JURNAL INFOTEL
Informatics - Telecommunication - Electronics
Website: http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN: 2085-3688; e-ISSN: 2460-0997

Prediction of patient length of stay using random forest method based on the Indonesian national health insurance

Aini Hanifa[1,*], Yogiek Indra Kurniawan[2], Jati Hiliamsyah Husein[3], Arief Kelik Nugroho[4], Ipung Permadi[5]
[1,2,4,5]Universitas Jenderal Soedirman
[3]Telkom University
[3]Waseda University
[1,2,4,5]Jl. Mayjen Sungkono KM. 5, Purbalingga 53371, Indonesia
[3]Jl. Telekomunikasi, No. 1, Bandung 40257, Indonesia
[3]Shinjuku, Tokyo 169-8050, Japan
[*]Corresponding email: ainihanifa@unsoed.ac.id

Received 10 May 2023, Revised 31 July 2023, Accepted 29 August 2023

Abstract — Inpatient care constitutes the most significant portion of healthcare service expenditure, making it crucial for healthcare management to focus on reducing costs and enhancing services. Therefore, identifying factors associated with patient length of stay and accurately predicting the duration of hospitalization become essential in supporting stakeholders' decision-making processes. This includes setting appropriate rates and optimizing resource allocation, ultimately leading to improved healthcare services and better patient outcomes. In addressing this issue, the study utilized the random forest method to predict the length of stay for patients utilizing BPJS (Indonesian healthcare and social security agency) insurance services while identifying the main determinant variables affecting patient length of stay. To assess the prediction model's effectiveness, an experiment was conducted, comparing various numbers of trees and candidate split attributes. The experimental results demonstrated that increasing the number of trees and candidate split attributes resulted in improved prediction performance and reduced error rates, thereby enhancing the overall accuracy of the predictions. The optimal value was found when the number of trees was 100 with the MSE/Variance value of 0.3805. These findings highlight the significance of the patient's disease diagnosis, participant segment, and healthcare facility type as the main determinant variables for predicting patient length of stay.

Keywords – BPJS, health insurance, length of stay, random forest

I. Introduction

The length of hospital stay refers to the number of days between a patient's admission date to a hospital and the date when the patient is discharged and has recovered. Inpatient care is the largest component of healthcare spending in hospitals. Healthcare service management continues to strive to improve the quality of healthcare services while reducing patient care costs. Therefore, identifying factors related to the length of hospital stay and accurately predicting how long patients will be hospitalized can assist healthcare service administrators and health insurance providers in developing better plans and strategies.

In Indonesia, there is a National Health Insurance managed by Badan Penyelenggara Jaminan Sosial (BPJS). The BPJS financing system uses the INA-CBG's casemix principle based on casemix from ICD X (International Classification of Disease X) which is used as a guideline for setting rates. The casemix system in INA-CBG's is a grouping of similar patient characteristics. Hospitals will get paid based on the average cost spent by a group of diagnoses [1]. Meanwhile, out of the top ten most frequent procedures in non-capitation primary healthcare facilities under BPJS, the majority are inpatient care in hospital rooms, accounting for a proportion of 37.2 [2].

To carry out its mission of serving the lower middle class, government hospitals are facing challenges due to limited financial resources and the presence

of bureaucracy and restrictive regulations. Therefore, hospitals need to have a prediction system that can help control costs by projecting the patient's length of stay (LoS). The results of this prediction can be used as a basis for determining patient group rates by BPJS. To improve the health services provided by BPJS, it is necessary to implement an efficient and effective service quality control system and payment system in health insurance [3]. By implementing such a system, hospitals can better manage their resources and budgets, making healthcare services more affordable and accessible for patients.

Accurately predicting the length of hospital stay for patients using health insurance, particularly through the implementation of data mining algorithms, holds significant value for healthcare service management and health insurance providers. The insights gained from these predictions enable informed decision-making, appropriate rate setting, and optimized resource allocation, ultimately leading to improved healthcare services and better patient outcomes. As the healthcare landscape continues to evolve, the use of predictive models becomes increasingly crucial for ensuring sustainable and efficient healthcare systems.

The practical implications of LoS prediction research using health insurance patient data are revolutionary for healthcare service management and health insurance practices. These predictions facilitate more precise reimbursement rates and premium adjustments, fostering data-driven decision-making and policy development. Furthermore, LoS predictions have the potential to elevate the quality of care, enhance health outcomes, and enhance fraud detection and prevention efforts. Overall, this research empowers the healthcare industry to operate with greater efficiency, deliver higher-quality care, and provide better experiences for patients. By harnessing the power of data mining algorithms in LoS prediction, the healthcare sector can adapt and thrive in an ever-changing environment, ultimately benefiting patients and stakeholders alike.

