Hospital Length of Stay Prediction with Ensemble Learning Methode

Dian Puspita*, Waras Lumandi, Arief Rachman
Institut Teknologi Adhi Tama Surabaya, *Surabaya*

Email: *dian.puspita@itats.ac.id

Received: 2023-04-10 Received in revised from: 2023-05-22 Accepted: 2023-06-05

Abstract

The hospital length of stay (LoS) is the number of days an inpatient will stay in the hospital. LoS is used as a measure of hospital performance so they can improve the quality of service to patients better. However, making an accurate estimate of LoS can be difficult due to the many factors that influence it. The research conducted aims to predict LoS for treated patients (ICU and non-ICU) with cases of brain vessel injuries by using the ensemble learning method. The Random Forest algorithm is one of the ensembles learning methods used to predict LoS in this study. The dataset used in this study is primary data at PHC Surabaya Hospital. From the results of the simulations performed, the random forest algorithm is able to predict LoS in a dataset of treated patients (ICU and non-ICU) with cases of brain vessel injuries. And the simulation results show a type II error value of 0.10 while the value of type I error is 0.16.

Keywords: The Hospital Length of Stay; Machine Learning; Ensemble Learning; Random Forest

1. Introduction

Private hospitals carry out intensive monitoring of patient care with the aim of minimizing losses and increasing the effectiveness of health services according to the established Clinical Pathway (CP). If you don't have CP for every case of disease, then you can predict the Length of Stay (LoS) of each patient who enters with a case of the disease for monitoring the patient being treated. Hospital length of stay (LoS) is the number of days an inpatient will stay in the hospital. The longer the patient is treated, the greater the costs incurred by the Hospital, while the cost of claims that will be received by the Hospital remains at the rate set. LoS is used as a measure of hospital performance so that they can better improve the quality of service to patients [1]. However, making an accurate LoS estimate can be difficult due to the many factors that influence it.

This study aims to predict LoS for treated patients (ICU and non-ICU) with cases of brain vessel injuries by using the ensemble learning method. The ensemble method is a supervised learning method and part of machine learning where this algorithm is used as a search for predictive solutions. Included in this ensemble method is the Random forest algorithm [2]. The Random Forest algorithm is one of the ensemble learning methods used to predict LoS in this study. The data group used is primary data at the PHC Surabaya hospital. PHC Surabaya Hospital is one of the private hospitals under the management of a BUMN Company (State Owned Enterprise) which provides comprehensive and integrated services [3].

Prediction activities begin by collecting data and data preparation, then building a model of the Random Forest classifier algorithm. It is continued at the model evaluation stage using training data. Then the model generated by the Random Forest classifier will be applied to the testing data [4]. Continues on the confusion matrix values that will be used to calculate performance measures such as Precision, Recall, MCC, and F1 Score. Performance measurement with Precision, the smaller the False Positive (FP), makes the precision value even greater. As for Recall, the smaller the False Negative (FN) makes the recall value bigger. The Matthews correlation coefficient (MCC) is a performance measure for classification analysis with binary targets/classes. MCC will produce a high score only if the binary predictor is able to predict correctly. The last measure is harmonic mean of precision and recall by using F1 score. The best value of F1-Score is 1.0 and the worst value is 0.

2. Method

According to WHO, ischemic heart disease is the most common cause of death worldwide. Since 2000, there has been an increase in cases of mortality due to this disease to 8.9 million in 2019. Ischemic heart disease is a disease that is the first topic of discussion for early prevention through various conventional approaches. Ischemic heart disease, also known as coronary heart disease, is a problem caused by narrowing of the heart arteries. As a result, blood and oxygen reaching the heart decreases. And can lead to blood vessel injuries in the brain [5].

Patients who have injured blood vessels in the brain require intensive care in a hospital [6]. This will be related to the classification of Tarif Standards for First Level Health Services and Advanced Health Facilities. In accordance with PMK number 64 of 2016, ischemic heart disease can be included in the category of Ina-CBG's G-4-14 code with a description of brain vessel injuries with INFARK. The INA-CBG's from the government stipulates a group of diagnostic procedures or service packages that are used as a financing basis for health service provider operators (Hospitals) as quality control for cost control [7].

