z^2\sigma_z^2 + 2w_x w_y \rho_{xy}\sigma_x\sigma_y + 2w_x w_z \rho_{xz}\sigma_x\sigma_z + 2w_y w_z \rho_{yz}\sigma_y\sigma_z \tag{7}$$

σ_p^2 : portfolio variance
w_i : weight of stock-i in the portfolio
σ_i : standard deviation of the returns of stock-i
ρ_{ij} : correlation of the returns of stocks i and j
where $i, j = x, y, z$.

RESULT AND DISCUSSION

This study used JII stock data from March 2018 to February 2023. The JII stock data consists of 30 types of stocks categorized under Sharia-based management. The stock prices used are the daily closing prices of each stock. As a comparison in the data analysis, the Jakarta Composite Index (IHSG) is used as the market portfolio, and Bank Indonesia's interest rate (BI rate) is used as the risk-free rate. By assigning scores in the analysis of the Sharpe, Treynor, and Jensen indices, a comprehensive understanding of each stock's relative performance and potential risks is expected to be gained. After scoring each index for the JII stocks, the top three stocks are identified, as shown in Table 2.

Table 2. Total score of the top three stocks

Stock code	Score			Total
	Sharpe	Treynor	Jensen	
MDKA	30	29	30	89
HRUM	29	28	28	85
BRIS	28	24	29	81

Table 2 shows that the selected stocks are the best performers in terms of the Sharpe and Jensen indices. BRIS, which is not the top performer in the Treynor index, can still be selected. We combine all three indices, so the selection is not based solely on one index. Next, we calculate the average returns and standard deviations of the three selected stocks, as shown in Table 3.

Tabel 3. Deskripsi saham MDKA, HRUM, BRIS

Stock code	Expected return	Standard deviation
MDKA	0.045941	0.128475
HRUM	0.036321	0.206434
BRIS	0.039199	0.227145

From this, MDKA ranked first and has the highest average returns with the lowest risk. Next in line are HRUM and BRIS, with lower returns and higher risks. Furthermore, to determine the weight of each stock in each portfolio, it is necessary to establish the correlation of returns between these three stocks, as indicated in Table 4.

Table 4. Correlation table of MDKA, HRUM, BRIS

	MDKA	HRUM	BRIS
MDKA	1		
HRUM	0.000540	1	
BRIS	0.323047	0.248750	1

The highest correlation between the stock BRIS and the other two stocks is observed. Meanwhile, MDKA and HRUM have a relatively low correlation. In forming a portfolio with only one stock, MDKA, with the highest combined score, the expected returns and risks can be seen in Table 5.

Table 5. Single stock portfolio of MDKA

Parameter	Value
w_x	1.000000
$E(r_p)$ (per month)	0.045940
$E(r_p)$ (per year)	0.551291
σ_p^2	0.0165057
σ_p	0.1284745

The parameters in Table 3 and Table 4 are used to determine the weights in portfolios with two or three stocks. A portfolio with a single stock, with a 100% allocation to MDKA, yields an expected return of 4.5% per month (55.12% per year) with a risk level (standard deviation) of 12.84%.

The optimal portfolio composition for two or three stocks in the JII can be calculated by minimizing the variance function in Equations (6) and (7). Furthermore, the expected return of the portfolio is determined by Equation (5).

Table 6. Portfolio of two stocks MDKA and HRUM

Parameter	Value
w_x	0.724924
w_y	0.275077
$E(r_p)$ (per monht)	0.043294
$E(r_p)$ (per year)	0.519536
σ_p^2	0.012114
σ_p	0.110063

Table 7. Three-stock portfolio of MDKA, HRUM, and BRIS

Parameter	Value
w_x	0.691715
w_y	0.265528
w_z	0.042757
$E(r_p)$ (per month)	0.043098
$E(r_p)$ (per year)	0.517179
σ_p^2	0.011824
σ_p	0.108739

Table 6 shows the composition of the two selected stocks, MDKA and HRUM. By diversifying the portfolio, a portfolio using the two selected stocks, MDKA at 72.49% and HRUM at 27.51%, is obtained, providing an expected return of 4.32% per month (51.95% per year) with a risk level (standard deviation) of 11%. Then, Table 7 displays the composition of the three selected stocks: MDKA, HRUM, and BRIS. The portfolio with the top three stocks, MDKA at 69.17%, HRUM at 26.55%, and BRIS at 4.28%, yields an expected return of 4.3% per month or 51.72% per year with a portfolio risk (standard deviation) of 10.87%. A comparison of the returns and risks of these three types of investments can be seen in Figure 1. Increasing the number of stocks in the portfolio will decrease the level of returns, but this also leads to a reduction in the level of risk.

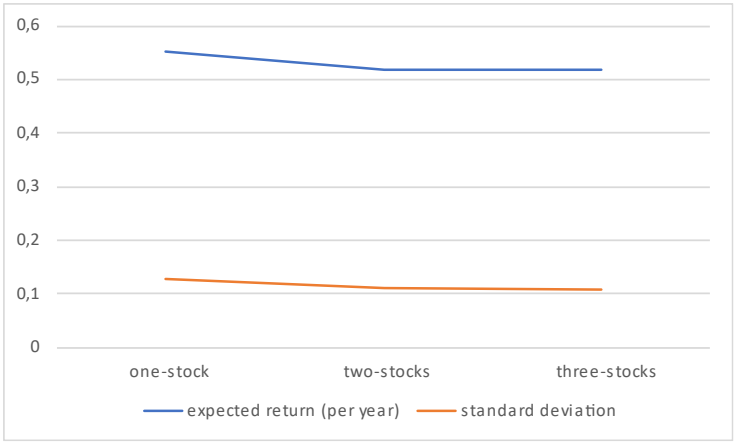


Figure 1. Comparison of single investment and diversification

CONCLUSION

This research aims to compare the expected profits and measured risks obtained from various types of investments. The focus of our research is on investments that diversify portfolios. We consider portfolios with one, two, and three different stocks. Our research results show that diversification significantly impacts the returns and risk levels that can be obtained. By separating our investments into several different stocks, we can control the risk that may arise from one of the lower-risk stocks. This diversification can be an essential strategy for investors, especially in the Islamic stock market.

This research can also be developed further by combining portfolios from various stock sectors so that other sectors can offset the risks associated with one sector. This can increase the effectiveness of diversification strategies and provide a better understanding of managing risk in Islamic stock investments. The findings of this research can be a valuable guide for investors interested in developing their portfolios efficiently in the context of the Islamic stock market.

REFERENCE

[1] L. Siti Aminah, "The Effect of Current Ratio, Net Profit Margin, And Return on Assets on Stock Return," *Journal of Management, Accounting, General Finance and International Economic Issues*, vol. 1, no. 1, pp. 1–9, Dec. 2021, doi: 10.55047/marginal.v1i1.8.

[2] N. Rahmawati, *Manajemen Investasi Syariah*. Mataram: Institut Agama Islam Negeri (IAIN) Mataram, 2015.

[3] Otoritas Jasa Keuangan, "Statistik Saham Syariah - Desember 2022," Dec. 2022.

[4] E. Pardiansyah, "Investasi dalam Perspektif Ekonomi Islam: Pendekatan Teoritis dan Empiris," *Economica: Jurnal Ekonomi Islam*, vol. 8, no. 2, pp. 337–373, Oct. 2017, doi: 10.21580/economica.2017.8.2.1920.

[5] N. latifa Hadi and A. Indrasetianingsih, "FORECASTING INDEKS HARGA SAHAM GABUNGAN (IHSG) DENGAN MENGGUNAKAN METODE ARIMA," *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, vol. 6, no. 1, Jul. 2014, doi: 10.36456/jstat.vol6.no1.a302.

[6] S. Febrianti, "Analisis Perbandingan Kinerja Indeks Saham Syariah dengan Indeks Saham Konvensional Periode 2015-2017 (Studi Kasus pada JII dan LQ45)," 2018. [Online]. Available: www.syariah.ojk.go.id

[7] Z. Bodie, A. Kane, and A. J. Marcus, *Investments and Portfolio Management*, 9th ed. New York: McGraw-Hill, 2010.

[8] A. Sodiqin, "Kinerja Portfolio dengan Metode Sharpe, Jensen dan Treynor pada Saham Industri Tekstil di Bursa Efek Indonesia," *Jurnal Manajemen Bisnis Krisnadwipayana*, vol. 8, no. 1, Apr. 2020, doi: 10.35137/jmbk.v8i1.382.

[9] N. Septiana, Primadina Hasanah, and Annisa Rahmita Soemarsono, "Analisis Volatilitas Harga Saham Sekor Minyak dan Gas di Indonesia pada Masa Pandemi Covid-19 dengan Metode ARIMA-GARCH," *J Statistika*, vol. 14, no. 2, pp. 99–109, Jan. 2022, doi: 10.36456/jstat.vol14.no2.a4497.