Previous research on predicting LoS in hospital using health insurance claims data has been conducted by [4] using ensemble method. This study recommends that future research should focus on exploring significant features that influence the performance of LoS prediction models, utilizing health examination data from different countries. Another study was conducted by [5] [6] [7] [8] to predict specific LoS for pediatric patients using the random forest method. The study compared the predictive performance using several methods and found that random forest had the best performance.

Based on the issues and solutions presented above, this research aims to predict the length of patient hospitalization using the random forest method. The data used in this study specifically focuses on patients who utilize health insurance services provided by the Social Security Administrating Body (BPJS). Prior to this research, there has been no specific study utilizing BPJS patient data to predict the length of patient hospitalization. This research is expected to assist BPJS and hospitals in formulating strategic policies related to patient hospitalization.

II. Related Work

There are several studies that have been conducted to predict patient LoS using health insurance claims data. Sato *et al.* [9] has made predictions which were provided by 170 regional public insurers in Gifu, Japan. Janwanishtaporn *et al.* [10] has done study to determine the national hospitalization rate in Thailand using public health security scheme data. While An *et al.* [4] using health checkup cohort DB data stored in the virtual server of National Health Insurance Service (NHIS) of Korea, while Srimannarayana *et al.* [11] used Health Insurance in India. There are no studies specifically using National Health Insurance (BPJS) data in Indonesia.

Numerous research studies have been undertaken to predict LoS using machine learning methods. Ayyoubzadeh *et al.* [5] conducted a study aimed at predicting LoS in Iran, stating that method Random Forest produced the best AUC. The study did not include disease categories as features in building the model. According to it, it is important for future research to incorporate disease codes to identify higher-risk disease categories requiring hospitalization. Thompson *et al.* [6] conducted a machine learning-based prediction on the LoS for newborn infants. Nine methods were compared, namely ZeroR, Naïve Bayes, Logistic Regression, Multi-layer Perceptron, Simple Logistic, Support Vector Machine (SVM), J48, Random Forest, and Random Tree. Among these methods, Random Forest produced the best performance with a significantly higher accuracy difference.

A similar study was conducted by Li *et al.* [7]. Five machine learning models, including naive Bayes, logistic regression, linear kernel support vector machine, random forest, and Gradient Boosted Decision Trees, were used to construct binary classifiers. The random forest model, which exhibited modest predictive capability, performed the best among all the models. Furthermore, Daghistani *et al.* [8] evaluated four classification techniques, namely Random Forest (RF), Artificial Neural Network (ANN), Support Vector Machine (SVM), and Bayesian Network (BN), to assess their performance. Random Forest exhibited the highest performance among these techniques. In addition, Gutierrez *et al.* [12] conducts research to predict LoS across various hospital departments and specialties. By comparing the performance of the Random Forest algorithm and neural networks, the researcher

finds that the Random Forest algorithm yields better outcomes.

Random forest as an ensemble method shows better performance than other methods to predict LoS which have also been shown by several studies. Including [13] which present the prediction of LoS at a tertiary referral hospital in Tasikmalaya, Indonesia. Ma *et al.* [14] using three different decision tree methods that is Bagging, Adaboost, and Random Forest. The findings suggest that all three approaches are successful in predicting the LoS. In other study conducted by Alsinglawi *et al.* [15], the objective was to predict the LoS for heart failure diagnoses. The findings indicated that the deep learning-based regressor did not outperform the traditional regression model (random forest) in terms of predictive performance. Another study was conducted by Wang *et al.* [16] which predicted specific LoS for patients in the pediatric category, compared to the prediction results using 3 ensemble methods, namely adaboost, bagging, and random forest; obtained random forest which has the best method performance.

Additionally, several studies also examine the most important factors affecting LoS. According to [5], the critical factors influencing the LoS include the quantity of para-clinical services, frequency of counseling, clinical wards, doctor's specialty and degree, and the reason for hospitalization. As stated by [11], there is a significant positive correlation between age and the duration of hospitalization. An *et al.*'s [4] findings indicate that key factors contributing to the LoS include demographic variables such as age, insurance deductible ratio, and information related to the primary diagnosis and sub-diagnoses. Furthermore, the researchers identified that the number of doctors and beds available in hospitals, cholesterol level, body mass index (BMI), and the month of admission were also found to be significant factors influencing the LoS. Whereas [9] found that the top important feature variables include indicators of current health status (such as current fitness level and age), risk factors for worsening healthcare status in the future (such as dementia), and preventive care services to improve healthcare status in the future (such as training and rehabilitation). Salmons *et al.* [17] created and internally validated a supervised machine learning algorithm to accurately detect factors influencing costs in ambulatory single-level lumbar decompression surgery. Their research discovered that key factors affecting costs included the type of anesthesia used, the length of time spent in the operating room, the patient's race, income and insurance status, the community type, worker's compensation status, and the comorbidity index. So, In this study, observations were made to find the main determinant variables for predicting patient LoS using random forest.

