

## Binary Logistic Regression Analysis on the Spread of Dengue Fever in Bali Province

Luh Putu Safitri Pratiwi<sup>(1)</sup>, I Made Pasek Pradnyana Wijaya<sup>(2)</sup>

<sup>1</sup>Program Studi Sistem Informasi, ITB STIKOM Bali

Jl. Raya Puputan No. 86 Renon –Denpasar –Bali. Tlpn (0361) 244445

e-mail: [putu\\_safitri@stikom-bali.ac.id](mailto:putu_safitri@stikom-bali.ac.id)<sup>(1)</sup>, [pasek\\_pradnyana@stikom-bali.ac.id](mailto:pasek_pradnyana@stikom-bali.ac.id)<sup>(2)</sup>

### ABSTRAK

Penyebaran demam berdarah melibatkan siklus kompleks antara manusia sebagai inang dan nyamuk sebagai vektor. Gejala demam berdarah dapat bervariasi dari demam ringan hingga bentuk berat yang dapat mengancam jiwa. Salah satu wilayah yang memiliki angka persebaran demam berdarah tertinggi di Provinsi Bali adalah wilayah Denpasar. Penelitian terus dilakukan untuk memahami faktor-faktor yang memengaruhi penyebaran demam berdarah menggunakan metode regresi logistik biner. Regresi logistik biner merupakan model regresi yang sering digunakan dalam pemodelan data kategori, dimana variabel dependen dalam penelitian ini adalah penyebaran kasus demam berdarah dengan angka penyebaran kasus di masing-masing wilayah akan diberikan kategori nol untuk jumlah kasus sedikit dan satu untuk jumlah kasus tinggi. Sehingga dalam penelitian ini dikembangkan strategi yang lebih efektif dalam pengendalian penyakit demam berdarah serta bagaimana model terbaik dari data penyebaran demam berdarah di Provinsi Bali. Hasil yang diperoleh dari penelitian ini adalah uji pengaruh variabel independent terhadap variabel dependen memperlihatkan bahwa variabel jumlah sarana sanitasi layak ( $X_5$ ) memiliki pengaruh yang signifikan terhadap jumlah penderita demam berdarah di Provinsi Bali yaitu sebesar 0,081.

**Kata kunci:** Demam Berdarah; Regresi Logistik; Provinsi Bali

### ABSTRACT

*The spread of dengue fever involves a complex cycle between humans as hosts and mosquitoes as vectors. Symptoms of dengue fever can vary from a mild fever to a severe form that can be life-threatening. One of the areas that has the highest spread of dengue fever in Bali Province is the Denpasar area. Research continues to be carried out to understand the factors that influence the spread of dengue fever using the binary logistic regression method. Binary logistic regression is a regression model that is often used in modeling categorical data, where the dependent variable in this study is the distribution of dengue fever cases with the number of cases spreading in each region being assigned a category of zero for a low number of cases and one for a high number of cases. So in this research a more effective strategy was developed in controlling this disease as well as the best model for data on the spread of dengue fever in Bali Province. The results obtained from this research were a test of the influence of the independent variable on the dependent variable, showing that the variable number of adequate sanitation facilities ( $X_5$ ) had a significant influence on the number of dengue fever sufferers in Bali Province, namely 0.081.*

**Keywords:** Dengue Fever; Logistic Regression; Bali Province

## INTRODUCTION

Dengue fever (DHF), also known as dengue hemorrhagic fever, is an infectious disease caused by the dengue virus, which is transmitted by the *Aedes aegypti* and *Aedes albopictus* mosquitoes. This disease is a significant public health problem in various tropical and subtropical regions throughout the world [1]. Dengue fever (DHF) has become a global concern because of its serious public health impacts. Indonesia is one of the countries affected by dengue fever, with a number of provinces, including Bali Province, reporting significant cases. Bali Province, with a high population and tourist visits, has its own challenges in controlling this disease. In addition, there are factors such as climate change, urbanization, and population mobility that can influence the spread of the dengue virus in this region [2].

