IAJ
Indonesian Actuarial Journal
Persatuan Aktuaris Indonesia (PAI)

Vol. 01, No. 01, July (2025)
e-ISSN: xxxx-xxxx
https://ojs.aktuaris.or.id/iaj/index

Analysis of the Health Social Security Administration (*BPJS Kesehatan*) Claim Amount using Random Forest Regression

Indah Gumala Andirasdini [1*], Desta Saputra [2], Muklas Rivai [3], Septia Eka Marsha Putra [4]

[1,2,3]*Actuarial Science, Science Faculty, Institut Teknologi Sumatera, Indonesia*
[4]*Physics Engineering, Industrial Technology Faculty, Institut Teknologi Sumatera, Indonesia*

*Corresponding email: indah.andirasdini@at.itera.ac.id

Abstract

Claims paid by hospitals need to be identified to verify the accuracy of health services, maintain service quality, and optimize services provided to the Health Social Security Administration (BPJS Kesehatan) participants. This aligns with the third goal of the Sustainable Development Goals (SDGs), which is to ensure healthy lives and promote well-being for all ages, particularly in the context of universal health coverage. The difference in tariffs set by BPJS Kesehatan (INA-CBGs) compared to the amount paid by hospitals has led to a problem that can harm health facilities, such as delayed claim payments. This study aims to analyze the amount of claims paid by a regional hospital to BPJS Kesehatan participants using machine learning with the Random Forest Regression method. Based on this modeling, it was found that the severity of patients, length of stay, and type of illness are the most significant factors in determining the amount of claims. This study has an accuracy value of 81.89%, an adjusted R-square value of 80.4%, and a Mean Absolute Percentage Error (MAPE) of 18.11% in estimating the amount of claims.

Article History:
Received: 12 January 2025
First Revised: 29 January 2025
Accepted: 28 February 2025
Published: 31 July 2025

Keywords:
BPJS Kesehatan , machine learning, random forest regression, claim amount,

1. INTRODUCTION

As a national health insurance provider in Indonesia, effective claims analysis is critical for the Health Social Security Administration (*BPJS Kesehatan*), particularly with the ongoing rise in claim volume. The increase in claims correlates with healthcare service utilization, reaching 1.4 million visits per day on December 31, 2022. This increase parallels the expansion of *BPJS Kesehatan*'s partnerships with healthcare facilities, which include 23,730 collaborations with Primary Healthcare Facilities (FKTP) and 2,963 partnerships with Advanced Referral Healthcare Facilities (FKRTL).

With the rising number of claims, traditional methods of claim verification often prove inadequate for efficiently handling large data sets. This situation requires the development of novel, faster, and more accurate claims analysis methods. The payment system employed by *BPJS Kesehatan*, which utilizes INA-CBGs as the basis for healthcare service reimbursement, frequently generates controversy among collaborating healthcare facilities. Tariff determination discrepancies are a key source of problem, ultimately impacting the profitability of these facilities. Additionally, delays in claim payments by *BPJS Kesehatan* are a common complaint from healthcare providers.

Research conducted by Sofi Romando *et al*. (2023) indicates that RSUD Kalideres incurred losses due to differences in claim tariffs, with an average INA-CBGs tariff of IDR 4,328,683, while the average hospital cost was IDR 7,938,303 [3]. Furthermore, a study by Devi *et al*. (2020) found that delays in *BPJS Kesehatan* claim payments were primarily due to the tardiness in collecting complete administrative documents and delayed claim coding in the casemix unit, causing claim payments to be postponed by approximately 7 to 45 days [4].

The presence of “disputes” and “pending claims” leads to delays in claim payments, affecting both outpatient and inpatient services. Therefore, INA-CBGs tariffs are considered to cause financial deficits for hospital because claim reimbursement rates are lower than the actual costs incurred. To solve this problem, a study was conducted to analyze the claim amounts submitted by hospital to *BPJS Kesehatan*, considering the influencing factors, using a machine learning approach, specifically Random Forest regression

Random Forest regression is an effective method for analyzing BPJS Health claim amounts. This machine learning method combines multiple decision trees to enhance accuracy and reduce the risk of overfitting. Random Forest works by constructing several decision trees from random subsets of the training data and then combining the results to produce more stable and accurate predictions. The advantages of this method include its ability to handle large and complex datasets, as well as its capacity to manage missing data [1-2].

