Stroke Prediction using Machine Learning Method with Extreme Gradient Boosting Algorithm

Abd Mizwar A. Rahim[1], Andi Sunyoto[2], Muhammad Rudyanto Arief [3]
Universitas Amikom Yogyakarta, Indonesia

Article Info

Article history:

Received 05 January, 2022
Revised 04 March, 2022
Accepted 23 April, 2022

Keywords:

Cardiovascular Disease,
Ensemble Learning,
Machine Learning,
Stroke Prediction,
Xtreme Gradient Boosting.

ABSTRACT

Based on data obtained from WHO, stroke is a disease that ranks as the second most deadly disease. The cause of a stroke is when a blood vessel is hit or ruptured, resulting in a part of the brain not getting the blood supply that carries the oxygen it needs, leading to death. By utilizing technology in the health sciences, especially in the health sector, machine learning models can adjust and make it easier for users to predict certain diseases. Previous studies have had problems with low accuracy when used in healthcare. The purpose of this research is to increase accuracy by proposing the application of one of the ensemble learning algorithms, namely the Xtreme Gradient Boosting algorithm. This stroke prediction research uses the Xtreme Gradient Boosting Algorithm; the application of this method with split data Training data and 70/30 test data, 70 % of the training data is 3582, 30 % of the test data is 1536, and the results are 96 % accuracy with these results having good results. This study increase accuracy in predicting stroke cases and get better accuracy than previous studies.

Corresponding Author:

Andi Sunyoto,
Department of Informatics Engineering,
Universitas Amikom Yogyakarta, Indonesia
Email: andi@amikom.ac.id

1. INTRODUCTION

The cause of death represents 32 % of all causes of death globallyaffected by cardiovascular disease (CVD). Causes of stroke when it occurs in the vessels while feeding the brain, risks such as an unhealthy diet, use, sedentary lifestyle, and unhealthy alcohol use are risk factors that can cause a person to have a stroke [1]. The World Stroke Organization (WSO) looks at the global impact of the COVID-19 pandemic on stroke care, as infection-related COVID-19 infection on the risk of having a stroke. Patients hospitalized when infected with Covid-19 can develop large blood vessels and increase mortality. The occurrence of stroke in about 1.4 % of patients [2], as for some risk of all-cause disease death, stroke, and bleeding in patients with atrial fibrillation (AF) and valvular heart disease (VHD) treated with vitamin K antagonists (VKA)), etc. [3].

Lately, there has been a lot of use of technology in the health sciences, using Machine Learning models to adjust and make it easier for users to make predictions on certain diseases by producing precise accuracy. This research is one step to help medical officers, etc. In the process of dealing with a patient who can have a stroke [4], not only being able to complete in the health sector but also being able to complete in various fields of agronomy, health, etc. as in [5–7]. The use of machine learning can solve problems that occur in fields that arise in human life. One example is in Business & Economic Statistics, where research conducted by machine learning can solve or assist in Predicting Inflation in a Data-Rich Environment. Assemble with a random forest algorithm [1].

Before entering into data processing, there are several methods of preprocessing data to get good accuracy results in research in [2–5], from these various studies that used machine learning before entering into processing. The preprocessing process is carried out using machine learning to produce quality data. An example of research that carried out these stages is the research conducted by Hakim B. This research aims to analyze and determine whether the preprocessing data in the text can affect or improve data mining results.

Previous research with the detailed case study was conducted by Ahmed H et al. (2019)) with the title Stroke Prediction using Distributed Machine Learning Based on Apache Spark. The result of this study is that machine learning plays an essential role in predicting stroke. The proposed hit prediction system is developed in Apache Spark. It consists of five stages: loading dataset, data preprocessing, cross-validation and hyperparameter tuning, classifier, and evaluation classifier. The results showed that the random forest classifier achieved the best accuracy of 90% [6].

Then carried out for comparison testing using the ANN and SVM algorithms in predicting stroke. This study was conducted by Colak C et al. (2015) with the title Application of knowledge discovery process on the prediction of stroke. The results given in this study are we use the model ANN and SVM to extract data patterns. In the end, the extracted models were evaluated and interpreted. Both models can predict stroke based on the predictor selected for early diagnosis as clinical. The findings of this study indicate that ANN has more predictive performance when compared to SVM, with a 92% result in predicting stroke [7].