Dataset treated patients (ICU and non-ICU) with cases of brain vessel injuries which is owned by the PHC Surabaya Hospital. In this prediction activity the dataset used has a binary target labeled loss claims and profit claims. There are 19 features in this dataset variable. Then, feature reduction is performed so that the features used are 7 features. These seven features include Sex, Age, LoS, rate, hospital rate, age group, INA-CBG.

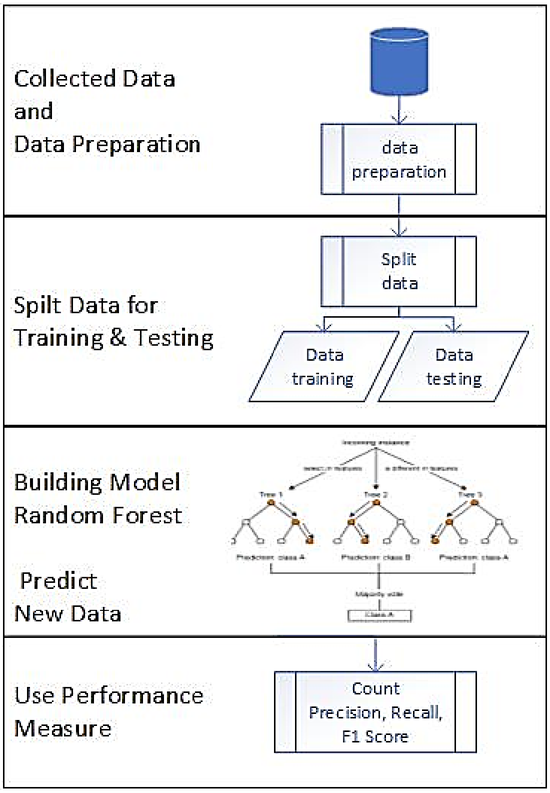


Figure 1. Classification analysis

The Random Forest algorithm is one of the ensembles learning methods used to predict LoS in this study. Random forest algorithm is also known as bootstrap aggregation, the parallel combining ensemble of multiple classifier models [8]. The computation of each model is important and there are basically two different techniques called bagging and boosting [9]. Random forest is one example for bagging, Bagging is also known as bootstrapping aggregation. Tree Bagging or bootstrap aggregation, is the process of selecting a sample of observations (data rows) randomly, determined by 3 key steps: first, Build n decision trees, by selecting n samples of observations at random; Train each decision tree; In order to make predictions on new data, each of the n trees must be used, and the majority is determined from the n prediction [10].

Feature sampling is a process of randomly choosing variables (columns of data). By default, n variables are selected for a problem with n variables in total from the root of the decision tree. This process makes it possible to weaken the correlation between the decision trees which could interfere with the quality of the results. In statistics, we say that the feature sampling makes it possible to reduce the variance of the data set created.

We can see figure 1 explain the classification analysis and it begin with collect data and preparation, then split data for training and testing. A training set D, Inducer I and the number of bootstrap samples m as input. Generate a classifier C^* as output
1. Create m new training sets D_i, from D with replacement
2. Classifier C_i is built from each set D_i using I to determine the classification of set D_i
3. Finally classifier $C^*$is generated by using the previously created set of classifiers C_i on the original data set D, the classification predicted most often by the sub-classifiers C_i is the final classification.

A decision tree creates sub-populations by successively dividing the leaves of a tree. There are different separation criteria for building a tree: The Gini criterion organizes the separation of the leaves of a tree by focusing on the class most represented in the data set. This must be separated as soon as possible. The entropy criterion is based on the measurement of the prevalent disorders (as in thermodynamics) in the studied population. The construction of the tree aims to lower the global entropy of the tree leaves at each stage [11][12].

In real-world classification problems, it is usually impossible for the model to be 100% correct. Therefore, the type of prediction activity carried out by the classifier algorithm to predict the target or class is not 100% possible for real problems [13]. Then the performance evaluation activity of the model produced by the classifier algorithm is carried out. What we want to know is how wrong the model is in predicting and how wrong the model is when predicting. In this study trying to optimize the model to work better in predicting data, it is done by using different performance metrics to choose the best model[14][15] .

