EKSAKTA
Journal of Sciences and Data Analysis

E-ISSN: 2720-9326
P-ISSN: 2716-0459

DOI : 10.20885/EKSAKTA.vol6.iss2.art3

Analysis of Gallstone Incidence Factors Using a Binary Logistic Regression Model.

Ihsan Fathoni Amri[1*], Febrian Hikmah Nur Rohim[2], M. Ilham Nurul Azka[3], Muji Silvi Rakhmawati[4]

[1*] Program Studi Sains Data, FSTP, Universitas Muhammadiyah Semarang, Jl. Kedungmundu No.18, Semarang, 50273, Jawa Tengah, Indonesia.
[2,3,4] Program Studi Sains Data, FSTP, Universitas Muhammadiyah Semarang, Jl. Kedungmundu No.18, Semarang, 50273, Jawa Tengah, Indonesia.

[*] *Corresponding author: ihsanfathoni@unimus.ac.id*

Abstract: Gallstone disease (cholelithiasis) is a digestive disorder with a globally increasing prevalence. This study aims to identify the risk factors contributing to the occurrence of gallstones using a binary logistic regression model. The dataset was obtained from the UC Irvine Machine Learning Repository and includes clinical records of 319 outpatient individuals treated at Ankara VM Medical Park Hospital, Turkey. The analysis was performed on 23 independent variables, covering demographic characteristics, body composition, medical history, and laboratory test results. The results showed that the binary logistic regression model was simultaneously significant based on the Likelihood Ratio Test (G^2) with a p-value of $0.000 < 0.05$. Furthermore, the Partial (Wald) Test indicated that six variables were individually significant (p-value = $0.000 < 0.05$), age, comorbidity, diabetes mellitus, visceral fat rating, visceral fat area, and vitamin D level. Diabetes mellitus emerged as the most dominant risk factor, with an odds ratio (OR) of 11.5, meaning that individuals with diabetes have approximately 11.5 times higher risk of developing gallstones compared to non-diabetic individuals. In contrast, higher levels of vitamin D showed a protective effect against gallstone formation. The logistic regression model achieved a classification accuracy of 77%, indicating good performance in predicting gallstone risk. The novelty of this study lies in the integration of both metabolic and nutritional variables within a single predictive model for gallstone risk assessment. By combining factors such as visceral fat indicators and vitamin D level, this research provides a more comprehensive understanding of the multifactorial nature of gallstone formation. The findings can support early detection and clinical decision-making for preventive interventions among high-risk individuals, potentially enhancing the effectiveness of prevention strategies and reducing the incidence of gallstone disease in the community.

Keywords: Gallstone, Binary logistic regression, Risk Factors, Clinical Prediction

Introduction

Gallstone disease (cholelithiasis) is one of the digestive system disorders with a globally increasing prevalence [1]. This condition occurs due to the formation of solid deposits from bile components, such as cholesterol or bilirubin, within the gallbladder [2]. Gallstone disease is a leading cause of emergency department visits and elective surgeries in many countries, particularly in developing nations [3]. Approximately 10–15% of the adult population in the United States has gallstones, with over 700,000 cholecystectomy procedures performed annually [4]. Meanwhile, population-based studies in Saudi Arabia, Pakistan, and India have also shown a similar increasing trend [5], primarily due to lifestyle changes and high-fat dietary patterns [6]. The clinical symptoms range from severe abdominal pain to serious complications such as cholecystitis, pancreatitis, and biliary obstruction [7], making this disease a significant public health burden. The consistent rise in prevalence, especially among adults and the elderly, indicates that gallstones are not only an individual clinical problem but also a global challenge within healthcare systems.

Early detection and identification of risk factors are key steps in preventing serious complications caused by gallstones [8]. Numerous studies have shown that individual characteristics such as older age, female gender,

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI :

a history of obesity, and a high-fat, low-fiber diet significantly contribute to the formation of gallstones [9]. In Mexico, it has been reported that the risk of gallstones increases significantly in individuals with a BMI ≥ 30 and elevated cholesterol levels [10]. Understanding these factors enables more accurate and evidence-based medical decision-making [11], particularly in determining population groups that require closer attention. Therefore, early identification of high-risk individuals is essential to support preventive strategies, continuous clinical monitoring, and the planning of more effective and efficient interventions.

Advancements in information technology and the availability of large-scale medical data have driven the adoption of statistical methods and machine learning algorithms in the healthcare domain [12]. Statistical data analysis not only aids in understanding the relationships between variables but also creates opportunities to develop predictive models capable of identifying disease risks more accurately and objectively [13]. One widely used method is logistic regression, which is effective in analyzing the relationship between predictor variables and binary outcomes [14], such as gallstone status. This approach enables the construction of models that are both interpretable and statistically robust, making them reliable tools for data-driven clinical decision-making [15]. In the context of gallstone disease, the application of logistic regression allows for a clear explanation of each risk factor's contribution to the probability of occurrence, thereby supporting more targeted and responsive patient management.