Stroke prediction is also carried out with a machine learning approach which was developed based on physiological data with incompleteness and class imbalance. This study was made by Liu T et al. (2019). The research title is a hybrid machine learning approach to cerebral stroke prediction based on the imbalanced medical dataset. The research results carried out are the imputation of missing values and the AutoHPO-based DNN prediction model. The NS result of our model is that the false-negative rate is only 19.1%, and the overall accuracy is 71.6%. Compared with the average results of other commonly used methods, the false-negative rate was reduced by 51.5%, and the general error increased by 1.7%. This change means that our approach can substantially reduce the false-negative rate without a high cost to overall accuracy. Therefore, this study's hybrid machine learning approach is practical and credible for stroke prediction [8].

This study was carried out by NA Syamsul (2020), with a study entitled Comparison of the Fuzzy K-Nearest Neighbor and Neighbor Weighted K-Nearest Neighbor Methods for Stroke Disease Detection. The results of the research were testing 50, 70, 90, 150, 200, and 17 to 21 as a membership value. The results obtained are 81.272% and 81.814% accuracy using the Fuzzy K-Nearest Neighbor and Neighbor Weighted K-Nearest Neighbor methods on balanced data, compared with the same amount of test data, which is 100. The data used is unbalanced data with a percentage of 40 stroke data: 60 no stroke data, 60 stroke data: 40 no stroke data, 70 stroke data: 30 data no stroke and 30 stroke data: 70 non-stroke data, respectively, namely 82.45% and 82.75% [9].

The subsequent research, namely applying the C4.5 method for stroke prediction and implementing the Particle Swarm Optimizing and Genetic Algorithm to increase disease prediction accuracy was carried out by Ramdhan Hospital (2020). Stroke, the results of this study are From the 11 attributes contained in the Kaggle dataset. Further, selected into only ten features were used in determining the prediction of hepatitis disease. These attributes are gender, age, hypertension, heart disease, ever married, work type, avg glucose level, BMI, residence status, and stroke. The application of particle swarm optimization techniques can select attributes on C4.5, resulting in a better level of accuracy in diagnosing hepatitis than using the unique method of the C4.5 algorithm [10].

And research conducted to be able to predict Predicting Stroke after predicting effective Endovascular Treatment (EVT) for stroke patients with large vessel occlusion (LVO) in the anterior circulation by applying machine learning algorithms such as Random Forests, Support Vector Machine, Neural Network, and Super Learner and compared its predictive value with classical logistic regression models using various variable selection methodologies. VAJ Hendrikus conducted this study et al. This study entitled Predicting Outcome of Endovascular Treatment for Acute Ischemic Stroke: Potential Value of Machine Learning Algorithms, the results of this study are included data on 1,383 EVT patients with good reperfusion in 531 (38%) and functional independence in 525 (38%) patients. Machine learning and logistic regression models all performed poorly in predicting good reperfusion (average AUC range: 0.53-0.57) and moderately in predicting 3-month functional independence (mean AUC: 0.77 - 0 .79) using only the base variable. All models predicted 3-month functional independence using baseline and treatment variables (mean AUC range: 0.88-0.91) [11].

There is a difference between the previous research and the research conducted during the preprocessing namely, the data used is missing. This study does not delete the data, but it does not lose the information contained in the dataset by adding the data. This study aims to improve accuracy in predicting cases of stroke and get better accuracy than previous studies. In this study, we propose the use of machine learning methods using the Xtreme Gradient Boosting algorithm, using the Xtreme Gradient Boosting algorithm because this method is included in the ensemble algorithm that uses increased predictor accuracy, and the way the gradient boost algorithm works are to build a tree to adjust the data, then the next tree is constructed to reduce errors. The confusion matrix is used to assess the performance of the methods used in the stroke prediction process.

