ILKOM Jurnal Ilmiah Vol. 14, No. 3, December 2022, pp. 186-193
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

Analysis of stroke classification using Random Forest method

Muhammad Firdaus Banjar [a,1,*]; Irawati [a,2]; Fitriyani Umar [a,3]; Lilis Nur Hayati [a,4]
[a] *Universitas Muslim Indonesia, Jl. Urip Sumoharjo KM.5 , Makassar and 90231, Indonesia*
[1] *agam5599@gmail.com;* [2] *irawan2804@gmail.com;* [3] *fitriyani.umar@umi.ac.id;* [4] *lilis.nurhayati@umi.ac.id*
** Corresponding author*

Article history: Received June 15, 2022; Revised August 20, 2022; Accepted November 15, 2022; Available online December 20, 2022

Abstract

Stroke is a disease in which the sufferer experiences or experiences a rupture of a blood vessel in the brain so that the brain does not get a blood supply that provides oxygen. Patients who suffer from stroke will experience cognitive disorders ranging from decreased consciousness, visuospatial disorders, non-verbal learning disorders, communication disorders, and reduced levels of patient attention. Data from the World Stroke Organization shows that there are 13.7 million new stroke cases every year, and about 5.5 million deaths occur due to stroke. This research aims to analyze the attributes of any variables that affect the classification of strike disease and to test the performance of stroke classification in the form of accuracy, precision, recall, and f-measure. The method used is a random forest using a tree, namely 50, 100, 200, and 500. The classification of stroke is divided into stroke and no stroke. The data used is 5110, divided into 70% training data and 30% testing data. The results showed that the performance of a random forest using 100 trees was better than using 50, 200, and 500 trees, with an accuracy value of 86.82%, a precision of 15.76%, a recall of 38.15%, and an f1-score 22.30% after doing SMOTE

Keywords: Attribute; Random Forest; SMOTE; Strokes

Introduction

Stroke is a disease in which the sufferer experiences a blockage or rupture of blood vessels in the brain so that the brain does not get the blood supply that carries oxygen [1]. Patients affected by stroke will experience cognitive disorders ranging from decreased consciousness, visuospatial disorders, non-verbal learning disorders, communication disorders, and decreased patient attention levels [2]. Stroke always attacks sufferers suddenly, regardless of age and gender. World Stroke Organization data shows that every year there are 13.7 million new stroke cases, and around 5.5 million deaths occur. Approximately 70% of strokes and 87% of deaths and disabilities due to strokes occur in low and middle-income countries [3].

With the high cases of mortality and disability due to stroke, it is necessary to analyze and study the possibility of a stroke based on the attributes contained in the dataset of stroke sufferers. The classification method used in this research is Random Forest. Random Forest is one of the methods used for classification and regression by using many trees in making a decision [4]. The Random Forest method is used to test performance in the form of accuracy, precision, recall, and f-measure, and the use of SMOTE (Synthetic Minority Oversampling Technique) method is used to overcome imbalance problems in the dataset [5].

Method

A. *Random Forest*

Random Forest is a development of the Classification and Regression Trees (CART) method [6]. Random Forest applies the concept of Bootstrap and Aggregation (BAGGING) in making models and decisions [7]. This classification method is a combination of each tree (tree) which is combined into one model during training. This classification's results are determined by voting based on the class resulting from the tree [8]. The flow of the random forest algorithm can be seen in **Figure 1**.

http://dx.doi.org/10.33096/ilkom.v14i3.1252.186-193

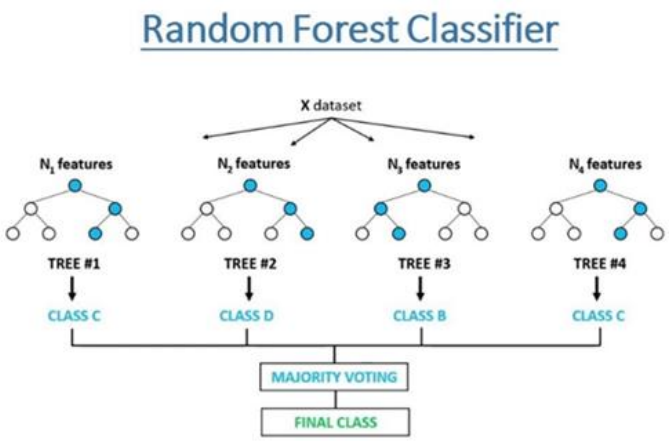

Figure 1. Classification Flow of the Random Forest Algorithm [15]