Measuring the performance of classifier models is generally indirect and highly dependent on the cases and available datasets [16]. Using Precision, Recall, F1 Score, and MCC is very important to understand the risk of error so that it can produce a truly useful model [17]. For LoS-related datasets, what needs to be considered is the type I error value and type II error value which can be seen from the confusion matrix.

The F1 score, we have this definition of precision and we have this definition of sensitivity through positivity or recall [18]. Remember actually counting how many actual positives our model captures via labelling them as true positives or true positives. Precision on the other hand talks about how precise is how accurate your model is from the predicted positives or actually how many of them are actually positive [19][20].

Memory and Precision are important and we cannot have high values of memory and precision. At the same time because we cannot have a decision boundary that separates the two classes perfectly. There is a trade-off between precision and recall, a higher level of recall can be obtained at the cost of a lower value of precision. To solve this, we want to define one mud that combines the two to evaluate the performance of the classifier. Some combined measures of precision and recall; F1 score, Matthew's Correlation Coefficient (MCC), 11-point average precision, and Breakeven point.

F1 Score is nothing but just a weighted harmonic means of recall and precision. Generalized f score is called f beta. If beta has a parameter and that is a weight assigned to in this recall.

$$F_\beta = \frac{1+\beta^2}{\frac{1}{Precision}+\frac{\beta^2}{Recall}}, \beta \geq 0 \quad (1)$$

For $\beta = 1$, we have harmonic means of precision and recall, that is

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

$$= \frac{2\,(Precision)(Recall)}{(Precision) + (Recall)}$$

$$= \frac{2TP}{2TP + FP + FN}$$

Beta measures the effectiveness of classification, so beta time as much importance to recall as precision because beta is coming with recall. We can say recall is considered beta times as important as precision. This is nothing but just a weighted harmonic means of recall and precision. When beta can be any value greater than equal to zero, when beta is equal to zero and beta score is precision only when beta is equal to infinity. So we only get recall, so commonly use the value of beta is equal to one and what we term as beta is F1 score. Why do we used harmonic means, because harmonic mean is preferred as it penalizes model from the classifier, or is less than or equal to geometric means and that's less than equal to arithmetic mean for two numbers.

Matthew's Correlation Coefficient - MCC that combine precision and recall [21][22]. We are already know that precision, recall and F1 score they are all asymmetric. Get a different result if the classes are switched. It means if we change or swap classes, we will get a different result or score for different value for this precision and different value for recall. MCC determines the correlation between true class and predicted class. The higher the correlation between true and predicted value, the better the prediction. MCC as an alternative measure that is unaffected by the problem of unequal data sets, the Matthews correlation coefficient is a contingency matrix between actual and predicted values.

Defines as

$$, MCC = \frac{TP\ TN - FP\ FN}{\sqrt{(TP+FN)(TP+FP)(TN+FN)(TN+FP)}} \qquad (2)$$

MCC=1 when FP=FN=0 is perfect classification, MCC=-1 when TP=TN=0 is perfect misclassification, and if MCC=0 performance of classifier is not better than a random classifier.

3. Result And Discussion

Simulation Result

In this study, a simulation was carried out on the data group treated patients (ICU and non-ICU) with cases of brain vessel injuries with a total of 14,455 patient data. The data consisted of sex features, 7890 female patients and 6565 male patients. Age group features include early adulthood, late adulthood, early age, elderly, and old. Then the LoS group features such long, medium, and fast.

The model generated by the random forest classifier gives a true positive result (TP) of 0.90 then 0.84 for a true negative (TN). For false positive (FP) or type I errors 0.10 and 0.16 for false negatives or type II errors.

In this case this value means that the RF classifier algorithm is able to predict 90% correctly for Los which benefit the hospital. RF classifier algorithm is able to predict 84% correct for LoS which is detrimental to the hospital. Type II error is positive data but is predicted as negative data. In the case of this study, patients needed LoS but the model made predicted that these patients did not need LoS.For the wrong predictive value of 10% for LoS that benefits the hospital. As well as the wrong prediction value of 16% for Los which is detrimental or a financial loss to the hospital.