[10] R. Setiawan, O. R. Putri, and A. C. Sukmawati, "Diversifikasi Portofolio Kredit, Risiko dan Return Bank," *Jurnal Akuntansi*, vol. 15, no. 1, pp. 189–199, May 2023, doi: 10.28932/jam.v15i1.6376.

[11] K. Alim, A. Rahmawati, and B. Matsaany, "Formation of Optimal Portfolio on JII Stock using Sharpe, Treynor, and Jensen Indices during the Period of 2018-2023," *Jurnal Matematika, Statistika dan Komputasi*, vol. 19, no. 3, pp. 593–601, May 2023, doi: 10.20956/j.v19i3.26354.

[12] W. Wihardi and A. Lutfi, "Pembentukan Portfolio Optimal Untuk Berinvestasi pada Saham Perusahaan Perbankan di Bursa Efek Indonesia dengan Metode Sharpe, Treynor dan Jensen Periode 2013-2017," *Jurnal Manajemen Bisnis dan Kewirausahaan*, vol. 4, no. 3, p. 74, May 2020, doi: 10.24912/jmbk.v4i3.7920.

[13] A. Priyanti, I. Nurhayati, R. S. Amind, and D. R. J. Manager, "Analisis Evaluasi Kinerja Portofolio Saham dengan Metode Sharpe," *2 Mei*, vol. 4, no. 2, pp. 174–181, 2021, [Online]. Available: http://ejournal.uika-bogor.ac.id/index.php/MANAGER

[14] S. Aeni, S. Siti, R. Handayani, and R. Hidayat, "Evaluasi Kinerja Investasi Portofolio dengan Menggunakan Model Treynor (Studi Pada Perusahaan Food & Beverages Yang Listing Di BEI Periode 2013)," *Jurnal Administrasi Bisnis (JAB)|Vol*, vol. 23, no. 1, 2015.

[15] S. Syam, A. Rauf, and C. I. Musa, "Analisis Pembentukan Portofolio Optimal dengan Single Index Model untuk Proyeksi Investasi pada Saham Index LQ45 di Bursa," *Jurnal Ilmu Manajemen Retail (JIMAT)*, vol. 2, no. 2, 2021, [Online]. Available: https://doi.org/

Forecasting PT Triputra Agro Persada Tbk (TAPG) Share Prices Using Multivariate Time Series Analysis

Dwi Sulistiowati [(1)], Maya Sari Syahrul [(2)], Iswan Rina [(3)]

[1]Departemen Statistika, Universitas Negeri Padang

Jalan Prof. Dr. Hamka, Air Tawar Barat, Kec. Padang Utara, Kota Padang, Sumatera Barat

[2,3]Program Studi Matematika, Universitas Dharma Andalas

Jalan Sawahan No. 103, Simpang Haru, Kec. Padang Timur, Kota Padang, Sumatera Barat

e-mail: dwisulistiowati@fmipa.unp.ac.id[(1)], maya@unidha.ac.id [(2)] , iswanrina@unidha.ac.id [(3)]

ABSTRAK

Peningkatan harga *crued palm oil* (CPO) menyebabkan pengaruh positif terhadap harga saham perusahaan yang bergerak di industri sawit. PT Triputra Agro Persada Tbk (TAPG) sepanjang tahun 2021 tercatat sebagai salah satu perusahaan dengan bisnis CPO yang memperoleh *capital gain* terbesar. Prediksi atau peramalan akan harga saham dimasa mendatang sangat perlu bagi investor sebagai pertimbangan sebelum keputusan untuk berinvestasi. Penelitian prediksi harga saham telah banyak dilakukan sebelumnya secara univariat. Pemodelan univariat tidak dapat mempresentasikan adanya pengaruh dari variabel lain terhadap harga saham. Peramalan dengan adanya pengaruh dari variabel lain dapat dilakukan dengan analisis *multivariate time series.* Penelitian ini bertujuan untuk menganalisis *multivariate time series* harga saham TAPG dan faktor yang mempengaruhinya. Berdasarkan hasil penelitian, data harga saham TAPG dan harga CPO terdapat kointegrasi, sehinga model *multivariate time series* yang digunakan adalah *vector error correction model* (VECM). Pada model VECM lag optimumnya yang digunakan adalah lag 11. Dalam jangka panjang harga CPO berpengaruh signifikan terhadap harga saham TAPG.

Kata kunci: *Kointegrasi*; *Multivariate Time Series*; *VAR*; *VECM*.

ABSTRACT

An increase in the price of crude palm oil (CPO) positively affects the share prices of companies engaged in the palm oil industry. PT Triputra Agro Persada Tbk (TAPG) 2021 was recorded as one of the companies with the CPO business that received the most significant capital gain. Prediction or forecasting of stock prices in the future is crucial for investors as a consideration before deciding to invest. Many kinds of research on stock price prediction have been carried out previously using univariate methods. Univariate modeling cannot represent the influence of other variables on stock prices. Forecasting with the influence of other variables can be done with multivariate time series analysis. This study aims to analyze the multivariate time series of TAPG stock prices and the factors that influence them. Based on the research results, data on TAPG stock prices and CPO prices are cointegrated, so the multivariate time series model used is the vector error correction model (VECM). In the VECM model, the optimum lag used is lag 11. In the long run, CPO prices significantly affect TAPG stock prices.

Keywords : *Cointegration; Multivariate Time Series; VAR; VECM.*

INTRODUCTION

From the end of 2021 until the first quarter of 2022, the price of cooking oil in traditional and modern markets has increased [1]. The leading cause of the increase in cooking oil prices in Indonesia is the increase in global demand for crude palm oil (CPO), which is used as a raw material, causing an increase in world CPO prices that has occurred since mid-2020 [2].

The increase in CPO prices positively influences the share prices of companies operating in the palm oil industry. PT Triputra Agro Persada Tbk (TAPG) 2021 was listed as one of the companies with a CPO business that obtained the most significant capital gain, where the company's shares provided a return of more than 50%. This company operates in the oil palm plantation and palm oil processing industry. TAPG recently joined BEI. TAPG conducted an initial public offering (IPO) on 12 April 2021 but can compete with large companies that have long been on the IDX. Investors are undoubtedly interested in investing in TAPG shares. However, predictions or forecasting of future share prices are necessary for investors to consider before investing.

Previous stock price prediction research conducted univariately has been widely carried out, such as stock price forecasting with Autoregressive Moving Average Generalized Autoregressive Conditional Heteroscedasticity. In this research, the data is under heteroscedasticity conditions. The best models to predict pre-pandemic conditions are GARCH (1, 1) and GARCH (1, 2) during pandemic conditions [3]. PT. Telkom Share Price Forecasting Using the Hybrid Time Series Regression Linear Model–Autoregressive Integrated Moving Average. As a result of this research, the best hybrid TSR linear-ARIMA (2, 1, 1) model was obtained [4]. Analysis of Oil and Gas Stock Price Volatility Analysis in Indonesia during the Covid-19 Pandemic using the ARIMA-GARCH Method. The results of this research for APEX, ELSA, and RUIS shares show there are symptoms of heteroscedasticity in the ARIMA model and the ARIMA(0, 1, 1) GARCH(1, 1) model for APEX, ELSA, and RUIS companies and the ARIMA(4, 1, 4) for MEDC companies [5]. Univariate modeling cannot represent the influence of other variables on share prices. Forecasting with the influence of other variables can be done using multivariate time series analysis. Multivariate time series analysis is a statistical technique used to analyze and model datasets that involve observing multiple variables over a series of time points.

Several studies that have been carried out previously have used multivariate time series, which can be considered in this research, including Multivariate Forecasting to Determine Global Gold Prices. Research results showed that the vector error correction model (VECM) could model the gold's price well and that all factors under investigation affected the gold's price [6]. Vector Autoregressive Integrated (VARI) Method for Forecasting the Number of International Visitors in Batam and Jakarta. Based on the research results, the model used is VARI (5, 1) [7]. Comparison of the Error Rate of Autoregressive Distributed Lag (ARDL) and Vector Auto-regressive (VAR). This research aimed to explain the application of the Autoregressive distributed-lag model and Vector Autoregressive (VAR) method for forecasting the export amount in DIY. It takes the export amount in DIY, inflation data, kurs, and Indonesia's foreign exchange reserve. Forecasting formation: After defining the Koyck and Almon distributed-lag dynamic model, the best model is chosen, and distribution-lag dynamic forecasting is performed [8]. The Influence of the United States Dollar Exchange Rate, Inflation, and Interest Rates on the Composite Stock Price Index using the Vector Error Correction Model. The VECM model obtained is VECM (2), which shows that changes in the US Dollar Exchange rate variable positively influence the IHSG. In contrast, Inflation and Interest Rates negatively influence changes in the IHSG [9]. Vector Error Correction Model Approach for Analysis of the

Relationship between Inflation, BI Rate, and United States Dollar Exchange Rate. Based on specifications, estimation, and model examination, the VECM (5) model is the best. [10].