The occurrence of dengue fever cases can be explained by several factors, including the rainy season. During the rainy season, environmental conditions create ideal conditions for the development of *Aedes* mosquitoes, which spread dengue fever [3]. Previous research shows that factors such as regional neighbourhoods, area size, and the presence of jumantik influence the spread of dengue fever in Denpasar City [4]. A deeper understanding of the pattern of dengue fever spread is key to effective control and prevention efforts. For this reason, it is necessary to use an effective method in predicting the probability of an event occurring with certain independent variables using a method, namely logistic regression [5]. Logistic regression is a mathematical modeling approach used to analyze the relationship of one or several independent variables with a categorical dependent variable that is dichotomous/binary [6]. Research using logistic regression was investigated by [7], based on the results of the analysis it was found that the number of applicants who were declared accepted was 50% of the participants, while the remaining 50% of applicants were declared not accepted. In binary logistic regression modeling, the influencing variables are obtained, namely education (length of last education) and experience (work experience). Other research on binary logistic regression was carried out by [8] regarding the stunting problem in West Java where after modeling there were three variables that had a significant effect at a real level of 0,10, namely complete basic immunization, food management places that met health requirements, and poor people. Similar research was also conducted by [9] regarding maternal knowledge about immunizations on toddlers' health where maternal knowledge about immunizations did not affect toddlers' health in general, seen from the Nagelkerke coefficient of determination of 0,148 or 148%.

With this approach, we can understand the relative influence of various factors on the results we want to predict, as well as identify the most significant factors in influencing the spread of dengue fever in Bali Province.

## METHOD

This research took the location of each sub-district in Bali Province, totaling 57 sub-districts. The data used in this research is secondary data obtained from the Bali Province Central Statistics Agency, Bali Province Bappeda and Bali Province Health Service in 2022 [10]. The data contains information about the spread of dengue fever cases and the factors that influence it, which will be used as research variables. The following are the response variables and predictor variables used.

Table 1. Research Variable

Variable	Variable Type	Notation
Number of dengue fever sufferers in categories: 0=low/few numbers of dengue sufferers 1=high number of dengue sufferers	Variable Response	Y
number of drinking water facilities	Variable Predictor	X <sub>1</sub>
population density	Variable Predictor	X <sub>2</sub>
number of doctors	Variable Predictor	X <sub>3</sub>
number of health workers, namely nurses	Variable Predictor	X <sub>4</sub>
number of adequate sanitation facilities	Variable Predictor	X <sub>5</sub>

### Logistic Regression

Binary logistic regression is a logistic regression model with a response variable (Y) on a binary category scale, namely having two categories of values 0 and 1 [11]. Binary logistic regression analysis is used to look for patterns of probability relationships between variables x and p (the probability of events caused by x). The value of the logistic function ranges between 0 and 1.

The following is the function of logistic regression

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{1}$$

$\pi(x) = E(Y|x)$  represents the conditional average for all x values

A transformation for  $\pi(x)$  the value called the logit transformation is carried out to obtain the assumption that the log odds ratio value has a linear relationship with x [12]. This transformation will produce a function g(x) which is linear in its parameters [13].

$$g(x) = \ln \left[ \frac{\pi(x)}{1 - \pi(x)} \right] = \beta_0 + \beta_1 x \tag{2}$$

Another difference between linear regression and logistic regression is the distribution of the response variable. In the linear regression model, the response variable is assumed to be  $Y = \pi(x) + \varepsilon$  where  $\varepsilon$  is the error following a normal distribution with a mean equal to zero and constant variance. But in binary logistic regression, the error value only consists of two possibilities, namely if then  $\varepsilon = 1 - \pi(x)$  with chance  $\pi(x)$  or if  $y = 0$  then  $\varepsilon = -\pi(x)$  with chance  $1 - \pi(x)$ . So the error has a distribution with mean equal to zero and variance  $[\pi(x)(1 - \pi(x))]$ .