Accurately predicting the LoS for patients is crucial for anticipating future demand and implementing appropriate measures. This is particularly relevant in the context of COVID-19, where studies [18], [19] have focused on predicting hospital LoS for COVID-19 patients. Additionally, analyzing the impact of a major pandemic like COVID-19 on LoS can provide valuable insights into how this factor is affected and its implications for hospital departments. Recent research [12] has highlighted a significant reduction in LoS. Given the high cost of treating COVID-19 patients, it becomes essential to identify admission-related factors that influence hospital LoS and establish a risk assessment for clinical management [20].

From several methods offered by the above-mentioned studies, it was found that ensemble methods performed the best, with random forest being the top-performing ensemble method. Ensemble methods are an extension of classification methods, where traditional classification techniques typically use a single classifier. However, Ensemble Methods or Classifier Combination Methods combine predictions from multiple classifiers. Ensembles are believed to improve the accuracy of predicted classes compared to single-classifiers [21], given they meet the following conditions:

1) The base classifiers must be independent or not dependent on each other. This means that if there is an error rate in one base classifier, it should not be correlated with the error rate of other base classifiers.
2) The selected base classifiers must perform better than a classifier with random guessing.

A number of researchers have devoted to the simulation of hospitalization or LoS predictions. But, no existing study has employed the random forest method to predict the LoS for patients utilizing BPJS health insurance. So in this study utilizing BPJS data that has been computerized and available for research needs.

III. Research Method

Random forest (RF) is a classification algorithm comprising multiple decision trees. Each decision tree is constructed using a random subset of features known as a random vector. To incorporate the random vector in tree construction, a random F value is chosen. This F value determines the number of input attributes (features) to be considered for splitting at each node in the decision tree. By randomly selecting F attributes, it is not necessary to evaluate all available attributes; only the selected F attributes are utilized. The control over the random forest's effectiveness lies in adjusting both the F value and the number of trees in the forest [4]. When the F value is too small, the trees generated will have minimal correlation, and the opposite is true as well. The first equation allows for the calculation of the value of F.

$$F = \log_2(M + 1) \tag{1}$$

Here, M denotes the total number of features. In addition to selecting attributes, randomization is employed in choosing the training set. Bagging, or bootstrap aggregating, is a technique used to create bootstrap samples. Each decision tree is constructed using a bootstrap sample of data and a random subset of attributes, which are considered for splitting at each node. These samples are generated from a random variable set of data produced through bagging. The random forest algorithm follows the flow outlined below:

1) Select a parameter n to determine the desired number of trees to be created within the forest.
2) Create n bootstrap samples by applying the bagging technique to the training dataset.
3) For each node in a tree, select the value of F based on the equation 1 to determine the number of features to be considered.
4) To build a decision tree, the first step is to select a group of F variables to be used as candidate split attributes at each node. Then, the tree is split using this group, and the attribute used as the next node is determined based on the criteria established according to the decision tree algorithm used. During the tree building process, the value of F remains constant.
5) The random forest is constructed without applying any pruning techniques to eliminate bias from the prediction outcomes.
6) The overall prediction result is obtained by averaging the error rates of all decision trees in the forest.

In this study, a system has been developed to predict the duration of patient stays using the random forest (RF) method. The system utilizes test data extracted from medical records of patients enrolled in the BPJS program. It assesses the effectiveness of the RF method in predicting LoS and evaluates the key variables that significantly impact the method's performance. Fig. 1 provides an overview of the system.

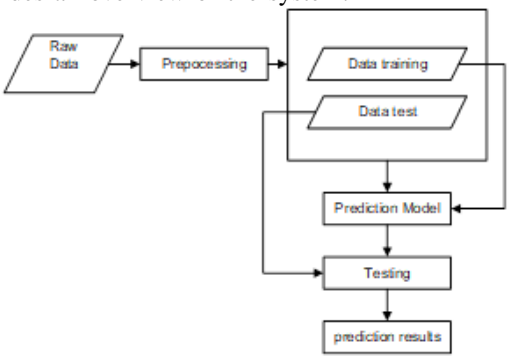

Fig. 1: System overview.

