Indonesian Journal for Health Sciences
Vol. 10, No. 1, March 2026, pp. 25-38
ISSN 2549-2721 (Print), ISSN 2549-2748 (Online)

APPLICATION OF LOGISTIC REGRESSION FOR HEALTH INSURANCE PREMIUM CLASSIFICATION

Celia Christy Merlinda Tantono[1*], Aqilla Dheanya Lucetta[2], Putri Felicia[3], Edwin Setiawan Nugraha[4]
[1,2,3,4]Study Program of Actuarial Science, President University, Bekasi, Indonesia

Article History:
Submitted: 02/08/2025
Accepted: 31/01/2026
Published: 25/03/2026

Keywords:
Logistic Regression,
Insurance Premium,
Risk Classification

ABSTRAK

Abstract:
This study analyzes individual health insurance data using logistic regression to classify premiums into high and low categories based on ten medical and demographic predictors. By transforming a continuous premium variable into binary classes, high and low premiums, this research evaluates the significant factors influencing premium pricing decisions. Logistic regression was selected for its ability to model binary outcomes and estimate the likelihood that the customer belongs to the high premium category. To evaluate the significance of the predictors and the overall model fit, the Likelihood Ratio Test and Wald test were performed, identifying Age (X_1) and Weight (X_7) as significant predictors affecting premium classification. The final logistic regression model has an excellent ability to predict, given the area under the curve (AUC) of 0.97 and a 95% accuracy. These results show how well logistic regression works to enhance risk classification and support data driven policy modifications in insurance underwriting procedures.

***Abstrak*:**
Studi ini menganalisis data asuransi kesehatan individu menggunakan regresi logistic untuk mengklasifikasikan premi ke dalam kategori tinggi dan rendah berdasarkan sepuluh prediktor medis dan demografis. Dengan mengubah variabel premi kontinu menjadi kelas biner, yaitu premi tinggi dan rendah, penelitian ini mengevaluasi faktor-faktor signifikan yang memengaruhi keputusan penetapan harga premi. Regresi logistik digunakan dalam studi ini karena kemampuannya dalam memodelkan hasil biner dan mengestimasi probabilitas seorang pelanggan termasuk dalam kategori premi tinggi. Untuk menilai signifikasi prediktor dan kesesuaian model secara keseluruhan, dilakukan Uji Rasio Kemungkinan dan Wald, yang mengidentifikasi Usia (X_1) dan Berat Badan (X_7) sebagai prediktor signifikan yang memengaruhi klasifikasi premi. Model regresi logistik akhir memiliki kemampuan prediksi yang baik dengan nilai area dibawah kurva (AUC) sebesar 0.97 dan akurasi 95%. Hasil ini menunjukkan seberapa baik regresi logistik bekerja untuk meningkatkan klasifikasi risiko dan mendukung penyesuaian kebijakan berbasis data dalam prosedur penjaminan asuransi.

Corresponding Author:
Celia Christy Merlinda Tantono
Study Program of Actuarial Science,
President University,
Bekasi, Indonesia.
Email: celia.tantono@student.president.ac.id

How to Cite:
C.C.M. Tantono, A.D. Lucetta, P. Felicia, and E.S. Nugraha, "Application of Logistic Regression for Health Insurance Premium Classification," Indonesia. J. Heal. Sci., vol. 10, no. 1, pp. 25-38, 2026.

INTRODUCTION

To determine the price of a health insurance policy premium, several health conditions of the policyholder must be taken into account. An insurance company must review various personal and medical attributes to determine the level of health risk associated with the individual, which directly affects the premium price. Variables such as age, history of chronic illness, previous surgeries, and genetic conditions play a crucial role in underwriting decisions [1]. This multifaceted assessment enables insurers to accurately determine risk pricing and maintain fairness and sustainability within the insurance system. Given the complexity of these variables, logistic regression is a statistical method that is often used to analyze and predict the likelihood of high or low premium groups.

Premium pricing in health insurance is not only part of underwriting, but also a key component in the actuarial structure that underpins the sustainability of the insurance industry. It reflects the balance between the risks covered and provided benefit and plays a role in maintaining the financial stability of insurance companies. According to Putri [2], premium determination in health insurance is essential for maintaining insurers' financial stability while ensuring fair pricing for policyholders through the assessment of individual risk characteristics. Similarly, Piontkowski [3] applies actuarial methods in pricing health insurance products with relatively small insured groups in Germany, providing empirical context actuarial pricing practices. Other studies show that errors in risk specification can distort premium determination and impact the financial stability of insurance companies [4]. In this context, a newer multitask network approach has been proposed to calculate premiums that are free from discrimination and more accurate from an actuarial perspective [5].

As a function with a financial protection mechanism that covers healthcare expenditures in return for constant premium payments, the health insurance underwriting process involves a comprehensive risk assessment based on the applicant's medical history, physical measurements, and current health conditions. Prisila *et al.*, [6] demonstrated that binary logistic regression is effective in identifying key factors influencing health insurance ownership in Indonesia, particularly employment status, which significantly increases the likelihood of individuals having health insurance coverage. Their study emphasized how demographic and socioeconomic factors, such as age, education, and employment status, contribute to insurance participation decisions. This supports the relevance of logistic regression in evaluating risk related behaviors in health coverage decisions and pricing strategies.