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)} \tag{3}$$

So the logit transformation form  $\pi(x)$  of equation (3) becomes:

$$g(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \tag{4}$$

$$g(x) = \sum_{j=0}^k \beta_j x_j, x_0 = 1$$

**Logistic Regression Model Parameter Estimator**

To estimate the parameters of the logistic regression model, use the Maximum Likelihood Estimator (MLE) method. Each observation has a distribution function [11]:

The ln likelihood function is as follows:

$$\begin{aligned}
 L(\beta) &= \ln[l(\beta)] = \ln \left[ \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \right] \\
 &= \sum_{i=1}^n y_i \beta^T x_i - \left( 1 + \exp(\beta^T x_i) \right)
 \end{aligned}
 \tag{5}$$

**Testing Logistic Regression Model Parameters**

After getting parameter estimates in a logistic regression model, the next step is to carry out tests to find out which predictor variables have a significant effect on the response variable. The significance test of the  $\beta$  coefficient in the model consists of a simultaneous test and a partial test.

**1. Simultaneous Test**

This test is used to check the significance of each  $\beta$  coefficient together in the model with the following hypothesis [14]:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{there is at least one } \beta_j \neq 0, j=1, 2, \dots, k$$

The test statistic used is the test statistic  $G_{hit}$  or Likelihood Ratio Test, namely [13].

$$G = -2 \ln \left[ \frac{\left( \frac{n_1}{n} \right)^{n_1} \left( \frac{n_0}{n} \right)^{n_0}}{\prod_{i=1}^n \hat{\pi}_i^{y_i} (1 - \hat{\pi}_i)^{1-y_i}} \right]
 \tag{6}$$

where

$$n_1 = \sum_{i=1}^n y_i ; n_0 = \sum_{i=1}^n (1 - y_i) ; n = n_0 + n_1$$

if  $G_{hit} > X^2_{(\alpha, p)}$  ; p is the number of parameters then Ho is rejected.

**2. Partial Test**

This test was carried out to find out whether the predictor variables had a significant effect on the response variable individually [15]. The hypothesis used is:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Wald test statistics:

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}
 \tag{7}$$

if  $|W| > Z_{\alpha/2}$  then Ho is rejected.

**RESULT AND DISCUSSION**

To determine the characteristics of each sub-district in Bali Province, descriptive statistical analysis was carried out on all the variables studied. These characteristics include exploration of

response and predictor variables so that broader information is obtained. The results of the descriptive analysis can be seen in Table 2 and Table 3.

**Table 2.** Percentage of DHF Categories in Bali Province

Group	Total	Percentage
Low DHF	49	85,96
High DHF	8	14,04

Based on Table 2 above, it can be seen that the majority of sub-districts in Bali are in the dengue fever group with a low/low number of dengue cases, namely 85,96 percent (49 sub-districts). Meanwhile, 14,04 percent (8 sub-districts) are classified as areas that have a high number of dengue cases.

The predictor variables in this study are factors that are thought to influence the dengue fever value in Bali Province. The descriptive analysis resulting from the predictor variables that are thought to have an influence on DHF is presented in the table as follows.

**Table 3.** Descriptive Statistics of Independent Variables

	Variabel	Mean	StDev	Min	Max
X <sub>1</sub>	Number of Drinking Water Facilities	26,60	43,993	1	279
X <sub>2</sub>	Population density	1438,33	1723,370	177	8620
X <sub>3</sub>	Number of Doctors	10,68	6,362	4	30
X <sub>4</sub>	Number of Health Personnel (Nurses)	30,51	19,418	12	114
X <sub>5</sub>	Number of Adequate Sanitation Facilities	19137,32	1043,216	753	49828

