Multidisciplinary Output Research For Actual and International Issue(MORFAI Journl) ISSN (e): 2808-6635
Volumes 5 No. 2 (2025)

ANALYSIS AND PREDICTION OF HEALTH INSURANCE PREMIUM VALUE USING MACHINE LEARNING ALGORITHM

Danica Recca Danendra[1]*, Januponsa Dio Firizqi[2]
Department of Informatics, Faculty of Science and Technology, Universitas Pradita, Indonesia [1,2]
E-mail: recca.danendra2@gmail.com

Received	: 17 March 2025	Published	: 19 May 2025
Revised	: 30 March 2025	DOI	: https://doi.org/10.54443/morfai.v5i2.2775
Accepted	: 16 April 2025	Link Publish	: https://radjapublika.com/index.php/MORFAI/article/view/2775

Abstract

Rising healthcare costs and administrative complexity in the health insurance sector underscore the need for an efficient predictive model to anticipate insurance premium prices. The study explores Machine Learning (ML) techniques to predict the value of health insurance premiums. Also, it aims to provide further insights to stakeholders to create strategies in premium pricing and risk management. This study uses the Kaggle.com datasets and a boosting regression algorithm to compare the accuracy and metric evaluation results in predicting the value of insurance premiums. Feature engineering techniques are applied to improve model performance, reduce over-fitting, and interpret the model to ensure the inclusion of relevant predictors by studying the strengths and limitations of each technique. They overcome this through feature selection, model interpret-ability, scalability, and generalization. Through this comprehensive review, the results of this study aim to provide valuable insights for practitioners, researchers, and policymakers, as well as facilitate informed decision-making in the context of determining the value of health insurance premiums through the use of ML methodologies.

Keywords: ***XGBoost, CatBoost, LGBM, Premiums of Insurance, Grid Search.***

INTRODUCTION

The insurance industry plays an important role in the economy. By definition, the function of insurance is to restore the financial position as it was before the risk occurred. Insurance also plays a role in providing encouragement for the direction of national economic development. Currently, the insurance industry in Indonesia consists of Conventional Insurance and Sharia Insurance. This industry continues to grow and adapt to developments in the era, such as digitalization, to provide protection against possible losses that will occur (Nugraha & Irawan, 2023; Tumbel & Ananto, 2024).

With the increasing number of insurance types, customer and insurance holder data is increasing every year. This allows for the complexity of administration and procedures that arise. The data generated is increasing, so it requires the role of technology to support it. Technology is expected to help speed up and simplify the insurance claim administration process is one of them (Hudori, 2020).

By implementing digital technologies in claims processes, policy management, and customer service, insurance companies can substantially optimize their operational efficiency (Pattipeilopy et al., 2017). Previously, the claims process was often time-consuming and required complicated documentation. However, with digital insurance, claims can be submitted online, reducing claim settlement time and increasing customer satisfaction (Groll et al., 2022).

Artificial Intelligence (AI) technology has changed the paradigm of the insurance industry, especially in risk analysis and improving personal services (Maier et al., 2020). Through AI algorithms, insurance companies are now able to explore data more deeply to assess risk more accurately and adjust premiums appropriately (Sushant K, 2020). AI technology in the claims process also helps detect activities that contain elements of fraud or deception faster and earlier (Dutt, 2020; Maier et al., 2020).

Leveraging blockchain technology can help combat insurance claims fraud by leveraging smart contracts for claims tracking, automating antiquated paper-based processes, and improving data security (Ahmad Nur Azam Ahmad Ridzuan et al., 2024). The immutable nature of blockchain provides transparency and accountability in the claims process. Every transaction recorded on the blockchain can be tracked and verified by all interested parties,

minimizing the possibility of fraud and error (Ahmad Nur Azam Ahmad Ridzuan et al., 2024; Reddy & Premamayudu, 2019).

Blockchain protects insurance policies from forgery and alteration. Because data is stored in a distributed and encrypted manner, it is very difficult for unauthorized parties to access or manipulate the stored information (Orji & Ukwandu, 2024; Reddy & Premamayudu, 2019). Insurance companies are using blockchain to automate claims processes, thereby reducing settlement times and increasing operational efficiency.

In addition, Machine Learning (ML) techniques are also believed to be able to help carry out the analysis and prediction process of existing instruments in the insurance world. By using ML techniques, AI can learn patterns from previous claims and identify certain risk indicators (Boodhun & Jayabalan, 2018; Manathunga & Zhu, 2022). This allows insurance companies to take early precautions or perhaps even deny potentially detrimental claims.

Several studies have shown that ML algorithms such as Extreme Gradient Boosting (XGBoost) can be used to predict insurance claim values with better accuracy compared to generalized linear models (Amor, 2023; Permai & Herdianto, 2023) . This algorithm is capable of handling complex and non-linear patterns in claims data (Pesantez-Narvaez et al., 2019). Tree-based ML techniques such as Random Forest (RF) and XGBoost have been used to predict claim frequency (Gupta et al., 2019; Putra et al., 2021; Yeo et al., 2003). This technique is able to overcome the inability of linear models to capture non-linear patterns and can adapt to the data, so that it does not need to fulfill any assumption tests.

In addition, ML techniques can also be used to detect insurance claim fraud (Roy & George, 2017). By studying patterns of legitimate and fraudulent claims, ML models can predict whether a particular claim is likely to be fraudulent (Gupta et al., 2019; Hanafy & Ming, 2021). Thus, ML techniques can help insurance companies in the process of analyzing and predicting claims, thereby increasing operational efficiency and reducing the risk of financial losses (Severino & Peng, 2021).

The purpose of this study is to provide an accurate estimate of the cost of health insurance premiums for individuals. The initial cost calculation can help individuals evaluate more carefully and ensure they choose the most appropriate coverage value. In its implementation, this study focuses on identifying the main factors that affect the size of the premium value, as well as exploring the most effective machine learning algorithm in predicting the premium. In addition to building a predictive model, this study also examines how the accuracy and reliability of the model can be measured through relevant evaluation metrics. Furthermore, the evaluation is carried out by comparing the performance of three different regression models using five non-statistical evaluation metrics, to obtain a comprehensive understanding of the quality of the predictions produced.

METHOD

Testing and evaluating ML models requires software, libraries and programming languages: a). Anaconda Application, b). Python Programming Language, c). Jupyter Notebook, d). Numpy, e). Pandas, f). Seaborn, g). Scikit-learn, h). Matplotlib, i). Scipy, j). Statsmodel. And Hardware with the following specifications: a). Windows Operating System, b). x86 64-bit CPU (Intel/AMD architecture), c). 8 GB RAM, d). 5 GB free disk space. While the data set for the experiment is sourced from Kaggle.com.