This method's tree construction consists of the roots, internals, and leaf nodes. The construction of this tree begins by calculating the entropy value as a determinant of the attribute's impurity level. The formula for finding entropy values can be seen in **Equation 1**[9].

$$Entropy(S) = \sum_{i}^{c} - p_i \, log_2 \, p_i \qquad (1)$$

Where c is the number of classes while pi represents the portion of data objects (samples) in class i to the total number of samples in the data set. After obtaining the entropy value, proceed with finding the information gained for the node-splitting process. Information gain is carried out to find variables that will be selected to grow trees [9]. The formula for finding the value of information gain can be seen in **Equation 2**.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \, Entropy(S_v) \qquad (2)$$

Where A is Attribute, V denotes a possible value for attribute A, |Sv| is the number of samples for the value v, |S| represents the total number of data samples, and Entropy(Sv) represents the entropy for samples that have a value of v. From **Figure 2** it can be seen that there are several stages in classifying the random forest method, while the stages in the random forest including [10]:

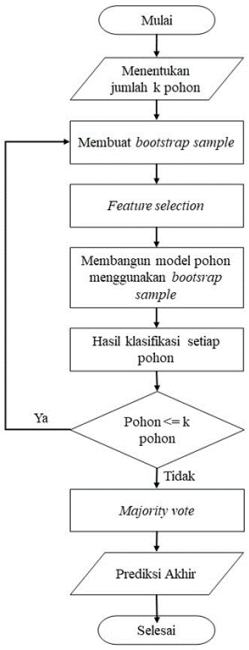

Figure 2. Flowchart of the random forest algorithm

1. Determine the number of k trees to be formed.
2. Create bootstrap samples by taking n samples from the initial dataset using random sampling with replacement techniques.
3. Choose m variables randomly from p variables, where m ≤ p. This stage is the feature selection stage.
4. Build a tree model (tree) using bootstrap samples and predict classification.
5. Repeating steps 1-4 until several trees are obtained. The repetition is done k times.
6. Determining the final prediction with a majority vote based on the prediction results of each tree model.

a. Misclassification Level (Out Of Bag)
The accuracy of the random forest is calculated using the Out-of-Bag (OOB). OOB is data in the dataset but not in the bootstrap sample. By testing the data for each tree, the percentage of correct prediction data will be seen by accumulating from each tree. Conversely, the error rate is known by calculating the percentage of wrong predictions from the aggregation of the entire tree [11].

b. Variable Importance
One of the most important random forest features is the output of Variable Importance. Variable Importance measures the degree of relationship between particular variables and the classification results. To estimate the Variable Importance for some variable j, an out-of-bag (OOB) sample is lowered into a tree, and the prediction accuracy is recorded. Then the value of the variable j is changed in the OOB sample, and the accuracy is measured again. This calculation is done tree by tree when the random forest is built. This average decrease in permutation accuracy is then averaged across all trees and used to quantify the Importance of variable j [12].

B. *Synthetic Minority Oversampling Technique* (SMOTE)

Synthetic Minority Oversampling Technique (SMOTE) is one of the derivatives of oversampling. SMOTE was first introduced by Nithes V. Chawla. This approach works by creating a replication of minority data. This replication is known as synthetic data. The SMOTE method works by finding the k-nearest neighbors (i.e. k closest data neighbors) for each data in the minority class, after which synthetic data is created as much as the desired percentage of duplication between the minority data and the k-nearest neighbors which are selected randomly [5].

