BAREKENG: Journal of Mathematics and Its Applications
September 2024 Volume 18 Issue 3 Page 1391–1402
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol18iss3pp1391-1402

LOGISTIC AND PROBIT REGRESSION MODELING TO PREDICT THE OPPORTUNITIES OF DIABETES IN PROSPECTIVE ATHLETES

Danang Ariyanto [1*], A'yunin Sofro[2], A'idah Nur Hanifah[3], Junaidi Budi Prihanto[4], Dimas A. Maulana[5], Riska W. Romadhonia[6]

[1,2,5] Department of Actuarial Science, Faculty of Mathematics and Natural Science, Universitas Negeri Surabaya
[3,6] Department of Mathematics, Faculty of Mathematics and Natural Science, Universitas Negeri Surabaya
Jl. Ketintang, Ketintang, Kec. Gayungan, Kota Surabaya, Jawa Timur 60231 Indonesia

[4] Department of Sport Education, Faculty of Sport Science, Universitas Negeri Surabaya
Jl. Lidah Wetan, Lidah Wetan, Kec. Lakarsantri, Kota Surabaya, Jawa Timur 60213 Indonesia

*Corresponding author's e-mail: * **danangariyanto@unesa.ac.id***

Article History:

Received: 9th September 2023
Revised: 6th Dec 2023
Accepted: 7th June 2024
Published: 1st September 2024

Keywords:

Anthropometric;
Athletic;
Diabetes;
Logistic Regression;
Probit Regression;
Socio-demographic.

ABSTRACT

Diabetes is among the most prevalent chronic diseases globally, posing significant health risks to individuals. The identification of individuals at risk of developing these conditions is of paramount importance, particularly in high-stress and physically demanding activities such as athletic training. To find out the chances of a prospective athlete suffering from diabetes or not, models for binary data can be used, including logistic regression and probit models. The data used is primary data from prospective athletes in East Java, including prospective athletes from the State University of Surabaya and East Java Koni Athletes. This study aimed to develop an early prediction model for diabetes in prospective athletic candidates using a bivariate logistic and probit regression approach while considering the influence of socio-demographic and anthropometric factors. To selecting the best model between logistic regression and probit regression using Akaike's Information Criterion (AIC) value, the smaller the AIC value gets means that the model is closer to the actual value or being the best model. Logistic regression has a smaller AIC value (129,85) than probit regression, this means that the logistic model is the best model. In this paper, an attempt is made to explore the use of logistic and probit regression to determine the factors which significantly influence the diabetes disease and we got that the logistic model as the best model because it has a smaller AIC value than the probit model. Based on the result of analysis and discussion, it can be concluded that there are two factors called mother's job and finance which are influenced to the response variable, diabetes disease at significance level of 5%.

How to cite this article:

D. Ariyanto, A. Sofro, A. N. Hanifah, J. B. Prihanto, D. A. Maulana and R. W. Romadhonia., "LOGISTIC AND PROBIT REGRESSION MODELING TO PREDICT THE OPPORTUNITIES OF DIABETES IN PROSPECTIVE ATHLETES," *BAREKENG: J. Math. & App.*, vol. 18, iss. 3, pp. 1391-1402, September, 2024.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Diabetes mellitus., commonly referred to as diabetes, represents a global health challenge of unparalleled proportions. Its insidious nature, with steadily increasing prevalence worldwide, poses a substantial threat to public health and the quality of life of millions of individuals. Diabetes is a chronic metabolic disorder characterized by hyperglycaemia resulting from either insufficient insulin production, insulin resistance, or a combination of both [1]. The consequences of uncontrolled diabetes are manifold and include cardiovascular disease, renal failure, blindness, neuropathy, and a host of other complications. Thus, early detection and intervention are crucial in managing this debilitating condition effectively [2].

The prevalence of diabetes is not confined to a particular demographic or geographic group; it affects individuals across diverse populations, regardless of age, gender, or physical activity level [3]. However, for prospective athletic candidates who engage in high-stress and physically demanding activities, the implications of diabetes can be especially profound. The demands of athletic training, coupled with the need for precise nutrition and energy balance, render athletes uniquely susceptible to diabetes-related challenges [4]. Consequently, it is imperative to develop early prediction models that can identify individuals at risk of diabetes, enabling tailored preventive measures and timely interventions within the athletic community. This study endeavors to contribute to the body of knowledge by focusing on the early prediction of diabetes in prospective athletic candidates. We employ a rigorous statistical approach, comparing the effectiveness of logistic and probit regression models in identifying individuals at risk. Furthermore, we explore the influence of socio-demographic and anthropometric factors on diabetes susceptibility within this specific population. Our chosen variables encompass a range of characteristics that have been previously associated with diabetes risk, including height, weight, BMI, waist circumference, age, gender, and socio-economic factors such as father's occupation, mother's occupation, and financial status.

The choice to employ both logistic and probit regression models in this study stems from the desire to offer a comprehensive assessment of predictive modelling techniques. Logistic regression and probit regression are frequent regression models used for modelling and analysis of categorical data. Both models included in the generalized linear model and differentiated by the logit and probit linking functions. Model logistic regression and probit regression are alternative approaches to models the relationship between the categorical response variable and the independent variable, where the response variable has a Bernoulli or multinomial distribution. While both logistic and probit regressions are commonly used in epidemiological and health-related studies, they have distinct mathematical underpinnings [5]. By comparing the performance of these models, we aim to identify the one that best fits our dataset and yields the most accurate predictions regarding diabetes risk among prospective athletic candidates.