This study focuses on the analysis of odds ratios to identify which factors most increase the risk of gallstone formation. In addition, the research applies a logistic regression model to predict the risk of gallstone occurrence based on patients' demographic and clinical characteristics. Logistic regression was chosen for its ability to estimate the likelihood of a binary event, in this case, the presence or absence of gallstones [16]. This model enables the identification of significantly influential factors while providing probability values that can be utilized in clinical decision-making [17].

The main objective of this study is to develop a predictive model capable of identifying individuals at high risk of developing gallstone disease [18], thereby supporting faster and more accurate medical decision-making [11]. Through the use of logistic regression, this research aims to reveal the clinical and demographic factors that significantly influence the occurrence of gallstones. The findings of this study are expected to provide practical value in planning preventive interventions and early detection strategies within healthcare facilities. Considering the significant increase in gallstone cases reported in the United States and several other countries, anticipatory measures are necessary should a similar trend emerge in Indonesia. Therefore, early detection and data-driven prevention efforts are crucial to mitigate the potential rise in the incidence of this disease within the community.

Materials and Methods

2.1 Research Variables

This study utilized a dataset from the UC Irvine Machine Learning Repository, containing clinical data from 319 patients at the Internal Medicine Outpatient Clinic of Ankara VM Medical Park Hospital, Ankara, Turkey, collected between June 2022 and June 2023. A total of 161 patients were diagnosed with gallstones. The dataset comprises 38 variables, including demographic data (age, gender, height, weight, BMI), bioimpedance measurements (body water, muscle and fat mass, protein, visceral fat, liver fat), as well as laboratory data (glucose, lipid profile, liver enzymes, creatinine, GFR, CRP, hemoglobin, and vitamin D). The dataset was approved by the Ethics Committee of Ankara City Hospital (approval number: E2-23-4632) and was officially published on April 19, 2025. The variables and data types used are presented in Table 1 below:

Table 1. Research Variables.

	Variabel	Tipe Data		Variabel	Tipe Data
Y	Gallstone Status	Kategorik (0=No, 1= Yes)	X_{12}	Body Protein Content (Protein) (%)	Numerik (float)
X_1	Age	Numerik (int)	X_{13}	Muscle Mass (MM)	Numerik (float)

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI :

	Variabel	Tipe Data		Variabel	Tipe Data
X_2	Gender	Kategorik (0=Male, 1=Female)	X_{14}	Total Body Water (TBW)	Numerik (float)
X_3	Comorbidity	Numerik (int)	X_{15}	Glucose	Numerik (float)
X_4	Diabetes Mellitus (DM)	Kategorik (0=No, 1=Yes)	X_{16}	Triglyceride	Numerik (float)
X_5	Coronary Artery Disease (CAD)	Kategorik (0=No, 1=Yes)	X_{17}	High Density Lipoprotein (HDL)	Numerik (float)
X_6	Hyperlipidemia	Kategorik (0=No, 1=Yes)	X_{18}	Low Density Lipoprotein (LDL)	Numerik (float)
X_7	Body Mass Index (BMI)	Numerik (float)	X_{19}	Total Cholesterol (TC)	Numerik (float)
X_8	Visceral Fat Rating (VFR)	Numerik (int)	X_{20}	Aspartat Aminotransferase (AST)	Numerik (float)
X_9	Visceral Fat Area (VFA)	Numerik (float)	X_{21}	Alanin Aminotransferase (ALT)	Numerik (float)
X_{10}	Total Body Fat Ratio (TBFR) (%)	Numerik (float)	X_{22}	Alkaline Phosphatase (ALP)	Numerik (float)
X_{11}	Lean Mass (LM) (%)	Numerik (float)	X_{23}	Vitamin D	Numerik (float)

2.2 Binary Logistic Regression

Binary logistic regression is a statistical technique used to explore the relationship between a dependent variable with two possible outcomes (e.g., yes or no, success or failure) and one or more independent variables. This method is applied to predict the likelihood of an event occurring based on influencing factors. The outcome of binary logistic regression typically yields a probability value ranging from 0 to 1 [19]. In this study, the dependent variable is dichotomous (with two possible values: positive (1) or negative (0)). A binary logistic regression model with a response variable taking values of 0 and 1 follows a Bernoulli distribution. Let y_k be the value of the response variable for the k-th observation, then the probability function of y is [20]:

$$f(y_k) = \pi(x_k)^{y_k}(1 - \pi(x_k))^{1-y_k} \quad (1)$$

Where the value of y_t= 0; 1, $\pi(x_t)$ represents the probability of success, $1 - \pi(x_t)$ represents the probability of failure, and t = 1,2,3 …, n denotes the observation index, with n being the total number of observations. The binary logistic regression model used is as follows Equation (2) [20]:

$$\pi(x) = \frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k}} \quad (2)$$

Where k is the number of independent variables. The logistic regression model in Equation (2) is transformed using the logit function, resulting in the following form Equation (3) [20]:

$$g(x) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k \quad (3)$$

The β coefficient in logistic regression represents the magnitude of the effect that the independent variable has on the odds of an event occurring in the dependent variable[21].