Results and Discussion

A. Dataset

The dataset was taken from the Kaggle repository with a total of 5110 data, 2 classes, and 12 attributes, including id, gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, bmi, smoking_status, stroke [14]. Class 0 (no stroke) in the stroke attribute is 4860, while class 1 (stroke) is 246. Based on this amount of data, it can be seen that the difference in distance between class 0 and class 1 is very large, so this dataset is imbalanced. **Figure 3** shows streaks of data after being visualized using a pie chart.

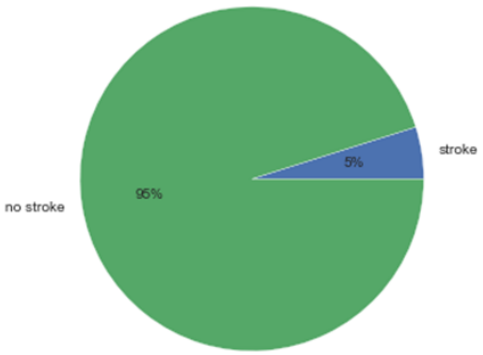

Figure 3. Visualization of the number of stroke and non-stroke patient data

B. Testing

1. Variable Importance

One of the outputs in the random forest is variable importance. This variable importance shows which variables influence the prediction of stroke. The following is a visualization of variable importance.

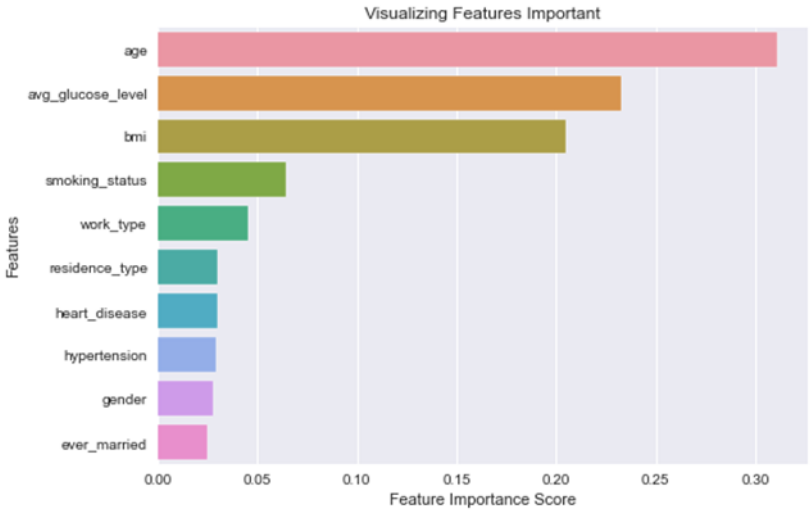

Figure 4. Variable importance of the stroke dataset

Based on **Figure 4**, three variables or attributes have a major influence on stroke: age, avg_glucose_level (average blood sugar level), and BMI (body mass index). The following is a description of the age, avg_glucose_level, and BMI attributes based on **Figure 4**.

a. Age

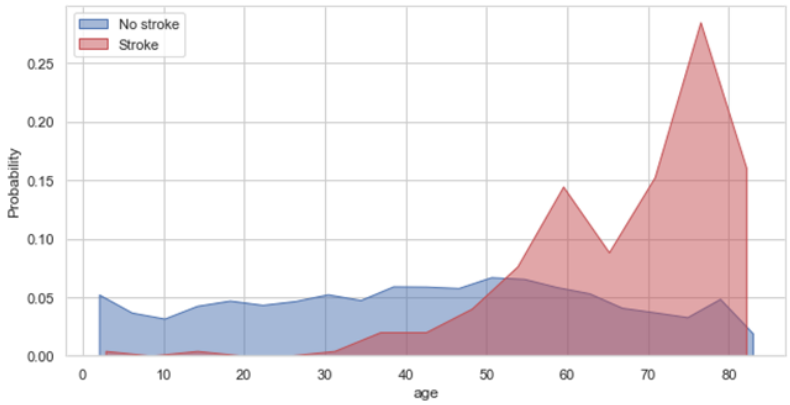

Figure 5. Probability of stroke by age

Figure 5 shows the relationship between stroke and age. The older you are, the more likely you will have a stroke. Based on **Figure 5**, at the age of 30, there is an increased chance of a stroke.