This research scrutinizes a suite of variables that encompass both socio-demographic and anthropometric dimensions. Understanding how these factors interplay and influence diabetes risk within the context of athletic candidates is a central focus of investigation. Height is an anthropometric measure that may correlate with diabetes risk. Studies have suggested that taller individuals may have a reduced risk of diabetes due to potential differences in body composition and insulin sensitivity [6]. Based on research conducted by Amalia et al (2007), body weight is a fundamental indicator of overall health and is closely associated with diabetes. Excess body weight, particularly in the form of adipose tissue, contributes to insulin resistance and increases the likelihood of developing diabetes [7]. Research conducted by Saputra et al (2020) shows that BMI is a derived measure that relates an individual's weight to their height. It serves as a practical proxy for assessing whether an individual is underweight, normal weight, overweight, or obese. BMI is a widely recognized risk factor for diabetes [8]. Waist circumference is a measure of abdominal obesity, which is strongly linked to insulin resistance and an increased risk of diabetes. It is considered a critical anthropometric measurement in assessing diabetes risk. Age is a well-established risk factor for diabetes. As individuals grow older, their risk of developing diabetes tends to increase due to natural physiological changes, including decreased insulin sensitivity. Gender can also play a role in diabetes risk. Studies have shown variations in diabetes prevalence between males and females, suggesting that hormonal differences may contribute to susceptibility [9]. Socio-economic status, often approximated by the occupation of the father, can influence lifestyle factors such as diet and access to healthcare. It is recognized as an important determinant of diabetes risk [10]. The mother's occupation, similar to the father's occupation, may impact the family's socio-economic status and, consequently, an individual's risk of diabetes [11]. The financial status of an individual can have far-reaching implications for their dietary choices, healthcare access, and overall lifestyle. Financial stability or instability may be associated with diabetes risk [12].

Research conducted by Fathurahman et al (2023) to look at the factors that influence the HDI of districts/cities on the island of Kalimantan in 2017. This research shows that the best model for modelling the HDI of districts/cities in Kalimantan in 2017 is the logistic regression model [13]. The comprehensive examination of these variables allows us to assess the multifaceted nature of diabetes risk among prospective athletic candidates. By exploring how socio-demographic factors intersect with anthropometric measures, we aim to provide valuable insights that can inform diabetes prevention strategies tailored to this unique population. The primary objectives of this study are as follows: To develop an early prediction model for diabetes among prospective athletic candidates, to compare the performance of logistic and probit regression models in predicting diabetes risk within this specific population, to identify significant socio-demographic and anthropometric factors that influence the likelihood of diabetes among athletic candidates and to provide insights that can inform targeted interventions and preventive measures for reducing diabetes risk in the athletic community.

The significance of this study extends beyond its academic contributions. It holds practical implications for athletes, coaches, trainers, healthcare professionals, and policymakers involved in the care and training of athletes. Early prediction of diabetes risk can facilitate proactive measures such as personalized dietary plans, lifestyle modifications, and early medical interventions. By understanding the interplay of socio-demographic and anthropometric factors, stakeholders can implement strategies aimed at minimizing diabetes risk within the athletic community, ultimately promoting the well-being and athletic performance of prospective candidates.

2. RESEARCH METHODS

2.1 Dataset

The research conducted is applied research with primary data taken from prospective athletes in East Java, including prospective athletes from State University of Surabaya and East Java Koni Athletes in 2023. The number of individuals in this study were all prospective athletes who took part in the selection of prospective athletes. In this study, two observations were made, namely socio-demographic and anthropometry. Socio-demographic observations involve collecting data on characteristics social and demographic characteristics of individuals or groups. This includes factors such as age, gender, father's job, mother's job, finance. Anthropometric observations on athletes are carried out with the aim of understanding body proportions and physical characteristics that influence an athlete's ability in sports certain. This includes factors such as height, weight, BMI, and waist.

2.2 Logistic Regression

Logistic regression is a statistical analysis method for describing relationships between response variable that is dichotomous (nominal or ordinal scale with two category) or polychotomous (nominal or ordinal scale with more than two category) with one or more predictor variables on continuous or categorical scale [14]. Logistic regression can be divided into binary, multinomial, and ordinal logistic regression [15]. In this paper, we will discuss about binary logistic regression that the response variable is categorized into 0 (fail) and 1 (success) [16]. The form of probability in binary logistic regression model can be expressed in the probability function [17]:

$$\boldsymbol{f(y_i) = \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}} \quad (1)$$

where $\boldsymbol{y_i = 0; 1, \pi(x_i)}$ is the probability of success, $\boldsymbol{1-\pi(x_i)}$ is the probability of failure and $\boldsymbol{i = 1, 2, \dots, n}$ is the observation index. Logistic regression model used are [5]:

$$\boldsymbol{\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1-e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}} \quad (2)$$

where $\boldsymbol{p}$ is the number of predictor variables. To estimate the regression parameter easier, $\boldsymbol{\pi(x)}$ in **Equation (2)** transformed. Then the logit transformation is:

$$g(x) = log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (3)$$

2.3 Parameter Estimation for Logistic Regression

Estimation of the regression model parameters is carried out using Maximum Likelihood Estimation (MLE) method [18]. Basically, the maximum likelihood method provides an estimated value of $\boldsymbol{\beta}$ to maximize the likelihood function [19]. Systematically, the likelihood function for the binary logistic regression model:

$$L(\beta) = \sum_{j=0}^{p} f(y_i;\beta) = \sum_{j=0}^{p} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \quad (4)$$