Previous studies have had problems with low accuracy when used in healthcare. The purpose of this research is to increase accuracy by proposing the application of one of the ensemble learning algorithms, namely the Xtreme Gradient Boosting algorithm. Several methods are used before classifying, namely dataset retrieval, data preprocessing, and dataset distribution (train/test split). After the previous process is complete, the classification process uses the Xtreme Gradient Boosting method and continues for the evaluation process using a confusion matrix to measure model performance in the form of accuracy, recall, and precision.

2. RESEARCH METHOD

In this study, the research method used is an experimental method, with the material and practical method described in the process or flow of the running of this research from the materials used, the process carried out, and the process result from data processing to the evaluation of classification methods. It can be seen in Figure 1, which describes the Materials and Methods of this Research.

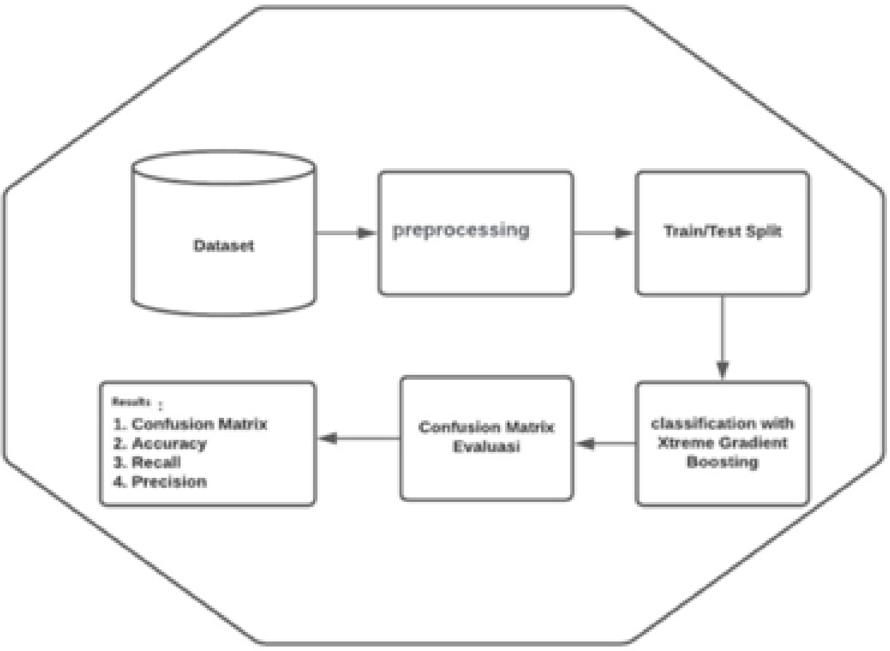


Figure 1. Research Materials and Methods

2.1. Dataset

Dataset is a collection of existing data from past experience, which is used for processing to become helpful information for an institution [12]. The Stroke Prediction Dataset is used in this research process. The data is taken from the Kaggle dataset https://www.kaggle.com/fedesoriano/stroke-prediction-dataset. This dataset is used to train and test the classification model, especially in the prediction of stroke. This dataset consists of 12 attributes, ten independent variables as features without the id variable being used, and one dependent variable as a class label used to predict stroke. Table 1 The dataset describes the feature information in the dataset used [13].

Table 1. Dataset

No	Fitur	Keterangan
1	Id	Unique Identifier
2	Gender	Female Male
3	Age	Age
4	Hypertensi	hypertension
5	heart_disease	1 Have heart disease 0 does not have heart disease
6	ever_married	1 means Married 0 means Not married
7	work_type	Children Personal Never work government work entrepreneur
8	Residence_type	Rural Urban
9	avg_glucose_level	Average glucose level
10	bmi	body mass index
11	smoking_status	Never smoked Used to smoke
12	Stroke	1 if the patient had a stroke or 0 if not

The ten independent variables in question are gender, age, hypertension, heart disease, ever_married, type_work, type_of residence, level_glucose, BMI, and smoking_status. The label class is the stroke attribute in this dataset has two values, namely 0 with the sign that there is no indication of stroke, while a value of 1 indicates an indication of a stroke.

2.2. Preprocessing

It was preprocessing data were for the method used to get good results in the classification process. There are several techniques in preprocessing according to the pattern of data that occurs cleaning, transformation, etc [14].