Initially, the raw data for predicting the patient's LoS underwent data preprocessing. This process involved several stages, such as feature selection and attribute value transformation. Once the preprocessing stage was completed, the data was divided into training data and testing data. A prediction model was then constructed using the training data. Subsequently, the generated prediction model was applied to classify the test data. This allowed for the prediction of the patient's LoS in the hospital, and the error value of the prediction model was computed.

A. Preprocessing

The data preprocessing stage as can be seen in Fig. 2, which includes data retrieval, feature selection, data cleaning, and transformation, is the process of organizing data into a standardized form that is ready to be processed in Data Mining. The data is retrieved from the source of raw patient hospitalization data using BPJS services. Feature selection involves the process of removing irrelevant attributes, and data cleaning includes handling noisy data. Meanwhile, transformation involves converting the data structure and normalizing the data. The output of this stage is observation data that is ready to undergo the prediction process.

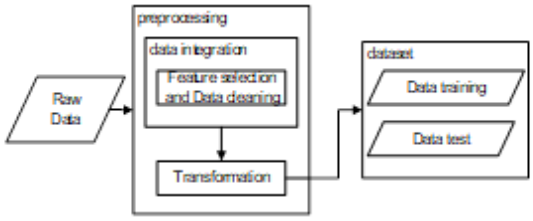

Fig. 2: Preprocessing stage.

B. Prediction Model

The model-building process is iterated based on the user-defined input for the number of trees (nbtree). Once the detection model is created, it undergoes testing using the testing data. During testing, the average prediction result is calculated from the formed trees, along with the average mean squared error (MSE) from the trees used to build the model. Further details of the system can be observed in Fig. 3.

The dataset, comprising n instances of data, is divided into multiple bootstraps with x instances and y attributes. Bootstrap formation is performed randomly with replacement, allowing for the presence of identical data within a bootstrap and potential variations in data attributes across different bootstraps.

The random forest ensemble stage, represented in Fig. 4, involves determining the best primary splitter among the independent variables. The primary splitter is identified as the splitter that yields the largest decrease in record set diversity. During this step, the Gini index value is computed for each attribute, and the top F attributes with the best Gini index value are selected.

Then for each bootstrap that has been formed, a tree is built using the above primary splitter. This prediction model produces a continuous valued class attribute, which is the length of patient stay, not a discrete class.

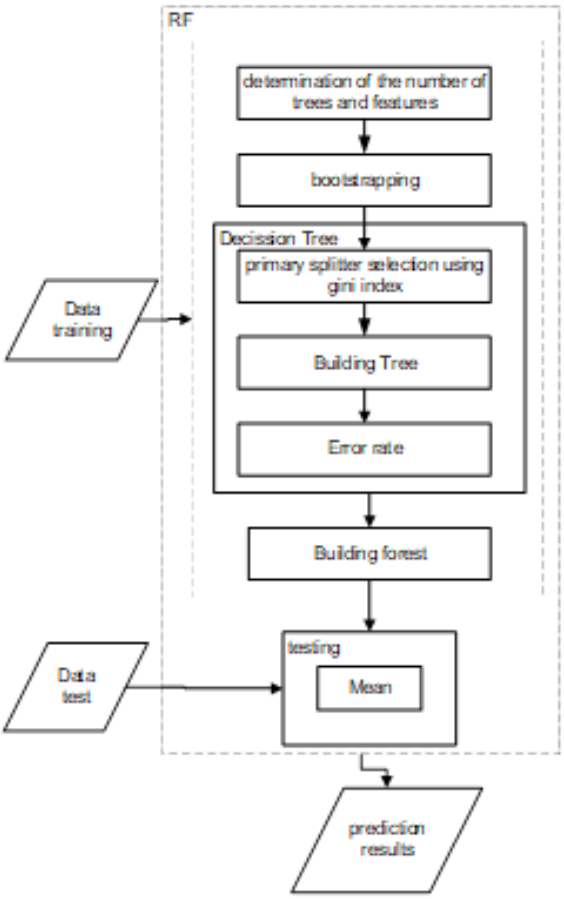

Fig. 3: Prediction model.

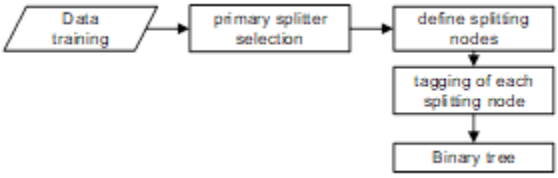

Fig. 4: Building tree.