Finding the right premium price is challenging for health insurance because numerous different and complex health profiles can affect it. Traditional methods of testing often fail to accurately measure the overall impact of various risk factors, such as age, chronic illness, and medical history. These limitations may result in erroneous pricing, which could harm insurance companies by underestimating risk or lead to unreasonably high costs for consumers. As a result, trustworthy analytical techniques like logistic regression are urgently needed. These techniques can predict the relationship between health factors and premium price classifications, allowing insurance companies to make more informed and precise underwriting decisions.

Statistical modeling methods, also known as logistic regression, can be used to estimate the probability of a binary outcome based on one or more predictor factors. Logistic regression is used to model the relationship between one or more predictor variables and a binary or dichotomous response variable, making it suitable for estimating the probability of occurrence of an event based on specific

categories [7]. Logistic regression may translate the linear combination of independent variables into probability values ranging from 0 to 1 by utilizing the logistic function, also known as the sigmoid curve. In the insurance sector, where decisions are influenced by the classification of individuals into different risk groups based on demographic and health-related characteristics, logistic regression provides an appropriate framework for categorizing binary outcomes and supporting risk-based decision making [7].

Several previous studies have applied logistic regression to model and predict insurance premium classifications based on medical and demographic variables. For instance, a study in [1] applied machine learning, including logistic regression, to predict premium levels based on customer health profiles. They developed an artificial neural network, ANN-based regression model using a dataset of over 1,300 entries with seven variables, such as charges, smoking status, area, number of children, BMI, gender, and age, to estimate individual insurance costs. Similarly, Narayana et al., [8] also analyzed medical insurance premium prediction using multiple machine learning models, including linear regression, ridge regression, support vector machine (SVM), and random forest regression, with variables such as age, sex, BMI, number of children, smoking status, region, and charges [8]. In another study Prisila *et al.*, [6] applied a two-level binary logistic regression model to examine how individual characteristics, especially employment status, affect health insurance ownership decisions in Indonesia, highlighting the suitability of logistic regression models for analyzing health insurance participation and related risk factors. Their studies identified employment, education, chronic conditions, marital status, and inpatient care as significan't factors affecting insurance ownership, while gender and general health condition were not significant.

While previous research has effectively applied logistic regression to analyze factors affecting medical insurance premiums or ownership, most studies focus on either large scale insurance schemes or datasets with limited demographic and health predictor variables [9]. In contrast, this study utilizes a distinct dataset of individuals with comprehensive health and demographic variables, including diabetes status, blood pressure problems, chronic diseases, allergies, family cancer history, transplants, and the number of major surgeries. By incorporating these comprehensive medical factors, this study provides a unique contribution in highlighting the relevance for premium classification. This allows for more precise analysis of the influence of individual health conditions on premium classification [10], [11]. Additionally, this study aims specifically at classifying premium levels, rather than predicting general risk or insurance eligibility, providing a more focused approach for premium pricing strategies. By including detailed health variables that are not commonly used in previous studies, this research offers deeper insights into the determinants of insurance premium classification [12], [13].

The objective of this research is to analyze and classify the premium price categories of medical insurance using logistic regression based on various health and demographic factors. By applying this method to a dataset of individuals, the study aims to identify which variables significantly influence premium pricing decisions. This analysis is intended to support insurance providers in making data driven and consistent underwriting decisions. Furthermore, the research seeks to demonstrate the effectiveness of logistic regression as a predictive tool in the context of medical insurance pricing.

RESEARCH METHOD

The secondary data used in this study came from Tejashvi's Medical Insurance Premium Prediction dataset, which was obtained on Kaggle. A total of 500 observations were chosen as sample data. Before conducting the analysis, data was cleaned up by checking for missing values, handling outliers, and ensuring that the dataset was balanced. The dataset was then divided into 80% for training and 20% for testing. It contains ten predictor variables and one response variable, which is the insurance premium price category. The response variable was created by converting the premium values into a binary outcome based on the median. Premiums below the median were classified as low, while premiums equal to or over the median were classified as high. Logistic regression models the relationship between independent variables and a binary response. Its utility is particularly evident in scenarios where the objective is to estimate the probability of an event occurring as a function of the explanatory variables [14].

1. Variable Description

The response variable in logistic regression is usually binary, resulting in a dependent variable with two categories, such as success or failure, the presence or absence of a particular event, which is commonly encoded as 0 or 1 [14], [15], [16].Predictors or explanatory variables, also called independent variables, can be either categorical or continuous. We use these variables to figure out how likely it is that a certain outcome will happen [14], [15].

2. Sigmoid Function

The sigmoid function is an S-shaped mathematical function used in logistic regression to map a linear function of the input variables into probabilities between 0 and 1. In this formulation, g denotes the sigmoid function, w represents the weights, and b the bias, making logistic regression a member of the generalized linear model family [17], [18]. In math, it is defined as:

$$P(y = 1|X) = sigmoid(z) = \frac{1}{1 + e^{-z}}$$

where:

$$z = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

$\hat{\beta}_k$ = coefficients (parameter) of the model.