Data Preprocessing is a step before data processing using machine learning methods, and several preprocessing steps are applied.

1. Using LabelEncoder, which can convert non-numeric features into numeric features, only some attributes are changed in this labeling, including gender, ever_married, work_type, residence_type, and smoking_status.
2. Replace missing values, replace empty values with the average BMI having a stroke and BMI not having a stroke, compared to deleting data rows with empty values, replacing empty values as an option because no data is wasted.

2.3. Train/Test Split

Train/Test Split distributes datasets into training data and testing data. In this study, training data and testing data are divided into 70/30, with the total of the test data being 1536. In general, machine learning models get good accuracy results with a small amount of testing data. So, in this study, we increase the test data and test whether the model gets good results or not. Table 2 Train/Test Split describes the data sharing that is carried out. Table 2 Train/Test Split describes the data sharing carried out.

Table 2. Train/Test Split

Description	Data Training	Data Testing	Total
Proportion	70%	30%	100%
Amount	3582	1536	5118

2.4. Classification using the Extreme Gradient Boosting Algorithm

XGBoost (Xtreme Gradient Boosting) is a combination method between boosting and gradient boosting. XGBoost with the boosting method is used to classify errors from the previous model. Because XGBoost uses gradient descent which helps narrow the occurrence of errors during the creation or formation of new models. In the XGBoost process, several parameters are needed to obtain an optimal model called a hyperparameter. The adjustment of various parameters can affect the performance of the methods in processing the dataset. The parameters used to increase the classification using the XGBoost (Xtreme Gradient Boosting) method. Some of the parameters that can be used in the classification can be seen in Table 3.

Table 3. Parameters in XGBoost Method

Parameter	Information
max_depth	Maximum depth of the tree.
eta (learning_rate)	Prevents overfitting by reducing size
min_child_weight	Minimum weight of child_node
n_estimators	Number of trees
subsample	Randomly Sampling From Training Data Before constructing the tree.
gamma	The minimum loss reduction value, which is required to create further partitions of the nodes in the tree
random_state	internal random number generator initialization

2.5. Evaluation Method

Confusion matrix to evaluate the method in this classification. A confusion matrix is an evaluation that is often used to assess the performance of a classification model based on research objects that have true and false prediction values. In measuring the performance of the model in this study, there are four points to identify a prediction with the method used, these four points including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Some of the evaluation results that are usually used are as follows :

1. Accuracy (ACC) is the overall effectiveness of the classification results

$$Accuracy\% = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

2. Precision (PREC) is the percentage of data labels with positive labels given by the classification

$$Precision\% = \frac{(FN)}{(FP + FN)} \tag{2}$$

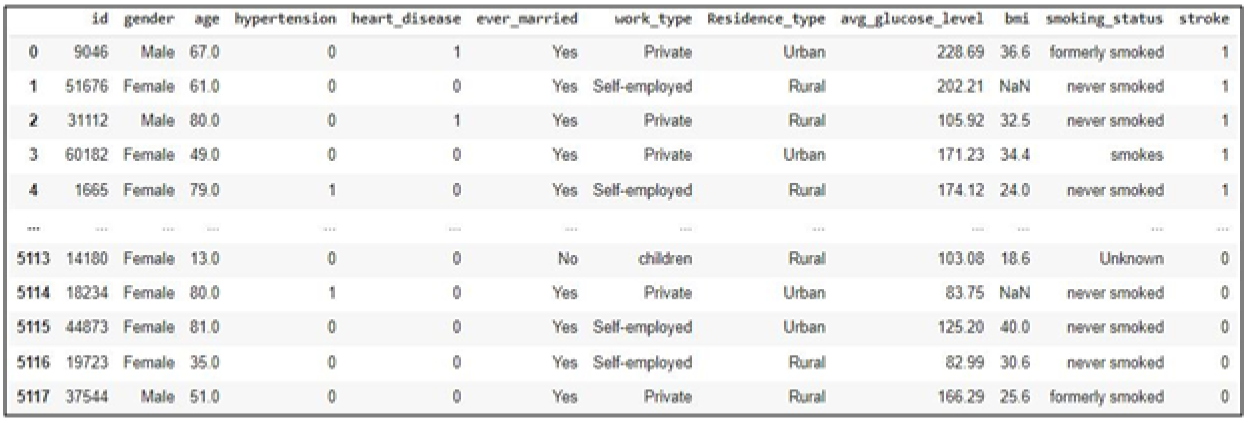
3. Recall (REC) or sensitivity is the effectiveness of the classifier in identifying positive labels