Maximize $L(\beta)$ on Equation (4) by transforming ln to the likelihood function [5]:

$$l(\beta) = \sum_{j=0}^{p}\left[\sum_{i=1}^{n} y_i x_{ij}\right]\beta_j - \sum_{i=1}^{n} ln\left[1 + exp\left(\sum_{i=1}^{n}\beta_j x_{ij}\right)\right] \quad (5)$$

Obtain an estimator for β by deriving the ln likelihood function to the parameter $\boldsymbol{\beta}$ and then equate it to zero:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij}\,\pi(x_i) = 0\,; j = 0, 1, 2, \ldots, p \quad (6)$$

where p is the number of predictor variables and

$$\pi(x_i) = \frac{e^{\sum_{j=1}^{p}\beta_j x_{ij}}}{1 - e^{\sum_{j=1}^{p}\beta_j x_{ij}}}; j = 0, 1, 2, \ldots, p$$

Based on the estimation results for parameter $\boldsymbol{\beta}$ with the maximum likelihood method above, an implicit function is obtained. So, the Newton-Raphson method is needed to optimize Equation (6) [20].

2.4 Probit Regression

Probit regression is a non-linear model with a normal distribution used to describing relationships between two or more variables with categorial data that have ε_i as error factor [21]. In some books and other references, the probit model is said to be with another name, namely the normit model. Since the response variable is binary, then the binomial distribution with the probability function [5]:

$$f(y_i;\pi_i) = \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \quad (7)$$

where $y_i = 0; 1, \pi_i$ is the probability occurance of the i for $y_i = 1$, and $1 - \pi_i$ is the probability occurance of the i for $y_i = 0$.

$$P(y_i = 1|x_i) = \Phi(x'_i\beta) = \int_{-\infty}^{x'_i\beta} \phi(z)dz \quad (8)$$

where $\Phi(.)$ is the cumulative function of the normal distribution and $\phi(.)$ is the normal distribution probability function [22].

$$\Phi(g(x)) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{g(x)} e^{\frac{z^2}{z}}dz \quad (9)$$

In general, the probit regression model can be expressed in the form[23]:

$$\pi_i = \Phi(Z_i) = \Phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i) \quad (10)$$

Since probit model is related to the cumulative function of Normal distribution, it can be written probit model [24]:

$$Z_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (11)$$

To obtain an expectation of the probit value (Z_i), then the inverse of the normal cumulative distribution function can be obtained [25]:

$$\pi_i = \Phi^{-1}(Z_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (12)$$

2.5 Parameter Estimation for Probit Regression

One of the parameter estimation in the probit regression is by using the Maximum Likelihood Estimation (MLE) method that estimates the parameter $\boldsymbol{\beta}$ by maximizing the likelihood function with the condition that the data follows a certain distribution (normal distribution) [26]. The likelihood function for the probit regression model [27]:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i; \pi_i) \quad (13)$$

The cumulative function of normal distribution function is then transformed into the probit regression function:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \Phi(x'_i\boldsymbol{\beta})^{y_i}(1 - \Phi(x'_i\boldsymbol{\beta}))^{1-y_i} \quad (14)$$

where $\Phi(.)$ is the cumulative function of the normal distribution. Then, maximize $L(\boldsymbol{\beta})$ on **Equation (14)** by transforming $\boldsymbol{ln}$ to the likelihood function, so we get:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \ln\left(\Phi(x'_i\boldsymbol{\beta})\right) + \sum_{i=1}^{n} (1 - y_i) \ln\left(1 - \Phi(x'_i\boldsymbol{\beta})\right) \quad (15)$$

Obtain an estimator for β by deriving the $\boldsymbol{ln}$ likelihood function of **Equation (15)** to the parameter $\boldsymbol{\beta}$ and then equate it to zero. However, it often does not get explicit result, so the Newton-Raphson method is needed to optimize the equation [20].

2.6 Testing of Parameter Significance

The model obtained needs to be tested, so the relationship between the response variable and the predictor variables can be clearly defined. The test is simultaneous and a partial test [5]. The simultaneous test is carried out to determine the effect of predictor variables on the response variable simultaneously with the following hypothesis [28]:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_j = 0$
H_1: at least one $\beta_j \neq 0; i = 1, 2, \ldots, j$
The form of G test statistic or likelihood ratio test [20]:

$$G = -2ln\left[\frac{\left(\frac{n_0}{n}\right)^{n_0}\left(\frac{n_1}{n}\right)^{n_1}}{\sum_{i=1}^{n} \hat{\pi}_i^{\,y_i}(1 - \hat{\pi}_i)^{1-y_i}}\right] \quad (16)$$

where $n_1 = \sum_{i=1}^{n} y_i$ or the number of observations with value $y = 1$, $n_0 = \sum_{i=1}^{n}(1 - y_i)$ or the number of observations with value $y = 0$, and $n = n_1 + n_0$. The G test statistic follow the Chi-Square distribution, so that reject H_0 if $G > X^2_{(\alpha,p)}$ or $p - value < \alpha$. The partial test is done by testing each parameter β_j individually [29]. It is conducted to determine the effect of each predictor variable on the response variable with the following hypothesis [19]:

$H_0: \beta_j = 0$
$H_1: \beta_j \neq 0; j = 1, 2, \ldots, p$
According to Hosmer and Lemeshow, the Wald test statistic [30]:

$$W = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)} \tag{17}$$

Where $\hat{\beta}_j$ is the estimator β_j and $\widehat{SE}(\hat{\beta}_j)$ is the standard error estimator β_j [19]. The test statistic W follows the standard normal distribution, so that reject H_0 if $|W| > Z_{(1-\alpha)/2}$ or $p - value < \alpha$.