3. Logistic Function

Logistic regression models the relationship between independent variables and the probability of an event in the dependent variable indirectly through a logistic function [19]. This function captures a nonlinear relationship between the predictor variables and the log-odds of the outcome. The parameters of the model are estimated in the form of a logit, defined as the natural logarithm of the odds that the dependent variable takes the value of one. This relationship can be mathematically represented as follows [19]:

$$logit(p) = log\left(\frac{\pi}{1-\pi}\right) = ln\left(\frac{\pi}{1-\pi}\right)$$

The probability value (π) ranges between 0 and 1, whereas the logit scale extends from negative infinity to positive infinity and is symmetric around 0.5. Therefore, the logistic regression model can be expressed as follows [19]:

$$logit(p) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

where:

$$\pi(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k}}$$

β_i= The change in log-odds of the outcome associated with one unit increase in the predictor x_i. By applying the logistic function, we convert linear predictions into probabilities, making the model interpretable and appropriate for classification tasks.

4. Likelihood Ratio Test (LRT)

Parameter estimation testing is used to determine whether there is a relationship between the response variable and the predictor variables. Simultaneous parameter estimation testing is conducted using the likelihood ratio test [20]. The likelihood function is the model's probabilities for each observation multiplied together. In logistic regression, this corresponds to the product of predicted probabilities for events and 1 minus the predicted probabilities for non-events [16]. Mathematically, the likelihood function is expressed as:

$$L(\theta|x_1, \dots, x_n) = L(\theta|x) = L(\beta) = \prod_{i=1}^{n} f(x_1|\theta) = \prod_{i=1}^{n} \pi_i^{yi}(1 - \pi_i)^{1-yi}$$

To facilitate computation, the log-likelihood function is commonly used:

$$Log\ L(\beta) = \sum_{i=1}^{n} [y_i \log(\pi_i)) + (1 - y_i)\log(1 - \pi_i)]$$

π = Predicted probability for observation
y_i = The actual outcome {0,1}

In addition, the Likelihood Ratio Test (LRT) is widely used to examine both the adequacy of the model and the contribution of individual predictors. The test operates by contrasting a comprehensive model containing the predictor with a nested reduced model that omits the predictor [16]. The contribution of each predictor to a specific model can be assessed using the Likelihood Ratio Test, which is also applied to evaluate the overall goodness-of-fit of the model. For a given parameter, this test compares the likelihood of the observed data under the full model that includes all parameters, denoted as L_1, with the likelihood of the data under a reduced model in which the parameter of interest is constrained to zero, denoted as L_0 [16]. The formula below is used for determining the test statistic:

$$G = -2\ ln\left(\frac{L_0}{L_1}\right) = -2 \ln (L_0 - L_1)$$

The degrees of freedom reflect the number of parameters included in the hypothesis test (commonly df = 1 when assessing a single predictor), and the test statistic is assumed to follow a chi-squared distribution. Through hierarchical model building and systematic comparison of nested models, the incremental effect of each predictor on model performance can be determined [16].

5. Wald Test

In logistic regression, inference about model parameters is commonly conducted through hypothesis testing of regression coefficients. Each coefficient can be evaluated using a Wald test, which assesses whether the estimated parameter differs significantly from zero. This test compares the null hypothesis that a coefficient equals zero (indicating no effect on the response) with the alternative hypothesis that the coefficient is nonzero. Additionally, the Wald test may be applied to compare a reduced model, in which certain parameters are constrained to zero, with a full model that includes all predictor variables, thereby determining the contribution of specific predictors to the model [21].

The Wald statistic is computed by dividing the squared estimated coefficient by the squared standard error of that coefficient. When evaluated with one degree of freedom, the resulting statistic follows an asymptotic chi-square (χ^2). The Wald test is widely applied in logistic regression to determine the statistical significance and individual contribution of predictor variables to the model [21]. The Wald statistic is distributed as a (χ^2):

$$W_i = \frac{\beta_i^2}{SE^2{}_{\beta i}}$$

β_i = The estimate regression coefficient
$SE\beta_i$ = The actual outcome {0,1}
W = Wald statistic

Once the Wald statistic (W) is computed, the corresponding p-value is obtained by calculating the tail probability of the chi-squared distribution:

$$p - value = 2\,x \left(1 - \Phi\left(\left|\frac{\hat{\beta}}{SE(\hat{\beta})}\right|\right)\right)$$
$$= P(X^2 > W)$$

$\hat{\beta}$ = The estimate regression coefficient
$SE(\widehat{\beta})$ = Standard error of the coefficient
ϕ = Cumulative distribution function (CDF) of the standard normal distribution

To determine statistical significance, the following rule is employed. The null hypothesis is rejected if the p-value is less than α, which means that the variable has a statistically significant impact on the dependent variable. On the other hand, the null hypothesis is not rejected if the p-value is less than α, indicating that there is no statistically significant effect of the variable [15].This approach allows the identification of the most influential predictors within the model and helps refine the regression to include only variables that contribute meaningfully to the prediction of the binary outcome.