$$Recall\% = \frac{(TP)}{(TP + FN)} \tag{3}$$

3. RESULT AND ANALYSIS

3.1. Dataset

The dataset used in this study is a dataset from Kaggle; the dataset is generally used for education research. This dataset consists of 12 attributes, ten independent variables as features without the id variable, and one dependent variable as a class label used to predict stroke. Figure 2 Results of the dataset used.



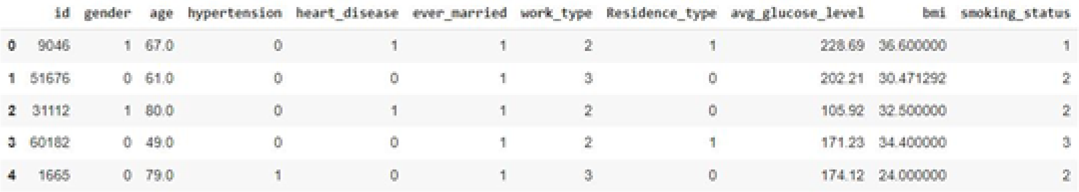
	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5113	14180	Female	13.0	0	0	No	children	Rural	103.08	18.6	Unknown	0
5114	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5115	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5116	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5117	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0

Figure 2. Dataset Results

in figure 2. it is explained that there are several variables in the dataset, namely from the id variable to the stroke or the number of attributes in the dataset, as many as 12 features, because id has no relationship that can determine the possibility of stroke patients or not. Therefore id is deleted in the dataset variable, and the number of dataset variables is only 11 attributes. Of the 11 attributes, 10 are independent attributes, namely from the gender variable to smoking status. and for the stroke variable to be a class/label in this classification process, the value contained in the class/label (stroke variable) is 0 which explains that the patient is not indicated for stroke, and one explains that the patient is indicated for stroke.

3.2. Preprocessing

Preprocessing data where in order for the method to get good results in the classification process, the preprocessing process in this classification has two steps, namely Label Encoder, which can convert non-numeric features into numeric features Replace missing values [15]. The results of this data preprocessing can be seen in Figure 3. the results of data preprocessing.



	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status
0	9046	1	67.0	0	1	1	2	1	228.69	36.600000	1
1	51676	0	61.0	0	0	1	3	0	202.21	30.471292	2
2	31112	1	80.0	0	1	1	2	0	105.92	32.500000	2
3	60182	0	49.0	0	0	1	2	1	171.23	34.400000	3
4	1665	0	79.0	1	0	1	3	0	174.12	24.000000	2

Figure 3. Results of data preprocessing

From the results of preprocessing, it can be seen that the dataset has previously been converted from non-numeric features into numeric features. Then replaces attributes that are empty with an average BMI having a stroke and BMI not having a stroke previously. There was a BMI value that was empty or contained on the Nan value dataset in the BMI column.

3.3. Train/Test Split

The division of the dataset into training data and testing data in this study, the results of the Train/Test Split of this study can be seen in Figure 4. Train/Test Split.

```
len(Xtrain)

3582

len(Xtest)

1536
```

Figure 4. Train/Test Split.

This study distributes training data and testing data to 70/30. The training data is 3582, and the test data is 1536.

3.4. Extreme Gradient Boosting Algorithm Classification

Using the Extreme Gradient Boosting (XGBoost) method for classification, the primary step taken is parameter tuning. The results of the tuning parameters are shown in table 3 of the tuning parameters.