b. Average glucose levels

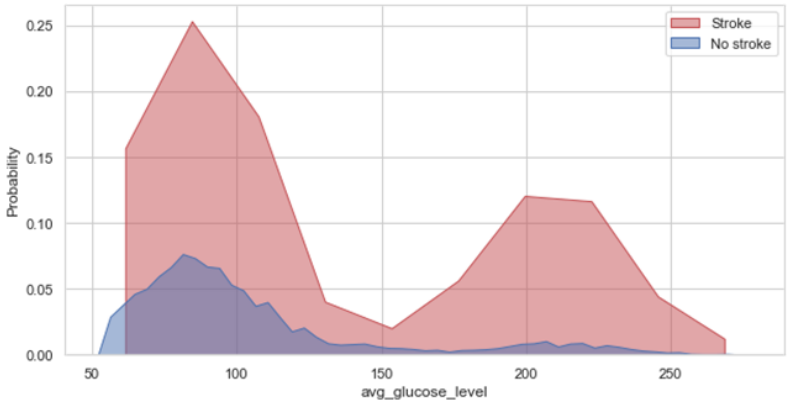

Figure 6. Probability of stroke based on average blood sugar levels

Figure 6 shows the relationship between blood sugar levels and stroke. Based on figure 6 blood sugar above 56 mg/dl has the possibility of having a stroke. Blood sugar above 100 mg/dl has decreased because the number of classes in the stroke patient dataset is not balanced.

c. BMI (*Body Mass Index*)

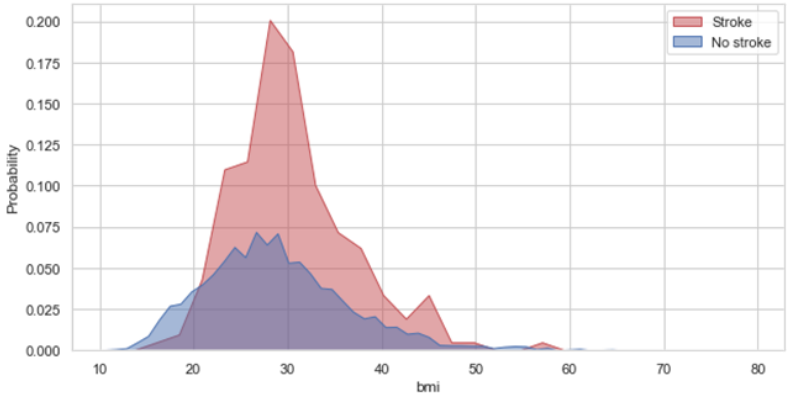

Figure 7. Probability of stroke by BMI

Figure 7 shows the relationship between Body Mass Index (BMI) and stroke. BMI is a measure used to assess the proportionality of the comparison between height and weight. Based on **Figure 7**, it can be concluded that BMI between 25-60 tends to get a stroke. BMI above 30 has a significant decrease. This is due to the unbalanced number of stroke patient data.

2. Performance

This discussion will use a confusion matrix to explain the performance results, namely accuracy, precision, recall, and f1-score. The confusion matrix is a method used in evaluating the classification model. In the calculation, four combinations of predicted and actual values exist. The four values are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [13]. The performance calculation begins by setting the TP, FP, and FN values.

a. If the testing data with class "stroke" is classified as "stroke," it is counted as TP.
b. If the testing data with class "stroke" is classified as "no stroke," it is calculated as FN.
c. If the testing data with class "no stroke" is classified as "stroke," it is counted as FP.
d. If the testing data with class "no stroke" is classified as "no stroke," then it is counted as TN