In this research, TAPG share prices will be modeled multivariate. Multivariate modeling was carried out to get a more accurate model and to know other variables on TAPG share prices. The variable analyzed in this research is the price of CPO and whether it can influence the TAPG share price. The multivariate time series model is the vector error correction model (VECM).

METHOD

The data used in this research is secondary data, namely stock price data (closing prices) with a daily period from PT Triputra Agro Persada Tbk (TAPG) and CPO prices starting 13 April 2021 to 13 April 2022. The multivariate time series model used in this research is the vector error correction model (VECM). The following are the steps that will be carried out in this research:

1. Carry out stationary tests using the Augmented Dickey-Fuller tests [11].
2. Determine the optimum lag length using the Akaike Information Criterion (AIC), Schwartz Information Criterion (SIC), Hannan-Quin Criterion (HQC), Likelihood Ratio (LR) and Final Prediction Error (FPE) [12].
3. Cointegration occurs, it is necessary to correct errors; for this reason, VECM is used. The Johansen cointegration test is used to determine the long-term relationship between variables [13].
4. Carrying out the Granger Causality test is used to see one-way or two-way relationships in the TAPG and CPO variables in the VECM model [14] .
5. Estimating the VECM model and forecasting TAPG share prices and CPO prices for the next ten days.
6. Conduct impulse response function (IRF) analysis to track the effect of a shock that occurs in one variable on other variables [15].

RESULT AND DISCUSSION

Observation Data

The type of data used in this research is secondary data, namely stock price data (closing price) with a daily period from PT Triputra Agro Persada Tbk (TAPG) and CPO prices starting 13 April 2021 to 13 April 2022, which consists of 248 data (Figure 1).

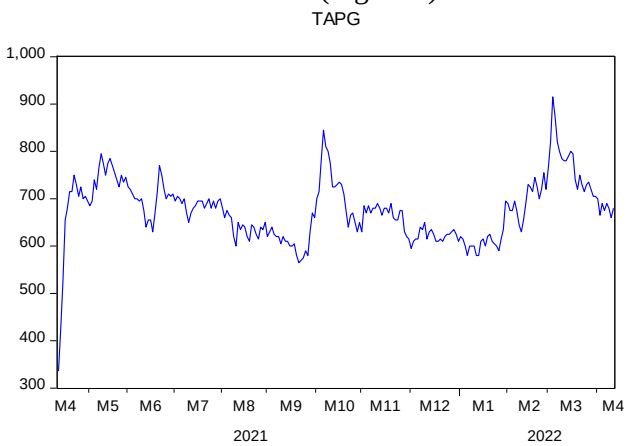


Figure 1. Daily closing price of TAPG shares (Rupiah**)**)

Figure 1 shows data that fluctuates and tends to be stationary on average. From October 2021 to December 2021, TAPG's share price tends to decline. From January 2022 to March 2022, TAPG's share price will increase (Figure 2).

Figure 2. Daily CPO price (US$/ton)

Figure 2 shows data that fluctuates and tends to be non-stationary on average. From June 2021 to March 2022, CPO prices tend to increase. The highest price occurred in March 2022.

Stationary Test

In time series analysis, the assumption of data stationarity is an important property. A time series data is said to be stationary if there is no increase or decrease in data over time; the data is spread around a constant mean and variance. Stationarity testing uses the Augmented Dickey-Fuller test. They are calculated using the t-statistic value of the Augmented Dickey-Fuller test (Table 1). The hypothesis used is:

Ho: $\phi_1 = 0$ (Y_t not stationary)

H_1: $\phi_1 < 0$ (Y_t stationary)

Table 1. Augmented Dickey Fuller test

Data	Augmented Dickey Fuller test	t-Statistic dengan α=5%	Probability
TAPG	-5.469873	-2.873045	0.0000
CPO level	-0.878600	-2.873045	0.7939
CPO lag-1	-14.70 104	-2.873045	0.0000

Table 1 shows that the TAPG data at the level has an Augmented Dickey-Fuller t-test value (-5.469873), which is smaller than the t-table (-2.873045) or the probability value (0.0000) is smaller than the significance level value (0.05), so reject Ho. The TAPG data is at a stationary level. The CPO data at the level has an Augmented Dickey-Fuller t-test value (-0.878600), which is greater than the t-table (-2.873045), or the probability value (0.7939) is greater than the significance level value (0.05) so accept Ho. The CPO data at the level is not stationary. The CPO data at lag-1 has an Augmented Dickey-Fuller t-test value (-14.70104), which is smaller than the t-table (-2.873045) or the probability

value (0.0000) is greater than the significance level value (0.05) so reject Ho. The CPO data at lag-1 is stationary.

Determine the Optimum Lag

Estimating the VECM model requires determining the optimal lag length, which will be used in subsequent analyses. Determining the optimal lag length is beneficial in eliminating autocorrelation in the VECM, which will be used as a stability analysis in the VECM. Determining the optimum lag length is known by looking at the lag with the most asterisks (*) in each of the Akaike Information Criterion (AIC), Schwartz Information Criterion (SIC), Hannan-Quin Criterion (HQC), Likelihood Ratio (LR) and Final Prediction Error (FPE) (Table 2).

Table 2. VAR lag order selection criteria

Lag	LogL	LR	FPE	AIC	SC	HQ
0	-2.464.913	NA	9427715	21.73492	21.76509	21.74709
1	-2.218.077	487.1480	1109746	19.59539	19.68592*	19.63192
2	-2.207.244	21.18927	1044917	19.53519	19.68607	19.59607*
3	-2.205.467	3.443746	1065614	19.55478	19.76601	19.64001
4	-2.200.999	8.581231	1061278	19.55066	19.82224	19.66024
5	-2.195.007	11.40407	1042882	19.53310	19.86503	19.66704
6	-2.191.619	6.388406	1048622	19.53849	19.93078	19.69678
7	-2.188.637	5.569342	1058215	19.54746	20.00010	19.73011
8	-2.186.096	4.700733	1072098	19.56032	20.07331	19.76732
9	-2.183.605	4.565634	1086687	19.57361	20.14695	19.80496
10	-2.177.806	10.52586	1069889	19.55776	20.19145	19.81346
11	-2.170.386	13.33605*	1038481*	19.52763*	20.22167	19.80768
12	-2.169.924	0.822104	1071790	19.55880	20.31319	19.86321
13	-2.168.824	1.937463	1100048	19.58436	20.39910	19.91312
14	-2.164.887	6.868992	1101253	19.58491	20.46000	19.93802
15	-2.159.756	8.859877	1091022	19.57494	20.51039	19.95241
16	-2.157.550	3.771143	1109193	19.59075	20.58655	19.99257
17	-2.153.946	6.096993	1113968	19.59424	20.65039	20.02041
18	-2.151.858	3.494726	1133928	19.61108	20.72759	20.06161
19	-2.151.468	0.646477	1171760	19.64289	20.81974	20.11777
20	-2.147.996	5.689666	1178556	19.64754	20.88475	20.14677

Table 2 above shows that lag-11 has the most signs (*), namely 3 with an LR value of 13.33605, an FPE value of 1038481, and an AIC criterion of 19.52763, so this shows that the optimum lag occurs at lag-11. So, the model will use lag-11.

Stability Test

Test the stability of the VAR model on TAPG shares and daily CPO prices to determine whether the VAR model used is stable (Table 3). The VAR model is considered stable if the characteristic inverse root has a modulus value < 1 or all the points are inside the circle (Figure 3).