Table 3 shows that the number of drinking water facilities in Bali Province has an average of 26,60, where Manggis and Kediri Districts have the smallest number of drinking water facilities and Kintamani District has the largest drinking water facilities. The fastest population density is in West Denpasar, while the lowest population density is in West Selemadeg at 177,2. The average number of doctors in Bali is 10,68 with the lowest number being 4 doctors (Payangan, Tembuku and Susut Districts) and the highest being 30 (Abiansemal District). The role of nurses is also no less important in handling dengue fever cases in Bali where the lowest number of nurses is in Sidemen District at 12 people and the highest is in Abiansemal District at 114 people with an average number of nurses of 30,51. Apart from that, another factor that causes the number of dengue fever sufferers to increase is the small number of adequate sanitation facilities, where the average household that has adequate sanitation is 19.137,32. The lowest number of households that have proper sanitation is 753, namely located in Petang District and the highest, namely located in East Denpasar, is 49.828, with the number of sub-districts in Bali Province being 57 sub-districts. Tabanan Regency is the district with the largest number of sub-districts, namely 10 sub-districts.

**Model Fit Test**

Hosmer and Lemeshow Test to see the suitability or FIT of the model, with the following hypothesis [16].

H0: Model FIT (p value>0,1)

H1: Model does not FIT

**Table 4.** Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4,461	8	0,813

From table 4, a sig value of 0,813>0.1 is obtained, so H<sub>0</sub> is accepted, which means the FIT model. The FIT model in question is a binary logistic regression model suitable for use for further analysis because there is no significant difference between the predicted classification and the observed classification.

**Coefficient of Determination Test**

Nagelkerke R Square is a test carried out to find out how much the independent variable is able to explain and influence the dependent variable. The Nagelkerke R Square value varies between 1 and 0. If the value is closer to 1, then the model is considered to have greater goodness of fit. Conversely, if the value is closer to 0, then the model is considered not to be good of fit [17].

**Table 5.** Coefficient of Determination Test

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	25,580 <sup>a</sup>	0,304	0,547

From the results of Table 5, the Nagelkerke R Square of 0,547 shows that the ability of the independent variable to explain the dependent variable, namely DHF, is 54,7% and the rest is explained by other variables not studied.

**Logistic Regression Model**

The Bali Province DBD data used in this research is binary category data, so a Logistic regression model is used. Logistic regression functions to determine the pattern of relationship between the independent variable (X) and the dependent variable (Y) in the form of categories.

**a. Simultaneous Testing**

The simultaneous formation of a regression model aims to obtain an appropriate and simple model based on factors that are considered to influence dengue fever. According to Vikaliana (2022), to test the influence of the independent variables together on the dependent variable by looking at the significant values of the Omnibus Tests of Model Coefficients table with the hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_1 : \text{there is at least one } \beta_j \neq 0$$

$$\alpha = 10\%=0,1$$

The test criterion is to reject H<sub>0</sub> if Sig. <10%

**Table 6. Simultaneous Testing**

		<b>Omnibus Tests of Model Coefficients</b>		
		Chi-square	df	Sig.
Step 1	Step	20,659	6	0,002
	Block	20,659	6	0,002
	Model	20,659	6	0,002

From table 6, a sig value of 0,02 < 0.1 is obtained, so H0 is rejected, which means the number of drinking water facilities (X<sub>1</sub>), population density (X<sub>2</sub>), number of doctors (X<sub>3</sub>), number of health workers, namely nurses (X<sub>4</sub>), number of facilities Proper sanitation (X<sub>5</sub>) influences the dependent variable, namely dengue fever.

**b. Individual Testing**

The following are the results of the parameter estimation of the Logistic regression model with. With the following hypothesis.

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

$$\alpha = 10\%$$

The test criterion is to reject H<sub>0</sub> if Sig. < 10%

**Table 7. Logistic Regression Model Parameter Estimation**

Variable	B	S.E.	Wald	Sig.	Exp(B)
X <sub>1</sub>	-0,036	0,049	0,556	0,456	0,964
X <sub>2</sub>	0,000	0,000	0,261	0,609	1,000
X <sub>3</sub>	0,085	0,180	0,225	0,635	1,089
X <sub>4</sub>	0,008	0,043	0,035	0,851	1,008
X <sub>5</sub>	0,000	0,000	3,036	<b>0,081</b>	1,000
Constant	-7,070	2,114	11,187	0,001	0,001