2.7 Best Model Selection Criteria

Akaike's Information Criterion (AIC) was introduced by Akaike as an information criteria that is used to obtain a model that fits to the data used. In addition, Konishi said that AIC is also being a comparison between statistical models considered as a basis for evaluating suitability model [31]. Akaike Information Criterion (AIC) equation [32]:

$$AIC = -2\log(L(\beta)) + 2p \tag{18}$$

where $L(\beta)$ is likelihood estimation and p is the number of variables.

3. RESULTS AND DISCUSSION

3.1 Descriptive Statistical Analysis

To be able to understand the characteristics of the data, descriptive statistical analysis is needed. This is to provide an overview of the distribution of data for each variable. In this research, the response variable is Diabetes Disease and the data representation, as follows:

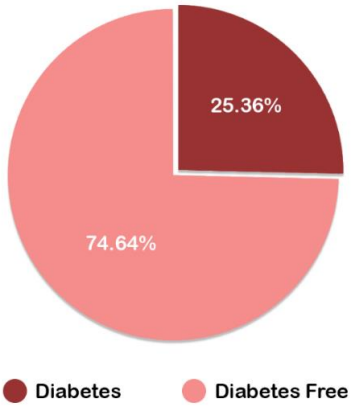

Figure 1. Percentage of the Response Variable

Based on the pie chart, there are only two categories of response variable. As much as 25.36% belong to category 1, state for a person suffering from diabetes, and 74.64% belong to category 0 state for someone who does not suffer from diabetes disease. There are 9 predictor variables that grouped into two, namely anthropometry (height/X_1, weight/X_2, Body Mass Index or BMI/X_3, waist/X_4) and socio-demographic (age/X_5, gender/X_6, father's job/X_7, mother's job/X_8, and finance/X_9). All of the predictor variables are also categorical.