Table 1. Insurance premium data set

Variables	Description	Type
Age	Policyholder age	Numeric
Diabetes	Diabetes status	Numeric
Blood Pressure Problems	Blood pressure status	Numeric
Any Transplants	Policyholder organ transplant history	Numeric
Any Chronic Diseases	Chronic disease status	Numeric
Height	Policyholder height	Numeric
Weight	Policyholder weight	Numeric
Known Allergies	Allergy status	Numeric
History of Cancer in Family	Family history of cancer	Numeric
Number of Major Surgeries	Number of major surgeries performed	Numeric

Premium Price	Insurance premium price	Numeric

Research Flow and Stages

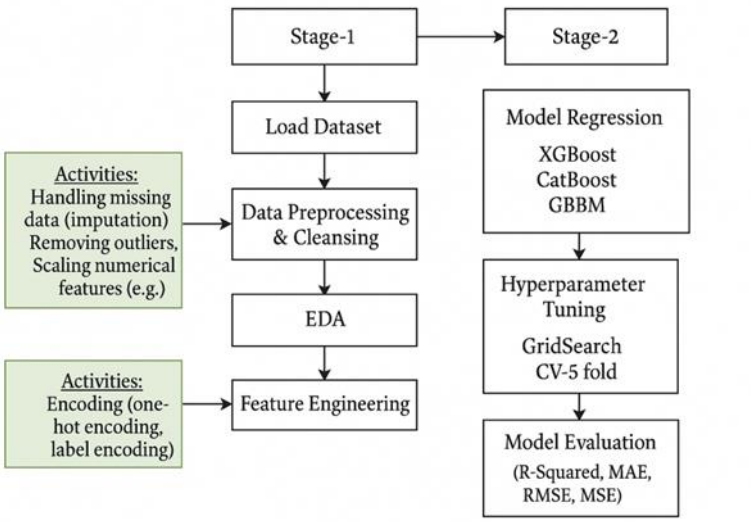

Figure 1. Proposed model structure and flow

In this study, the model design to predict health insurance premiums by applying ML based on customer data is shown in table (1). The study was conducted with the process flow in Figure 1. and the stages are as follows:

1. Data Collection
   The first step is to collect relevant data. This dataset, sourced from Kaggle.com, includes information on age, diabetes, blood pressure, transplants, chronic diseases, height, weight, allergies, family history of cancer, surgery and insurance premiums.
2. Preprocessing Data
   Collected data often requires cleaning and transformation before it is used for model training. This process involves addressing missing values, removing outliers, checking for duplication, and transforming inappropriate data formats so that the data is properly structured and ready to use.
3. Feature Engineering:
   At this stage, the process of identifying relevant features to predict premium value.
4. Selection of Regression Model
   The ML algorithms used are Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Categorical Boosting (CatBoost). The selection of this model is by looking at and considering the characteristics of the data and the objectives of the study.
5. Model Training
   The ML model is trained using historical data. This data is divided into two parts: training data (to train the model), and testing data (to measure the model performance), with a ratio of 80%:20% of the total data of 986 rows.
6. Model Evaluation:
   After training, model performance measurements were performed using evaluation metrics such as: Accuracy, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and R-Squared (R^2).
7. Model Optimization:
   If the model performance is not satisfactory, parameter tuning is performed or techniques such as oversampling or undersampling are used to overcome data imbalance. The hyperparameter tuning technique used is Grid Search.
8. Premium Value Prediction:

Once the model is optimized, the model can be used to predict insurance premium values based on new customer data.

ML Techniques in Insurance Industry

Several ML techniques that have been applied in conducting analysis and predictions related to insurance are as in Table 2.

Table 2. Research related to the application of machine learning

No	Author - Year	Title	ML Algorithm
1	(Billa & Nagpal, 2024)	*Medical Insurance Price Prediction Using Machine Learning*	Comparison of regression algorithms: LR, XGBoost, Lasso, RF, Ridge, CART, KNN, SVR, and Gradient Boosting, with evaluation matrices MAE, RMSE, R-squared and MSE.
2	(Orji & Ukwandu, 2024)	*Machine Learning for an Explainable Cost Prediction of Medical Insurance*	Comparison of regression algorithms: XGBoost, GBM and RF with a comparison of performance evaluation results based on MAE, RMSE, MAPE and R-squared values and the application of Explainable Artificial Intelligence (Xai), Shapley Additive exPlanations (SHAP) and Individual Conditional Expectation (ICE) methods.
3	(Vijayalakshmi et al., 2023)	*Implementation of Medical Insurance Price Prediction System using Regression Algorithms*	Comparison of regression algorithms: LR, DT, Lasso, Ridge, RF, ElasticNet, SVM, KNN and NN in predicting insurance costs and evaluation matrices: MSE, RMSE, MAE, MAPE, R-squared, Adjusted R-squared and Explained Variance Score (EVS) for evaluating the performance of each model.
4	(Kofi Immanuel Jones & Swati Sah, 2023)	*The Implementation of Machine Learning in The Insurance Industry with Big Data Analytics*	Using classification algorithms: AdaBoost, Naïve Bayes, KNN, and DT to predict insurance claims using evaluation matrices: Accuracy, Precision, Recall, F-measure, and AUC.
5	(Narayana et al., 2023)	*Medical Insurance Premium Prediction Using Regression Models*	Using regression algorithms: LR, Ridge, SVM, and RG for health insurance cost prediction with metric evaluation using RMSE, R-squared and Cross Validation score.
6	(Sahai et al., 2023)	*Insurance Risk Prediction Using Machine Learning*	Using algorithms by comparing tree-based classifier, RF, DT and XGBoost models for life insurance risk prediction with confusion matrix, AUC and Accuracy evaluation.
7	(Sahu et al., 2023)	*Health Insurance Cost Prediction by Using Machine Learning*	Using a comparison of regression algorithms: LR, SVM, RF, and Gradient Boosting with a comparison of performance evaluation results based on Accuracy values.
8	(Permai & Herdianto, 2023)	*Prediction of Health Insurance Claims Using Logistic Regression and XGBoost Methods*	Using classification algorithms: LR and XGBoost with a comparison of the confusion matrix evaluation results, for accuracy, precision and recall values.