6. Odds Ratio

In logistic regression, the impact of an independent variable on a binary outcome is interpreted using the odds ratio (OR). When all other variables are held constant, it measures the change in the probabilities of the outcome with a one-unit increase in the predictor variable.

$$OR = e^{-\beta_i}$$

β_i = The estimate regression coefficient

The odds ratio (OR) can be interpreted as follows. An OR greater than 1, shows that the likelihood of the outcome happening has increased. An OR less than 1, shows a decline in the likelihood that the result will materialize. While an OR equal to 1, implies that the outcome's probabilities are unaffected by the predictor [3].

7. Confusion Matrix

An overview of the classification problem's forecast outcomes is called a confusion matrix. It shows the total number of correct and incorrect predictions that the model generated compared to the actual outcomes (target values) of the data [14].In a binary classification task, the n a binary classification task, the instances are typically predicted to be either positive or negative, where a positive label represents the presence of a condition, illness, abnormality, or some deviation from the baseline, and a negative label represents the absence of such a condition. Each predicted binary label can therefore fall into four possible designations [22].

The relationship between actual and predicted classifications in a binary classification task is organized into a confusion matrix. This table serves as the basis for calculating different evaluation metrics that are used in evaluating the model's performance. The structure of the confusion matrix is shown below [14]:

Table 1.
Confusion Matrix

Actual	Predicted			
	Yes		No	
Yes	True (TP)	Positive	False (FN)	Negative
No	False (FP)	Positive	True (TN)	Negative

From this table, we can calculate various classification performance evaluation metrics such as accuracy,

sensitivity/recall, specificity, and precision [23].

Accuracy represents the total percentage of accurate predictions the model makes is known as accuracy. Both accurately detected positive and negative cases are taken into account [23]. The following is the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity, also known as recall or the true positive rate, indicates the model's sensitivity is its capacity to accurately detect real positive cases. It is important in situations like illness detection where false negative results might be harmful [23]. The formula is:

$$Sensitivity = \frac{TP}{TP + FN}$$

A high sensitivity means that the model is effective in capturing most of the positive cases, minimizing the number of missed detections.

Specificity or the true negative rate is the key metric that determines how well the model correctly identifies actual negative cases. When false positives are undesirable, it can be helpful [23]. The formula is:

$$Specificity = \frac{TN}{TN + FP}$$

The high specificity of the model is clear as it filters out true negatives and avoids false alerts.

Lastly, precision measures the percentage of expected positive cases that are actually accurate. The formula is:

$$Precision = \frac{TP}{TP + FP}$$

A high precision model lowers the possibility of incorrect positive classifications by guaranteeing that the majority of positive predictions are right [24].

8. ROC & AUC

The ROC curve, also called the Receiver Operating Characteristic curve, is a graphical tool for evaluating the performance of a binary classification model [25]. Over a range of threshold values, it compares the True Positive Rate (sensitivity) against the False Positive Rate (1 − specificity). Although the ROC curve is typically visualized as a graph, in this study it is described textually without including the actual plot [25]. This allows us to understand how well the model separates the two classes at various classification cutoffs. A curve that bends toward the upper-left corner, indicating high sensitivity and low false positive rate, reflects a model with strong discriminatory ability. The ROC curve is particularly useful for comparing models or assessing classifier behavior under different threshold settings [14], [15].

The Area Under the Curve (AUC) is derived from the ROC curve and provides a scalar metric summarizing the model's overall performance [25]. Its value ranges from 0.5, indicating no discriminative ability (equivalent to random guessing), to 1.0, representing perfect classification. A higher AUC means that, across all thresholds, the model more successfully ranks positive examples above negative ones. Unlike the ROC curve, which gives a visual representation, AUC offers a threshold-independent measure, making it especially useful for comparing classifiers when only one summary statistic is desired [14], [15].

RESULT AND ANALYSIS

Table 2 presents the predictor variables in the dataset along with their frequencies and percentage. For continuous variables such as age, height, and weight, the mean and standard deviation are provided, while categorical variables are summarized by counts and percentages.

Table 2.
The Frequency of Predictor Variables

Variable		n	(%)
X_1	Age		
	Mean	36.2	
	Standard deviation	14.3	
X_2	Diabetes		
	No	241	60.2
	Yes	159	39.8
X_3	BP Problems		
	No	231	57.8
	Yes	169	42.2
X_4	Transplants		
	No	384	96
	Yes	16	4
X_5	Chronic Diseases		
	No	333	83.2
	Yes	67	16.8
X_6	Height		
	Mean	167.5	
	Standard Deviation	10.1	
X_7	Weight		
	Mean	75.8	
	Standard deviation	12.8	
X_8	Allergies		
	No	314	78.5
	Yes	86	21.5
X_9	Family Cancer History		
	No	363	90.8
	Yes	37	9.25
X_{10}	Number of Major Surgeries		
	0	226	56.5
	1	143	35.8
	2	27	6.75
	3	4	1
	Total	400	100

Given the variables shown in Table 1, logistic regression analysis was used to analyze the relationship between the predictors and the likelihood of a customer having a high or low insurance premium.