Using Random Forest Regression for *BPJS Kesehatan* claim analysis aims not only to improve the efficiency of claim amount determination but also to identify the factors influencing claim amounts. These factors can include the location of healthcare facilities, the age of participants, initial diagnoses, and the type of healthcare facilities. By understanding these factors, *BPJS Kesehatan* can take more appropriate steps in managing claims and improving overall healthcare services [2].

This research investigates *BPJS Kesehatan* claim amounts using the Random Forest Regression method. The study seeks to provide deeper insights into claim patterns, taking into account factors that influence claim amounts such as Length-of-Stay, age, gender, patient discharge method, INA-CBGs code, class, and severity level. The study leverages claim data for inpatients insured by *BPJS Kesehatan* at a regional public hospital (RSUD) in Lampung Province, collected from January to September 2022. The expected outcomes are expected to help *BPJS Kesehatan* more accurately predict claim risks and support efforts to enhance the effectiveness of healthcare services provided by both *BPJS Kesehatan* and hospital [1][3].

2. METHODS

The data used in this study is secondary data, which is derived from inpatient claims submitted by *BPJS Kesehatan* participants for hospital costs at a District General Hospital (RSUD) in Lampung Province from January 2022 to September 2022. The research dataset includes 1233 patients, with the claim amount as the dependent variable. The independent variables consist of BPJS inpatient class, gender, age, INA-CBGs code, severity of the patient's illness, discharge method, and length of stay. Each variable involved is detailed in Table 1.

Table 1. Research Variables

Variables	Description	Frequency (n)	Percentage (%)
Claim amount	The amount of money disbursed by *BPJS Kesehatan* to patients for hospital claims (in IDR).	-	-
BPJS Class	0 = Class-1	398	32.28%
	1 = Class-2	410	33.25%
	2 = Class-3	425	34.47%
Gender	0 = Men	643	52.15%
	1 = Woman	590	47.85%
Age	Age of patients (from 1-98 years old)	-	-
INA-CBGs Codes	The case-based disease classification system used by BPJS to classify and calculate the cost of patient care consists of:		
	0 = Endokrin and Metabolic system	212	17.19%
	1 = Hematopoietik and Immune system	132	10.71%
	2 = Digestive and Hepatobiliary system	178	14.44%
	3 = Respiratory and Cardiovascular System	189	15.33%
	4 = Reproductive system	184	14.92%
	5 = Nervous and Health Mental system	172	13.95%
	6 = Sight and Hearing system	166	13.46%
Level-severity	0 = Mild	502	40.71%
	1 = Moderate	349	28.31%
	2 = Severe	382	30.98%

Table 1. Research Variables (continued)

Patient discharge status	0 = Healed	784	63.58%
	1 = Passed away	90	7.30%
	2 = Patient Discharged Against Medical Advice (PDAMA)	213	17.27%
	3 = Referred to Another Hospital	146	11.84%
Length of stay	The number of days the patient was hospitalized from admission to discharge	-	-

Based on Table 1, the average claim amount submitted was IDR 4,668,065, with the lowest claim value at IDR 643,103 and the highest at IDR 29,372,956. The average age of patients submitting claims was 46 years, with the youngest patient being 1 year old and the oldest 98 years old. Additionally, the average length of stay for patients was 4 days, with the shortest stay being 1 day and the longest 24 days.

In terms of inpatient classes, Class 3 had the largest number of patients at 425, followed by Class 2 with 410 patients. Class 1 had the smallest number of patients at 398. The gender distribution showed that there were more female patients at 644 compared to male patients at 589.