\*) Parameter yang signifikan pada α=10%

The results of the test for the influence of the independent variable on the dependent variable in Table 7 are that the variable number of adequate sanitation facilities. Meanwhile, the odds ratio or Exp(B) model is selected from only significant variables, so the model is formed as follows.

$$\ln \left( \frac{P(y-1)}{1-P(y-1)} \right) = \alpha + \beta_1 X_1$$

$$= -7,070 + 0,040 X_1$$

From the significant model, it can be interpreted that the odds ratio value or Exp(B) of variable X<sub>5</sub>, which is the number of adequate sanitation facilities, is 1,000. This indicates that with an increase in the number of adequate sanitation facilities, it is estimated to have a 1 times greater chance of reducing the number of cases of dengue fever.

**Probability Estimation**

$$P(y=1) = \frac{e^{\alpha + \beta_1 X_1}}{1 + e^{\alpha + \beta_1 X_1}}$$

$$P(y=1) = \frac{e^{-7,070 + 0,040 X_1}}{1 + e^{-7,070 + 0,040 X_1}}$$

$$\hat{\pi}(x) = \frac{\exp(2,931x_3 - 0,142x_5 + 1,18x_6)}{1 + \exp(2,931x_3 - 0,142x_5 + 1,18x_6)}$$

$$\hat{g}(x) = -7,070 + 0,040 X_1$$

Based on the odds ratio value above, it is known that the greater the amount of adequate sanitation, a sub-district will have a higher number of dengue sufferers than low dengue fever, or in other words, the greater the amount of adequate sanitation, the greater the tendency for a sub-district to have low dengue fever.

Based on the Logistic regression model above, it is possible to reduce the dengue fever value for each sub-district in Bali by increasing the amount of adequate sanitation for each sub-district in Bali.

**Calculation of Logistic Regression Classification Accuracy**

Based on probability calculations, prediction results can be obtained, so that the truth of the logit model can be seen based on the results of classification between predictions and observations.

**Table 8.** Classification of DHF Results Logistic Regression Model

Observation	Prediction		Accuracy percentage
	Low DHF	High DHF	
Low DHF	48	1	98,0
High DHF	3	5	62,5
Overall Accuracy Percentage			93

Table 8 informs that the number of low category dengue fever is known in 49 sub-districts. The prediction results for areas with the lowest DHF were correctly classified in the low category as many as 48 areas, while 1 area was incorrectly classified (from low to high), namely East Denpasar District with a classification accuracy of 98%. Meanwhile, the number of high DBD categories is 8 sub-districts, where the areas with high DBD change to low DBD are 3 areas, namely Klungkung, Karangasem and South Kuta sub-districts, and the high DHF areas remain in the upper high DBD category in 5 areas with an accuracy percentage of 62,5% and the overall percentage of accuracy is 93%, which means that the percentage of accuracy that the model can predict correctly is 93%.

**CONCLUSION**

The results obtained from this research are a test of the influence of the independent variable on the dependent variable shows the variable number of adequate sanitation facilities ( $X_5$ ) has a value of 0,081 which is smaller than the significance level of 0,1, so the decision taken is that  $H_0$  is

rejected, which means that the number of adequate sanitation facilities influences the number of dengue fever sufferers in the province of Bali.

So from a significant model it can be interpreted that the value of the odds ratio or  $\text{Exp}(B)$  variable ( $X_5$ ), namely the number of adequate sanitation facilities, is 1,000. This indicates that if there is an increase in the number of adequate sanitation facilities, it is estimated that the number of dengue fever will decrease by 1 time.

## ACKNOWLEDGMENTS

Thank you to ITB STIKOM Bali for providing research funding grants to individual lecturers for this research so that it can be completed well.