Table 1. Confusion Matrix

	TP	TN	FP	FN
RF	0,900	0,840	0,100	0,160
SVM	0,672	0,697	0,388	0,308
NN	0,865	0,833	0,135	0,167

We also make comparisons between the RF ensemble learning algorithm and other classifier algorithms such as Neural Net and SVM. Table 1 shows the simulation results of these comparisons. And it can be seen in the table that ensemble learning has the lowest FP and FN values.

$$Precision = \frac{TP}{TP + FP} = 0{,}90$$

$$Recall = \frac{TP}{TP + FN} = 0{,}84$$

$$F1\ Score = \frac{2TP}{2TP + FP + FN} = 0{,}87$$

$$MCC = \frac{TP \times TN - FP\ \times FN}{\sqrt{(TP + FP)}(TP + FN)(TN + FP)(TN + FN)} = 0{,}56$$

Table 2 Performance Measure

	Precision	Recall	F1 Score	MCC
RF	0,872	0,87	0,87	0,56
SVM	0,682	0,68	0,676	0,328
NN	0,865	0,838	0,85	0,999

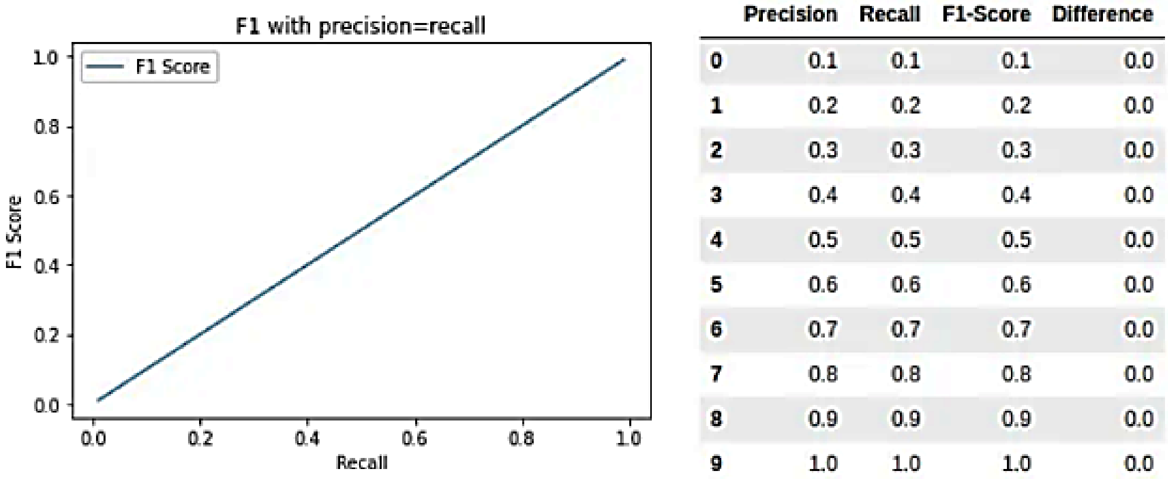
Furthermore, we also calculate the performance of the algorithm classifier with performance measures such as Precision, Recall, F1 Score, and MCC or Matthew's Correlation Coefficient. Our goal is to use a performance measure to find out how far the classifier algorithm's performance is in correctly predicting new data.

Analysis and Discussion

Performance measure precision based on table 2, is an indicator of how many positive predictions are made true (true positive). How precise or how accurately the model predicts true positives. The precision should ideally be 1 (high) for a model to produce a good classifier algorithm. The precision becomes 1 only if the numerator and denominator are the same, namely TP = TP+FP, this also means that FP is zero. As FP increases, the denominator becomes larger than the quantifier and the precision decreases. For Ensemble Learning with the RF algorithm it has a precision value close to one with a value of 0.87. For other classifier algorithms, NN has a precision value that is close to one with a value of 0.86 and finally SVM has a value farthest from one, namely 0.68.