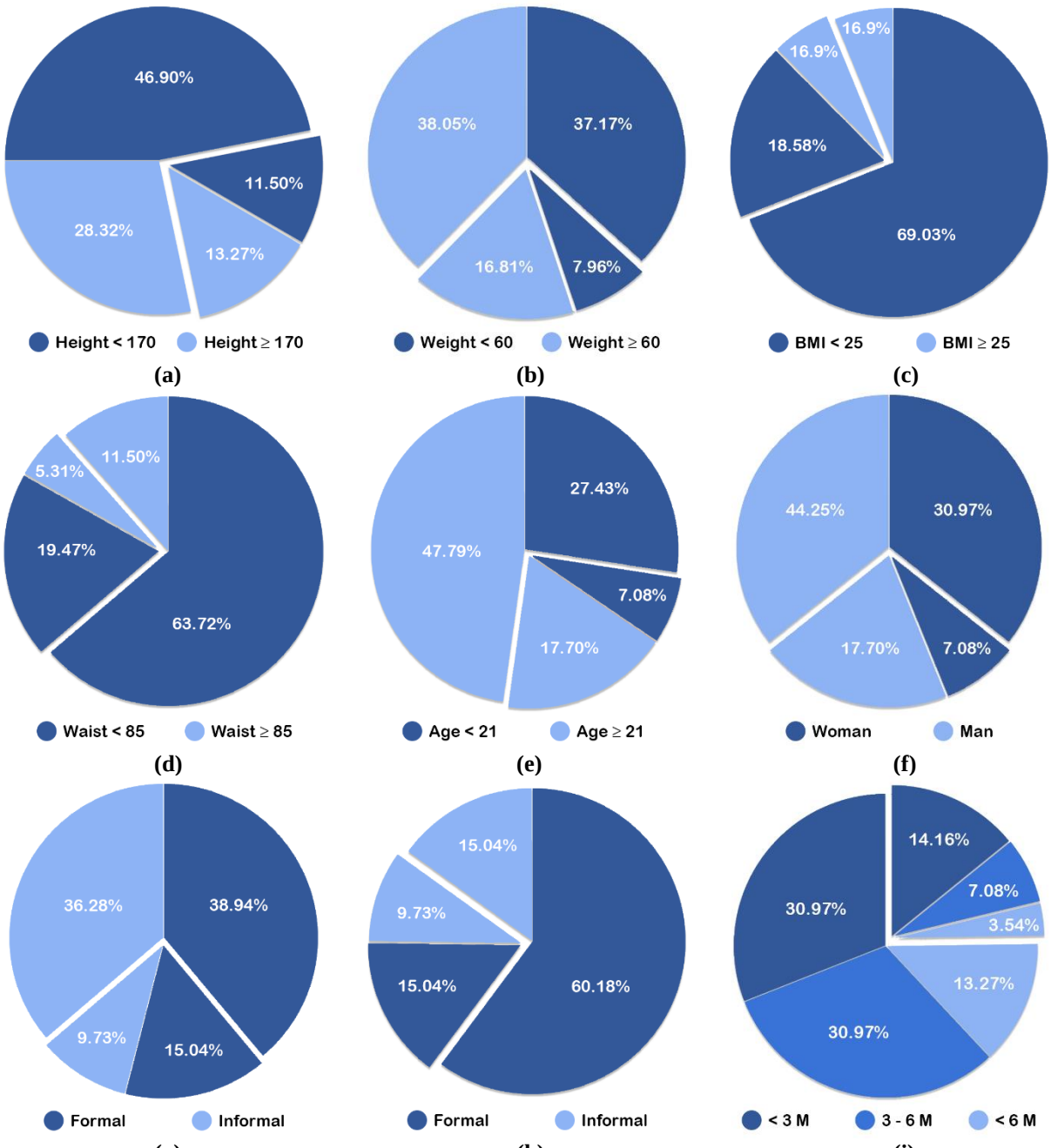

Figure 2. Variable Characteristics of (a) Height, (b) Weight, (c) Body Mass Index or BMI, (d) Waist, (e) Age, (f) Gender, (g) Father's job, (h) Mother's job, and (i) Finance.

Based on Figure 2, height is the measurement from head to foot that is divided into 2 categories: 0 for height under 170 cm, and 1 for height greater than or equal to 170 cm, in which there are 11.50% and 13.27% respondents who suffered from diabetes from categories 0 and 1, respectively. Weight is body's relative mass that is divided into 2 categories: 0 for weight under 60 kg, and 1 for weight greater than or equal to 60 kg, in which there are 7.96% and 16.81% respondents who suffered from diabetes from categories 0 and 1, respectively. Body Mass Index is a person's weight in kilograms divided by the square of the height in meters that is divided into 2 categories are 0 for BMI under 25, and 1 for BMI greater than or equal to 25, in which there are 18.58% and 6.19% respondents who suffered from diabetes from categories 0 and 1, respectively. The person's waist size is divided into 2 categories: 0 for waist size under 85 cm, and 1 for waist size greater than or equal to 85 cm, in which there are 19.47% and 5.31% respondents are suffered from diabetes from categories 0 and 1, respectively. Age is the length of the respondent's life calculated from birth to the last birthday that is divided into 2 categories: 0 for age under 21 years old, and 1 for age greater than or equal to 21 years old.

Gender is divided into 2 categories: 0 for woman, and 1 for man. Age and Gender has the same percentage for the respondent who suffer from diabetes, in which there are 7.08% and 17.70% from categories 0 and 1, respectively. Father's job and mother's job are divided into 2 categories: 0 for formal job, and 1 for informal job. Both also has the same percentage for the respondent who suffer from diabetes, in which there are 15.04% and 9.73% from categories 0 and 1, respectively. The last is finance, there are 3 categories of finance, 0 for parents' salary under 3 million rupiahs, 1 for parents' salary in the range of 3-6 million rupiahs and 2 for parents' salary above 6 million rupiahs, in which there are 14.16%, 7.07%, and 3.54% respondents who suffered from diabetes from categories 0, 1, and 2, respectively. The minimum, maximum, mean, variance, and standard deviation values of predictor variables are presented in the table of descriptive statistical analysis for predictor variables.

Table 1. Descriptive Statistical Analysis for Predictor Variables

Predictor Variables	Min	Max	Mean (μ)	Variance (σ^2)	Standard Deviation (σ)
Height	0	1	0.4159	0.2451011	0.4950769
Weight	0	1	0.5487	0.2498420	0.4998419
BMI	0	1	0.1239	0.1095133	0.3309279
Waist	0	1	0.1681	0.1411188	0.3756579
Age	0	1	0.6549	0.2280341	0.4775292
Gender	0	1	0.6195	0.2378319	0.4876801
Father's Job	0	1	0.4602	0.2506321	0.5006317
Mother's Job	0	1	0.2478	0.1880531	0.4336509
Finance	0	2	0.7168	0.5440898	0.7376244

Based on **Table 1**, all the minimum value is 0 and the maximum value is 1 except for finance variable, because it is divided into 3 categories. Height has a mean value of 0.4159 and variance of 0.2451. Weight has a mean value of 0.5487 and variance of 0.2498. Body Mass Index has a mean value of 0.1239 and variance of 0.1095. Waist has a mean value of 0.1681 and variance of 0.1411. Age has a mean value of 0.6549 and variance of 0.2280. Gender has a mean value of 0.6195 and variance of 0.2378. Father's Job has a mean value of 0.4602 and variance of 0.2506. Mother's Job has a mean value of 0.2478 and variance of 0.1880. Lastly, Finance has a mean value of 0.7168 and variance of 0.5440.

3.2 Factors Affecting of Diabetes Disease Using Logistic Regression

There are two steps to analyze the relationship between the response variable and the predictor variables. First step is simultaneous test using G test or likelihood ratio test. Simultaneous parameter testing produces the value of test statistic G is 16.69, which is more than critical value from Chi-square with alpha 0.05 and degree of freedom 9, $\mathcal{X}^2{}_{(9,0.05)} = 3.32$. In addition, the likelihood ratio test obtained a p-value of 0.006036 less than α (0.05). The decision is to reject H_0, means that there is at least one predictor variable that influence the response variable. The Second step is partial test using Wald test.