9	(Hanafy & Ming, 2022)	*Classification of the Insureds Using Integrated Machine Learning Algorithms: A Comparative Study*	Using three different datasets and four binary classification algorithms (KNN, RF, CART, LR) with resampling method, feature selection, and feature discretization techniques as well as evaluation matrices: accuracy, sensitivity (Recall), specificity and AUC.
10	(Kulkarni et al., 2022)	*Medical Insurance Cost Prediction using Machine Learning*	Using regression algorithms: LR, DT and Gradient Boosting for health insurance cost prediction and matrix evaluation using R-squared.

Boosting Regressor Algorithm

The selection of the boosting regressor algorithm in this study is:

1. XGBoost is a widely used ML method for classification problems and can handle missing values without imputation pre-processing (Aydin & Ozturk, 2021). XGBoost is a very popular and powerful decision tree-based ML algorithm.
2. LightGBM is a boosting algorithm designed for efficiency and speed, especially on large datasets, as well as robustness against overfitting (Wang, 2020). LightGBM is a decision tree-based ML algorithm developed by Microsoft.
3. CatBoost is a boosting algorithm that is excellent at handling categorical data without the need for extensive preprocessing (So, 2024). CatBoost is a decision tree-based ML algorithm developed by Yandex.

Table 3. Comparison of boosting algorithms

Algorithm	Excess	Lack
XGBoost	High Performance: Has excellent speed and performance in handling large datasets (Bentéjac et al., 2021) . Regularization: Able to reduce overfitting by using regularization techniques (Daoud, 2019). Flexibility: Supports a wide range of loss functions and can be used for both classification and regression (So, 2024).	Parameter Settings: Requires more complex parameter tuning to get optimal results (So, 2024). Training Time: May be slower compared to LightGBM for very large datasets (Daoud, 2019) .
LightGBM	Speed: Very fast in training, especially for large datasets thanks to the histogram algorithm (Awan et al., 2022). Memory Efficiency: Enables more efficient use of memory (So, 2024). Support for Category Data: Able to handle category features directly without the need for encoding (So, 2024).	*Overfitting: Sometimes more prone to overfitting if not properly tuned* (Wang, 2020). Complexity: Requires a good understanding of parameter settings to use effectively (Bentéjac et al., 2021).
CatBoost	Category Data Handling: Automatically handle category features without the need for manual encoding (Hancock & Khoshgoftaar, 2020). *Robust against Overfitting: Has a mechanism to reduce overfitting* (So, 2024). Simple in Setup: easier to setup compared to XGBoost and LightGBM (Wang, 2020).	Speed: May be slower than LightGBM in training for very large datasets (Wang, 2020). Model Complexity: Sometimes produces more complex models, which can be difficult to interpret (Hancock & Khoshgoftaar, 2020).

Table 3. shows a comparison between the algorithms used based on the results of previous researchers in terms of their advantages and disadvantages.

Grid Search

Hyperparameters are parameters that are determined before the ML model training process begins. Hyperparameters are different from ML model parameters that can be found through the training process. Hyperparameters affect how ML models learn data and how they make decisions. Hyperparameter tuning is an important part of the ML model development process to ensure that the ML model has the right hyperparameters so that it can provide good results and performance as expected (Bischl et al., 2023; Putatunda & Rama, 2019). Grid Search is a commonly used technique and chosen in this study, to perform parameter tuning in ML models.

Evaluation Metrics

Evaluation of model performance is one of the important aspects in building a predictive model. Choosing the right evaluation metrics for a regression model is very important to accurately assess model performance with several considerations and evaluation metrics used in this study:

1. MAE

   MAE is an evaluation method used to calculate the average of the absolute difference between the predicted value and the actual value. The smaller the MAE value, the better the quality of the model (Quan & Valdez, 2018). MAE is more intuitive and gives equal weight to all errors.

   The advantage of MAE is that it is easy to understand and produces absolute values so it does not depend on low or high predictions. However, MAE does not take into account the weight of large and small prediction errors, so predictions that are far from the true value and predictions that are close to the true value are considered equally important.

   MAE formula:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

   Where:
   n is the number of samples in the data
   y_i is the actual value
   ŷ_i is the predicted value

2. MSE

   MSE is an evaluation method by calculating the average of the squared differences between the predicted and actual values. In other words, MSE calculates the average squared error in the prediction. The smaller the MSE value, the better the quality of the model (Quan & Valdez, 2018). MSE calculates the average of the squared differences between the predicted and actual values in the dataset by measuring the variance of the residuals.
   Formula:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

   Where:
   n is the number of samples in the data
   y_i is the actual value
   ŷ_i is the predicted value

   The advantage of MSE is that it gives greater weight to predictions that are far from the actual value and is therefore more sensitive to large prediction errors. However, MSE produces squared values that are difficult to understand and depend on low or high predictions. MSE is more sensitive to outliers than MAE due to squared errors.

3. RMSE

   RMSE is the derivative of MSE or the square root of MSE. RMSE calculates the average of the squared differences between the predicted and actual values and then takes the square root. The smaller the RMSE value, the better the quality of the model (Pesantez-Narvaez et al., 2019). RMSE measures the standard deviation of the residuals. RMSE provides a more intuitive interpretation because it is on the same scale as the target and is useful when comparing models across datasets. RMSE is more sensitive to outliers.

Formula:

$$\text{RMSE} = \left(\frac{\Sigma(y_i - \hat{y}_i)}{\text{n}}\right)^{1/2}$$

Where:
n is the number of samples in the data
y_i is the actual value
ŷ_i is the predicted value

The advantage of RMSE is that it produces absolute values like MAE and takes into account weights like MSE, so it is the most commonly used evaluation method. However, RMSE is more difficult to understand than MAE and requires additional calculations because it involves square roots (Pesantez-Narvaez et al., 2019).

4. R-squared (R^2)

Definition: The proportion of variance in the target variable that can be explained by the model. R-squared: Measures the proportion of variance in the data that can be explained by the model. R-squared is the proportion of variance in the dependent variable that is explained by the linear regression model. The R-squared value is always between 0 and 1. The higher the value, the better the regression model explains the variation in the data (Quan & Valdez, 2018).

Formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Where:
n is the number of samples in the data
y_i is the actual value
ŷ_i is the predicted value

5. Mean Absolute Percentage Error (MAPE)

MAPE is an evaluation method used to calculate the average of the percentage difference between the predicted value and the actual value. In other words, MAPE calculates how much the average error in prediction is as a percentage of the actual value. The smaller the MAPE value, the better the quality of the model (Quan & Valdez, 2018). MAPE is useful because it provides information about the relative error in percentage form, but it is important to note its limitations.