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI :

2.3 G Likelihood Ratio Test

In this study, the simultaneous test was conducted using the G statistic or Likelihood Ratio Test [22]. The G Likelihood Ratio Test is used to assess the extent to which a more complex model (full model) provides a significantly better fit compared to a simpler model (reduced model). This test is calculated based on the difference between the log-likelihood values of the two models. The simultaneous test is performed to determine the overall significance of the parameters with respect to the response variable. The G test statistic follows a Chi-Square distribution. Mathematically, the G test can be expressed by the following Equation (4) [23]:

$$G = 2\ (ln\ L_1 - \ ln\ L_0) \tag{4}$$

Explanation:
L_0: the maximum likelihood value of the function without predictor variables (reduced model)
L_1: the maximum likelihood value of the function with all predictor variables (full model)
Hypotheses used:
H_0: $\beta 1 = \beta 2 = \cdots = \beta k = 0$, all predictor variables have no effect on the dependent variable
H_1: $\beta k \neq 0$; $k = 1, 2, \ldots, k$, at least one predictor variable has an effect on the dependent variable

The G statistic follows a Chi-Square distribution with p degrees of freedom, where p is the number of predictor variables (excluding the intercept) in the model. The null hypothesis H_0 is rejected if $G > \chi^2_{kritis}$, or equivalently, if the $p - value < \alpha$. A larger G value indicates that the more complex model provides a significantly better fit to the data. Conversely, a smaller G value suggests that the simpler model cannot be rejected, meaning it adequately describes the data.

2.4 Wald Test

The partial test using the Wald statistic is employed to examine the individual effect of each parameter coefficient β_k on the resulting model [24]. The Wald test statistic is calculated by dividing the estimated coefficient (β_k) by its standard error ($SE(\hat{\beta}_k)$), which yields a test statistic that follows a normal distribution [25]. A large Wald statistic value indicates that the corresponding independent variable contributes significantly to the model.

The results of this partial or individual test can be used to assess whether a predictor variable should be included in the model. Mathematically, the Wald statistic can be expressed by the following Equation (5)-(6):

$$W = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \tag{5}$$

$$SE(\hat{\beta}_k) = \sqrt{(\sigma^2(\hat{\beta}_k))} \tag{6}$$

Explanation:
$\hat{\beta}_k$: estimated value of the coefficient for the k-th predictor variable
$SE(\hat{\beta}_k)$: standard error of the estimate for the k-th predictor variable
Hypotheses used:
H_0: $\beta_k = 0$, $k = 1,2, \ldots, k$ the k-th predictor variable has no significant effect on the response variable
H_1: $\beta_k \neq 0$; $k = 1,2, \ldots, k$, the k-th predictor variable has a significant effect on the response variable

H_0 will be rejected when the Wald test statistic $W > Z_{kritis}$, or when the p-value < α, indicating that the k-th predictor variable has a significant effect on the response variable.

2.5 Odds Ratio

The odds ratio (OR) represents the ratio between the probability of an event occurring and the probability of the event not occurring[26]. In logistic regression, the OR is used to interpret the model results by measuring how much the odds of the event change with a one-unit increase in the independent variable. Mathematically, the odds ratio can be expressed by the following Equation (7)[27]:

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI :

$$\hat{\theta}_i = e^{\hat{\beta}_i} \tag{7}$$

OR value greater than 1 indicates that the examined factor increases the likelihood of the event occurring, while an OR value less than 1 suggests a decreasing effect on the event's probability [24]. In this study, odds analysis is used to identify significant risk factors and evaluate the extent of their influence on the likelihood of developing gallstones, which aids in determining priorities for public health interventions.

2.6 Classification Accuracy

Classification accuracy is a measure of how accurately the model predicts data correctly, calculated as the percentage of correctly classified observations out of the total number of observations[29]. Total accuracy represents the overall percentage of correct classifications the higher the total accuracy, the better the model's performance. APER (Apparent Error Rate) is a value used to assess the probability of misclassification. Mathematically, it can be expressed by the following (8) [30]:

$$APER = \frac{n_{1m} + n_{2m}}{n_1 + n_2} \times 100\% \tag{8}$$

Explanation:
n_{1c}: Number of observations with y = 0 correctly classified as y = 0
n_{1m}: Number of observations with y = 0 incorrectly classified as y = 1
n_{2c}: Number of observations with y = 1 correctly classified as y = 1
n_{2m}: Number of observations with y = 1 incorrectly classified as y = 0

2.7 Research Steps

The analysis steps used in this study are as follows:

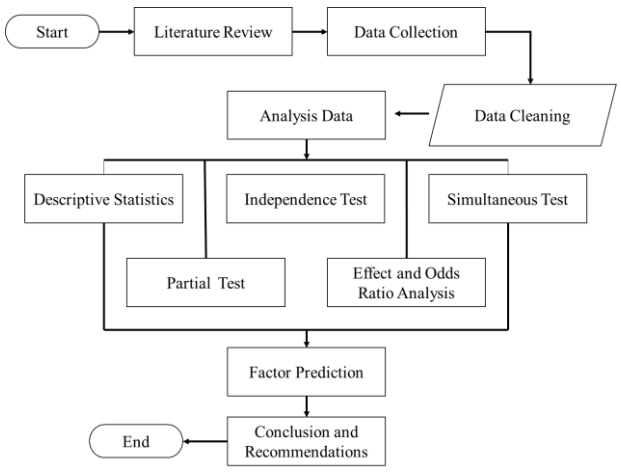

Figure 1. Research Steps Flow Chart.