Table 2. Partial Test using Logistic Regression using R

Predictor Variables	Coef.	SE	z_{value}	$p-value$	
Height	0.40317	0.67393	0.598	0.54968	
Weight	0.34662	0.72062	0.481	0.63052	
BMI	1.06226	0.76512	1.388	0.16503	
Waist	-0.17123	0.75113	-0.228	0.81968	
Age	0.33628	0.54582	0.616	0.53783	
Gender	-0.09838	0.75882	-0.130	0.89684	
Father's job	-0.24320	0.54280	-0.448	0.65412	
Mother's job	1.82118	0.63826	2.853	0.00433	**
Finance	-0.84703	0.41987	-2.017	0.04366	*
Intercept	-1.62751	0.65011	-2.503	0.01230	*

*) significant at $\alpha = 5\%$
**) significant at $\alpha = 1\%$

Based on the result of partial test logistic parameter, the value of test statistic W for mother's job is 2.853 and the value of test statistic W for finance is -2.017, which are more than critical value from $Z_{(1-\alpha)/2}$ with α (0.05). In addition, the p-value for mother's job is 0.0043 and p-value for finance is 0.04366, which

are less than α (0.05). The decision is to reject H_0, means that the predictor variables mother's job and finance significant to the response variable diabetes disease. In logistic regression analysis, it was found that there was no significant relationship between anthropometric factors and diabetes. This is because almost all athletes have good body condition in terms of height, weight, BMI and waist. So, it is possible that diabetes can be caused by factors other than anthropometry.

The logistic regression model formed is as follows:

$$\pi(x) = \frac{\exp\left(-1.62751 + 0.40317x_1 + 0.34662x_2 + 1.06226x_3 - 0.17123x_4 + 0.33628x_5 - 0.09838x_6 - 0.24320x_7 + 1.82118x_8 - 0.84703x_9\right)}{1 + \left(\exp\left(-1.62751 + 0.40317x_1 + 0.34662x_2 + 1.06226x_3 - 0.17123x_4 + 0.33628x_5 - 0.09838x_6 - 0.24320x_7 + 1.82118x_8 - 0.84703x_9\right)\right)}$$

From the logistic model that was formed, the prediction results were obtained, namely that of the 85 respondents who were diabetes free, there were 82 respondents who were categorized correctly and 3 people who were categorized incorrectly. There were 28 respondents who were indicated to be diabetic, with 9 people classified correctly and 19 people categorized incorrectly. The prediction results from the model above are presented in the contingency **Table 3** below:

Table 3. Prediction Results with Logistic Regression

Observed	Prediction	
	Diabetes Free	Diabetes
Diabetes Free	82	3
Diabetes	19	9

The prediction accuracy level with the logistic regression model was 80.53%.

3.3 Factors Affecting of Diabetes Disease Using Probit Regression

As previously explained, analysis using probit regression has the same steps as logistic regression analysis. The only thing that differentiates these two methods is the link function. In the simultaneous test, the G test or likelihood ratio test is used. The statistical test value in simultaneous testing produces a G value of 16.18, this value is greater than the Chi Square critical point value with α 0.05 and degree of freedom 9, $\mathcal{X}^2{}_{(9,0.05)} = 3.32$. Based on the likelihood ratio test, it produces a p-value of 0.01335, which is less than α (0.05). Therefore, the decision is to reject H_0 means that there is at least a predictor variable that influenced the response variable. Simultaneous tests give results that there are variables that have an influence in the model. For variables that have a significant effect, a partial test is carried out using the Wald test.

Table 4. Partial Test using Probit Regression using R

Predictor Variables	Coef.	SE	z_{value}	$p-value$	
Height	0.27481	0.39219	0.701	0.48348	
Weight	0.20267	0.41254	0.491	0.62324	
BMI	0.61471	0.45278	1.358	0.17457	
Waist	-0.08545	0.42673	-0.200	0.84129	
Age	0.16518	0.30694	0.538	0.59047	
Gender	-0.07885	0.43362	-0.182	0.85571	
Father's job	-0.10886	0.30733	-0.354	0.72318	
Mother's job	1.04701	0.36210	2.892	0.00383	**
Finance	-0.47030	0.23379	-2.012	0.04426	*
(*intercept*)	-0.96681	0.35985	-2.687	0.00722	**

*) significant at $\alpha = 5\%$
**) significant at $\alpha = 1\%$

The result of the analysis probit regression using the Wald test gave results that were not much different from the logistic regression analysis, in the anthropometric factors there was not a single significant variable. In the socio-demographic factors, there are maternal employment and income variables which show a significant influence. These two variables show opportunity values that are smaller than α (0,05). The probit regression model formed is as follows:

$$\pi_I = \Phi(Z_i) = -0.96681 + 0.27481x_1 + 0.20267x_2 + 0.61471x_3 - 0.08545x_4 + 0.16518x_5 - 0.07885x_6 - 0.10886x_7 + 0.47030x_8 - 0.47030x_9$$

Based on the probit model obtained, categories of model prediction results and actual data were grouped. This is to show the accuracy of the predictions of the logit regression model. Based on the category results, it was found that of the 85 respondents who were free of diabetes, there were 82 respondents in the correct category and 3 people in the incorrect category. There were 28 respondents who were indicated to be suffering from diabetes, with details of 8 people in the correct category and 20 people in the wrong category. The prediction results from the model above are presented in contingency **Table 5** below:

Table 5. Prediction Results with Logistic Regression

Observed	Prediction	
	Diabetes Free	Diabetes
Diabetes Free	82	3
Diabetes	20	8

The prediction accuracy level with the probit regression model was 79.65%.