1. Likelihood Ratio Test (LRT)

This test determines whether the model with all predictor variables (full model) are significantly better fitted to the data than the model with only the intercept (null model). The difference in deviance between the models is represented by the test statistic, G^2. The anova() function in R software was used to conduct this test with the result of the test presented in Table 3.

Table 3.
Likelihood Ratio Test

Model	Df	Deviance	p-value
Null Model			
Full Model	10	404.35	<2.2e-16

Table 3 shows that the difference in deviance between the two models results in a test statistic G^2 of 404.35 and a p-value less than 2.2e-16. This suggests that at the level of significance of 5%, the full model fits the data significantly better than the null model. In other words, at least one predictor variable has significant effects on the response variable or premium price.

2. Wald Test

Wald test is used to assess the significance of each independent variable (predictor variable) on the dependent variable (response variable). This procedure helps determine whether each variable has a meaningful influence in the model. In addition, this test is useful for evaluating the model's ability to explain the response variable. The test statistics are obtained using the glm(formula, family = , data =) function in R with the results for each variable presented in Table 4.

Table 4.
Wald Test

Variable	β	S. E	Wald	-value
Intercept	-18.54	4.40	-4.21	<0.01
Age	0.426	0.05	7.98	<0.01
Diabetes	-0.045	0.48	-0.09	0.925
BP Problems	0.275	0.45	0.61	0.542
Transplants	0.128	1.16	0.11	0.912
Chronic Diseases	-0.252	0.52	-0.48	0.630
Height	-0.002	0.02	-0.07	0.942
Weight	0.068	0.02	3.78	<0.01
Allergies	-0.133	0.56	-0.24	0.812
Family Cancer History	0.235	0.76	0.31	0.759
Number of Major Surgeries	-0.016	0.53	-0.03	0.976

Based on the Wald test shown in Table 4, it can be concluded that the predictor variables that significantly affect the premium price are Age (X_1) and

Weight (X_7). The p-values of these variables are less than 0.01, which is below the 5% significance level. The remaining predictor variables have a p-values above the significance level and therefore excluded from the final logistic regression model.

Table 5.
Wald Test for Final Model

Variable	β	S. E	Wald	-value
Intercept	-18.75	2.48	-7.57	<0.01
Age	0.424	0.05	8.13	<0.01
Weight	0.069	0.02	3.96	<0.01

Table 5 shows the Wald test results for the significant predictor variables identified from the full model shown in Table 4. Based on these results, the final logistic regression model can be expressed as follows:

$$z = \beta_0 + \beta_1 X_1 + \beta_7 X_7$$

$$z = -18.75 + 0.424 \times Age + 0.069 \times Weight$$

Thus, the probability of having a high premium price is given by:

$$p = \frac{1}{1 + e^{-(-18.75 + 0.424 \times Age + 0.069 \times Weight)}}$$

3. Odds Ratio

Based on the significant predictor variables identified from the Wald test, the odds ratio is calculated to interpret the effect of each variable on the probability of being categorized as having a high premium price. The results are presented in Table 6.

Table 6.
Odds Ratio

Variable	β	Odds Ratio
Age (X_1)	0.424	1.528
Weight (X_7)	0.069	1.071

Table 6 shows the odds ratios for variable Age (X_1) and Weight (X_7) are 1.528 and 1.071, respectively. These results indicate that for every one-year rise in age, the odds of having a high premium price increase by 52.8% and for each one-kilogram increase in weight, the odds increase by 7.1%.

4. Confusion Matrix

The confusion matrix is used to summarize the number of correct and incorrect classifications based on the predictions and actual outcomes from the test data. Predictions were classified as high premium (Yes) if the predicted probability exceeded 0.5, and as low premium (No) otherwise. This evaluation helps measure how well the model performs in distinguishing between high and low premium categories. Additionally, it provides insight into potential misclassifications, allowing further assessment of model reliability and accuracy. The confusion matrix results are presented in Table 7.

Table 7.
Confusion Matrix

Actual	Predicted	
	Yes	No
Yes	48	2
No	3	47

Based on the confusion matrix results in Table 7, the model performance metrics are calculated and presented in Table 8. This calculation is carried out to provide a clearer understanding of the model's ability to classify the data accurately.

Table 8.
Model Performance Metrics

Metric	p-Value
Accuracy	0.95
Sensitivity	0.94
Specificity	0.96
Precision	0.96

Table 8 shows the model performance metrics for the test data. The model has an accuracy of 95%, meaning

that the majority of the predictions are right. The sensitivity of 94% indicates that the model accurately identifies 94% of clients with high premiums while a 96% specificity means the model correctly identified 96% of clients with low premiums. Furthermore, the precision of 96% indicates that the model correctly predicts 96% of clients with a high premium.

5. ROC & AUC

The ROC curve was generated to visualize how sensitivity and specificity change across different classification thresholds. This curve illustrates the comparison between the model's ability to correctly identify customers with high premium and its accuracy in rejecting low premium customers, providing a comprehensive evaluation of the model's discriminative ability as shown in Figure 1.