According to the flowchart in Figure 1, the detailed explanation of the research steps is as follows:

1. Conducting a literature review to gather relevant references and sources related to the research topic.
2. Collecting data related to the research variables for further analysis.
3. Performing data cleaning to ensure the dataset is free from duplicates, errors, or invalid values.
4. Carrying out descriptive statistical analysis on gallstone occurrence data.
5. Conducting an independence test between variables using the Pearson Chi-Square test.
6. Applying the G test for simultaneous testing to perform multivariate analysis on the available data.

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI :

7. Using the Wald test to test hypotheses partially (individually).
8. Analyzing the influence of research variables by calculating the odds ratio.
9. Building a predictive model using logistic regression to classify suspected cases of gallstone occurrence.
10. Drawing conclusions based on the analytical results obtained throughout the study.

Result and Discussion

3.1 Descriptive Analysis

The first step in this study is descriptive analysis to understand the characteristics of the data. The analysis includes calculating the minimum, maximum, mean, and standard deviation (std) for each variable. The resulting descriptive statistics provide insights into the distribution and variability of the data, as shown in Figure 2 and Table 2 below:

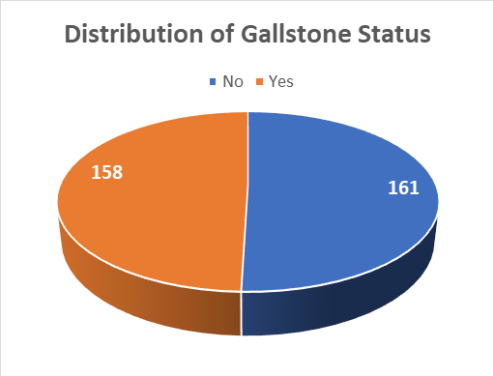

Figure 2. Distribution of Gallstone Status.

Tabel 2. Descriptive Statistics.

Variable	Mean	Std	Min	Max	Variable	Mean	Std	Min	Max
Gallstone Status	0.50	0.50	0.00	1.00	Body Protein Content (Protein) (%)	15.94	2.33	5.56	24.81
Age	48.07	12.11	20.00	96.00	Muscle Mass (MM)	54.27	10.60	4.70	78.80
Gender	0.49	0.50	0.00	1.00	Total Body Water (TBW)	40.59	7.93	13.00	66.20
Comorbidity	0.34	0.52	0.00	3.00	Glucose	108.69	44.85	69.00	575.00
Diabetes Mellitus (DM)	0.13	0.34	0.00	1.00	Triglyceride	144.50	97.90	1.39	838.00
Coronary Artery Disease (CAD)	0.04	0.19	0.00	1.00	High Density Lipoprotein (HDL)	49.48	17.72	25.00	273.00
Hyperlipidemia	0.03	0.16	0.00	1.00	Low Density Lipoprotein (LDL)	126.65	38.54	11.00	293.00
Body Mass Index (BMI)	28.88	5.31	17.40	49.70	Total Cholesterol (TC)	203.50	45.76	60.00	360.00
Visceral Fat Rating (VFR)	9.08	4.33	1.00	31.00	Aspartat Aminotransferaz (AST)	21.68	16.70	8.00	195.00
Visceral Fat Area (VFA)	12.17	5.26	0.90	41.00	Alanin Aminotransferaz (ALT)	26.86	27.88	3.00	372.00

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI :

Total Body Fat Ratio (TBFR) (%)	28.27	8.44	6.30	50.92	Alkaline Phosphatase (ALP)	73.11	24.18	7.00	197.00
Lean Mass (LM) (%)	71.64	8.44	48.99	93.67	Vitamin D	21.40	9.98	3.50	53.10

Based on Figure 2 and Table 2, the average values for gallstone status and gender are 0.5 and 0.49 respectively, indicating a balanced distribution of respondents. The average age of the respondents is 48.07 years, with an age range of 20–96 years, reflecting a diverse age group. The mean BMI of 28.88 suggests that the majority of respondents fall into the overweight category. The average visceral fat measurements (VFR 9.08; VFA 12.17 cm²), total body fat mass (28.27%), and muscle mass (54.27 kg) indicate a notable body composition, supported by an average total body water (TBW) of 40.97 liters and body protein of 9.06%. The mean blood glucose level of 108.69 mg/dL is close to the prediabetic threshold, while triglycerides (144.5 mg/dL), LDL (126.63 mg/dL), and total cholesterol (205.6 mg/dL) reflect a risk of dyslipidemia, although HDL levels (49.48 mg/dL) remain within the normal range. Liver enzyme values (AST, ALT, ALP) fall within normal limits, whereas the average vitamin D level is only 21.4 ng/mL, indicating a deficiency. Overall, most respondents exhibit metabolic and nutritional risk factors that warrant further analysis.

3.2 Chi-Square Independent Test

Following the descriptive analysis, the next step is to perform the Chi-Square independence test to evaluate whether there is a statistically significant relationship between the independent variables. The results of this test indicate the extent to which the predictor variables are associated with the response variable, as presented in Table 3 below:

Table 3. Independent Test.