Performance measure recall based on table 2, is an indicator of how many positive cases the classifier predicts correctly, for all positive cases in the data. Recall will be good if it has a value that must be 1 for the model to produce a good classifier algorithm. Actually calculates how many of the actual positives our model capture through labelling it as positive or true positive. Recall becomes 1 only if the numerator and denominator are the same, namely TP = TP +FN, meaning FN is zero. When FN increases the value of the denominator becomes larger than the numerator and the value of the withdrawal decreases.

So ideally the model can classify properly, precision and recall have a value of one which means that FP and FN have a value of zero. When the precision and recall values are the same and close to one, a performance measure is needed that is able to describe the harmonic average values of precision and recall. Learning with the RF algorithm has a recall value close to one with a value of 0.87. For other classifier algorithms, the NN recall value is close to one with a value of 0.83 and finally the SVM value is farthest from one, namely 0.68.



	Precision	Recall	F1-Score	Difference
0	0.1	0.1	0.1	0.0
1	0.2	0.2	0.2	0.0
2	0.3	0.3	0.3	0.0
3	0.4	0.4	0.4	0.0
4	0.5	0.5	0.5	0.0
5	0.6	0.6	0.6	0.0
6	0.7	0.7	0.7	0.0
7	0.8	0.8	0.8	0.0
8	0.9	0.9	0.9	0.0
9	1.0	1.0	1.0	0.0

Figure 2. The F1-score if precision equals recall

Performance measure F1 score is an indicator that combines precision and recall values. The F1 score is a harmonic average and is just another way of calculating the average of values, generally described as more appropriate with ratio designations such as precision and recall. As seen in Figure 2, the figure shows that the precision and recall values have the same value, so the F1 score appears as a linear line on the graph. F1 score is the harmonic mean of precision and recall and is a better measure than accuracy. Learning with the RF algorithm has an F1 score close to one with a value of 0.87. For other classifier algorithms, the NN F1 score is the closest to one with a value of 0.85 and finally the SVM has the farthest value from one, namely 0.67.

For the last performance measure MCC is the only binary classification rate that generates a high score only if the binary predictor was able to correctly predict the majority of positive data instances and the majority of negative data instances. Learning with the RF algorithm has an MCC close to one with a value of 0,56. For other classifier algorithms, the NN MCC is the closest to one with a value of 0,99 and finally the SVM has the farthest value from one, namely 0.32.

Overall, of the four performance measures used for ensemble learning simulations with the RF algorithm, they produce good predictive values. Comparisons have been made with other algorithm classifiers such as NN and SVM, the RF algorithm still produces the best predictive value.

4. Conclusions

The performance of the Random forest algorithm in predicting patient care dataset (ICU and non-ICU) with cases of brain vessel injuries reported by the studies, identified in this review supports the need for further research on the usefulness of machine learning in particular the ensemble method in predicting LoS in medical patient. Future studies should develop a methodology that involves optimization methods to minimize the value of type I error and type II error value. The complexity of the LoS data group and the amount of electronic data collected is very large at this time, it really requires a prediction method that is reliable and can be widely applied.

Acknowledgements

We thank you for the permission given to use the patient care data group (ICU and non-ICU) with cases of BRAIN VESSEL INJURIES from PHC Surabaya Hospital.

Referensi

[1] S. Bacchi, Y. Tan, L. Oakden-Rayner, J. Jannes, T. Kleinig, and S. Koblar, “Machine learning in the prediction of medical inpatient length of stay,” *Internal Medicine Journal*. 2022. doi: 10.1111/imj.14962.

[2] A. B. Shaik and S. Srinivasan, “A brief survey on random forest ensembles in classification model,” in *Lecture Notes in Networks and Systems*, 2019. doi: 10.1007/978-981-13-2354-6_27.

[3] V. Lequertier, T. Wang, J. Fondrevelle, V. Augusto, and A. Duclos, “Hospital Length of Stay Prediction Methods: A Systematic Review,” *Medical Care*, vol. 59, no. 10. 2021. doi: 10.1097/MLR.0000000000001596.