The advantage of MAPE is that it provides relative values, making it useful in situations where predictions depend on the percentage error. However, MAPE cannot be used for data that has zero values or data that has large variations in values (Quan & Valdez, 2018).

$$\text{MAPE} = \left(\frac{1}{\text{N}}\right) * \text{E|i} = 1|^{n} \left|\frac{y_i - y_i}{y_i}\right| * 100\%$$

Where:
n is the number of samples in the data
y_i is the actual value
ŷ_i is the predicted value

MAPE value limits are used as a guide (Novita et al., 2022):

a. MAPE values of 10% or less indicate that the average difference between the forecasted and actual values is very small. Models with MAPE in this range are considered very accurate.
b. MAPE values between 10% and 20% indicate good forecasting results. Models with MAPE in this range provide adequate predictions.
c. MAPE values between 20% and 50% are considered decent or good enough. Despite errors, the model still provides acceptable results.

d. MAPE values above 50% indicate that the model has significant errors in prediction. Models with high MAPE need to be improved or further evaluated.

RESULTS AND DISCUSSION

The data pre-processing stage is carried out before the model formation is carried out using the selected algorithm. This stage is very important, because it will affect the training process and the final results. Table 4. shows the statistical data from the dataset used.

Table 4. Statistical data

	count	mean	std	min	25%	50%	75%	max
Age	986.0	41.745436	13.963371	18.0	30.0	42.0	53.0	66.0
Diabetes	986.0	0.419878	0.493789	0.0	0.0	0.0	1.0	1.0
BloodPressureProblems	986.0	0.468560	0.499264	0.0	0.0	0.0	1.0	1.0
AnyTransplants	986.0	0.055781	0.229615	0.0	0.0	0.0	0.0	1.0
AnyChronicDiseases	986.0	0.180527	0.384821	0.0	0.0	0.0	0.0	1.0
Height	986.0	168.182556	10.098155	145.0	161.0	168.0	176.0	188.0
Weight	986.0	76.950304	14.265096	51.0	67.0	75.0	87.0	132.0
KnownAllergies	986.0	0.215010	0.411038	0.0	0.0	0.0	0.0	1.0
HistoryOfCancerInFamily	986.0	0.117647	0.322353	0.0	0.0	0.0	0.0	1.0
NumberOfMajorSurgeries	986.0	0.667343	0.749205	0.0	0.0	1.0	1.0	3.0
PremiumPrice	986.0	24336.713996	6248.184382	15000.0	21000.0	23000.0	28000.0	40000.0

The dataset consists of 11 variables and each variable has a contribution to predict the value of health insurance premiums where the dependent variable (target) is Premium Price and the others are independent variables (predictors). The experiment uses regression analysis, a predictive method that explores the relationship between targets and predictors, and finds causal effect relationships between these variables.

1. The minimum age of the insurance participants is 18 years old, and the maximum is 66 years old, and the average age of all insurance participants is 41.7 years old.
2. The largest insurance policy value is 40,000, the lowest is 15,000, and the average insurance policy value is 24,337.

Statistical analysis was conducted to determine the relationship between Premium Price and the variables Age, Diabetes, Blood Pressure Problems, Any Transplants, Any Chronic Diseases, Height, Weight, Known Allergies, History of Cancer in Family, and Number of Major Surgeries using the variable importance technique as in Figure 2. Where the largest value is the one that has the greatest effect.

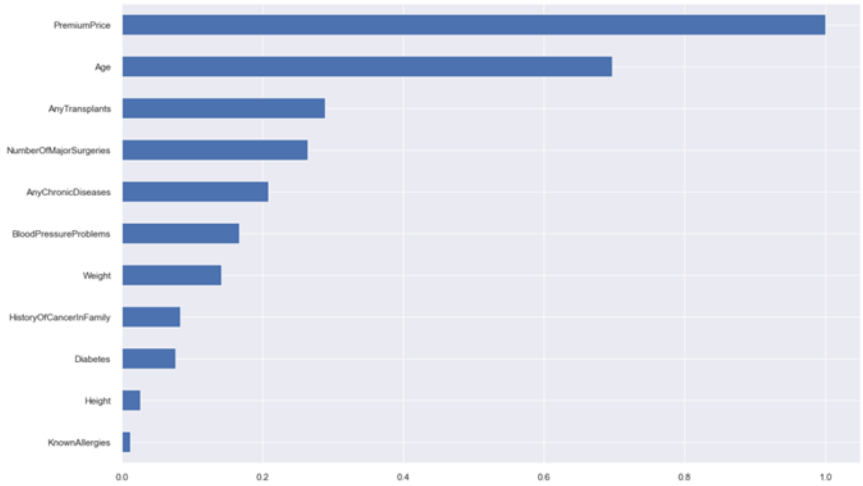

Figure 2. Variable importance

In addition, to detect multicollinearity in this study, the Correlation Coefficient, Variance Inflation Factor (VIF), and Scatterplot are used as graphical methods that indicate a linear relationship between pairs of independent variables. VIF is used to measure how much the estimated regression coefficient variance increases if the independent variables are correlated (Shrestha, 2020).

In the previous step, the dataset was cleaned so that the model can be trained and visualized. In this step, the data is visualized to obtain information, namely a line plot diagram, as illustrated in Figure 3. The line plot diagram is used to observe the pattern or trend of the relationship between Age and Premium Price, where as age increases, the insurance premium price also increases, and there is an anomaly at the age limit of 30 years, the insurance premium price increases significantly (anomaly) which is possibly caused by external factors or insurance company policies.

Figure 3. Relationship between Agent and Premium Price.

Correlation between variables is also shown using heatmap diagram in Figure 4. It can be seen that the correlation between PremiumPrice target with Age, AnyTransplants, NumberofMajorSurgery, AnyChronicDesease, BloodPressureProblems, Weight variables has correlation compared to the variables of HistoryofCancerinFamily, Height, Diabetes, and AnyAllergics. Correlation matrix is plotted to see positive and negative relationships between several factors. After observing the correlation matrix in Figure 4, it can be concluded that PremiumPrice value is positively related to Age, AnyTransplants and NumberofMajorSurgery.