A number of trees will be constructed, and the error value of each tree will be calculated.

In Fig. 5 it is explained that the average calculation is carried out on the results of the decision tree predictions with test data that has been prepared beforehand. Then the best classifier will be produced. From the best classifier, the main variable that determines the predictive length of hospitalization of BPJS users can be determined.

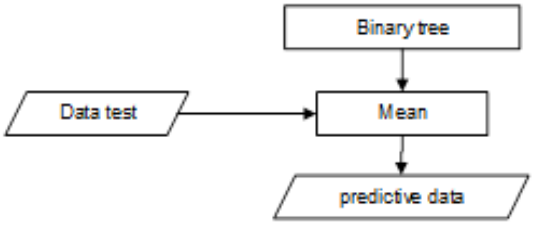

Fig. 5: Mean.

The final step is to evaluate the selected subtree by applying it to the test data as can be seen in Fig. 6. In this step, the predicted data is compared to the test data, which will result in the percentage of error from the created classifier model and determine the main predictor variables for the length of inpatient stay prediction. The study [22] analyzed several models' prediction accuracy. Based on this research, the selected error measurement in this study is Scale-Dependent Error because the predicted value and the actual value for the inpatient stay prediction have the same data scale.

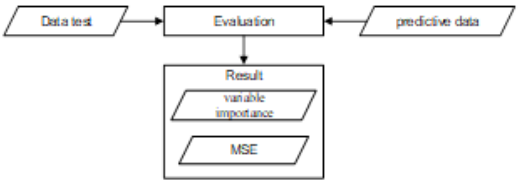

Fig. 6: Evaluation of predictive results.

If y_i is the i-th value and $\hat{y}_i$ is the predicted value for y_i, then the prediction error value is $e_i = y_i - \hat{y}_i$, which has the same scale as the data. The most commonly used scale-dependent accuracy measures are mean absolute error (MAE) and mean squared error (MSE). MAE and MSE are used as relative measures to compare the performance of the same prediction with different models. If the value of MSE/σ^2 approaches 1, then the prediction error is large, if it approaches 0, then the prediction error is small.

$$MSE = \frac{\sum_{i=1}^{n} e_i^2}{n} \tag{2}$$

IV. Result

This section discusses the dataset, test scenario, and result analysis.

A. Dataset

The data used in this study is data obtained from the "2015-2016 BPJS Health sample data" which has been provided by the BPJS for researchers. The sample data presents 111 variables that can be processed, consisting of 15 membership variables, 23 capitation service variables for Primary Health Facilities (FKTP), 20 non-capitation service variables for FKTP, and 53 referral health facility service variables (FKRTL) that are interconnected through the participant card number variable [23]. The sample data for Non-Capitation FKTPs is 104,456 rows of data, which includes information about the characteristics of the health facility, type of service, diagnosis code, type of procedures, rate paid, etc. While the membership data is 1,697,451 rows of data.

B. Test Scenario

The preprocessed sample data consisted of 25,903 rows. Then, the system testing was performed on 75 % of the data as training data and 25% as testing data. The testing was performed on the given data, utilizing varying numbers of trees and splitting attributes. One of the parameters in the Random Forest method is the

number of trees or classifiers used in the model. A study conducted by [21] examined the impact of the number of trees through testing. The findings indicated that increasing the number of trees did not lead to overfitting, but rather resulted in a general limit of error values. In this study, an experiment will be conducted to verify whether adding trees only establishes a general limit for error values and does not significantly affect the model's performance.

In random forest, there exists an optimal number of trees where accuracy and error rate stabilize while considering the required time. However, this minimum number of trees is not universally fixed for all dataset characteristics. This variability stems from the utilization of random functions, which randomly select data rows and attributes. Therefore, in this test, four models will be created with 10, 50, 100, and 150 trees. From each model, the MSE value and variable importance will be recorded for analysis.

Concerning the random function, when applied to a dataset with identical input parameters, consecutive executions will yield different performance values. This variability arises from the random function's influence on the model's predictions. The MSE obtained during the second execution may either improve or worsen compared to the previous execution. Therefore, in this test, observations are made by conducting five trials or experiments.

C. Results Analysis

The experiment involved testing different numbers of trees (10, 50, 100, and 150), as described earlier in the testing scenario. The testing phase aimed to assess the extent of prediction error and determine whether the number of trees had an impact on the error value of the tested data. The error value reflects the performance level of the prediction. The results obtained from a series of experiments conducted on varying numbers of trees are depicted in Fig. 7, illustrating the performance outcomes.