Table 3. VAR model stability test results on TAPG shares and CPO prices

Root	Modulus
0.902280 + 0.080243i	0.905841
0.902280 - 0.080243i	0.905841
-0.743166 - 0.516016i	0.904748
-0.743166 + 0.516016i	0.904748
-0.309314 + 0.847774i	0.902439
-0.309314 - 0.847774i	0.902439
0.194752 + 0.862690i	0.884399
0.194752 - 0.862690i	0.884399
0.624524 - 0.598305i	0.864869
0.624524 + 0.598305i	0.864869
0.314865 - 0.802390i	0.861957
0.314865 + 0.802390i	0.861957
-0.584701 - 0.620009i	0.852225
-0.584701 + 0.620009i	0.852225
-0.841190 + 0.036708i	0.841990
-0.841190 - 0.036708i	0.841990
0.775165 - 0.296340i	0.829879
0.775165 + 0.296340i	0.829879
-0.038162 - 0.798688i	0.799599
-0.038162 + 0.798688i	0.799599
0.683852	0.683852
-0.619084	0.619084

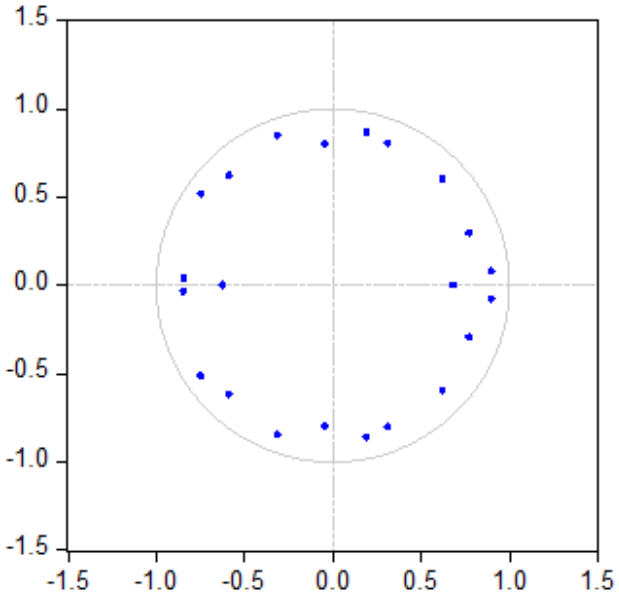


Figure 3. Inverse roots of AR characteristic polynomial

Table 3 shows that all moduli have values < 1. Figure 3 shows that the points are inside the circle; this means that the VAR model is in a stable condition.

Cointegration Test

The Johansen cointegration test is used to determine the long-term relationship between variables. At the stationary test stage, Augmented Dickey-Fuller (ADF) shows that the CPO variable is stationary at the First Difference. So, carrying out a Johansen cointegration test is necessary to see the long-term relationship between the TAPG share price variables and daily CPO prices (Table 4).

Table 4. Cointegration Test

Hypothesized No. of CE(s)	Eigen value	Trace Statistic	0.05 Critical Value	Prob.**
None *	0.095	30.940	15.495	0.0001
At most 1 *	0.031	7.448	3.842	0.0063

Table 4 shows that the probability value is smaller than the significance level α=5%; this means there is cointegration in the model. So, the model that will be used is the VECM model.

Granger Causality Test

The Granger causality test is used to see one-way or two-way relationships in the TAPG and CPO variables in the VECM model. Based on the Granger causality test in Table 5, it shows that TAPG significantly influences CPO, with a probability value of 0.0287 < 0.05, and CPO significantly influences TAPG, with a probability value of 0.031 < 0.05. So, there is a two-way causality between TAPG and CPO.

Table 5. Granger Causality Test

Null Hypothesis:	Obs	F-Statistic	Prob.
TAPG does not Granger Cause CPO	237	2.01071	0.0287
CPO does not Granger Cause TAPG	237	2.67288	0.0031

VECM Model Estimation

The model between variables in this study is VECM with lag 11, using a deterministic trend with the assumption of no intercept, no trend, and one cointegration (Table 6). The VECM model if TAPG is the dependent variable is as follows:

$$D(TAPG) = C(1)*(TAPG(-1) - 0.427789766837*CPO(-1)) + C(2)*D(TAPG(-1)) + C(3)*D(TAPG(-2)) + C(4)*D(TAPG(-3)) + C(5)*D(TAPG(-4)) + C(6)*D(TAPG(-5)) + C(7)*D(TAPG(-6)) + C(8)*D(TAPG(-7)) + C(9)*D(TAPG(-8)) + C(10)*D(TAPG(-9)) + C(11)*D(TAPG(-10)) + C(12)*D(TAPG(-11)) + C(13)*D(CPO(-1)) + C(14)*D(CPO(-2)) + C(15)*D(CPO(-3)) + C(16)*D(CPO(-4)) + C(17)*D(CPO(-5)) + C(18)*D(CPO(-6)) + C(19)*D(CPO(-7)) + C(20)*D(CPO(-8)) + C(21)*D(CPO(-9)) + C(22)*D(CPO(-10)) + C(23)*D(CPO(-11)) \quad (1)$$

Table 6. VECM coefficient with TPAG as dependent variable

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-0.010098	0.010964	-0.920984	0.3581
C(2)	0.001572	0.071985	0.021840	0.9826
C(3)	-0.157124*	0.071903	-2.185.241	0.0300
C(4)	-0.035250	0.071947	-0.489945	0.6247
C(5)	-0.023473	0.071562	-0.328012	0.7432
C(6)	0.003092	0.069659	0.044381	0.9646
C(7)	-0.018209	0.069332	-0.262643	0.7931
C(8)	-0.125067	0.068826	-1.817.147	0.0706
C(9)	0.057415	0.069323	0.828229	0.4085
C(10)	-0.105243	0.065083	-1.617.049	0.1073
C(11)	-0.014492	0.063280	-0.229014	0.8191
C(12)	-0.025057	0.061990	-0.404203	0.6865
C(13)	0.093138*	0.033902	2.747.247	0.0065
C(14)	0.134615*	0.035886	3.751.175	0.0002
C(15)	0.037349	0.036781	1.015.439	0.3110
C(16)	0.085114*	0.036552	2.328.595	0.0208
C(17)	-0.006210	0.036726	-0.169083	0.8659
C(18)	-0.065943	0.036999	-1.782.311	0.0761
C(19)	-0.059160	0.035891	-1.648.317	0.1008
C(20)	0.019201	0.035730	0.537385	0.5916
C(21)	0.023702	0.035485	0.667941	0.5049
C(22)	0.045722	0.034225	1.335.931	0.1830
C(23)	0.053892	0.031144	1.730.430	0.0850

Based on Table 6, it can be seen that the coefficients C(3), C(13), C(14), and C(16) are significant because they have a probability value smaller than the value α=5%. So, it can be interpreted that in the short term, the change in TAPG's share price two days ago significantly influenced changes in TAPG's current share price. If the change in TAPG's share price two days ago increased by 1 rupiah, it would cause the current change in TAPG's share price to fall by 0.157124 rupiah. In the short term, changes in CPO prices one day ago significantly influenced TAPG's current share prices. If the change in the CPO price one day ago increased by 1 dollar, it would cause the change in the current TAPG share price to increase by 0.093138 rupiah. In the short term, changes in CPO prices two days ago significantly influenced TAPG's current share prices. If the price of CPO 2 days ago increased by 1 dollar, it would cause the change in the current TAPG share price to increase by 0.134615 rupiah. In the short term, changes in the CPO price four days ago significantly influenced changes in the current TAPG share price. If the CPO price increased by 1 dollar four days ago, it would cause the change in the current TAPG share price to increase by 0.085114 rupiah (Table 7).

The VECM model if CPO is the dependent variable is as follows:

D(CPO) = C(24)*(TAPG(-1) - 0.427789766837*CPO(-1)) + C(25)*D(TAPG(-1)) + C(26)*D(TAPG(-2)) + C(27)*D(TAPG(-3)) + C(28)*D(TAPG(-4)) + C(29)*D(TAPG(-5)) +

C(30)*D(TAPG(-6)) + C(31)*D(TAPG(-7)) + C(32)*D(TAPG(-8)) + C(33)*D(TAPG(-9)) + C(34)*D(TAPG(-10)) + C(35)*D(TAPG(-11)) + C(36)*D(CPO(-1)) + C(37)*D(CPO(-2)) + C(38)*D(CPO(-3)) + C(39)*D(CPO(-4)) + C(40)*D(CPO(-5)) + C(41)*D(CPO(-6)) + C(42)*D(CPO(-7)) + C(43)*D(CPO(-8)) + C(44)*D(CPO(-9)) + C(45)*D(CPO(-10)) + C(46)*D(CPO(-11)) (2)

Table 7. VECM coefficient with CPO as the dependent variable

	Coefficient	Std. Error	t-Statistic	Prob.
C(24)	0.012587	0.023387	0.538189	0.5910
C(25)	-0.172913	0.153545	-1.126.139	0.2614
C(26)	-0.232738	0.153369	-1.517.503	0.1306
C(27)	0.013543	0.153464	0.088252	0.9298
C(28)	-0.139134	0.152643	-0.911501	0.3631
C(29)	0.019711	0.148583	0.132659	0.8946
C(30)	-0.009082	0.147885	-0.061413	0.9511
C(31)	0.256917	0.146807	1.750.028	0.0816
C(32)	0.037776	0.147867	0.255470	0.7986
C(33)	0.276100*	0.138824	1.988.847	0.0480
C(34)	-0.400692*	0.134978	-2.968.580	0.0033
C(35)	-0.059817	0.132226	-0.452384	0.6515
C(36)	-0.301166*	0.072314	-4.164.670	0.0000
C(37)	-0.106542	0.076546	-1.391.874	0.1654
C(38)	0.005086	0.078454	0.064830	0.9484
C(39)	0.182824*	0.077965	2.344.947	0.0199
C(40)	0.242396*	0.078337	3.094.270	0.0022
C(41)	0.129729	0.078919	1.643.822	0.1017
C(42)	0.006538	0.076557	0.085400	0.9320
C(43)	-0.090709	0.076214	-1.190.192	0.2353
C(44)	-0.090522	0.075689	-1.195.967	0.2330
C(45)	0.139044	0.073002	1.904.672	0.0582
C(46)	-0.081330	0.066430	-1.224.290	0.2222