	Chi2	p-value	Results		Chi2	p-value	Results
X_1	48.480	0.720	Not Significant	X_{13}	224.992	0.358	Not Significant
X_2	6.913	0.009*	Significant	X_{14}	202.990	0.389	Not Significant
X_3	2.987	0.394	Not Significant	X_{15}	76.393	0.498	Not Significant
X_4	2.910	0.088	Not Significant	X_{16}	181.588	0.391	Not Significant
X_5	2.068	0.150	Not Significant	X_{17}	72.956	0.161	Not Significant
X_6	6.419	0.011*	Significant	X_{18}	134.469	0.472	Not Significant
X_7	167.949	0.338	Not Significant	X_{19}	172.454	0.092	Not Significant
X_8	33.440	0.042*	Significant	X_{20}	65.583	0.015*	Significant
X_9	217.458	0.312	Not Significant	X_{21}	70.492	0.330	Not Significant
X_{10}	224.658	0.531	Not Significant	X_{22}	108.081	0.188	Not Significant
X_{11}	294.998	0.456	Not Significant	X_{23}	249.861	0.271	Not Significant
X_{12}	260.662	0.442	Not Significant				

Based on the results of the independence test presented in Table 3, it is shown that out of the 23 variables analyzed in relation to gallstone occurrence, 4 variables demonstrate a statistically significant association with the condition: Gender (X_2), Hyperlipidemia (X_6), Visceral Fat Rating/VFR (X_8), and Aspartate Aminotransferase/AST (X_{20}), with respective p-values of 0.009, 0.011, 0.042, and 0.015 (p-value < 0.05). These findings indicate that these four variables are not independent of gallstone occurrence and can be considered as factors that significantly contribute to the condition.

In contrast, the remaining 19 variables did not show a significant association (p-value > 0.05) and are therefore statistically considered not meaningfully related to gallstone occurrence. These results provide an initial indication of the importance of considering gender, history of hyperlipidemia, visceral fat level, and AST enzyme levels in risk analysis and early detection of gallstones.

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI :

3.3 Simultaneous Test

After conducting the independence test using the Chi-Square method to determine the relationships between each independent variable, the next stage of analysis is to implement the logistic regression test simultaneously. This test aims to assess whether all the independent variables included in the model collectively contribute significantly to the dependent variable. The results of this analysis are presented in Table 4 below:

Table 4. Simultaneous Test.

G-Statistics	p-value
75.03100933	0.000*

Based on the results of the simultaneous test presented in Table 4, the G statistic is 75.03100933 with a p-value of 0.000. Using a significance level of α = 5% and degrees of freedom (*df*) = 23, the critical value from the Chi-Square distribution table is $\chi^2_{Critical}$ = 35.172. Since G > $\chi^2_{Critical}$ and p-value < α = 0.05, the null hypothesis H_0 is rejected. This indicates that at least one parameter $\beta_i \neq 0$, meaning that one or more predictor variables have a statistically significant effect on the response variable in this study.

3.4 Partial Test (Wald Test)

After conducting the logistic regression model test simultaneously to evaluate the combined effect of the variables, the next step is to perform a partial test using the Wald statistic for each predictor variable. This test aims to identify the individual contribution of each predictor to the model. The results of this partial analysis are presented in Table 5 below:

Table 5. Partial Test.

Variable	Coefficient (β)	Standard Error (SE)	Wald Statistic	p-value	EXP (β)
Age (X_1)	0.895714	0.372	5.783	0.016*	2.449
Comorbidity (X_3)	0.592725	0.271	4.774	0.029*	0.553
Diabetes Mellitus (DM) (X_4)	2.442964	0.784	9.718	0.002*	11.507
Visceral Fat Rating (VFR) (X_8)	1.978433	0.753	6.896	0.009*	0.138
Visceral Fat Area (VFA) (X_9)	1.597411	0.713	4.946	0.026*	4.940
Vitamin D (X_{23})	-0.759343	0.185	16.844	0.000*	0.468

Based on the results of the Wald partial test presented in Table 5, out of the 23 variables analyzed, 6 variables were found to have a statistically significant effect on gallstone occurrence. These variables are Age (X_1; p-value = 0.016), Comorbidity (X_3; p-value = 0.029), Diabetes Mellitus (X_4; p-value = 0.002), Visceral Fat Rating/VFR (X_8; p-value = 0.009), Visceral Fat Area/VFA (X_9; p-value = 0.026), and Vitamin D (X_{23}; p-value = 0.000). The p-values for these variables are all below the 0.05 significance threshold, indicating that they play a statistically significant role in the occurrence of gallstones and can be considered important factors for further analysis. Based on this analysis, the logistic regression model for gallstone occurrence is as follows:

$$\pi(x) = \frac{e^{\begin{array}{c}0.137548153+0.895713738x_1+(-1.146045467)x_2+0.592724651x_3+2.442963536x_4+\\ 0.138878196x_5+20.17742058x_6+0.843338595x_7+1.978433227x_8+1.59741098x_9+(-3.205279835)x_{10}\\ +(-2.586098678)x_{11}+0.039899382x_{12}+(-0.210193402)x_{13}+(-0.430058844)x_{14}+(-0.291544115)x_{15}+(-0.090033557)x_{16}\\ +0.278187439x_{17}+(-0.267897789)x_{18}+0.139314884x_{19}+(-0.92048127)x_{20}+0.73480066x_{21}+0.120772085x_{22}\\ +(-0.75934338)x_{23}\end{array}}}{1+e^{\begin{array}{c}0.137548153+0.895713738x_1+(-1.146045467)x_2+0.592724651x_3+2.442963536x_4+\\ 0.138878196x_5+20.17742058x_6+0.843338595x_7+1.978433227x_8+1.59741098x_9+(-3.205279835)x_{10}\\ +(-2.586098678)x_{11}+0.039899382x_{12}+(-0.210193402)x_{13}+(-0.430058844)x_{14}+(-0.291544115)x_{15}+(-0.090033557)x_{16}\\ +0.278187439x_{17}+(-0.267897789)x_{18}+0.139314884x_{19}+(-0.92048127)x_{20}+0.73480066x_{21}+0.120772085x_{22}\\ +(-0.75934338)x_{23}\end{array}}}$$

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI :

3.5 Interpretation of Odds Ratio Values

The calculation of odds ratio values in logistic regression analysis can be derived from the EXP(β) values found in the output of the partial test. Based on Table 5, the interpretation of the odds ratios for the predictor variables is as follows:

- The Age variable (X_1) has an odds ratio of 2.449, indicating that as individuals age, their likelihood of developing gallstones increases by approximately 2.45 times compared to younger individuals.
- Comorbidity (X_3) has an odds ratio of 0.553, suggesting that the presence of comorbid conditions is associated with a 44.7% increased likelihood of gallstone occurrence.
- Diabetes Mellitus (X_4) has an odds ratio of 11.507, meaning individuals with diabetes have a risk approximately 11.5 times higher of developing gallstones compared to non-diabetics.
- For Visceral Fat Rating (VFR) (X_8), the odds ratio is 0.138, implying an 86.2% increased likelihood of gallstones with higher VFR values.
- Visceral Fat Area (VFA) (X_9) shows an odds ratio of 4.940, indicating that individuals with higher visceral fat area have an almost fivefold increased risk of gallstone occurrence.
- Lastly, Vitamin D (X_{23}) has an odds ratio of 0.468, meaning higher levels of vitamin D are associated with a 53.2% reduction in the risk of gallstones.

Diabetes Mellitus (X_4) stands out as the variable with the highest odds ratio among all predictors, highlighting its major contribution to increased gallstone risk. These findings reinforce the significance of these variables in explaining variations in gallstone risk and underscore their relevance for both preventive strategies and early detection efforts. This is illustrated in Figure 3, the bar chart of odds ratios below:

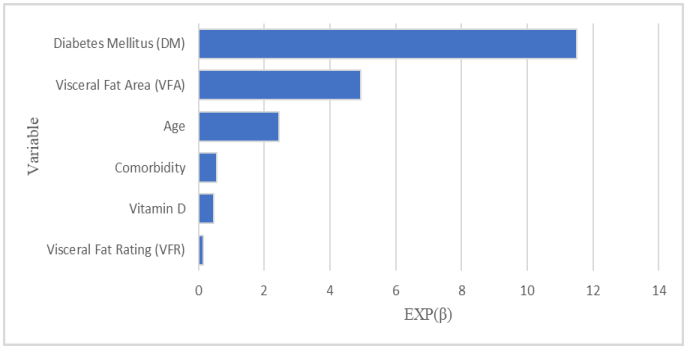

Figure 3. Bar Odds Ratio Chart.

3.6 Prediction, Classification, and Model Evaluation

The first step in this study is descriptive analysis to understand the characteristics of the data. The analysis includes calculating the minimum, maximum, mean, and standard deviation (std) for each variable. The resulting descriptive statistics provide insights into the distribution and variability of the data, as shown in Figure 2 and Table 2 below:

Based on the evaluation of the classification model for gallstone occurrence using the testing dataset, the model achieved an accuracy score of 0.77, indicating that it correctly classified 77% of the total 96 test observations. This suggests a good generalization capability of the model when applied to unseen data. For class 0 (negative / no gallstones), the model achieved a precision of 0.79, recall of 0.76, and an F1-score of 0.78, based on 50 actual cases in the test set. For class 1 (positive / with gallstones), it recorded a precision of 0.75, recall of 0.78, and an F1-score of 0.77, showing the model's fairly balanced ability to correctly identify positive cases from 46 actual observations. Furthermore, the macro average and weighted average values for precision, recall, and F1-score were all 0.77, reflecting consistent and robust performance across both classes, even on the test data. This consistency also indicates that the model does not suffer from overfitting and can handle class distributions

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI :

effectively without a significant bias toward either the majority or minority class. These results are illustrated in Figure 4 and Table 6 below:

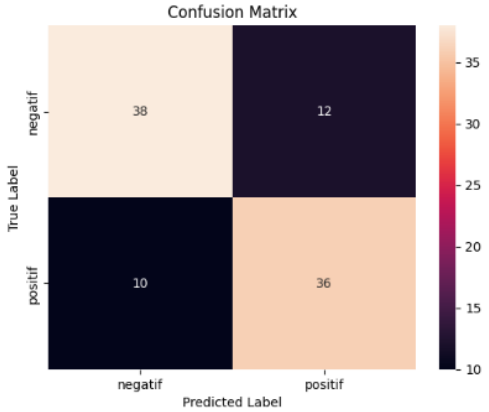

Figure 4. Confusion Matrix Prediction Classification.