The INA-CBGs code breakdown revealed that there were 212 claims related to endocrine and metabolic disorders, which was more than the 132 claims for hematopoietic and immunological disorders. It was followed by gastrointestinal and hepatobiliary disorders, respiratory and cardiovascular disorders, reproductive disorders, neurological and mental health disorders, visual and auditory disorders, each accounting for 178, 189, 184, 172, and 166 claims respectively. Regarding the severity level, most patients were classified as having mild severity at 502, followed by severe severity at 382, and moderate severity at 349. The discharge status showed that most patients were discharged in a recovered state at 784, while 101 patients passed away. Additionally, 202 patients were transferred to another hospital, and 146 patients were discharged against medical advice.

Random Forest is a method that offers several advantages, such as the ability to improve accuracy in the presence of missing data, resistance to outliers, and efficiency in data storage[6]. The Random Forest regression algorithm does not require specific assumptions to be met, making it suitable for various types of data and situations. The Random Forest process involves randomizing the training process by selecting a subset of features from all available features each time a tree is trained. The selected features are then used to construct the optimal branches of the tree.

The training data is used to train the machine learning model and develop optimal parameters. The primary goal of this process is to build a model that can predict outcomes with high accuracy. The testing data is employed to evaluate the accuracy and reliability of the trained machine learning model. In this study, the training and testing data were tested in ratios of 60:40, 70:30, 80:20, and 90:10, allowing for the observation of the number of data points trained and tested, as shown in Table 2.

Table 2. Training and Testing Data

Rasio	Training	Testing	Total
60:40	741	492	1233
70:30	865	368	1233
80:20	989	244	1233
90:10	1113	120	1233

The term "ntree" (number of trees) represents the number of decision trees used in a Random Forest model. As the value of ntree increases, the complexity and accuracy of the predictive model also improve. However, this increase in ntree can also lead to higher computational time and memory requirements. Therefore, the value of ntree should be adjusted according to the data and the analysis goals.

The term "mtry" (number of features to consider at each split) indicates the number of features considered for each split in a decision tree model. In a Random Forest, each tree can select different features for each split, thereby enhancing model diversity and reducing overfitting. The optimal value of mtry should be chosen using cross-validation or grid search to achieve the best results. In practice, the optimal mtry value is often around 1/3 of the total number of features, but this can vary depending on the dataset and analysis goals. The selection of appropriate values for ntree and mtry can significantly impact the performance of the Random Forest model [7-8]. This study will build models with ntree values of 500, 800, 1000, and 1500 for the four data partition ratios used. In Random Forest regression, determining the best value for mtry can be done using $m_{try} = \frac{p}{3}$, where p represents the number of variables involved.

The Out-Of-Bag (OOB) error method is an evaluation technique used in machine learning, particularly in Random Forest regression. OOB samples utilize data not used in the training process to assess the model's performance. This approach provides a more accurate estimation of how well the model will perform on unseen data. The OOB error is calculated as the average prediction error on OOB samples for each tree. The estimation error of the Random Forest regression is predicted through the Root Mean Square Error (RMSE), which is obtained by calculating the average estimation error on OOB data for each decision tree in the Random Forest[7]. Given that n represents the number of samples in the OOB data, Y_i is the i-th sample value in the OOB data, and $\hat{Y}_i$ is the estimated value of the i-th sample using OOB data, the RMSE equation is expressed as[5]:

$$RMSE_{OOB} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{1}$$

$$\hat{Y}_i = \frac{1}{N_{tree}}\sum_{n=1}^{N_{tree}} \hat{Y}_n \tag{2}$$

The evaluation of the Random Forest regression model is measured using the coefficient of determination (R-Square), adjusted R-square, and Mean Absolute Percentage Error (MAPE). The coefficient of determination (R^2) measures how much of the variation in the dependent variable can be explained by the independent variables in the model, as per equation (3)[5]. On the other hand, the adjusted R-squared measures how much of the variation in the dependent variable can be explained by the independent variables in the model, while also accounting for potential biases or overfitting that may arise from adding more independent variables to the model, as per equation (4) [5]. The adjusted R-squared value ($R^2{}_{adj}$) is corrected to address these biases and overfitting issues.