Based on Table 7, it can be seen that the coefficients C(33), C(34), C(36), C(39), and C(40) are significant because they have a probability value smaller than the value α=5%. So, it can be interpreted that in the short term, changes in TAPG share prices nine days ago significantly influenced changes in current CPO prices. If the TAPG share price increased by 1 rupiah 9 days ago, it would cause the change in the current CPO price to increase by 0.276100 dollars. In the short term, changes in TAPG share prices ten days ago significantly influenced current CPO prices. If the TAPG share price increased by 1 rupiah 10 days ago, it would cause the change in the current CPO price to decrease by 0.400692 dollars. In the short term, changes in CPO prices one day ago significantly influenced current CPO prices. If the CPO price increased by 1 dollar one day ago, it would cause the current CPO price to decrease by 0.301166 dollars. In the short term, changes in CPO prices four days ago significantly influenced current CPO prices. If the CPO price increased by 1 dollar four days

ago, it would cause the current CPO price to increase by 0.182824 dollars. In the short term, changes in CPO prices five days ago significantly influenced current CPO prices. If the CPO price increased by 1 dollar five days ago, it would cause the current CPO price to increase by 0.242396 dollars. In the long term, CPO prices significantly affect TAPG share prices.

Forecasting TAPG share prices and CPO prices

Based on the results of VECM modeling, forecasting of TAPG share and CPO prices for the next ten days can be predicted, as in Table 8.

Table 8. Forecasting TPAG share prices and CPO prices for the next ten days

Date	TAPG	CPO
14/4/2022	687,353	1678,825
15/4/2022	712,861	1672,707
16/4/2022	718,301	1657,182
17/4/2022	711,784	1671,627
18/4/2022	713,584	1662,646
19/4/2022	713,092	1671,799
20/4/2022	708,384	1656,295
21/4/2022	710,194	1657,632
22/4/2022	705,026	1679,577
23/4/2022	708,717	1674,195

Based on Table 8, the forecast for the TAPG share price in the next three days will experience an increase, while on the fourth day, it will experience a decrease; on the fifth day, it will increase. It will decrease again on days 6 to 7, then increase on the eighth day, then decrease again. On day nine and day ten, there was an increase. Forecasting results show that prices fluctuate and tend to rise. Meanwhile, the forecast price of CPO shows a contradictory relationship with the TAPG share price.

Impulse Response Function

The impulse response function (IRF) is an approach to viewing relationships between variables. IRF is a dynamic function that tracks the influence of a shock that occurs in one variable on other variables (Figure 4).

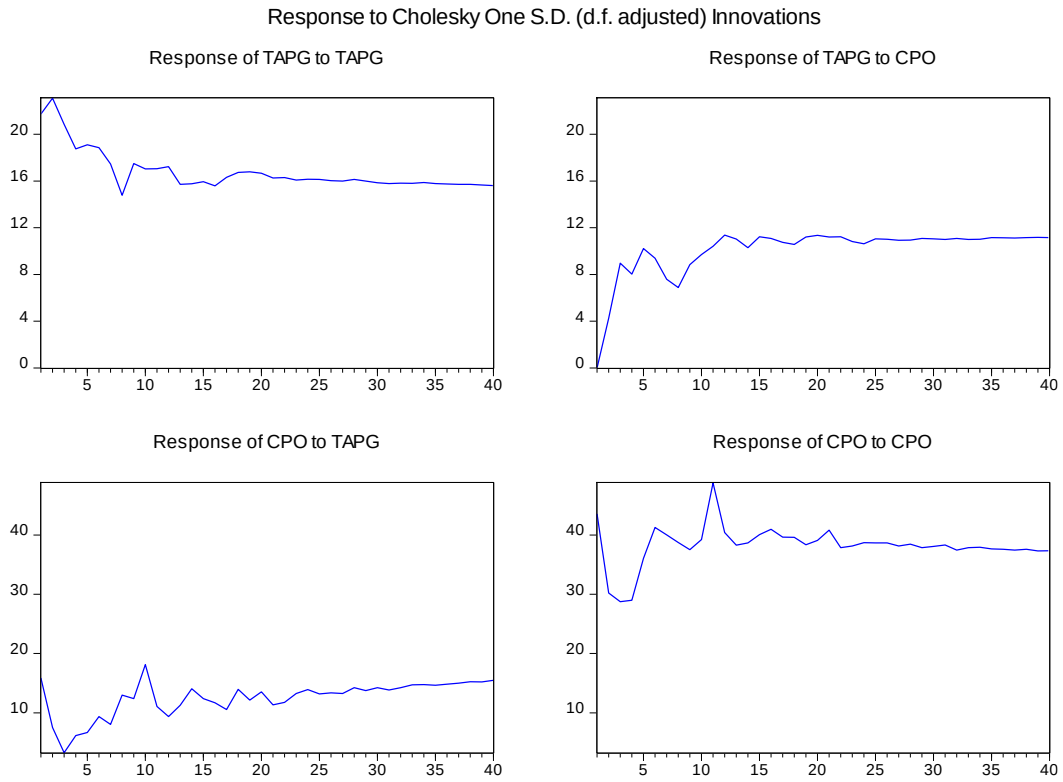


Figure 4. Impulse Response Function

Figure 4 shows the IRF analysis for the next 40 days. The horizontal axis shows the period, where one time period represents one day. Meanwhile, the vertical axis shows changes in variables due to shocks to certain variables, where these changes are expressed in a standard deviation.

TAPG's response to itself decreased until day 10. After day 10, TAPG's reaction to shocks from itself tended to stabilize. TAPG's response to shocks from CPO increased until day five, decreased until day 8, and then increased until day 16. After day 16, TAPG's reaction to shocks from CPO tended to be stable. CPO's response to shocks from TAPG decreased until day two, increased until day 10, and then decreased until day 15. After day 15, CPO's reaction to shocks from TAPG tended to be stable. CPO's response to itself decreased until day five, increased until day 11, and then decreased until day 13. After the 13th day, CPO's reaction to shocks from himself tended to stabilize.

CONCLUSION

Based on the research results, TAPG share price data and CPO prices are cointegrated, so the VECM model is used. In the VECM model, the optimum lag used is lag 11. Based on the results of the estimation of the Vector Error Correction Model (VECM) equation system, in the short term, changes in the current TAPG share price are significantly influenced by changes in the TAPG share price two days ago, changes in CPO prices. One day ago, the CPO price changed two days ago, and the CPO price changed four days ago. In the short term, changes in current CPO prices are significantly influenced by changes in TAPG share prices nine days ago, changes in TAPG share prices ten days ago, changes in CPO prices one day ago, changes in CPO prices four days ago and changes in CPO prices four days ago. Meanwhile, in the long term, CPO's price significantly affects the TAPG

share price. This study only used two-time series data to examine the model. In further research, other variables can be added, such as inflation, dollar exchange rates, and share prices of other palm oil companies.

REFERENCE

[1] BPS, " Minyak Goreng Murah, Apa Kata Data," 2022. https://bigdata.bps.go.id/documents/datain/feb_2022_edit.pdf

[2] CPOPC, " Palm Oil Supplay and Demand Outlook Report 2021,"2021. https://www.cpopc.org

[3] N. N. Layla and E. Kurniati," Peramalan Indeks Harga Saham dengan Autoregressive Moving Average Generelized Autoregressive Conditional Heteroscedasticity (ARMA-GARCH), " *Jurnal Riset Matematika,* vol. 1, no.1, pp. 7-12, 2021.

[4] K. Ramadani, S. Wahyunigsih, and M.N. Hayati, "Forecasting Stock Price PT. Telkom Using Hybrid Time Series Regression Linear– Autoregressive Integrated Moving Average Model," *Jurnal Matematika, Statistika dan Komputasi, v*ol. 18, no. 2, pp. 293-307, 2022.

[5] N. Septiana, P. Hasanah and A. R. Soemarsono,"Analisis Volatilitas Harga Saham Sekor Minyak dan Gas di Indonesia pada Masa Pandemi Covid-19 dengan Metode ARIMA-GARCH," *J Statistika,* vol 14, no. 2, pp. 99-109, 2021.