Table 6. Classification Accuracy

	Precision	Recall	F1-Score	Support
0	0.79	0.76	0.78	50
1	0.75	0.78	0.77	46
Accuracy			0.77	96
Macro Avg	0.77	0.77	0.77	96
Weighted Avg	0.77	0.77	0.77	96

Through the confusion matrix, it can be observed that out of 50 negative cases, 38 were correctly predicted as negative (true negatives), while 12 were incorrectly classified as positive (false positives). Among the 46 positive cases, 36 were accurately classified as positive (true positives), and the remaining 10 were misclassified as negative (false negatives). These results indicate that the model has a fairly good predictive capability in identifying both classes, with a slightly better tendency in recognizing positive cases, as reflected by the recall score of 0.78 for class 1. This is particularly important in the context of diagnosis or early detection of gallstones, where accurately identifying positive cases can directly impact clinical decision-making and patient outcomes.

Conclusion

The results of this study conclude that logistic regression effectively identified six variables that significantly influence the occurrence of gallstones: age, comorbidity, diabetes mellitus, visceral fat rating (VFR), visceral fat area (VFA), and vitamin D levels. Older age, the presence of comorbidities, diabetes, elevated VFR, and increased VFA were found to significantly raise the risk of gallstone formation, with diabetes mellitus emerging as the most dominant risk factor (OR = 11.5). In contrast, higher levels of vitamin D were associated with a reduced risk of developing gallstones.

The predictive model was statistically significant overall and achieved an accuracy of 77%, highlighting its potential for real-world application. These findings can be utilized in clinical practice for the early identification of high-risk individuals and serve as a foundation for preventive interventions. Lifestyle modifications such as diabetes management, visceral fat reduction, and vitamin D optimization may offer effective strategies to reduce the risk of gallstone disease.

For future research, it is recommended to include environmental, genetic, and dietary pattern variables to provide a more comprehensive understanding of the risk factors associated with gallstone disease. This approach is expected to enhance the predictive model and produce more accurate analyses in identifying the multifactorial causes of gallstone formation.

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI :

Acknowledgment

Thank you to Bapak. Ihsan Fathoni Amri, S.Si., M.Sc. as the supervisor for his guidance and direction during this research process. Gratitude was also conveyed to my teammates, M. Ilham Nurul Azka, and Muji Silvi Rakhmawati for their contributions in the preparation of this article.

References

[1] X. Wang *et al.*, "Global epidemiology of gallstones in the 21st century: a systematic review and meta-analysis," *Clinical Gastroenterology and Hepatology*, vol. 22, no. 8, pp. 1586–1595, 2024.

[2] I. Biantara, V. R. Dewi, L. N. Kharomah, G. P. Dwikijayanti, Y. T. Hidayat, and S. Supriyanto, "Case Study: Application of Perioperative Care in the Diagnosis of Cholelithiasis with Surgical Action of Cholecystectomy Laparatomy," *Jurnal Sains dan Kesehatan*, vol. 7, no. 1, pp. 39–48, 2023.

[3] A. Unalp-Arida and C. E. Ruhl, "Burden of gallstone disease in the United States population: Prepandemic rates and trends," *World J Gastrointest Surg*, vol. 16, no. 4, p. 1130, 2024.

[4] D. Arimbi, A. Novella, and T. Nurina, "Case Report: Acute Cholelitiasis," *Quantum Wellness: Jurnal Ilmu Kesehatan*, vol. 1, no. 3, pp. 34–42, 2024.

[5] T. R. Y. Meidina, N. Mudjihartini, D. R. Gunarti, Y. Yulhasri, S. Dewi, and N. S. Hardiany, "Analysis of Composition and Distribution of Gallstones in the Laboratory of the Faculty of Medicine, University of Indonesia (FKUI) Jakarta," *Jurnal Biotek Medisiana Indonesia*, vol. 9, no. 1, pp. 19–26, 2020.

[6] I. P. Suiraoka, "Penyakit degeneratif," *Yogyakarta: Nuha Medika*, vol. 45, no. 51, 2012.

[7] A. Raflyno, "Radiological Overview of Gallbladder Disorders," *GALENICAL: Jurnal Kedokteran dan Kesehatan Mahasiswa Malikussaleh*, vol. 3, no. 6, pp. 51–62, 2024.

[8] S. Sheth, A. Bedford, and S. Chopra, "Primary gallbladder cancer: recognition of risk factors and the role of prophylactic cholecystectomy," *Official journal of the American College of Gastroenterology| ACG*, vol. 95, no. 6, pp. 1402–1410, 2000.