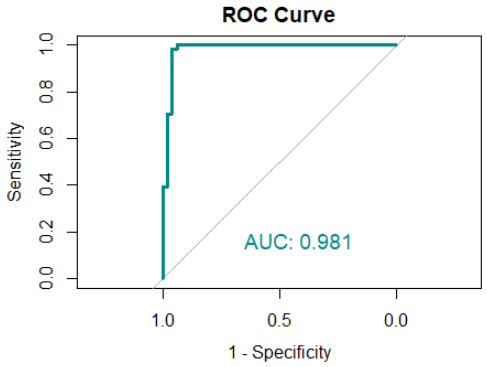

Figure 1. The Receiver Operating Characteristic (ROC) curve with Area Under the Curve (AUC)

Based on Figure 1, the area under the curve (AUC) is 0.97, indicating excellent model performance in distinguishing between customers with high and low premiums since it is close to 1

DISCUSSION

The findings of this study indicate that Age and Weight are significant predictors of health insurance premium classification. The Wald test results show that both variables have p-values below the 5% significance level, indicating that these variables have a statistically significant effect on premium classification.

The significance of Age and Weight in the model is consistent with the foundational logic of insurance underwriting. Age is a key determinant of healthcare expenditures, as older individuals tend to experience higher medical costs due to the increased prevalence of chronic conditions [26]. Similarly, Weight, which is closely associated with BMI, represents an important indicator of health risk in insurance models. Excess weight is linked to a higher incidence of chronic diseases, such as diabetes, cardiovascular and respiratory disorders, osteoarthritis, and premature death. Higher BMI is also associated with increased healthcare expenditures, which are expected to rise alongside the growing prevalence of obesity [27]. In health insurance, obesity has been formally identified as one of the underwriting risk classification factors that can be used to estimate the mortality driven benefits, further supporting the Weight as a significant variable in the model [28].

In contrast, the findings indicated that several comprehensive health variables such as Diabetes, BP Problems, Transplants, Chronic Diseases, Allergies, Family Cancer History, and Number of Major Surgeries are not statistically significant. This could be attributed to factors such as a small sample size or the presence of other variables in the model that reduces its apparent impact. However, these variables may still have a significant impact that could become apparent with a larger or different sample, even though the statistical significance was not found [21].

The final model obtained 95% accuracy with an AUC of 0.97, showing that the model is able to distinguish between high and low premium categories accurately. Additionally, the model achieved 94% sensitivity, 96% specificity and precision, showing that it can reliably recognize customers whether they have high premium levels as well as those with

low premium levels. These results indicate that the logistic regression model has high values for all evaluation measures, indicating great predictive ability. The high performance across all evaluation metrics indicates that logistic regression models can effectively distinguish between high- and low-risk groups, consistent with findings in previous studies on health insurance classification models [29].

The strong performance of the model, which uses only two variables, highlights its practical utility. The insurance industry's ongoing shift toward data driven underwriting priorities not only accuracy but also interpretability and operational efficiency in risk assessment tools [30]. This reflects a recognized challenge where even highly accurate advanced machine learning models face limited application in actuarial science due to a lack of transparency and strict regulatory demands [31]. Furthermore, this demand for transparency is reinforced by empirical evidence of regulatory challenges, where models lacking transparency have faced compliance hurdles in insurance applications [32]. Therefore, a simple yet powerful model like this one is particularly valuable. This efficiency makes it a practical starting point for initial premium assessment. This directly addresses a key industry demand for interpretable solutions as underwriting becomes more dependent on data and algorithms.

These findings are aligned with the study by Patil and Saravanan [33], who identified Age and BMI as significant determinants of insurance premium pricing. Although our study uses Weight instead of BMI, this aligns with previous findings, since BMI is established as a standard clinical and public health measure for weight related health risk and is fundamentally calculated from an individual's weight and height [34]. Similarly, the association of Age with premium price in our study aligns with the research by Kaushik et al. [1] and Kongyir and Agbemade [26], who found that Age is a significant factor influencing insurance charges, supporting the relevance of Age as an indicator predictor variable in premium classification. Despite differences in the models used for the regression, these studies suggest that Age and Weight are significant predictors of health insurance premiums.

However, despite the promising results, this study has several limitations. The dataset is secondary and relatively small (500 observations), which may restrict the generalizability of the findings. Moreover, since the data originates from Kaggle and not from actual underwriting records, there may be potential biases that affect the real-world applicability of the model.

CONCLUSION

This study applied logistic regression to classify health insurance premium levels based on demographic and medical factors. The analysis identified Age (X_1) and Weight (X_7) as the significant predictors of premium classification, with the final model achieving high predictive accuracy, demonstrating 95% accuracy and an AUC of 0.97. The probability of having a premium price is expressed by the final model below.

$$p = \frac{1}{1 + e^{-(-18.75+0.424\times Age+0.069\times Weight)}}$$

From a practical standpoint, the findings offer valuable insights for insurers in early risk assessment and premium pricing. Specifically, Age and Weight can be used as preliminary screening variables to guide underwriting decisions, optimize pricing strategies, and develop data driven policy adjustments.