[6] D.Christian and S. Halim, "Peramalan Multivariat untuk Menentukan Harga Saham Global," *Jurnal Teknik Industri,* vol. 18, no. 2, pp. 137-144, 2016.

[7] S. Wulandary," Vector Autoregressive Integrated (VARI) Method for Forecasting the Number of Internasional Visitor in Batam and Jakarta, *Jurnal Matematika, Statistika dan Komputasi*, vol. 17, no. 1, pp. 94-108, 2020.

[8] D. Kusumaningrum and S. Surono," Comparison the Error Rate of Autoregressive Distributed Lag (ARDL) and Vector Autoregressive (VAR) (Case study: Forecast of Export Quantities in DIY)", *Eksakta,* vol 18, no.2, pp. 168-177, 2018.

[9] N. L. P. D. Wikayanti, Q. Aini, and N. Fitriyani," Pengaruh Kurs Dolar Amerika Serikat, Inflasi, dan Tingkat Suku Bunga Terhadap Indeks Harga Saham Gabungan dengan Vector Error Correction," *Eigen Mathematics Journal*, vol. 3, no. 1, pp. 64-72, 2020.

[10] L. J. Sinay, "Pendekatan Vector Error Correction Model Untuk Analisis Hubungan Inflasi, BI Rate Dan Kurs Dolar Amerika Serikat," *Jurnal Barekeng,* vol. 8, no. 2, pp. 9–18, 2014.

[11] F. Fitriyani, S. Fasya A, M. Irfan R, and T. Amar T, " Peramalan Indek Harga Saham PT Verena Multi Finance Tbk Dengan Metode Pemodelan ARIMA dan ARCH-GARCH," *J Statistika,* vol. 14, no. 1, pp. 11-23, 2021.

[12] L. Loves, M. Usman, Warsono, Widiarti, and E. Russel, " Modeling Multivariate Time Series by Vector Error Correction Models (VECM) (Study: PT Kalbe Farma Tbk. and PT Kimia Farma (Persero) Tbk)," *Journal of Physics: Conference Series*, . **175**1 012013, 2021, pp. 1-9.

[13] A. A. Ikudayisi and K. K. Salman, "Spatial Integration of Maize Market in Nigeria- A Vector Error Correction Model," *International Journal of Food and Agricultural Economics*, vol. 2, no. 3, pp. 71-80, 2014.

[14] W. S. Nugroho, A. B. Astuti and S. Astutik, "Vector Error Correction Model to Forecasting Spot Prices for Coffee Commodities During Covid-19 Pandemic," *Journal of Physics: Conference Series*, **1811** 012076, 2021, pp. 1-9.

[15] I. Fahria and I. Sulistiana, "Vector error correction model to analyze energy uses, environmental quality and economic growth during Covid-19 Pandemic," *IOP Conference Series: Earth and Environmental Science*, **926** 012066, 2021,pp. 1-8.

Identifying Factors that Influence Life Expectancy in Central Java Using Spatial Regression Models

Prizka Rismawati Arum[(1)], Rahmad Putra Gautama[(2)], Indah Fitriani[(3)], Fellya Naza Nurvahyani[(4)].

[1,2,3,4] Departements of Statistics, Universitas Muhammadiyah Semarang, 20273, Semarang, Indonesia

e-mail: prizka.rismawatiarum@unimus.ac.id[(1)], rahmadgautama15@gmail.com[(2)], indahfitriyani110@gmail.com[(3)], fellya2181@gmail.com[(4)]

ABSTRAK

Angka Harapan Hidup adalah rata-rata harapan hidup penduduk dalam beberapa tahun, dengan asumsi angka kematian tetap konstan seiring bertambahnya umur. Fungsinya sebagai alat pengukur keberhasilan pembangunan kesehatan penduduk di perkotaan dan kesejahteraan umum, terutama dalam aspek kesehatan. Tinggi rendahnya angka harapan hidup dipengaruhi oleh beberapa indikator, seperti kondisi sosial ekonomi, lingkungan, dan kesehatan. Penelitian ini bertujuan mengidentifikasi komponen-komponen penting yang memengaruhi angka harapan hidup di 35 kabupaten dan kota di Provinsi Jawa Tengah melalui pendekatan proses dalam menganalisis metode regresi spasial. Selain itu, penelitian ini mencari persamaan regresi spasial terbaik untuk pemodelan angka harapan hidup di Provinsi Jawa Tengah. Regresi spasial adalah metode pengembangan regresi linier yang tergolong dalam elemen titik model. Menggunakan dua variabel independen terpilih dari tujuh variabel independen, penelitian ini mempelajari persamaan regresi spasial dengan pendekatan wilayah SAR, SEM, dan SARMA. Hasilnya menunjukkan bahwa model SAR terpilih dengan nilai p-value 0,02183 yang sesuai untuk mengidentifikasi ketergantungan efek spasial terhadap angka harapan hidup di Jawa Tengah. Tingkat Pengangguran Terbuka (X4) dan Persentase Penduduk Miskin (X6) adalah faktor signifikan yang memengaruhi angka harapan hidup di Jawa Tengah secara spasial.

Kata Kunci: Angka Harapan Hidup, Regresi Spasial, Jawa Tengah, Log Range Multiplier (SAR)

ABSTRACT

Life Expectancy is an average calculated over several years, assuming that mortality remains constant as age increases. It serves as a metric to gauge the success of population health development at the urban level and overall well-being, particularly in terms of health. Various indicators, including socioeconomic conditions, environmental factors, and health indicators, influence the highs and lows of life expectancy. This study in Central Java Province's 35 districts and cities aims to identify crucial components impacting life expectancy through a process-oriented spatial regression analysis. Additionally, the research endeavors to determine the optimal spatial regression equation for modeling life expectancy in the province. Spatial regression, a linear regression development method falling under the point element model, is employed. Utilizing two independent variables selected from seven, the study explores spatial regression equations using SAR, SEM, and SARMA area approaches. Data sourced from BPS in 2020 reveals that the SAR model, with a p-value of 0.02183, is apt for identifying spatial effects on Central Java's life expectancy. The Open Unemployment Rate (X_4) and the Percentage of Poor Population (X_6) emerge as significant spatial factors influencing life expectancy in Central Java.

Keywords: *Life Expectancy, Spatial Regression, Central Java, Lagrange Multiplier Log (SAR)*

INTRODUCTION

The Life Expectancy Rate is an estimate of the average lifespan of a population over several years, assuming that mortality does not change with age. In developing countries like Indonesia, the Life Expectancy Rate is considered low due to factors such as inadequate healthcare facilities, making the population susceptible to clinical disorders. As a result, mortality rates in developing countries are much higher compared to those in developed countries, leading to a lower life expectancy in these nations. The Life Expectancy Rate can be evaluated by examining the effectiveness of healthcare development for the population, measuring urbanization levels, and promoting the general well-being of the population, particularly in terms of health status. Several indicators, including health indicators, environmental conditions, and socio-economic factors, play a crucial role, with the most important indicators reflecting the level of life expectancy seen from the development of the health sector in a region, which has a significant impact on determining the standard of life expectancy [1].

According to data released by the Population Reference Bureau, Indonesia's Life Expectancy (LE) has been consistently below the global average each year, ranging from 69 to 71 years in 2018-2020 and increasing slightly to 72-73 years in subsequent year [21] s. In the period of 2018-2020, Indonesia held the 7th position in Life Expectancy among ASEAN countries. Based on statistics released by the Central Statistics Agency (BPS), Indonesia's Life Expectancy reached 71.6 years in 2021, reflecting a slight increase of 0.1 years compared to the previous year when it was 71.5 years in 2020. It's noteworthy that the Life Expectancy for females in Indonesia is higher compared to males [21].

Java Island is one of the large island clusters in Indonesia. According to the 2022 data from BPS, Indonesia has a population of 275.77 million people, with 56.05% of the population residing on Java Island. The island's high population is attributed to the various attractions it possesses, and it is also the economic center of Indonesia. The LE in Central Java Province ranks third, with a 2021 LE score of 74.5 years, indicating a 0.1-year increase from 2020. However, despite this improvement, Central Java couldn't maintain its position as the second-highest life expectancy province as it had in the previous year. In 2021, the average Life Expectancy in Central Java Province showed that residents in Sukoharjo Regency had the highest LE at 77.73 years, while residents in Brebes Regency had the lowest LE at 69.54 years [2].