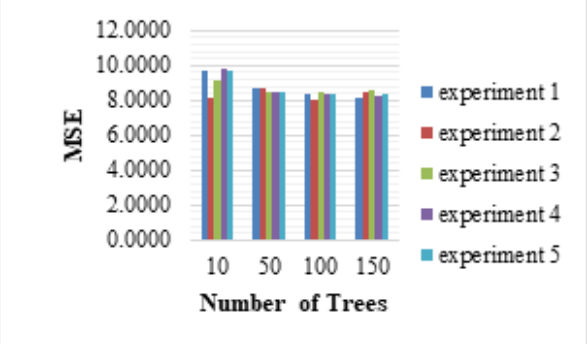

Fig. 7: Test results analyzed the relationship between the number of trees and MSE value.

According to [22], the predictive error was evaluated by comparing MSE with the value of variance (σ^2). If the value of MSE/σ^2 approaches 1, then the predictive error is large, and if it approaches 0, then the predictive error is small. The results of the comparison between the MSE value and the variance of the data that the model formed are shown in Table 1. It can be seen that the comparison values approach 0 for all tests.

Table 1: Comparison of MSE Value with Variance

Testing	Number of Trees			
	10	50	100	150
experiment 1	0.444	0.399	0.383	0.370
experiment 2	0.371	0.398	0.367	0.389
experiment 3	0.417	0.388	0.387	0.389
experiment 4	0.448	0.389	0.384	0.377
experiment 5	0.440	0.387	0.382	0.382

Additionally, the study analyzed variables that affect hospitalization duration prediction. Fig. 8 presents the average importance values of each predictor variable, derived from testing scenarios with 100 numbers of trees. Importance value refers to the significance of a variable in predicting the outcome of a model. The importance value is calculated based on the decrease in accuracy of the model when a variable is randomly permuted, which measures the contribution of the variable to the model's predictive power. The higher the importance value, the more important the variable is in predicting the outcome of the model. The importance value level for each variable is presented in Fig. 8 as a bar chart, where the height of the bar represents the importance value, and the variables are listed on the x-axis. The main determinant variables for predicting patient LoS were found to be the patient's disease diagnosis, participant segment, and healthcare facility type, which had the highest importance values in the model.

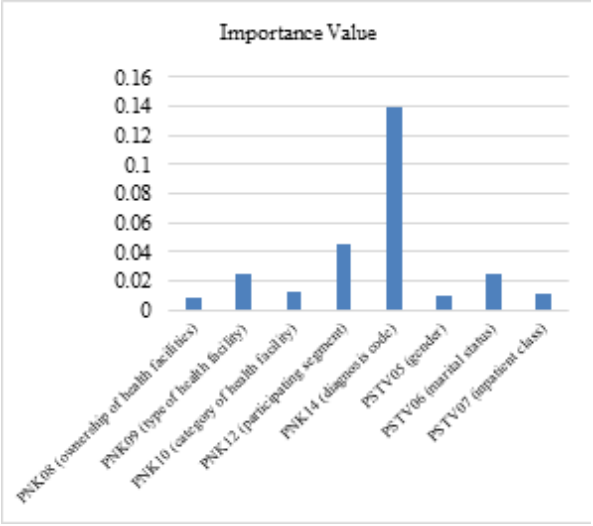

Fig. 8: Importance value level for each variable.

V. Discussion

This research provides valuable insights into predicting patient LoS using the random forest method based on the Indonesian National Health Insurance. The study highlights the importance of prediction performance in healthcare management decision-making

Table 2: The Variable Importance Values in the Model with 100 Trees

Variables	Importance Value				
PNK08 (ownership of health facilities)	0.007	0.013	0.009	0.01	0.008
PNK09 (type of health facility)	0.024	0.027	0.028	0.021	0.026
PNK10 (category of health facility)	0.013	0.012	0.012	0.009	0.008
PNK12 (participating segment)	0.049	0.046	0.042	0.041	0.038
PNK14 (diagnosis code)	0.143	0.152	0.142	0.149	0.144
PSTV05 (gender)	0.013	0.015	0.012	0.006	0.009
PSTV06 (marital status)	0.023	0.02	0.025	0.023	0.029
PSTV07 (inpatient class)	0.008	0.009	0.014	0.013	0.012
MSE/σ^2	**0.383**	**0.367**	**0.386**	**0.384**	**0.382**

and provides a framework for future research in this area.