This study focused on binary classification of health insurance premiums based on a limited number of predictors. Future research could explore several extensions. First, incorporating larger and more diverse datasets, including behavioral and socioeconomic factors, could improve

the model's generalizability. Second, comparing logistic regression with other machine learning algorithms such as Random Forest or Gradient Boosting could reveal performance trade-offs and improve decision support. Third, applying cross-validation and regularization techniques (e.g., LASSO) could improve model robustness and prevent overfitting. Finally, applying this framework to a real-world insurance underwriting context could validate the model's practical utility and assess its economic impact.

REFERENCES

[1] K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, "Machine Learning-Based Regression Framework to Predict Health Insurance Premiums," *Int. J. Environ. Res. Public Health*, vol. 19, no. 13, p. 7898, Jun. 2022, doi: 10.3390/ijerph19137898.

[2] A. Putri, "Estimation Model of Pure Health Insurance Premiums in Southeast America Using Generalized Linear Model (GLM) with Gamma Distribution," *International Journal of Business, Economics, and Social Development*, vol. 6, no. 1, pp. 44–51, Feb. 2025, doi: 10.46336/ijbesd.v6i1.873.

[3] J. Piontkowski, "Pricing German health insurance products with only few insured persons," *Eur. Actuar. J.*, vol. 15, no. 3, pp. 831–857, Dec. 2025, doi: 10.1007/s13385-025-00427-6.

[4] D. Finger, H. Albrecher, and L. Wilhelmy, "On the cost of risk misspecification in insurance pricing," *Japanese Journal of Statistics and Data Science*, vol. 7, no. 2, pp. 1111–1153, Nov. 2024, doi: 10.1007/s42081-024-00256-z.

[5] M. Lindholm, R. Richman, A. Tsanakas, and M. V. Wüthrich, "A multi-task network approach for calculating discrimination-free insurance prices," *Eur. Actuar. J.*, vol. 14, no. 2, pp. 329–369, Aug. 2024, doi: 10.1007/s13385-023-00367-z.

[6] M. A. Prisila, A. Islamiyati, and A. K. Jaya, "Model Data Kepemilikan Asuransi Kesehatan di Indonesia Berdasarkan Status Pekerjaan Melalui Analisis Regresi Logistik Biner Dua Level," *Contemporary Mathematics and Applications (ConMathA)*, vol. 4, no. 2, pp. 125–133, Oct. 2022, doi: 10.20473/conmatha.v4i2.39354.

[7] R. S. Sari, Zahedi, P. Sembiring, and A. Manurung, "Penerapan Model Regresi Logistik Untuk Mengidentifikasi Potensi Faktor Risiko Malaria Di Sumatera Utara," *Leibniz: Jurnal Matematika*, vol. 4, no. 2, pp. 56–70, Jul. 2024, doi: 10.59632/leibniz.v4i02.415.

[8] K. L. Narayana, Yogesh, and P. Kowshik, "Medical Insurance Premium Prediction Using Regression Models," *International Journal for Research Trends and Innovation*, vol. 8, no. 4, pp. 1512–1517, 2023, [Online]. Available: www.ijrti.org

[9] S. D. Permai and K. Herdianto, "Prediction of Health Insurance Claims Using Logistic Regression and XGBoost Methods," *Procedia Comput. Sci.*, vol. 227, pp. 1012–1019, 2023, doi: 10.1016/j.procs.2023.10.610.

[10] M. Kapse, V. Sharma, R. Vidhale, and V. Vellanki, "Customization of health insurance premiums using machine learning and explainable AI," *International Journal of Information Management Data Insights*, vol. 5, no. 1, p. 100328, Jun. 2025, doi: 10.1016/j.jjimei.2025.100328.

[11] K. Vootukuri, V. K. S. Kumar, and V. S. Naik, "Social and demographic determinants of health insurance status in India: Evidence from a nationally representative cross-sectional survey," *J. Educ. Health Promot.*, vol. 13, no. 1, Apr. 2024, doi: 10.4103/jehp.jehp_1006_23.

[12] F. EL KASSIMI, A. SAOUDI, and J. ZAHI, "Health insurance risk classification using multinomial logistic regression," *IJCEDS*, vol. 3, no. 2, pp. 13–18, Aug. 2023, [Online]. Available: https://ijceds.com/ijceds/article/view/54/26

[13] N. Rambeli, N. Abdul Jalil, N. Katmon, A. Hashim, E. Hashim, and M. R. Ramdan, "The Distribution of Failure in Medical and Health Insurance Claim: The Implications of Multiple Binary Logistic Regressions," *Journal of Distribution Science*, vol. 21, no. 11, pp. 23–30, Nov. 2023, doi: https://doi.org/10.15722/jds.21.11.202311.23.

[14] G. James, D. Witten, T. Hastie, and R. Tibshirani, "Classification," in *An Introduction to Statistical Learning*, 2nd ed., 2021, ch. 4, pp. 133–152. doi: 10.1007/978-1-0716-1418-1_4.

[15] J. M. Hilbe, *Practical Guide to Logistic Regression*. Chapman and Hall/CRC, 2016. doi: 10.1201/b18678.