Table 4. Tuning Parameters

No	max_depth	Eta (learning_rate)	min_child_weight	n_estimators	Random_State	Accuracy
1	5	0.1	1	50	-	0.91
2	10	0.2	1	60	5	0.92
3	10	0.3	1	70	50	0.93
4	12	0.4	2	100	4	0.95
5	15	0.09	-	100	-	0.96

Table 4 of the tuning parameters shows that the parameters used for the classification process of the experiments carried out include max_depth, learning_rate, min_child_weight, n_estimator random_state. In the first experiment, fifty trees were used, then each tree had five branches, learning_rate used a value of 0.1 which was used as a learning level that affected the XGBoost algorithm in tree-shaped classification, then the minimum amount of weight used was 1 where if the tree partition resulted in a node leaves with the number of instance weights less than min_child_weight, the development process will stop further partitioning, the accuracy of the first try is 91%. in the second experiment has sixty trees used, then each tree has ten branches, learning_rate uses a value of 0.2 which is used as a learning level that affects the XGBoost algorithm in tree-shaped classification, then the minimum amount of weight used is 1 where if the tree partition produces nodes leaf with the number of instance weights less than min_child_weight, then the development process will stop further partitioning, and use a random state that is worth five, the accuracy results obtained in the second try is 92%.

In the third experiment by adding the number of trees used, namely seventy, then each tree has ten branches, adding learning_rate with a value of 0.3 which is used as the learning level that affects the xgboost algorithm in tree-shaped classification, then the minimum amount of weight used is 1 where if the tree partition produces leaf nodes with the number of instance weights less than min_child_weight, then the development process will stop further partitioning, and use a random state that is worth fifty, the accuracy results obtained in the third try is 93%. in the fourth experiment by adding the number of trees used, namely one hundred, then each tree has ten branches, adding learning_rate with a value of 0.4 which is used as a learning level that affects the xgboost algorithm in tree-shaped classification, then the minimum amount of weight used is 2 where if tree partitioning produces leaf nodes with the number of instance weights less than min_child_weight, then the development process will stop further partitioning, and reduce the random state which is worth four, the accuracy results obtained in the fourth experiment are 95%, with the parameters used and the number is capable of increase the accuracy rate to 2%. And from the five experiments carried out, the best results were obtained in the 5th experiment process by having an accuracy value of 96% using only three parameters, namely max_depth, learning_rate, n_estimator. The best classification results are that one hundred trees are used, and each tree makes fifteen branches. learning_rate uses a value of 0.09, which is used as the level of learning that affects the Extreme Gradient Boosting algorithm in a tree-shaped classification.