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \underline{y}_t)^2} \tag{3}$$

$$R^2{}_{adj} = 1 - \left(\frac{n-1}{n-j-1}\right)(1 - R^2) \tag{4}$$

The Mean Absolute Percentage Error (MAPE) is a measure that calculates the average absolute percentage error in estimates. MAPE determines the absolute difference between estimated and actual values based on equation 5. A model is considered very good if the MAPE is less than 10%, good if it falls within the range of 10% to 20%, acceptable if it falls within the range of 20% to 50%, and poor if it exceeds 50% [6].

$$MAPE = \frac{1}{n}\left(\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|\right) \times 100\% \tag{5}$$

3. RESULT AND DISCUSSION

Based on the values of mtry and ntree obtained, several Random Forest models will be presented for each training and testing data set, as shown in Table 3 below.

Table 3. Random Forest Model for each ratio Data Training-Testing set.

Ratio	model	mtry	ntree	Ratio	model	mtry	ntree
60 :40	1	2	500	80 :20	17	2	500
	2	3	500		18	3	500
	3	2	800		19	2	800
	4	3	800		20	3	800
	5	2	1000		21	2	1000
	6	3	1000		22	3	1000
	7	2	1500		23	2	1500
	8	3	1500		24	3	1500
	9	2	500		25	2	500
70 :30	10	3	500	90 :10	26	3	500
	11	2	800		27	2	800
	12	3	800		28	3	800
	13	2	1000		29	2	1000
	14	3	1000		30	3	1000
	15	2	1500		31	2	1500
	16	3	1500		32	3	1500

Table 3 demonstrates the models constructed from all data partition ratios and involves training 32 models. After the training phase, all models will be evaluated for performance using the RMSE method with OOB data. The Random Forest model with the lowest $RMSE_{OOB}$ value will be selected as the best model and further analyzed for estimating claim amounts. The overall $RMSE_{OOB}$ calculations are presented in Table 4.

Table 4. Random Forest Model for each ratio Data Training-Testing set.

Ratio	mtry	ntree	RMSE OOB (IDR)	Ratio	mtry	ntree	RMSE OOB(IDR)
60 :40	2	500	986.686	80 :20	2	500	739.111
		800	990.909			800	738.034
		1000	994.797			1000	742.019
		1500	986.958			1500	736.841
	3	500	959.003		3	500	723.747
		800	954.979			800	728.220
		1000	960.454			1000	728.281
		1500	960.326			1500	728.488
70 :30	2	500	792.998	90 :10	2	500	640.562
		800	787.690			800	642.530
		1000	793.595			1000	637.206
		1500	794.045			1500	642.741
	3	500	779.915		3	500	629.612
		800	778.974			800	630.183
		1000	783.602			1000	631.653
		1500	779.522			1500	630.918

Based on Table 4, the performance of the Random Forest model can be observed across the four data partition ratio. This comparison aims to determine the best model with the smallest $RMSE_{OOB}$value. The optimal model selected in this case is the 90:10 ratio with mtry = 3 and ntree = 500.

Using this optimal model, we identified the most important variables and ranked the independent variables based on their influence on claim amounts, as shown in Figure 1. This ranking reveals which variables most significantly impact claim amount, enabling refined analysis and improved predictive accuracy.

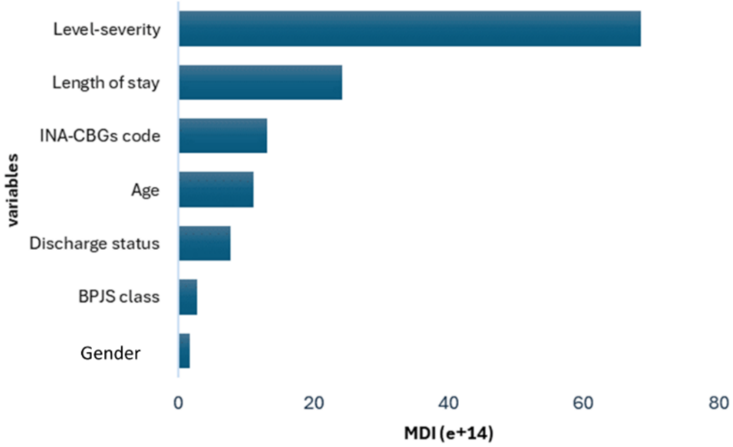


Figure 1. Plot Mean Decrease in Impurity

Based on Figure 1, the level-severity variable exhibits the highest significance among all independent variables. The level-severity often drives significant splits in Random Forest decision trees, substantially improving claim amount estimation. In contrast, gender was the least significant independent variable. The results indicate gender is rarely a splitting variable in the Random Forest model. Consequently, gender does not significantly influence the estimation of claim amounts. All variables will be modeled in the Random Forest regression simulation for calculating claim amounts.