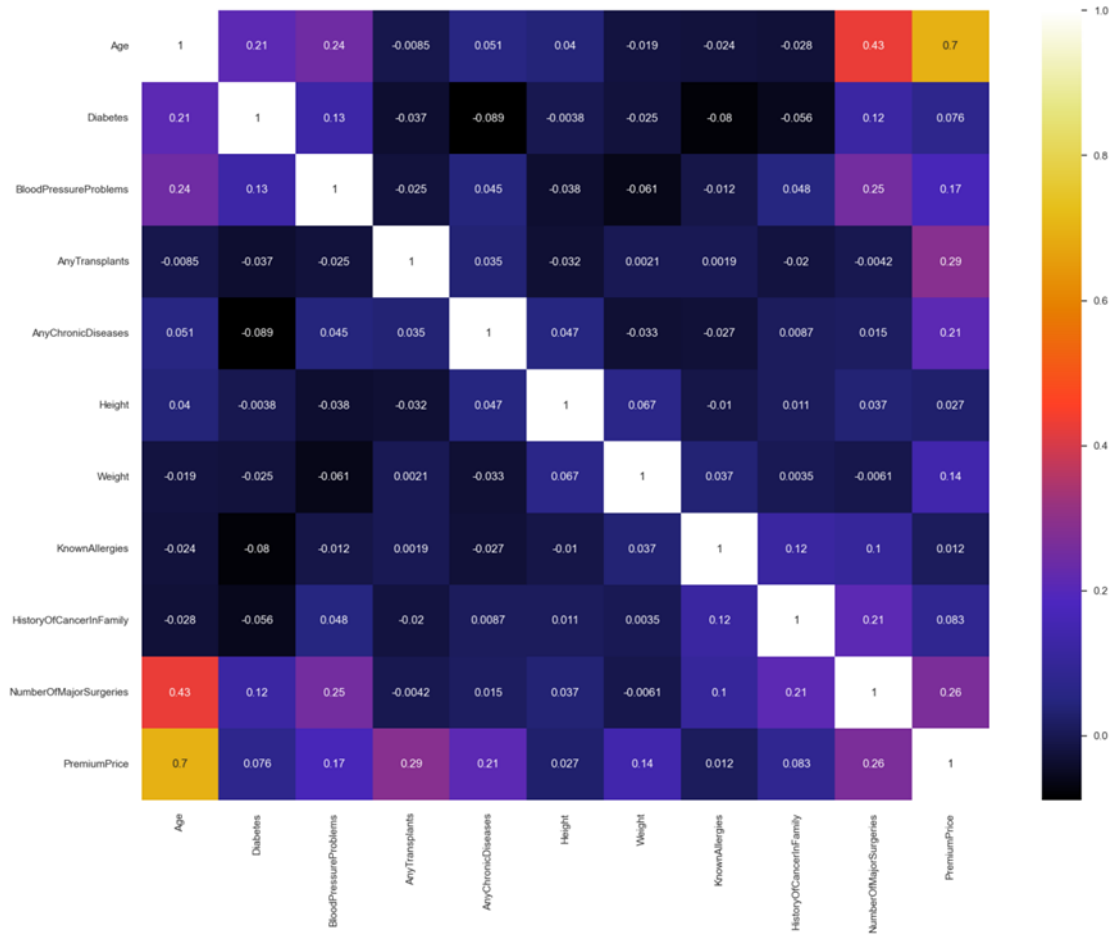

Figure 4. Correlation diagram between variables

The outlier identification process is carried out in addition to using visualization with boxplots and using descriptive statistics Z-Score, namely calculating the Z-score for each value, where values with a Z-score greater than 3 or less than -3 can be considered outliers.

IQR (Interquartile Range): Outliers can be determined by calculating Q1 (first quartile) and Q3 (third quartile): IQR = Q3 − Q1

Values that are below Q1−1.5 × IQR or above Q3+1.5 × IQR are considered outliers.

Modeling with Grid Search

Table 5. Grid search hyperparameters

Model	Hyperparameter	Parameter	Best Parameter
XGboost	*n_estimators*	[40, 50, 60, 70, 80]	60
	learning_rate	[0.06, 0.07, 0.08, 0.09, 1.0]	0.09
	subsample	[0.3, 0.4, 0.5, 0.6, 0.7]	0.4
	max_depth	[7, 8, 9, 10, 11]	11
	gamma	[0, 1]	0
LightGBM	*n_estimators*	[5, 7, 9, 11, 13]	3
	learning_rate	[0.3, 0.5, 0.7, 0.9, 1.1]	0.3
	max_depth	[3, 5, 7, 9, 11]	9
	num_leaves	[3, 5, 7, 9, 11]	11
	num_interations	[5, 7, 9, 11, 13]	13

Catboost	*learning_rate*	[0.07, 0.08, 0.09, 0.1, 0.2]	0.09
	depth	[1, 3, 5, 7, 9]	7
	l2_leaf_reg	[0.09, 0.1, 0.2, 0.3, 0.4]	0.4
	iterations	[130, 140, 150, 160, 170]	170

XGBoost Performance

The parameters used are as per Table 5., with a focus on optimizing n_estimators, max_depth and learning_rate by increasing them to speed up the model without reducing performance so that it becomes faster without losing performance. To overcome higher memory usage, researchers try to optimize parameters such as max_depth to control tree depth and reduce memory usage. However, the training time of the XGBoost model can be longer compared to LightGBM and CatBoost as shown in Table 8. Although tuning has been done on parameters such as n_estimators and learning_rate.

LightGBM Performance

The parameters used are as per Table 5., the same as the XGboost model, namely focusing on optimizing n_estimators, max_depth and learning_rate by increasing them to speed up the model without reducing performance so that it becomes faster without losing performance. To overcome the lower sensitivity compared to XGBoost, researchers try to tune parameters such as num_leaves and max_depth. And the result is that the training time of the LightGBM model can be faster compared to XGboost and CatBoost as shown in Table 8.

CatBoost Performance

The parameters used are according to Table 5., which focuses on optimizing iterations, depth and learning_rate by increasing them to speed up the model without reducing performance so that it becomes faster without losing performance. CatBoost shows similar performance to XGBoost in the training stage in terms of accuracy of 97.604. However, CatBoost requires a longer training time of 1574.055775 sec compared to LightGBM - 799.028072 sec and faster than XGBoost - 2548.521826 sec as shown in Table 8. Although CatBoost has advantages in handling categorical features, in this experiment, the prediction accuracy of the CatBoost model is lower than XGBoost and LightGBM as shown in tables 6 and 7.

Model Performance Evaluation.