Fig. 7 illustrates the relationship between the number of trees and the MSE of the prediction model, revealing a trend where increasing the number of trees leads to a decrease in MSE, signifying enhanced model accuracy. The testing phase confirms this observation, as the MSE value consistently reduces from 10 to 100 trees, but the decrease between 100 and 150 trees becomes insignificant. Consequently, the study identifies the optimal model with 100 trees. This valuable insight can be leveraged to develop more accurate prediction models for patient LoS, ultimately contributing to the improvement of healthcare management strategies.

Refer to Table 1, when building the model with 100 trees in experiment 2, the lowest error value was obtained compared to other experiments. This can be analyzed based on the variable importance values shown in Table 2. Table 2 shows the variable importance values in the model with 100 trees. The variable importance values in the model indicate the relative importance of each predictor variable in predicting the outcome variable. The variable importance values are calculated based on the mean decrease impurity (MDI) method, which measures the total reduction of impurity across all trees in the forest when a particular variable is used for splitting. The higher the variable's importance value, the more important the variable is in predicting the outcome variable. The variable importance values can be used to identify the most important predictor variables and to remove the less important ones, which can improve the prediction performance of the model.

Based on Table 2, it can be observed that in the second experiment with the lowest MSE/σ^2 value, the importance value of variable PNK14 showed a significant difference compared to the importance values in other experiments. This can strengthen the conclusion that the type of diagnosis is the main variable in determining the prediction of patient's LoS using BPJS. As mentioned in section IV of this study, the variable importance value of PNK14 (diagnosis code) always occupies the highest value followed by PNK12 (participant segment) and PNK9 (health facility type). Which means that these variables are the most important variables in predicting patient's LoS. These findings can help healthcare stakeholders make informed decisions to reduce expenditure costs and improve healthcare services.

VI. Conclusion

This study discusses the prediction of patient's LoS using the random forest method based on the Indonesian National Health Insurance. The study found that increasing the number of trees and candidate split attributes can improve prediction performance and reduce the resulting error rate. The optimal value was found when the number of trees was 100 with the MSE/Variance value of 0.3805. This suggests that the random forest method is effective in predicting patient's LoS based on the Indonesian National Health Insurance. In this study also resulted that, The patient's diagnosis is identified as the most significant variable in determining the length of hospital stay, accompanied by supporting variables such as the participant segment and the type of healthcare facility.

Overall, the study provides valuable insights into the factors that impact patient's LoS and how accurate predictions can be made using the random forest method. The results of this study can be used to develop more effective healthcare management strategies and improve patient outcomes. For further research, the model can be applied to one type of patient diagnosis and additional variables related to the patient's medical records, such as blood pressure, BMI, *etc*.

Acknowledgment

We would like to thank the LPPM Universitas Jenderal Soedirman for funding the research, and we also extend our thanks to our colleagues who provided assistance throughout the study.

References

[1] L. Ardini, D. Maryam, and N. Munaa, eds., *Fraud Detection in Indonesia National Health Insurance Implementation: A Phenomenology Experience from Hospital*, 1st International Conference on Business & Social Sciences (ICOBUSS), 2020.

[2] BPJS, "Data sampel bpjs kesehatan tahun 2015-2016," tech. rep., BPJS, Indonesia, 2016.

[3] H. Fahlevi, I. Irsyadillah, M. Indriani, and R. S. Oktari, "DRG-based payment system and management accounting changes in an Indonesian public hospital: exploring potential roles of big data analytics," *Journal of Accounting & Organizational Change*, vol. 18, pp. 325–345, Mar. 2022.

[4] J. An, M. Jung, S. Ryu, Y. Choi, and J. Kim, "Analysis of length of stay for patients admitted to Korean hospitals based on the Korean National Health Insurance Service Database," *Informatics in Medicine Unlocked*, vol. 37, p. 101178, 2023.

[5] S. M. Ayyoubzadeh, M. Ghazisaeedi, S. Rostam Niakan Kalhori, M. Hassaniazad, T. Baniasadi, K. Maghooli, and K. Kahnouji, "A study of factors related to patients' length of stay using data mining techniques in a general hospital in southern Iran," *Health Information Science and Systems*, vol. 8, p. 9, Dec. 2020.

[6] B. Thompson, K. O. Elish, and R. Steele, "Machine Learning-Based Prediction of Prolonged Length of Stay in Newborns," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, (Orlando, FL), pp. 1454–1459, IEEE, Dec. 2018.