[4] R. T. Disler *et al.*, “Factors impairing the postural balance in COPD patients and its influence upon activities of daily living,” *Eur. Respir. J.*, vol. 15, no. 1, 2019.

[5] L. Su *et al.*, “Early Prediction of Mortality, Severity, and Length of Stay in the Intensive Care Unit of Sepsis Patients Based on Sepsis 3.0 by Machine Learning Models,” *Front. Med.*, vol. 8, Jun. 2021, doi: 10.3389/fmed.2021.664966.

[6] 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,” *Int. J. Cardiol.*, vol. 288, 2019, doi: 10.1016/j.ijcard.2019.01.046.

[7] J. Chrusciel, F. Girardon, L. Roquette, D. Laplanche, A. Duclos, and S. Sanchez, “The prediction of hospital length of stay using unstructured data,” *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, 2021, doi: 10.1186/s12911-021-01722-4.

[8] T. K. Ho, “Random decision forests,” *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, vol. 1, pp. 278–282, 1995, doi: 10.1109/ICDAR.1995.598994.

[9] A. Murugan, S. A. H. Nair, and K. P. S. Kumar, “Detection of Skin Cancer Using SVM, Random Forest and kNN Classifiers,” *J. Med. Syst.*, vol. 43, no. 8, 2019, doi: 10.1007/s10916-019-1400-8.

[10] T. Ho, Kam, “The Random Subspace Method for Constructing Decision Forests,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, 1998, [Online]. Available: %3CGo%0Ato

[11] G. Nanfack, P. Temple, and B. Frénay, “Constraint Enforcement on Decision Trees: A Survey,” *ACM Comput. Surv.*, 2022, doi: 10.1145/3506734.

[12] B. Charbuty and A. Abdulazeez, “Classification Based on Decision Tree Algorithm for Machine Learning,” *J. Appl. Sci. Technol. Trends*, 2021, doi: 10.38094/jastt20165.

[13] I. M. De Diego, A. R. Redondo, R. R. Fernández, J. Navarro, and J. M. Moguerza, “General Performance Score for classification problems,” *Appl. Intell.*, 2022, doi: 10.1007/s10489-021-03041-7.

[14] M. Wever, A. Tornede, F. Mohr, and E. Hullermeier, “AutoML for Multi-Label Classification: Overview and Empirical Evaluation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2021. doi: 10.1109/TPAMI.2021.3051276.

[15] R. G. Patel, N. A. Trask, M. A. Gulian, and E. C. Cyr, “A block coordinate descent optimizer for classification problems exploiting convexity,” in *CEUR Workshop Proceedings*, 2021. doi: 10.2172/1859695.

[16] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, “Assessing the accuracy of prediction algorithms for classification: An overview,” *Bioinformatics*, vol. 16, no. 5. 2000. doi: 10.1093/bioinformatics/16.5.412.

[17] X. Zhang, Y. Li, P. Wang, and X. Tan, “ClassificationAlgorithm of speech data of Parkinson’s disease based on convolution sparse kernel transfer learning with optimal kernel and parallel sample/feature selection,” *arXiv*, 2020.

[18] J. Mohajon, “Confusion Matrix for Your Multi-Class Machine Learning Model,” *Towardsdatascience.Com*, 2020.

[19] D. J. Hand, P. Christen, and N. Kirielle, “F*: an interpretable transformation of the F-measure,”

Mach. Learn., 2021, doi: 10.1007/s10994-021-05964-1.

[20] A. Chan and J. A. Tuszynski, “Automatic prediction of tumour malignancy in breast cancer with fractal dimension,” *R. Soc. Open Sci.*, 2016, doi: 10.1098/rsos.160558.

[21] M. Singh, R. Divakaran, L. S. K. Konda, and R. Kristam, “A classification model for blood brain barrier penetration,” *J. Mol. Graph. Model.*, 2020, doi: 10.1016/j.jmgm.2019.107516.

[22] A. Al-Ramini *et al.*, “Machine Learning-Based Peripheral Artery Disease Identification Using Laboratory-Based Gait Data,” *Sensors*, 2022, doi: 10.3390/s22197432.