In order to simplify the explanation of Random Forest regression simulation to calculation of claim amounts, we show the regression simulation only the first tree model as seen in Figure 2. Because of the complexity of the three in random forest regression simulation, the first tree model can be used as a representation of another tree model, hence the random forest can easily be understood. As seen in Figure 2, we consider a male patient (code 0), aged 42 years with mild disease severity (code 0), INA-CBGs (code 5), BPJS class 3, discharged in a recovered state (code 0), and a 6-day hospital stay. The patient's path through the tree will proceed to the next node with a true statement according to Figure 2. Based on Figure 2, the process will check the severity level again and direct the path to true because the patient has mild disease severity. The path will then proceed to false because the patient had a 6-day hospital stay. Subsequently, the age of the patient will be checked, as the patient is 42 years old. The path resolves to false and terminates at the final node, providing an estimated claim amount of IDR 2,897,080.967 for the patient, calculated using equation (2) with mtry = 3 and ntree = 500.

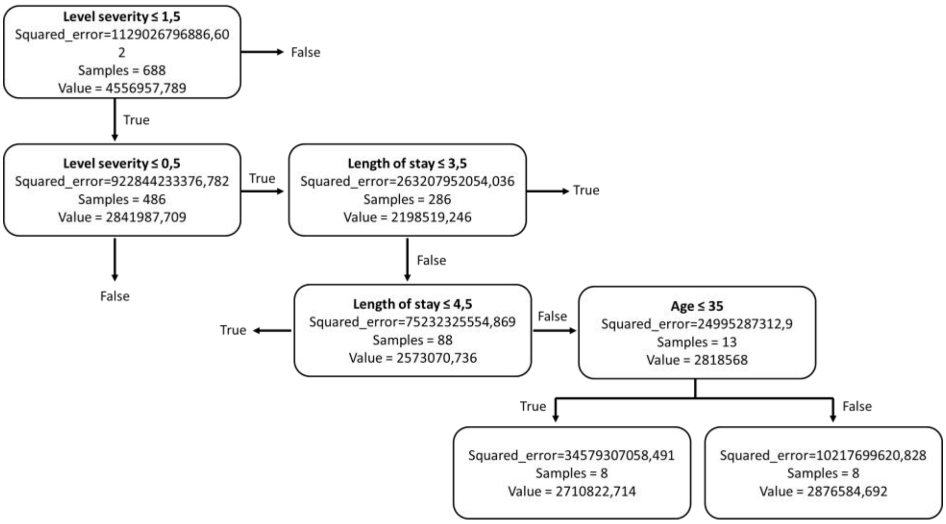
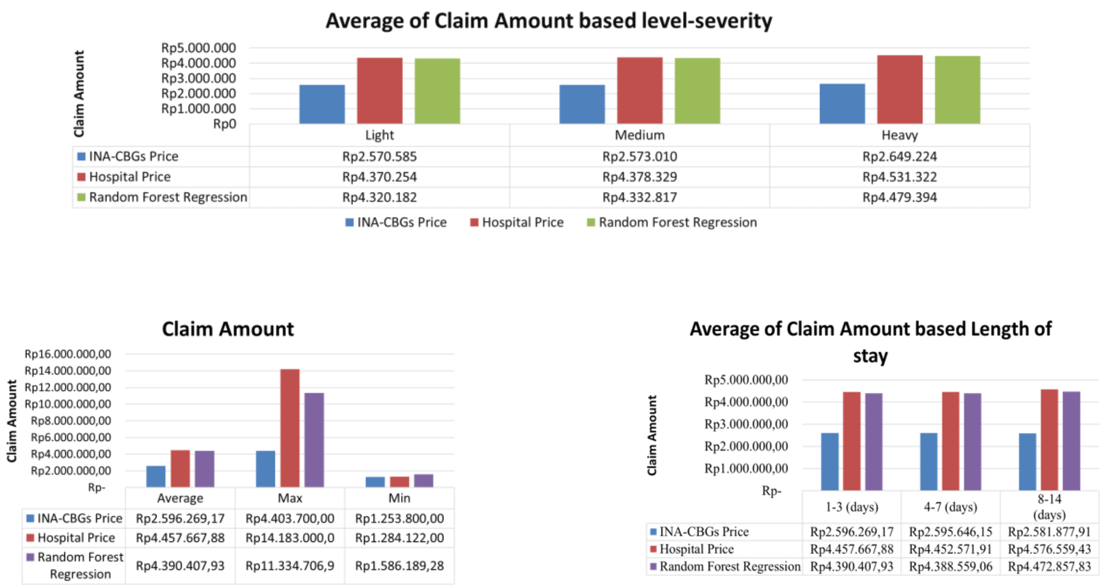