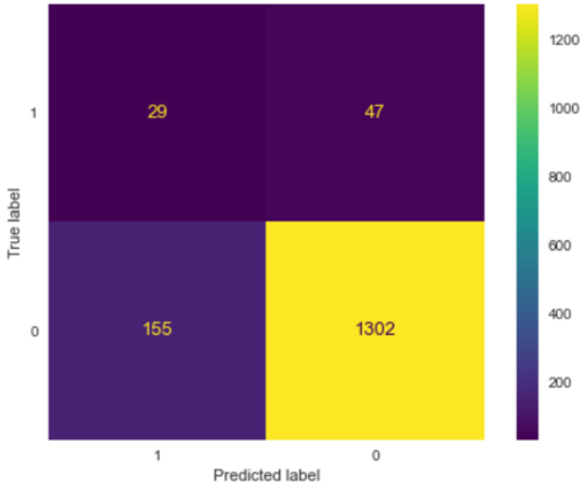

Figure 8. Confusion matrix random forest using 100 trees.

Based on the results of the confusion matrix in **Figure 8**, the calculations for accuracy, precision, recall, and f-measure are as follows.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \text{ x } 100\%$$

$$\text{Accuracy} = \frac{29 + 1302}{29 + 47 + 155 + 1302} \text{ x } 100\%$$

$$\text{Precision} = \frac{TP}{TP + FP} \text{ x } 100\%$$

$$\text{Precision} = \frac{29}{29 + 155} \text{ x } 100\%$$

$$\textit{Recall} = \frac{TP}{TP + FN} \text{ x } 100\%$$

$$\textit{Recall} = \frac{29}{29 + 47} \text{ x } 100\%$$

$$\textit{f-measure} = 2 \text{ x} \frac{\text{precision x } recall}{\text{precision} + \text{recall}} \text{ x } 100\%$$

$$\textit{f-measure} = 2 \text{ x} \frac{14{,}50 \text{ x } 36{,}84}{14{,}50 + 36{,}84} \text{ x } 100\%$$

The performance results from a random forest using 50, 100, 200, and 500 trees can be seen in **Table 1**.

Tree	**Performance**			
	Accuracy (%)	**Precision (%)**	***Recall* (%)**	***F1-Score* (%)**
50	86.49	14.59	35.52	20.68
100	86.82	15.76	38.15	22.30
200	86.30	13.58	32.89	19.23
500	86.49	14.97	36.84	21.29

Based on **Table 1**, the best performance is obtained by using 100 trees where the accuracy reaches 86.82%, precision 15.76%, recall 38.15% and f1-score 22.30%. Even though it has quite high accuracy, the model cannot be said to be good because it has a low f1-score. The f1-score value is obtained from the results of precision and recall calculations and the cause of the low precision and recall values is due to the large number of incorrectly predicted data.

3. *Out-Of-Bag*

Out-Of-Bag is obtained from testing data that is not included in the bootstrap sample and is used in determining the OOB. The smaller the OOB value, the better the model. The following is a visualization of the OOB value.

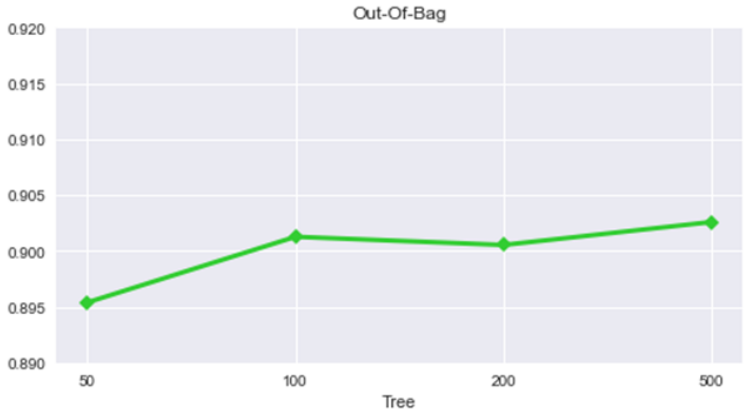

Figure 9. Graph of out-of-bag scores

Based on **Figure 9**, OOB gets a high score of 89.53% -90.25%. **Figure 6** shows that the OOB score does not decrease with an increase in the number of trees. So it can be concluded that the random forest performance in classifying is still not good.