From what has been explained, this research employs spatial regression analysis to examine the significant factors influencing Life Expectancy (LE) in Central Java and obtain the best model for modeling Life Expectancy in the Central Java Province. Linear regression analysis is a type of statistical analysis used to explain the correlation between independent and dependent variables. The method used to analyze relationships between different variables is called spatial regression. This method considers the spatial influence at different locations, making it the focus of the research. This issue emphasizes the possibility that spatial dependence in data, spatial dependence, and spatial variability can impact spatial data. [3]. Spatial modeling of this kind uses line and area methods. Spatial Autoregressive (SAR) is based on spatial lag effects, Spatial Error Model (SEM) is based on spatial error effects, and Spatial Autoregressive Moving Average (SARMA) are three crucial spatial methods for area analysis based on a combination of lag and spatial error effects [4].

The previous research, cited from the Journal of Data Analysis and written by Evi Ramadhani, Nani Salwa, and Madina Suha Mazaya (2022), regarding the analysis of factors using spatial regression, concludes that the factors influencing Life Expectancy in Sumatra in 2018 can be analyzed

using the best area approach, namely the Spatial Error Model (SEM) [5]. As a reference and another research source, a study by Fatkhurokhman Fauzi (2016) on the Best Spatial Regression Model for the Human Development Index in Central Java Province yielded the conclusion that Life Expectancy reflects the level and degree of health of a community [6].

This research aims to utilize spatial regression analysis to examine the significant factors influencing Life Expectancy (LE) across various regencies and cities in the Central Java Province. Additionally, the study strives to identify the best spatial regression model for accurately simulating LE, forecasting LE values in the future, and pinpointing the variables that influence LE in the province of Central Java. It is expected that insights into the LE model for each regency and city can assist communities and governments in promoting LE values, especially in less common areas.

METHODOLOGY

Data

The data used in this study is secondary data sourced from the Central Statistics Agency (Badan Pusat Statistik) of Central Java Province, consisting of 35 regencies/cities. The variables employed include Life Expectancy (Y), Proportion of Malnourished and Undernourished Toddlers (X_1), Proportion of Households with Sanitation Access (X_2), Proportion of Households with Access to Clean Water (X_3), Open Unemployment Rate (X_4), Real Per Capita Expenditure (X_5), Percentage of Poor Population (X_6), and Labor Force Participation Rate (X_7). These variables are observed across all 35 regencies/cities in Central Java Province.

Data Analysis Procedure

The following are the analysis steps conducted in the study:

1. Creating data based on the research variables used.
2. Performing a descriptive exploration of the research data.
3. Using Pearson correlation analysis to select independent variables.

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i) - (\sum_{i=1}^{n} Y_i)}{\sqrt{[n\sum_{i=1}^{n} {X_i}^2 - (\sum_{i=1}^{n} X_i)^2][n\sum_{i=1}^{n} {Y_i}^2 - (\sum_{i=1}^{n} Y_i)^2]}}$$

4. Conducting multiple linear regression analysis on life expectancy (LE) in Central Java, involving 35 regencies and cities and 7 independent variables.

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ki} + \varepsilon_i$$

5. Creating a spatial weighting matrix ($\boldsymbol{W}$) using the *Queen Contiguity* method.

$$W = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{bmatrix}$$

Formation of the result matrix used in the model, where:

$$W_{ij}^{*} = \frac{W_{ij}}{\sum_{i=1}^{n} w_{ij}}$$

6. Testing spatial dependence and heterogeneity using:
   a. *Moran's I* Test
      The spatial autocorrelation of residuals is tested using Moran's I test. Spatial autocorrelation is employed to assess whether there is correlation among members of a series and observations decomposed according to time and space.
   b. Spatial Heterogeneity Test
      Breusch-Pagan Test, testing spatial heterogeneity.
7. Determining the spatial regression model using the *Lagrange Multiplier (LM)* test.

$$LM_{SARMA} = E^{-1}\left\{\left(R_y\right)^2 T - 2R_y R_\varepsilon T + (R_\varepsilon)^2 (D+T)\right\} \sim \chi^2_{(k)}$$

8. Calculating and testing the parameters of the regression analysis model.
9. Analyzing the spatial regression model and drawing conclusions.

RESULTS AND DISCUSSION

Figure 1 illustrates the distribution map of Life Expectancy (LE) in Central Java, divided into five interval groups. This division aims to provide a more detailed overview of LE variations in the region. Each interval group encompasses a specific range of LE values representing categories within each section of the map. Visual observations on the map provide information about LE levels in Central Java, where an increasing color intensity indicates higher LE values. The use of different color scales offers clear visual cues about LE variations across the region. The transition of colors from light to dark reflects an increase in LE values. The presentation of information through this map is designed to enhance understanding of the geographic distribution of LE in Central Java. This approach provides a comprehensive insight into the health and life expectancy disparities in the region. Therefore, this map can serve as a valuable tool in the context of public health analysis and decision-making for related policy considerations.

Figure 1. LE data distribution map for Central Java Province

In Figure 1, the map illustrates the distribution of Life Expectancy (LE) in Central Java divided into five interval groups. The increasing intensity of color on the map indicates higher LE values. Summary statistics can provide information about the central tendency, dispersion, and trend patterns of the data. The average LE value in regencies/cities in Central Java in 2021 is 75.012 years, with a median value of 74.84 years. A higher mean value compared to the median suggests that the distribution of LE data in Central Java in 2018 tends to be right-skewed (positive skewness). The highest LE achievement, with a value of 77.73 years, is in Sukoharjo Regency, while the lowest LE achievement, with a value of 69.54 years, is in Brebes Regency.

Multiple Linear Regression Model LE in Central Java

In the initial stage of building the regression model, to identify which independent variables are related to LE, the following are the results of Pearson correlation analysis:

Table 1. Pearson Correlation

Variable	Value	p-value
X_1	-0.4303829	0.009859
X_2	0.4564822	0.005845
X_3	0.3004232	0.07951
X_4	-0.2875033	0.094
X_5	0.4412123	0.007975
X_6	-0.6186169	7.466×10^{-5}
X_7	0.2891654	0.09203

In Table 1, it is indicated that there are seven selected independent variables, namely: X_1, X_2, X_3, X_4, X_5, X_6, and X_7.

Table 2. Partial Significance Test of Linear Regression

	Estimate	t-value	p-value
Intercept	8.924×10^{1}	10.137	1.06×10^{-10}
X_1	-9.689×10^{-6}	-1.169	0.25266
X_2	2.103×10^{-2}	0.971	0.34015
X_3	-1.131×10^{-2}	-0.250	0.80431
X_4	-4.800×10^{-1}	-3.111	0.00437
X_5	1.031×10^{-4}	1.191	0.24389
X_6	-2.870×10^{-1}	-3.611	0.00123
X_7	-1.520×10^{-1}	-1.698	0.10101

The independent factors influencing life expectancy (LE) need to be identified. Identification of independent variables can be done by examining the reliability interval of parameters and the p-value from the t-test. If the parameter estimates include zero, the variable is not significantly influential. In this case, the influence of independent variables is examined using the p-value. In Table 2, it is shown that at a significant level (α) of 0.01 (10%), two variables independently significantly influence LE in Regencies and Cities in Central Java. The significant variables consist of the open unemployment rate per 1000 population (X_4) and the percentage of the population living in poverty (X_6). It can be observed that all p-values for the variables are less than α (0.1).

Formation of Spatial Weight Matrix

The Queen Contiguity weight matrix, also known as the WQ matrix (Queen Weight matrix), is utilized in spatial analysis to model the spatial relationships between neighboring areas or units. This matrix is employed to map the presence or intensity of neighbor relationships in spatial data. In the Queen Contiguity weight matrix, each area or unit is assigned a weight based on its neighbors. If two areas or units are adjacent, then the weight is 1, indicating the presence of a neighborly relationship. If two areas or units are not adjacent, then the weight is 0, indicating no neighborly relationship. This matrix defines Wij = 1 for entities that share a common side or vertex with the area of interest and Wij = 0 for other areas. An illustration of the spatial weight matrix is as follows:

$$W = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

The standardization of matrix results used in the model is:

$W_{ij}^{*} = \frac{W_{ij}}{\sum_{i=1}^{n} w_{ij}}$, Where $W_{ij(std)}$ is a weight matrix element that has been standardized

Testing Spatial Data Aspects

Spatial dependency test on residuals from linear regression using Moran's I. Moran's I is used to test the correlation between residuals and the regression line. The generated Moran's I value is 4.3773 (p-value = 0.000). The regression residual line indicates positive spatial autocorrelation at a significant level of 0.01. The Breusch-Pagan (BP) method is employed to test for heterogeneity. The resulting Breusch-Pagan (BP) value is 0.94643 (p-value = 0.3306), indicating no spatial heterogeneity or detected spatial homogeneity variance at the 0.01 level. The spatial regression model is performed after fulfilling the two spatial effect tests. Subsequently, Lagrange Multiplier (LM) testing is conducted to determine the appropriate spatial regression model. The LM test results are presented in Table 3 as follows:

Table 3. *Lagrange Multiplier* Test

	Value	p-value
Lagrange Multiplier lag (SAR)	5.2595	0,02183
Lagrange Multiplier error (SEM)	1.4317	0,23148
Lagrange Multiplier SARMA	6.5985	0,03691

The LM test results indicate that the SAR model is the most appropriate for testing the spatial dependence of LE in Central Java, based on the lowest p-value of 0.02183 among the three models. Therefore, LE in Central Java can be modeled using the Spatial Autoregressive (SAR) regression model.