3.4. Best Model Selecting Criteria

To selecting the best model between logistic regression and probit regression using Akaike's Information Criterion (AIC) value, the smaller AIC value means closer to the actual value or being the best model. The result of AIC value is shown in the table model selection with AIC:

Table 4. Model Selection Using AIC

Regression Model	AIC
Logistic	129.85
Probit	130.35

To get the best model in predicting diabetes in prospective athletes, look at the model with the lowest AIC value. Logistic regression has smaller AIC value than probit regression, means that the logistic model is the best model. This result is in accordance with the accuracy results of the two models, the logistic regression model provides a higher accuracy value. So, to model the anthropometric and socio-demographic factors of prospective athletes on the chance of diabetes, it is better to use a logistic regression model.

4. CONCLUSIONS

Prediction results using logistic regression analysis provide higher accuracy values when compared to prediction accuracy values using the probit regression model. Our analysis unequivocally demonstrated that the logistic regression model outperformed the probit regression model, as evidenced by its lower AIC value (129.85). This indicates that the logistic regression model provides a more accurate representation of the data for predicting diabetes risk among prospective athletic candidates. Furthermore, our investigation unveiled the pivotal role of two specific factors in influencing the likelihood of diabetes within this population: mother's occupation and financial status. The implications of these findings are profound. Coaches, trainers, healthcare professionals, and policymakers can now draw from this study's insights to develop targeted programs aimed at mitigating diabetes risk among athletes. Strategies may include specialized dietary plans, lifestyle modifications, and enhanced access to healthcare resources, especially for athletes from economically disadvantaged backgrounds.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to the Directorate of Research and Development of Science and Technology and Higher Education, Ministry of Education, Culture, Research, and Technology (DRTPM Kemendikbudristek) dan Universitas Negeri Surabaya for their generous support and funding of our research project under contract number B/51196/UN38.III.1/LK.04.00/2023. This support has been instrumental in the successful completion of our research, allowing us to make valuable contributions to the field. We also extend our appreciation to all those who contributed to this project and helped us achieve our research goals.