Figure 2. The 1st Tree Model

Based on the results of the simulation in Figure 2 and using the optimal Random Forest regression model, we can estimate claim amounts based on variable importance as presented. Figure 3 illustrates the contribution of each independent variable to the overall claim estimation, emphasizing their individual significance in the predictive model. The visualization helps elucidate the relationships and impacts of various factors on estimated claim amounts.



Average of Claim Amount based level-severity

	Light	Medium	Heavy
INA-CBGs Price	Rp2.570.585	Rp2.573.010	Rp2.649.224
Hospital Price	Rp4.370.254	Rp4.378.329	Rp4.531.322
Random Forest Regression	Rp4.320.182	Rp4.332.817	Rp4.479.394

Claim Amount

	Average	Max	Min
INA-CBGs Price	Rp2.596.269,17	Rp4.403.700,00	Rp1.253.800,00
Hospital Price	Rp4.457.667,88	Rp14.183.000,0	Rp1.284.122,00
Random Forest Regression	Rp4.390.407,93	Rp11.334.706,9	Rp1.586.189,28

Average of Claim Amount based Length of stay

	1-3 (days)	4-7 (days)	8-14 (days)
INA-CBGs Price	Rp2.596.269,17	Rp2.595.646,15	Rp2.581.877,91
Hospital Price	Rp4.457.667,88	Rp4.452.571,91	Rp4.576.559,43
Random Forest Regression	Rp4.390.407,93	Rp4.388.559,06	Rp4.472.857,83

Figure 3. Claim Amounts based on Random Forest Regression. Note that INA-CBGs Price (INA-CBGs Tariff) and Hospital Price (Hospital Cost)

Based on Figure 3, the claim amounts based on the INA-CBGs tariff tend to be lower, reflecting the limitations set by the government tariff. In contrast, the hospital cost represents the actual costs incurred by patients during their hospital stay. The Random Forest regression model provides an estimate of the possible claim amounts, with the model estimating values closer to the hospital cost and demonstrating greater sensitivity to data variability and complexity. The results of the Random Forest regression model calculation yield R-Square, Adjusted R-Square, and MAPE (Mean Absolute Percentage Error) values of 83%, 82%, and 17%, respectively. These results indicate that our calculated model performs well.

Our results indicate differences between *BPJS Kesehatan* claim amounts compared to hospital cost and INA-CBGs tariff. BPJS can recalculate the funds allocated to cover claim amounts and ensure these claims align with hospital cost. This calculation is crucial to ensure the adequacy of *BPJS Kesehatan*'s claim reserves, preventing adverse financial consequences for stakeholders, including insurance participants, hospitals, and *BPJS Kesehatan* itself. Moreover, it is important to accurately determine claim amounts to avoid loss ratio for *BPJS Kesehatan*. These results are not a definitive benchmark for *BPJS Kesehatan* claim calculations in general. Due to the regional nature of the data, applying it to other regions may produce different results. Further investigation is needed to determine *BPJS Kesehatan*'s overall claim payments.