Model performance evaluation is one of the important aspects in building a predictive model. The performance evaluation of the model used can be seen in tables 6, 7 and 8.

Table 6. Results of model training and testing

Model	Train - R2	Test - R2	CV - R2
XGBoost	97.227	88.997	79.309
LightGBM	82.601	88.565	78.027
CatBoost	97.604	83.198	76.125

Table 7. Comparison of model performance

Model	MAE	RMSE	MSE	MAPE (%)
XGBoost	1145.411	2166.113	4692047.59	5.035
LightGBM	1393.345	2208.225	4876258.11	5.732
CatBoost	1523.578	2676.678	7164603.75	6.371

Table 8. Processing time and memory usage

Model	Elapsed time (Sec)	Memory used (MB)
XGBoost	2548.521826	-22.589844
LightGBM	799.028072	7.722656

CatBoost	1574.055775	-292.058594

The second approach uses SHAP values to assess the average marginal contribution of each input to the model. SHAP (Shapley Additive Explanations) is a game theory-based method for explaining the results of machine learning models. Figures 5a, 6a and 7a, calculate the average contribution of each feature to the model prediction (PremiumPrice). This value indicates how much influence a particular feature has on the prediction result, where the 3 (three) models produce the order of features with the greatest impact consistently in the same order. Positive or negative signs in Figures 5a, 6a, and 7a mean that a positive value means the feature improves the prediction, while a negative value means the feature decreases the prediction.

The mean SHAP chart values in Figures 5b, 6b and 7b, show the average contribution of each feature to the model prediction (PremiumPrice). These results help understand which features have the greatest impact overall. Features with high mean SHAP values, namely Age, Weight and AnyTransplants, have a significant impact on the model with the provision that the higher the mean SHAP value, the greater the influence of the feature on the prediction.

By comparing between features, if a feature has a mean SHAP value that is much larger than other features, then that feature becomes the dominant factor in the model decision (Age).

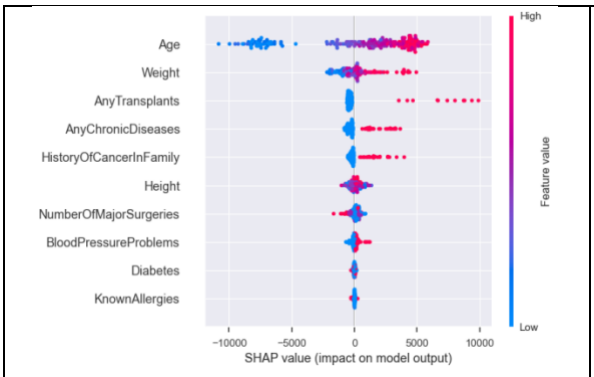

Figure 5a. SHAP model of LightGBM

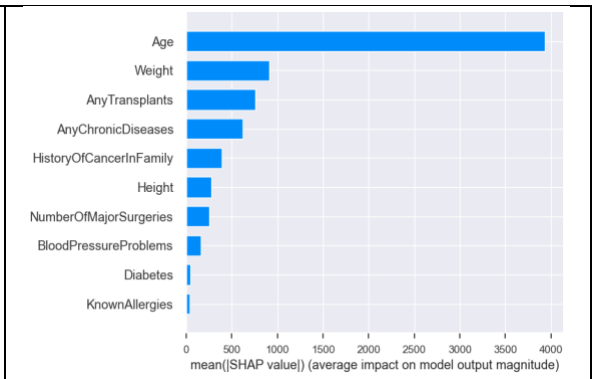

Figure 5b. Mean SHAP LightGBM model

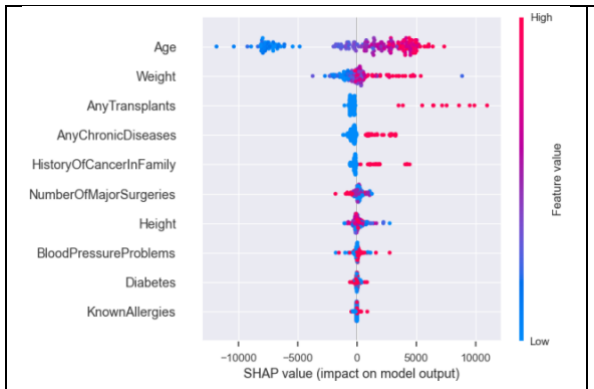

Figure 6a. SHAP model of XGboost

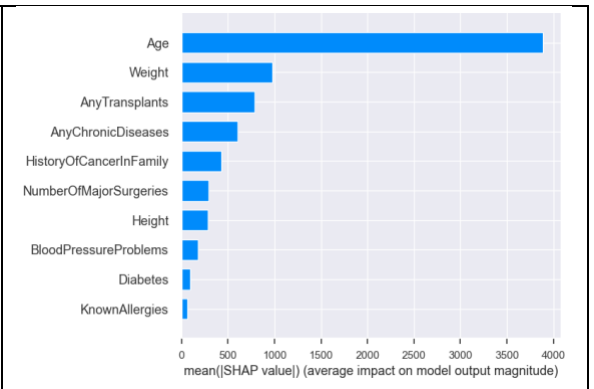

Figure 6b. Mean SHAP model XGBoost

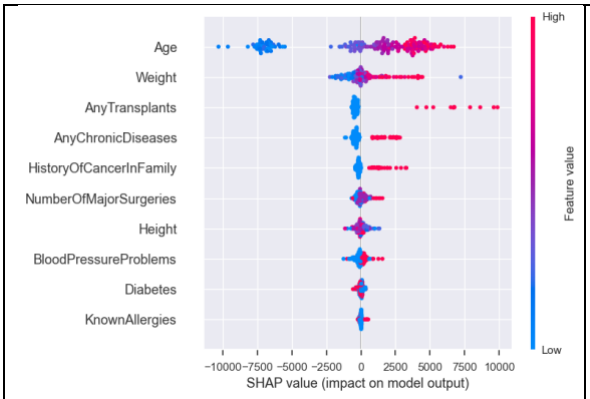

Figure 7a. SHAP model of Catboost

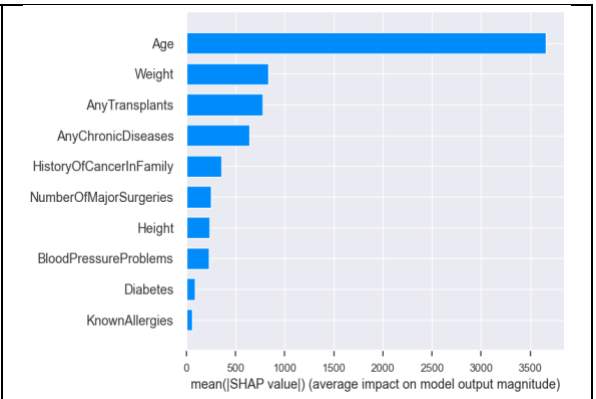

Figure 7b. Mean SHAP of CatBoost model

CONCLUSION

The use of ML techniques in the insurance industry has significant benefits. The application of ML can provide accurate estimates of the value of health insurance premiums for individuals. Although the predictions may not always follow a consistent pattern, they help in making informed decisions regarding the selection of appropriate health insurance premiums. In addition, this study can provide insights to identify factors that influence the value of health insurance premiums. Factors such as age, medical history, and lifestyle can affect the amount of insurance premiums.