[9] S. Ujani, "The relationship between age and sex and cholesterol levels of obese people at Abdul Moeloek Hospital, Lampung Province," *Jurnal kesehatan*, vol. 6, no. 1, 2016.

[10] N. M. Parra-Landazury, J. Cordova-Gallardo, and N. Méndez-Sánchez, "Obesity and gallstones," *Visc Med*, vol. 37, no. 5, pp. 394–402, 2021.

[11] A. E. M. Rahayu, A. S. Fauziyah, H. S. Desky, I. Febriyanti, R. H. Artameysia, and P. Haryeti, "Factors in Nurse Clinical Decision Making in Emergency Patients: Literature Review: The Factors In Nurses Clinical Decision-Making In Emergency Department Patients: A Literature Review," *Jurnal Mitra Kesehatan*, vol. 7, no. 2, pp. 183–195, 2025.

[12] M. F. Abdillah, "Healthcare Digital Revolution: Improving Services With Artificial Intelligence.," *Journal of Syntax Literate*, vol. 9, no. 10, 2024.

[13] S. P. Hastono, "Synergy of Biostatistics and Artificial Intelligence for Data-Based Decision-Making in Public Health", 2025.

[14] S. Daruyani, Y. Wilandari, and H. Yasin, "Factors that affect the achievement index of FSM students of Diponegoro University in the first semester with binary logistics method," in *prosiding seminar nasional statistika universitas diponegoro 2013*, Jurusan Statistika Undip, 2013, pp. 185–194.

[15] D. Fabiyanto and Z. P. Putra, "Validation of the Effectiveness of Logistic Regression for Heart Disease Diagnosis through Machine Learning Approaches", 2024.

[16] A. Sheibani, H. Reihani, A. Shoja, M. M. Gharibvand, and M. G. Hanafi, "Gallstones increase the risk of nonalcoholic fatty liver: A case-control study," *Health Sci Rep*, vol. 7, no. 11, p. e70068, 2024.

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI :

[17] A. N. Hapsari, M. S. Chamid, and N. Azizah, "Factors Affecting Low Birth Weight Using Binary Logistic Regression," *Jurnal Sains dan Seni ITS*, vol. 11, no. 1, pp. D50–D56, 2022.
[18] R. Rahman, "Application of Decision Tree for Heart Attack Risk Prediction Based on Digital Lifestyle," *Journal of Information System*, vol. 3, no. 1, pp. 137–142, 2024.
[19] N. K. Hasibuan, S. Dur, and I. Husein, "Factors Causing Diabetes Mellitus with Logistic Regression Method," *G-Tech: Jurnal Teknologi Terapan*, vol. 6, no. 2, pp. 257–264, 2022.
[20] D. Kartikasari, "Analysis of Factors Affecting Air Pollution Levels with Binary Logistics Regression Method," *Mathunesa: Jurnal Ilmiah Matematika*, vol. 8, no. 1, pp. 55–59, 2020.
[21] P. Schober and T. R. Vetter, "Logistic regression in medical research," *Anesth Analg*, vol. 132, no. 2, pp. 365–366, 2021.
[22] T. Pentury, S. N. Aulele, and R. Wattimena, "Ordinal Logistic Regression Analysis," *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, vol. 10, no. 1, pp. 55–60, 2016.
[23] O. Haloho, P. Sembiring, and A. Manurung, "Application of Logistic Regression Analysis on the Use of Female Contraceptives (Case Study in Dolok Mariah Village, Simalungun Regency)," 2013.
[24] F. Febyanti, "Modeling of factors that affect house prices in Greater Jakarta using the probit regression method," *Jurnal Riset Statistika*, pp. 51–57, 2022.
[25] R. Hariyani and M. Martini, "Factors That Affect The Interest Of Accounting Students In Participating In Accounting Professional Education (PPAk)(Case Study on Accounting Students of Budi Luhur University)," *Jurnal Akuntansi dan Keuangan*, vol. 3, no. 1, 2017.
[26] F. Lapenangga and K. B. Ginting, "Application of Multiple Logistics Regression in the Case of Stunting Causative Factors (Case Study: Eimadake Health Center, Sabu Raijua Regency)," *Jurnal Diferensial*, vol. 3, no. 1, pp. 28–37, 2021.
[27] A. Ahmadi, "Analysis Of The Provision Of Health Insurance By Agencies/Companies/Workplace Businesses To Workers In Pariaman City Using Logistic Regression," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 13, no. 3, 2025.
[28] H. Pangestika, D. Ekawati, and N. S. Murni, "Factors associated with the incidence of type 2 diabetes mellitus," *Jurnal'Aisyiyah Medika*, vol. 7, no. 1, pp. 27–31, 2022.
[29] R. R. Adhitya, W. Witanti, and R. Yuniarti, "Comparison of Cart and Naïve Bayes Methods for Customer Churn Classification," *INFOTECH journal*, vol. 9, no. 2, pp. 307–318, 2023.
[30] M. Hadijati and I. Irwansyah, "Comparison of the Classification and Regression Trees (CART) Method with the Naïve Bayes Classification (NBC) in the Classification of Nutritional Status of Toddlers in West Pagesangan Village," *EIGEN MATHEMATICS JOURNAL*, pp. 9–22, 2020.