[16] R. R. HashunatilMar'ah and Y. Setiadi, "Penerapan Regresi Logistik Biner Terhadap Faktor-Faktor yang Memengaruhi Pemanfaatan Jaminan Kesehatan Pasien Rawat Jalan di Provinsi Nusa Tenggara Barat Tahun 2020," *Seminar Nasional Official Statistics*, vol. 2022, no. 1, pp. 947–956, Nov. 2022, doi: 10.34123/semnasoffstat.v2022i1.1295.

[17] S. M, S. J. R, K. K, and Ganjendran, "SPAM DETECTION USING MACHINE LEARNING," *International Advanced Research Journal in Science, Engineering and Technology (IARJSET)*, vol. 11, no. 4, pp. 278–284, 2024, doi: 10.17148/IARJSET.2024.11440.

[18] C. B.S., A. R, V. A, and D. P. M, "Movie Profit Prediction," *International Journal of Modern Agriculture*, vol. 10, no. 1, pp. 915–927, 2021, [Online]. Available: http://www.modern-journals.com/index.php/ijma/article/view/689/596

[19] I. Azagi, Erfiani, Indahwati, Anwar Fitrianto, and Reni Amelia, "Pemodelan Regresi Logisitik Biner pada Faktor-Faktor yang Memengaruhi Indeks Pembangunan Manusia di Pulau Jawa," *Jurnal Statistika dan Aplikasinya*, vol. 6, no. 1, pp. 1–11, Jun. 2022, doi: 10.21009/JSA.06101.

[20] Alfi Indah Nurrizqi, Erfiani, Indahwati, Anwar Fitrianto, and Reni Amelia, "Pemodelan Regresi Logistik Berbasis Backward Elimination Untuk Mengetahui Faktor yang Mempengaruhi Tingkat Kemiskinan di Indonesia Tahun 2021," *Jurnal Statistika dan Aplikasinya*, vol. 6, no. 2, pp. 160–170, Dec. 2022, doi: 10.21009/JSA.06202.

[21] Alan. Agresti, *An introduction to categorical data analysis*, 3rd ed. John Wiley & Sons, 2019.

[22] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[23] J. C. Obi, "A comparative study of several classification metrics and their performances on data," *World Journal of Advanced Engineering Technology and Sciences*, vol. 8, no. 1, pp. 308–314, Feb. 2023, doi: 10.30574/wjaets.2023.8.1.0054.

[24] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *African Journal of Biomedical Research*, pp. 4023–4031, Nov. 2024, doi: 10.53555/AJBR.v27i4S.4345.

[25] M. Imani, M. Joudaki, A. Bagheri, and H. R. Arabnia, "Why ROC-AUC Is Misleading for Highly Imbalanced Data: In-Depth Evaluation of MCC, F2-Score, H-Measure, and AUC-Based Metrics Across Diverse Classifiers," *Technologies (Basel).*, vol. 14, no. 1, p. 54, Jan. 2026, doi: 10.3390/technologies14010054.

[26] B. Kongyir and E. Agbemade, "Modeling Health Insurance Premium Using Bayesian Hierarchical Models," 2024, [Online]. Available: http://library.ucf.edu

[27] A. I. Taloba, R. M. Abd El-Aziz, H. M. Alshanbari, and A.-A. H. El-Bagoury, "Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning," *J. Healthc. Eng.*, vol. 2022, pp. 1–10, Mar. 2022, doi: 10.1155/2022/7969220.

[28] T. Woods and T. Miljkovic, "Modeling the Economic Cost of Obesity Risk and Its Relation to the Health Insurance Premium in the United States: A State Level Analysis," *Risks*, vol. 10, no. 10, pp. 197–225, Oct. 2022, doi: 10.3390/risks10100197.

[29] V. P. Thakre, R. D. Poul, and A. D. Sawarkar, "Predictive Precision: Unraveling Health Insurance Claim Patterns With Logistic Regression and Decision Trees," *Cureus Journal of Computer Science*, Mar. 2025, doi: 10.7759/s44389-025-03010-y.

[30] N. Bishop, "Application of Machine Learning Techniques in Insurance Underwriting," *Journal of Actuarial Research*, vol. 2, no. 1, pp. 1–13, Mar. 2024, doi: 10.47941/jar.1756.

[31] C. Lozano-Murcia, F. P. Romero, and M. C. Gonzalez-Ramos, "Protocol for Evaluating Explainability in Actuarial Models," *Electronics (Basel).*, vol. 14, no. 8, p. 1561, Apr. 2025, doi: 10.3390/electronics14081561.

[32] H. S. B. Gummadi, "Explainable AI-Enhanced Underwriting Automation for Personalized Insurance Policy Recommendations," *European Journal of Computer Science and Information Technology*, vol. 13, no. 19, pp. 24–40, Apr. 2025, doi: 10.37745/ejcsit.2013/vol13n192440.

[33] S. Patil and S. R, "The Role of Actuarial Models in Risk Assessment: A Study on Insurance Pricing and Underwriting," *International Journal of Research Publication and Reviews Journal homepage: www.ijrpr.com*, vol. 6, pp. 839–845, 2025, [Online]. Available: https://ijrpr.com/uploads/V6ISSUE4/IJRPR41434.pdf

[34] *Health at a Glance 2025: OECD Indicators*. Paris: OECD Publishing, 2025. doi: 10.1787/8f9e3f98-en.