4. CONCLUSIONS

The selected Random Forest Regression model is based on the 90:10 training-testing data split, with mtry = 3 and ntree = 500. The results of the model calculations indicate that the variables influencing the calculation of claim amounts in order of significance are: Patient Severity Level, Length of Stay, INA-CBGs Code, Age, Discharge Status, BPJS Class, and Gender. Based on our study, the average estimated claim amount from the Random Forest regression model is higher than the average claims paid by *BPJS Kesehatan*. We then suggested that further evaluation regarding the claims paid by *BPJS Kesehatan* and the tariffs that must be paid to hospitals are needed in the future investigation.

5. REFERENCES

[1] Raihananda, Q. , Putra, I. W. E. D. ., Sijabat, M. S. ., Rofatunnisa, S. ., Maruf, I. ., Hermarwan, H., & Nooraeni, R. (2020). *Application of Random Forest Method Classification to Predict BPJS Kesehatan Card Users Who Receive Contribution Assistance in Karangasem District, Bali Province 2017*. *Jurnal Matematika, Statistika Dan Komputasi*, *17*(2), 178-188. https://doi.org/10.20956/jmsk.v17i2.11710

[2] Rofiq, H. N. (2023).“-[Translation] Detection of Inefficiency in *BPJS Kesehatan* Claims using Machine Learning : Deteksi Inefisiensi pada Klaim BPJS Kesehatan dengan menggunakan Machine Learning”. *Jurnal Jaminan Kesehatan Nasional*. *3*(1), 83-98. https://doi.org/10.53756/jjkn.v3i1.134

[3] S. R. Putri, N. A. Rumana, L. Widjaja and D. Sonia. (2023)."-[Translation] Differences Between Hospital Tariffs & INA-CBGs Tariffs for Inpatient Sectio Caesarea Cases at RSUD Kalideres in 2022 : Perbedaan Tarif Rumah Sakit & Tarif INA-CBGS Pelayanan Rawat Inap Kasus Sectio Caesarea di RSUD Kalideres Tahun 2022." *Jurnal Health Tambusai.* 4(4),5211-5216.

[4] D. Anyaprita, K. N. Siregar, B. Hartono, M. Fachri and F. Ariyanti. (2020)."-[Translation] Impact of Delays in *BPJS Kesehatan* Claim Payments on the Quality of Service at Jakarta Sukapura Islamic Hospital : Dampak Keterlambatan Pembayaran Klaim BPJS Kesehatan Terhadap Mutu Pelayanan Rumah Sakit Islam Jakarta Sukapura," *Muhammadiyah Public Health Journal,.* 1(1),1-77.

[5] Forests, R. (2001). By Leo Breiman. Mach Learn, 45(1), 5-32.

[6] S. Fachid and A. Triayudi. (2022). "-[Translation] Comparison of Linear Regression and Random Forest Algorithms in Predicting Positive COVID-19 Cases : Perbandingan Algoritma Regresi Linier dan Regresi Random Forest Dalam Memprediksi Kasus Positif Covid-19," Jurnal Media Informatika Budidarma. 6(1), 68-73.

[7] Hadi, Nicholas and Benedict, Jason. (2024).“-[Translation] Implementation of Machine Learning for House Price Prediction Using the Random Forest Algorithm : Implementasi Machine Learning untuk Prediksi Harga Rumah Menggunakan Algoritma Random Forest”. Computatio: Journal of Computer Science and Information Systems. 8(1),50-61.

[8] Hakim, Arif Rahman. dkk.(2023).Implementation of Random Forest Algorithm on Palm Oil Price. *Journal Tech-E.6(2)*,34-42.

[9] E. S. Lestari and I. Astuti.(2022)."-[Translation] Application of Random Forest Regression for Predicting House Selling Prices and Cosine Similarity for House Recommendations in West Java Province : Penerapan Random Forest Regression Untuk Memprediksi Harga Jual Rumah Dan Cosine Similarity Untuk Rekomendasi Rumah Pada Provinsi Jawa Barat.*Jurnal Ilmiah FIFO.* 14(2),131-146.

[10] M. Mercadier and J.-P. Lardy.(2019). Credit Spread Approximation and Improvement using Random Forest Regression. *European Journal of Operational Research.* 277 (1), 351-365.