The results of the experiment show that the accuracy of health insurance premium predictions is greatly influenced by several important factors. First, data quality is a crucial aspect—models trained with clean, complete, and relevant data tend to produce more accurate predictions. In addition, feature selection also plays a significant role, where variables such as age (Age), weight (Weight), and history of organ transplants (AnyTransplants) are proven to affect the amount of premiums. Model accuracy is also determined by the configuration of parameters in each algorithm; parameters such as n_estimators, max_depth, and learning_rate in XGBoost and LightGBM, as well as iterations and depth in CatBoost, are shown to affect predictive performance. Data processing techniques such as normalization and standardization also strengthen model accuracy. Based on the evaluation results, the XGBoost model showed the best performance with the highest R-squared value, as well as lower MAPE, RMSE, MSE, and MAE values compared to LightGBM and CatBoost. This indicates that XGBoost is the most accurate model for the case of health insurance premium prediction in this study.

REFERENCES

Ahmad Nur Azam Ahmad Ridzuan, Aina Zafirah Azman, Fatin Alya Marzuki, Wan Shazmien Danieal Mohamed Faudzi, Siti Hajar Abd Aziz, & Norida Abu Bakar. (2024). Health Insurance Premium Pricing Using Machine Learning Methods. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *41*(1), 134–141. https://doi.org/10.37934/araset.41.1.134141

Amor, E. N. (2023). Analisis Klasifikasi Dengan Metode Random Forest, LogitBoost, dan XGBoost untuk Memprediksi Status Klaim Asuransi. *Repository UGM*.

Awan, M. J., Mohd Rahim, M. S., Salim, N., Rehman, A., & Nobanee, H. (2022). Machine Learning-Based Performance Comparison to Diagnose Anterior Cruciate Ligament Tears. *Journal of Healthcare Engineering*, *2022*(Mcl), 1–18. https://doi.org/10.1155/2022/2550120

Aydin, Z. E., & Ozturk, Z. K. (2021). XGBoost Feature Selection on Chronic Kidney Disease Diagnosis. *International Conference on Data Science and Applications*, *June*, 7.

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A Comparative Analysis of Gradient Boosting Algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Billa, M. M., & Nagpal, T. (2024). Medical Insurance Price Prediction Using Machine Learning. *Journal of Electrical Systems*, *20*(7s), 2270–2279. https://doi.org/10.52783/jes.3962

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A., Deng, D., & Lindauer, M. (2023). Hyperparameter Optimization: Foundations, Algorithms, Best Practices, and Open Challenges. *WIREs Data Mining and Knowledge Discovery*, *13*(2).

https://doi.org/10.1002/widm.1484

Boodhun, N., & Jayabalan, M. (2018). Risk Prediction in Life Insurance Industry Using Supervised Learning Algorithms. *Complex & Intelligent Systems*, *4*(2), 145–154. https://doi.org/10.1007/s40747-018-0072-1

Daoud, E. Al. (2019). Comparison between XGBoost, Light GBM and CatBoost Using a Home Credit Dataset. *World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering*, *13*(1), 6–10. https://doi.org/https://doi.org/10.5281/zenodo.3607805

Dutt, R. (2020). The Impact of Artificial Intelligence on Healthcare Insurances. In *Artificial Intelligence in Healthcare* (pp. 271–293). Elsevier. https://doi.org/10.1016/B978-0-12-818438-7.00011-3

Groll, A., Wasserfuhr, C., & Zeldin, L. (2022). Churn Modeling of Life Insurance Policies via Statistical and Machine Learning Methods Analysis of Important Features. *ArXiv*, *1*(2202), 1–35.

Gupta, R. Y., Sai Mudigonda, S., Kandala, P. K., & Baruah, P. K. (2019). Implementation of a Predictive Model for Fraud Detection in Motor Insurance using Gradient Boosting Method and Validation with Actuarial Models. *2019 IEEE International Conference on Clean Energy and Energy Efficient Electronics Circuit for Sustainable Development (INCCES)*, 1–6. https://doi.org/10.1109/INCCES47820.2019.9167733

Hanafy, M., & Ming, R. (2021). Using Machine Learning Models to Compare Various Resampling Methods in Predicting Insurance Fraud. *Journal of Theoretical and Applied Information Technology*, *99*(12), 2819–2833.

Hanafy, M., & Ming, R. (2022). Classification of the Insureds Using Integrated Machine Learning Algorithms: A Comparative Study. *Applied Artificial Intelligence*, *36*(1). https://doi.org/10.1080/08839514.2021.2020489

Hancock, J., & Khoshgoftaar, T. M. (2020). Medicare Fraud Detection using CatBoost. *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, 97–103. https://doi.org/10.1109/IRI49571.2020.00022

Hudori, H. (2020). Resampling Neural Network Untuk Penanganan Class Imbalance Pada Prediksi Klaim Asuransi. *Teknois : Jurnal Ilmiah Teknologi Informasi Dan Sains*, *10*(1), 57–64. https://doi.org/10.36350/jbs.v10i1.78

Kofi Immanuel Jones, & Swati Sah. (2023). The Implementation of Machine Learning in The Insurance Industry with Big Data Analytics. *International Journal of Data Informatics and Intelligent Computing*, *2*(2), 21–38. https://doi.org/10.59461/ijdiic.v2i2.47

Kulkarni, M., Meshram, D. D., Patil, B., More, R., Sharma, M., & Patange, P. (2022). Medical Insurance Cost Prediction using Machine Learning. *International Journal for Research in Applied Science and Engineering Technology*, *10*(12), 449–456. https://doi.org/10.22214/ijraset.2022.47923