Conclusion

The conclusion is that the attributes or variables that influence the most in classifying stroke are age, average sugar levels, and BMI. And the random forest performance in classifying stroke based on the number of trees used obtains an accuracy of 86.82%. , 15.76% precision, 38.15% recall, and 22.30% f1-score using 100 trees. From these results, although it has excellent accuracy, it has a low f1-score, so it cannot be said to be good. The suggestions are to try other ensemble learning methods, such as Bagging and Boosting, and test the method's performance by applying different neighbor values to the SMOTE method.

References

[1] V. Adelina, D. E. Ratnawati, and M. A. Fauzi, "Klasifikasi tingkat risiko penyakit stroke menggunakan metode GA-Fuzzy Tsukamoto," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 2, no. 9, pp. 3015–3021, 2018.

[2] M. S. Fadlilah, R. C. Wihandika, and B. Rahayudi, "Klasifikasi penurunan fungsi kognitif pasien stroke menggunakan metode klasifikasi Random Forest," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 3, pp. 3005–3013, 2019.

[3] R. Aprianda, "Stroke: Don't be the One," *Pusat Data dan Informasi Kementrian Kesehatan RI*, pp. 1–10, 2019.

[4] L. Breiman, "Random Forests," *Kluwer Academic Publishers*, pp. 6–32, 2021.

[5] R. Siringoringo, "Klasifikasi data tidak seimbang menggunakan algoritma SMOTE dan K-Nearest Neighbor," *Journal Information System Development (ISD)*, vol. 3, no. 1, pp. 44–49, 2018.

[6] L. Andiani, Sukemi, and D. P. Rini, "Analisis penyakit jantung menggunakan metode KNN dan Random Forest," *Prosiding Annual Research Seminar 2019 Computer Science and ICT*, vol. 5, no. 1, pp. 165–169, 2019.

[7] M. R. Amiarrahman and T. Handhika, "Analisis dan Implementasi algoritma klasifikasi Random Forest dalam pengenalan Bahasa Isyarat Indonesia (BISINDO)," *Seminar Nasional Inovasi Teknologi*, pp. 83–88, 2018.

[8] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, "Prediksi kemungkinan diabetes pada tahap awal menggunakan algoritma klasifikasi Random Forest," *SISTEMASI: Jurnal Sistem Informasi*, vol. 10, no. 1, pp. 163–171, 2021.

[9] L. Fadilah, "Klasifikasi Random Forest pada data imbalanced," Skripsi, Universitas Islam Negeri Syarif Hidayatullah , Jakarta, 2018.

[10] Nidhomuddin and B. W. Otok, "Random Forest dan Multivariate Adaptive Regresion Spline (MARS) binary response untuk klasifikasi penderita HIV/AIDS di Surabaya," *Jurnal Statistika Universitas Muhammadiyah Semarang*, vol. 1, no. 3, pp. 59–57, 2015.

[11] A. W. A. Ruslam, "Analisis keberlanjutan pengguna jala menggunakan factor analysis," Skripsi, Universitas Islam Indonesia, Yogyakarta, 2021.

[12] Mustikasari and ST. A. D. Ghani, "Analisis performa klasifikasi algoritma pada pendeteksian penyakit kanker dengan partition membership," *In SISITI: Seminar Ilmiah Sistem Informasi dan Teknologi Informasi*, vol. 10, no. 1, pp. 117–126, 2021.

[13] M. D. Purbolaksono, M. I. Tantowi, A. I. Hidayat, and Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam deteksi pasien penyakit diabetes," *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 2, pp. 393–399, 2021.

[14] Fedesario, "Stroke Prediction Dataset,", Kaggle, 7 juni 2021, [Online]. Tersedia: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset [Diakses: 2 September 2021]

[15] A. Cintami, "Mengenal Random Forest dengan Rstudio,", Medium, 18 juli 2020, [Online]. Tersedia: https://medium.com/@17611104_/mengenal-random-forest-dengan-rstudio-39e9f2c0c9df [Diakses: 5 Agustus 2021]