***Spatial Autoregressive* (SAR)**

Table 4. Results of SAR Test

	Estiamate	z-value	p-value
Intercept	8.5080×10^{1}	11.6254	$<2.2 \times 10^{-16}$
X_1	-6.5694×10^{-6}	-0.9626	0.33573
X_2	4.6743×10^{-2}	2.2979	0.02157
X_3	1.2138×10^{-3}	0.0329	0.97378
X_4	-4.9205×10^{-1}	-3.9370	8.252×10^{-5}
X_5	1.0451×10^{-4}	1.4922	0.13564
X_6	-2.8512×10^{-1}	-4.4328	9.303×10^{-6}
X_7	-1.5232×10^{-1}	-2.1020	0.03555

Table 4 indicates that at a significance level of 0.01, there are 2 independent variables that significantly influence Life Expectancy (LE) in Central Java. The variables with significant values are the open unemployment rate (X_4) and the percentage of the population living in poverty (X_6).

Table 5. Parameter Estimation Results in SAR Model

	Estimate	Standar Error	z-value	p-value
Intercept	80.847296	0.998248	80.9892	$<2.2 \times 10^{-16}$
X_4	-0.334508	0.102310	-3.2696	0.001077
X_6	-0.356391	0.061661	-5.7799	7.475×10^{-9}

Based on table 5, the obtained SAR model equation for LE data in Central Java is as follows:

$$\hat{y}_i = 80.847296 - 0.334508X_4 - 0.356391X_6$$

to $$u_i = 0.55166 \sum_{j=1, i \neq j}^{n} w_{ij} u_j + \varepsilon_i$$

The results of the spatial autocorrelation coefficient (λ) being statistically significant spatially indicate the presence of spatial interdependence in the spatial residuals among one district/city with its neighboring district/city, amounting to 0.55166 multiplied by the average residual in its neighboring regions.

CONCLUSION

In this study, it can be concluded that based on the best spatial regression analysis, specifically the Spatial Autoregressive (SAR) model with Lagrange Multiplier (LM) testing, it was found that the Open Unemployment Rate (X_4) and Percentage of Poor Population (X_6) re factors influencing Life Expectancy (LE) in Central Java. These factors have a significant real impact at a level of 0.1 (10%). The Open Unemployment Rate (X_4) and Percentage of Poor Population (X_6) negatively affect Life Expectancy (LE) in Central Java.

REFERENCES

[1] Santika, N. Hanum, Safuridar, and Asnidar, "Pengaruh Jumlah Penduduk, Angka Harapan Hidup dan Rata-Rata Lama Sekolah terhadap Indeks Pembangunan Manusia di Kabupaten Aceh Tamiang," *Jurnal Ekonomi dan Manajemen*, vol. 2, no. 4, 2022.

[2] BPS, "Badan Pusat Statistik Kabupaten Semarang," Jun. 2023. Accessed: Jun.08, 2023. [Online]. https://semarangkab.bps.go.id/indicator/40/161/1/angka-harapan-hidup-ahh-menurutkabupaten-kota-dan-jenis-kelamin-di-jawa-tengah.html

[3] Z. Niaz Mahmud and K. Asif, "A Spatial Regression Modeling Framework for Examining Relationships Between the Built Environment and Pedestrian Crash Occurrences at Macroscopic Level: A Study in A Developing Country Context," *Geography and Sustainability*, vol. 3, no. 4, pp. 312–324, Dec. 2022, doi: 10.1016/j.geosus.2022.09.005.

[4] J. Olmo and M. Sanso-Navarro, "A Nonparametric Spatial Regression Model Using Partitioning Estimators," *Econom Stat*, Feb. 2023, doi: 10.1016/j.ecosta.2023.02.003.

[5] R. Evi, S. Nany, and M. Medina Suha, "Identifikasi Faktor-Faktor yang Memengaruhi Angka Harapan Hidup di Sumatera Tahun 2018 Menggunakan Analisis Regresi Spasial Pendekatan Area," 2020.

[6] F. Fatkhurohman, "Model Regresi Spasial Terbaik Indeks Pembangunan Manusia Provinsi Jawa Tengah," 2016.

[7] K. Suryowati, R. D. Bekti, and A. Faradila, "A Comparison of Weights Matrices on Computation of Dengue Spatial Autocorrelation," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Apr. 2018. doi: 10.1088/1757-899X/335/1/012052.

[8] W. S. Tarigan, "Analisis Regresi Spasial pada Indeks Pembangunan Manusia di Provinsi Sumatera Utara Tahun 2020 (Spatial Regression Analysis on the HDI in North Sumatera Province in 2020)," 2020.

[9] L. Ni Made Lasti, S. I Wayan, and S. I Komang Gede, "Pemodelan Jumlah Tindak Kriminalitas Di Provinsi Jawa Timur Dengan Analisis Regresi Spatial Autoregressive And Moving Average," *E-Jurnal Matematika*, vol. 7, no. 4, p. 346, Dec. 2018, doi: 10.24843/mtk.2018.v07.i04.p224.

[10] Firmansyah, Rangga Hadi. (2022). 5 Provinsi dengan Angka Harapan Hidup Tertinggi di Indonesia. *https://goodstats.id/article/5-provinsi-dengan-angka-harapan-hidup%20tertinggi-di-indonesia-xz4cQ*.

[11] Halicioglu, F. (2011). Munich Personal RePEc Archive Modelling life expectancy in Turkey. *Economic Modelling*, *28*(5), 2075–2082. https://doi.org/10.1016/j.econmod.2011.05.002.

[12] World Population Data Sheet. (2020). *Demographic Trends May Make Us Vulnerable to Pandemics Data Table*. 22. https://www.prb.org/wpcontent/uploads/2020/07/letter-booklet-2020-world-population.pdf

[13] Dindas Kesehatan Jayapura. Diakses pada tanggal 7 Mei 20023. *https://dinkes.jayapurakab.go.id/2933-2/#:~:text=H*.

[14] A. Yasir *et al.*, "Model Regresi Spasial untuk Analisis Presentase Penduduk Miskin di Provinsi Nanggroe Aceh Darussalam," *Jurnal Statistika Industri dan Komputasi*, vol. 1, no. 1, pp. 53-61, 2016.

[15] R. Faizatun Nisa and A. Rachman Hakim, "Pemodelan Mixed Geographically Weighted Regression dengan Adaptive Bandwidth untuk Angka Harapan Hidup (Studi Kasus: Angak Harapan Hidup di Jawa Tengah)," vol. 11, no. 1, pp. 67-76, 2022, [Online]. Available: https://ejournal3.undip.ac.id/index.php/gaussian/

[16] Y. Wardani, "Estimasi Parameter *Spatial Error Model* yang Memuat Pecilan," 2019.

[17] Chotimah Husnul and I. Rinjani, "Pemodelan Spasial Konsumsi Pemerintah dalam Perekonomian Jawa Timur: Spatial Autoregressive and Moving Average," *Jurnal Ilmiah Komputasi dan Statistika*, vol. 2, pp. 2087-3657, 2022

[18] Alfiani, S., Arum, P. R., & Arum, R. (2022). Pemodelan Pertumbuhan Ekonomi di Jawa Barat Menggunakan Metode Geographically Weighted Panel Regression. In *Universitas Muhammadiyah Semarang Jl. Kedungmundu* (Vol. 15, Issue 2). www.unipasby.ac.id

[19] Huriyatullah Rona Nabila, N., Fitri, Y., Rismawati Arum, P., Studi Statistika, P., & Matematika dan Ilmu Pengetahuan Alam, F. (2023). *Analisis Faktor-Faktor Yang Mempengaruhi Indeks Pembangunan Manusia Berdasarkan Kabupaten/Kota Di Jawa Tengah* (Vol. 16, Issue 1).

[20] Oktaviana, E., Arum, P. R., & Al Haris, M. (n.d.). *Pemodelan Spatial Autoregressive Quantile Regression (SARQR) Menggunakan Pembobot Queen Contiguity Pada Kasus Stunting Balita di Indonesia Spatial Autoregressive Quantile Regression (SARQR) Modeling Using Queen Contiguity Weights in Toddler Stunting Cases in Indonesia*.

[21] *Demographic Trends May Make Us Vulnerable to Pandemics Data Table*. (n.d.).