Maier, M., Carlotto, H., Saperstein, S., Sanchez, F., Balogun, S., & Merritt, S. (2020). Improving the Accuracy and Transparency of Underwriting with Artificial Intelligence to Transform the Life-Insurance Industry. *AI Magazine*, *41*(3), 78–93. https://doi.org/10.1609/aimag.v41i3.5320

Manathunga, V., & Zhu, D. (2022). Unearned Premium Risk and Machine Learning Techniques. *Frontiers in Applied Mathematics and Statistics*, *8*, 16. https://doi.org/10.3389/fams.2022.1056529

Narayana, K. L., Yogesh, & Kowshik, P. (2023). Medical Insurance Premium Prediction Using Regression Models. *International Journal for Research Trends and Innovation*, *8*(4), 1512–1517. https://doi.org/https://www.ijrti.org/viewpaperforall?paper=IJRTI2304248

Novita, R., Yani, I., & Ali, G. (2022). Sistem Prediksi untuk Penentuan Jumlah Pemesanan Obat Menggunakan Regresi Linier: Prediction System for Determine The Number of Drug Orders using Linear Regression. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *2*(1), 62–70.

Nugraha, A. C., & Irawan, M. I. (2023). Komparasi Deteksi Kecurangan pada Data Klaim Asuransi Pelayanan Kesehatan Menggunakan Metode Support Vector Machine (SVM) dan Extreme Gradient Boosting (XGBoost). *Jurnal Sains Dan Seni ITS*, *12*(1), 7. https://doi.org/10.12962/j23373520.v12i1.107032

Orji, U., & Ukwandu, E. (2024). Machine Learning for an Explainable Cost Prediction of Medical Insurance. *Machine Learning with Applications*, *15*(July 2023), 100516. https://doi.org/10.1016/j.mlwa.2023.100516

Pattipeilopy, W. F., Wibowo, A., & Utari, D. R. (2017). Pemodelan Dan Prototipe Sistem Informasi Untuk Prediksi Pembaharuan Polis Asuransi Mobil Menggunakan Algoritma C.45. *Prosiding SNATIF*, *4*(1), 791–799.

Permai, S. D., & Herdianto, K. (2023). Prediction of Health Insurance Claims Using Logistic Regression and XGBoost Methods. *Procedia Computer Science*, *227*, 1012–1019. https://doi.org/10.1016/j.procs.2023.10.610

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression. *Risks*, *7*(2), 70. https://doi.org/10.3390/risks7020070

Putatunda, S., & Rama, K. (2019). A Modified Bayesian Optimization Based Hyper-Parameter Tuning Approach for Extreme Gradient Boosting. *2019 Fifteenth International Conference on Information Processing (ICINPRO)*, 1–6. https://doi.org/10.1109/ICInPro47689.2019.9092025

Putra, T. A. J., Lesmana, D. C., & Purnaba, I. G. P. (2021). Penghitungan Premi Asuransi Kendaraan Bermotor Menggunakan Generalized Linear Models dengan Distribusi Tweedie. *Jambura Journal of Mathematics*, *3*(2), 115–127. https://doi.org/10.34312/jjom.v3i2.10136

Quan, Z., & Valdez, E. A. (2018). Predictive Analytics of Insurance Claims using Multivariate Decision Trees. *Dependence Modeling*, *6*(1), 377–407. https://doi.org/10.1515/demo-2018-0022

Reddy, T., & Premamayudu, B. (2019). Vehicle Insurance Model Using Telematics System with Improved Machine Learning Techniques: A Survey. *Ingénierie Des Systèmes d Information*, *24*(5), 507–512. https://doi.org/10.18280/isi.240507

Roy, R., & George, K. T. (2017). Detecting Insurance Claims Fraud using Machine Learning Techniques. *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT)*, 1–6. https://doi.org/10.1109/ICCPCT.2017.8074258

Sahai, R., Al-Ataby, A., Assi, S., Jayabalan, M., Liatsis, P., Loy, C. K., Al-Hamid, A., Al-Sudani, S., Alamran, M., & Kolivand, H. (2023). Insurance Risk Prediction Using Machine Learning. In *Lecture Notes on Data Engineering and Communications Technologies* (Vol. 165, Issue June, pp. 419–433). https://doi.org/10.1007/978-981-99-0741-0_30

Sahu, A., Sharma, G., Kaushik, J., Agarwal, K., & Singh, D. (2023). Health Insurance Cost Prediction by Using Machine Learning. *SSRN Electronic Journal*, 1381–1384. https://doi.org/10.2139/ssrn.4366801

Severino, M. K., & Peng, Y. (2021). Machine Learning Algorithms for Fraud Prediction in Property Insurance: Empirical Evidence Using Real-world Microdata. *Machine Learning with Applications*, *5*(June), 100074. https://doi.org/10.1016/j.mlwa.2021.100074

Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, *8*(2), 39–42.

So, B. (2024). Enhanced Gradient Boosting for Zero-inflated Insurance Claims and Comparative Analysis of CatBoost , XGBoost , and LightGBM. *Scandinavian Actuarial Journal*, *June*, 1–23. https://doi.org/10.1080/03461238.2024.2365390

Sushant K, S. (2020). A Commentary on the Application of Artificial Intelligence in the Insurance Industry. *Trends in Artificial Intelligence*, *4*(1), 75–79. https://doi.org/10.36959/643/305

Tumbel, N. J., & Ananto, N. (2024). Identification on Financial Fraud by Companies Using the Logistic Regression Model. *YUME : Journal of Management*, *7*(2), 167–179. https://doi.org/https://doi.org/10.37531/yum.v7i2.6611

Vijayalakshmi, V., Selvakumar, A., & Panimalar, K. (2023). Implementation of Medical Insurance Price Prediction System using Regression Algorithms. *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, *January*, 1529–1534. https://doi.org/10.1109/ICSSIT55814.2023.10060926

Wang, H. D. (2020). Research on the Features of Car Insurance Data Based on Machine Learning. *Procedia Computer Science*, *166*, 582–587. https://doi.org/10.1016/j.procs.2020.02.016

Yeo, A. C., Smith, K. A., Willis, R. J., & Brooks, M. (2003). A Comparison of Soft Computing and Traditional Approaches for Risk Classification and Claim Cost Prediction in the Automobile Insurance Industry. In *Springer* (pp. 249–261). https://doi.org/10.1007/978-3-540-36216-6_17