[7] Y. Li, H. Wang, and Y. Luo, "Improving Fairness in the Prediction of Heart Failure Length of Stay and Mortality by Integrating Social Determinants of Health," *Circulation: Heart Failure*, vol. 15, Nov. 2022.

[8] T. A. Daghistani, R. Elshawi, S. Sakr, A. M. Ahmed, A. Al-Thwayee, and M. H. Al-Mallah, "Predictors of in-hospital length of stay among cardiac patients: A machine learning approach," *International Journal of Cardiology*, vol. 288, pp. 140–147, Aug. 2019.

[9] J. Sato, N. Mitsutake, M. Kitsuregawa, T. Ishikawa, and K. Goda, "Predicting demand for long-term care using Japanese healthcare insurance claims data," *Environmental Health and Preventive Medicine*, vol. 27, no. 0, pp. 42–42, 2022.

[10] S. Janwanishstaporn, K. Karaketklang, and R. Krittayaphong, "National trend in heart failure hospitalization and outcome under public health insurance system in Thailand 2008–2013," *BMC Cardiovascular Disorders*, vol. 22, p. 203, Dec. 2022.

[11] G. Srimannarayana, P. V. R. Kumar, and P. Dhanavanthan, "A Study on Days of hospitalization of insured with the claims data," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, pp. 5230–5238, Apr. 2021.

[12] J. M. P. Gutierrez, M.-A. Sicilia, S. Sanchez-Alonso, and E. Garcia-Barriocanal, "Predicting Length of Stay Across Hospital Departments," *IEEE Access*, vol. 9, pp. 44671–44680, 2021.

[13] D. Barsasella, S. Gupta, S. Malwade, Aminin, Y. Susanti, B. Tirmadi, A. Mutamakin, J. Jonnagaddala, and S. Syed-Abdul, "Predicting length of stay and mortality among hospitalized patients with type 2 diabetes mellitus and hypertension," *International Journal of Medical Informatics*, vol. 154, p. 104569, Oct. 2021.

[14] F. Ma, L. Yu, L. Ye, D. D. Yao, and W. Zhuang, "Length-of-Stay Prediction for Pediatric Patients With Respiratory Diseases Using Decision Tree Methods," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, pp. 2651–2662, Sept. 2020.

[15] B. Alsinglawi, F. Alnajjar, O. Mubin, M. Novoa, M. Alorjani, O. Karajeh, and O. Darwish, "Predicting Length of Stay for Cardiovascular Hospitalizations in the Intensive Care Unit: Machine Learning Approach," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, (Montreal, QC, Canada), pp. 5442–5445, IEEE, July 2020.

[16] C. Wang, X. Dong, L. Yu, L. Ye, W. Zhuang, and F. Ma, "Prediction of days in hospital for children using random forest," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, (Shanghai), pp. 1–6, IEEE, Oct. 2017.

[17] H. I. Salmons, Y. Lu, R. R. Reed, B. Forsythe, and A. S. Sebastian, "Implementation of Machine Learning to Predict Cost of Care Associated with Ambulatory Single-Level Lumbar Decompression," *World Neurosurgery*, vol. 167, pp. e1072–e1079, Nov. 2022.

[18] A. Orooji, M. Shanbehzadeh, E. Mirbagheri, and H. Kazemi-Arpanahi, "Comparing artificial neural network training algorithms to predict length of stay in hospitalized patients with COVID-19," *BMC Infectious Diseases*, vol. 22, p. 923, Dec. 2022.

[19] A. Rivera-Sepulveda, T. Maul, K. Dong, K. Crate, T. Helman, C. Bria, L. Martin, K. Bogers, J. W. Pearce, and T. F. Glass, "Effect of the COVID-19 Pandemic on the Pediatric Emergency Department Flow," *Disaster Medicine and Public Health Preparedness*, vol. 17, p. e83, 2023.

[20] A. Guo, J. Lu, H. Tan, Z. Kuang, Y. Luo, T. Yang, J. Xu, J. Yu, C. Wen, and A. Shen, "Risk factors on admission associated with hospital length of stay in patients with COVID-19: a retrospective cohort study," *Scientific Reports*, vol. 11, p. 7310, Mar. 2021.

[21] L. Breiman, "[No title found]," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell, "NextPlace: A Spatio-temporal Prediction Framework for Pervasive Systems," in *Pervasive Computing* (K. Lyons, J. Hightower, and E. M. Huang, eds.), vol. 6696, pp. 152–169, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[23] BPJS, "Laporan pengelolaan program dan laporan keuangan jaminan sosial kesehatan," tech. rep., BPJS, Indonesia, 2019.