REFERENCES

[1] N. Chaudhary and N. Tyagi, "Diabetes mellitus: An Overview," *Int. J. Res. Dev. Pharm. Life Sci.*, vol. 7, no. 4, pp. 3030–3033, 2018.
[2] D. A. Naqvi, D. A. K. Shadani, and D. S. Gupta, "Study of Serum Magnesium Level in Type 2 Diabetes Mellitus in Relation to Its Microvascular Complication," *Int. J. Sci. Healthc. Res.*, vol. 7, no. 2, pp. 108–117, 2022.
[3] L. Mbuagbaw, R. Aronson, A. Walker, R. E. Brown, and N. Orzech, "The LMC Skills, Confidence & Preparedness Index (SCPI): development and evaluation of a novel tool for assessing self-management in patients with diabetes," *Health Qual. Life Outcomes*, vol. 15, no. 1, pp. 1–9, 2017.
[4] C. C. Jimenez, "Diabetes and exercise: the role of the athletic trainer," *J. Athl. Train.*, vol. 32, no. 4, p. 339, 1997.
[5] D. I. Ruspriyanty and A. Sofro, "Analysis of Hypertension Disease using Logistic and Probit Regression," *J. Phys. Conf. Ser.*, vol. 1108, no. 1, 2018.
[6] D. K. Mulyantoro, "Tinggi badan usia dewasa dan risiko penyakit diabetes melitus," Ph.D. dissertation, Universitas Indonesia, Jakarta, Indonesia, 2013.
[7] L. Amalia, Y. Mokodompis, and G. A. Ismail, "Hubungan Overweight Dengan Kejadian Diabetes Mellitus Tipe 2 Di Wilayah Kerja Puskesmas Bulango Utara," *Jambura J. Epidemiol.*, vol. 1, no. 1, pp. 11–19, 2022.
[8] I. Saputra, F. Esfandiari, E. Marhayuni, and M. Nur, "Indeks massa tubuh dengan kadar Hb-A1c pada pasien diabetes melitus tipe II," *J. Ilm. Kesehat. Sandi Husada*, vol. 9, no. 2, pp. 597–603, 2020.
[9] K. A. Putri, D. N. Kahanjak, and R. A. A. H. S. Putra, "Literature Review: Hubungan Lingkar Pinggang Dengan Kadar Gula Darah Pada Dewasa Muda," *J. Kedokt. Univ. Palangka Raya*, vol. 10, no. 1, pp. 18–23, 2022.
[10] K. Komariah and S. Rahayu, "Hubungan usia, jenis kelamin dan indeks massa tubuh dengan kadar gula darah puasa pada pasien diabetes melitus tipe 2 di klinik pratama rawat jalan proklamasi, Depok, Jawa Barat," *J. Kesehat. Kusuma Husada*, pp. 41–50, 2020.
[11] R. Arania, T. Triwahyuni, T. Prasetya, and S. D. Cahyani, "Hubungan Antara Pekerjaan Dan Aktivitas Fisik Dengan Kejadian Diabetes Mellitus Di Klinik Mardi Waluyo Kabupaten Lampung Tengah," *J. Med. Malahayati*, vol. 5, no. 3, pp. 163–169, 2021.
[12] W. P. Aji, "Hubungan status sosial ekonomi dengan kejadian diabetes melitus tipe 2 diwilayah kerja puskesmas patrang." *Ph.D* Ph.D. dissertation, Universitas dr. SOEBANDI, Jember, Indonesia, 2022.
[13] M. Fathurahman, M. N. Hayati, and N. A. Rizki, "Pemodelan Indeks Pembangunan Kesehatan Masyarakat Kabupaten/Kota di Pulau Kalimantan dengan Regresi Spasial," *Pros. SNMSA*, pp. 183–191, 2019.
[14] A. Agresti, *Categorical data analysis*, vol. 792. John Wiley & Sons, 2012.
[15] Z. A. A. Al-Bairmani and A. A. Ismael, "Using Logistic Regression Model to Study the Most Important Factors Which Affects Diabetes for the Elderly in the City of Hilla / 2019," *J. Phys. Conf. Ser.*, vol. 1818, no. 1, pp. 0–10, 2021.
[16] D. R. S. Saputro and P. Widyaningsih, "Algoritme Pendugaan Parameter Model Regresi Logistik Biner (RLB) dengan Maksimum Likelihood dan Broyden-Fletcher-Goldfarb-Shanno (BFGS)," in *Seminar Nasional Matematika Dan Pendidikan Matematika UNY*, pp. 97–104, 2016.
[17] S. Yuhadisi and S. Suliadi, "Penerapan Metode Modifikasi Hosmer-Lemeshow Test pada Model Regresi Logistik Data Penderita Penyakit Hipertensi," *Pros. Stat.*, vol. 7, no. 1, pp. 50–55, 2021.
[18] Y. Tampil, H. Komaliq, and Y. Langi, "Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado," *d'CARTESIAN J. Mat. dan Apl.*, vol. 6, no. 2, pp. 56–62, 2017.
[19] W. Alwi, E. Ermawati, and S. Husain, "Analisis Regresi Logistik Biner Untuk Memprediksi Kepuasan Pengunjung Pada Rumah Sakit Umum Daerah Majene," *J. Mat. dan Stat. serta Apl.*, vol. 6, no. 1, p. 20, 2018.
[20] E. Wulandari, "Model regresi probit untuk mengetahui faktor-faktor yang mempengaruhi jumlah penderita diare di Jawa Timur," *MATHunesa J. Ilm. Mat.*, vol. 1, no. 1, 2013.
[21] P. McCullagh, *Generalized linear models*. Routledge, 2019.
[22] A. E. Review and L. Document, "Analyzing the Determinants of Project Success : A Probit Regression Approach 2 Analyzing the Determinants of Project Success : A Probit Regression Approach," vol. 1990, pp. 1–5, 2016, [Online]. Tersedia: https://www.adb.org/sites/default/files/linked-documents/4-Analyzing-the-Determinants-of-Project-Success-A-Probit-Regression-Approach.pdf [Diakses: 2 Agustus 2023].
[23] P. L. Ayunda, "Analisis Perbandingan Regresi Logistik Model Logit dan Probit untuk Menentukan Variabel yang Mempengaruhi Fluktuasi Harga Beras pada Daerah Surplus dan Defisit," Institut Teknologi Sepuluh November, 2018.
[24] M. Zaidi and A. Amirat, "Forecasting Stock Market Trends By Logistic Regression and Neural Networks Evidence From Ksa Stock Market," *Int. J. Econ. Commer. Manag. United Kingdom*, vol. IV, no. 6, pp. 220–234, 2016.

[25] A. Suwardi, "Modul LPM, Logit, dan Probit Model." Fakultas Ekonomi. Universitas Indonesia. Depok, 2011.
[26] M. Astsaqofi, "Analisis Regresi Probit dengan Efek Interaksi untuk Memodelkan Indeks Pembangunan Manusia di Indonesia," *Inst. Teknol. Sepuluh Nop.*, 2016.
[27] N. A. Salsabila, S. Andriani, M. Mirisda, and D. A. Nohe, "Analisis Pengaruh Tingkat Partisipasi Angkatan Kerja Dan Indeks Pembangunan Manusia Terhadap Tingkat Pengangguran Terbuka Menggunakan Regresi Probit Dan Logit," in *Prosiding Seminar Nasional Matematika dan Statistika*, 2022, vol. 2.
[28] D. L. W. Putri, S. Mariani, and S. Sunarmi, "Peningkatan Ketepatan Klasifikasi Model Regresi Logistik Biner dengan Metode Bagging (Bootstrap Aggregating)," *Indones. J. Math. Nat. Sci.*, vol. 44, no. 2, pp. 61–72, 2021.
[29] S. B. Sari, N. Mukhtar, and A. Anisa, "Analisis faktor-faktor yang mempengaruhi keseringan mahasiswa unhas mengikuti program gumsb (gerakan unhas mengaji dan sholat berjamaah) dengan model regresi logistik," *J. Mat. Stat. dan Komputasi*, vol. 15, no. 1, pp. 104–113, 2018.
[30] D. W. Hosmer, S. Lemeshow, and E. Cook, "Applied logistic regression 2nd edition," *New York Jhon Wiley Sons Inc*, 2000.
[31] S. Konishi and G. Kitagawa, *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.
[32] H. de-G. Acquah, "Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship," 2010.