## REFERENCE

- [1] Ibrahim, A. Irviani, H. Habibi, and E. Latifahanun, "Faktor-Faktor Yang Berhubungan dengan Ketepatan Waktu Pelaksanaan Penyelidikan Epidemiologi Demam Berdarah Dengue ( DBD ) Tingkat Puskesmas Di Kota Makassar Tahun 2015." *Al-Sihah : Public Health Science Journal* 7(2), 2015.
- [2] World Health Organization (2022) Dengue and severe dengue. Available at: <https://www.who.int/news-room/factsheets/detail/dengue-and-severe-dengue>
- [3] M. Andani, " Hubungan Pengetahuan Dan Sikap Kepala Keluarga Dengan Pelaksanaan 4m Plus Dalam Mengatasi DBD Di Wilayah Kerja Puskesmas Basuki Rahmad Bengkulu Tahun 2016", Skripsi, Program Studi Ilmu Keperawatan ,Stikes Dehasen Bengkulu,2016.
- [4] Malela, "Asuhan Keperawatan pada Anak dengan Dengue Hamorrhagic Fever (DHF) di RSUD dr. Wahid Sudiro", Asuhan Keperawatan Hemorrhagic.
- [5] N. M. S. Jayanti, I. W. Sumarjaya, dan M. Susilawati, "Pemodelan Penyebaran Kasus Demam Berdarah Dengue (DBD) di Kota Denpasar dengan Metode SpatialAutoregressive (SAR)", *E-Jurnal Matematika*, Vol. 6 (1), Januari 2017, pp.37-46, 2017
- [6] A. C. Kurniawan, A. Maharani, F.E. Zulvia, M. Lusiana, R. Vikaliana, Y. A. Iskandar dan N. Ruswandi, Penerapan Statistik dalam Logistik, Jakarta: Media Sains Indonesia, 2023.
- [7] A. I. Sofiyat, A. Tjalla, & Mahdiyah, "Pemodelan Regresi Logistik Biner Terhadap Penerimaan Pegawai Di PT XYZ Jakarta". *Matematika Sains*, 1(1), 1–11, 2023.
- [8] S. A. R. Mana, Erfiani, Indahwati, A. Fitrianto, & R. Amelia, "Faktor –Faktor yang Memengaruhi Permasalahan *Stunting* di Jawa Barat Menggunakan Regresi Logistik Biner". *J Statistika*, 15(2), 265 -274, 2022.
- [9] R. Oktavianingsih dan W. Pramesti, "Pengetahuan Ibu tentang Imunisasi terhadap Kesehatan Balita di Desa Sumpat Kecamatan Driyorejo Kabupaten Gresik". *J Statistika*, 10(2), 27-34, 2017.
- [10] Kementerian Kesehatan Republik Indonesia. Profil Kesehatan Indonesia. Profil Kesehatan Provinsi Bali. 2022.
- [11] A. Agresti, Categorical Data Analysis, John Willey and Sons, New York,1990.
- [12] Zainal M, M., A. J. Rindengan, dan W. Ch. D. Weku, "Penggunaan Association Rule Data Mining Untuk Menentukan Pola Lama Studi Mahasiswa F-MIPA UNSRAT", *Jurnal de Cartesian*, 3(1):1-8, 2014.
- [13] D. W. Hosmer, & S. Lemeshow, Applied Logistic Regression, John Wiley and Son, New

York, 2000.

- [14] Misna, Rais, L. T. Utami, "Analisis Regresi Logistik Biner untuk Mengklasifikasi Penderita Hipertensi Berdasarkan Merokok di RSUD Mokopido Tolitoli" dalam Natural Science : *Journal Of Science And Technology*, Vol 7 (3) : 341-348. Palu : Universitas Tadulako, 2018.
- [15] R. D. Haloho & S. I. Santoso, "Analisis profitabilitas pada usaha peternakan sapi perah di Kabupaten Semarang". *Ragam*, 13(1), 2013
- [16] Lai, X., & Liu, L, "A simple test procedure in standardizing the power of HosmerLemeshow test in large data sets", *Journal of Statistical Computation and Simulation*, 1. 2018.
- [17] I. Ghozali. Aplikasi Analisis Multivariate Dengan Program SPSS. Semarang: Badan Penerbit Universitas Diponegoro, 2011.