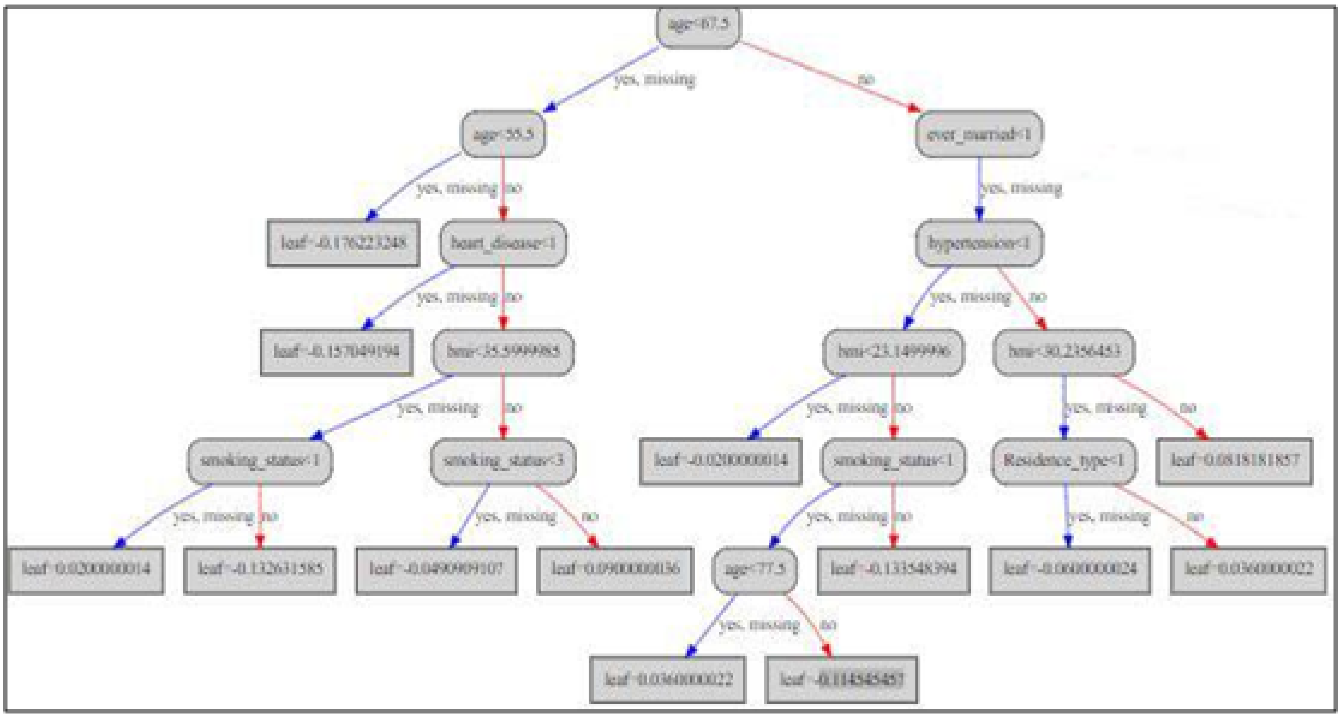


Figure 5. Example of an Extreme Gradient Boosting Tree

Figure 2 is the results of the Extreme Gradient Boosting Tree Example. The following is an example of a tree from Extreme Gradient Boosting (XGBoost) from this research. The results obtained from the tree above are as follows.

1. If age is less than 67.5, and age is less than 55.5 it will return -0.176223248.
2. If age is less than 67.5, age is more than 55.5, and heart disease is less than 1, it returns -0.157049194.
3. If age is less than 67.5, age is more than 55.5, heart disease is more than 1, BMI is less than 35.59, smoking status is less than 1 then it returns 0.0200000014 for the next three and the rest will return -0.132631585.
4. 4. If age is less than 67.5, age is more than 55.5, heart disease is more than 1, BMI is more than 35.59, smoking status is more than 3, then it returns -0.0490909107 in the next tree and returns 0.0900000036 for the others.
5. 5. If age is more than 67.5, ever_married is less than 1, hypertension is less than 1, and BMI is less than 23.1499 it will return -0.0200000014.
6. 6. If age is more than 67.5, ever_married is less than 1, hypertension is less than 1, and BMI is more than 23.1499, smoking status is more than 1, it will return -0.133548394 for other trees.
7. If age is more than 67.5, ever_married is less than 1, hypertension is less than 1, and BMI is more than 23.1499, smoking status is more than 1, and age is less than 77.5, it will return -0.0360000022 and reverse -0.114545457 for the other trees.
8. 8. If age is more than 67.5, ever_married is less than 1, hypertension is more than 1, BMI is more than 30.23 it will return 0.0818181857 for other trees.
9. 9. If age is more than 67.5, ever_married is less than 1, and hypertension is more than 1, BMI is more than 30.23, Residence_type is less than 1, it will return -0.0600000024 in the next tree and return 0.0360000022 for the other trees.

The best results from this study were an increase in the value of several parameters, namely the number of trees (n_estimator) and the maximum number of trees (max_depth). Then decreased the learning rate parameter used and did not set some parameters in this classification process such as Random_State, and min_child_weight, which was previously used in the previous classification process experiment. The resulting accuracy is 96%.

3.5. Confusion Matrix Evaluation and Classification Report

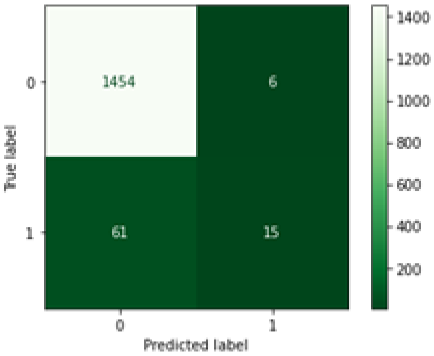
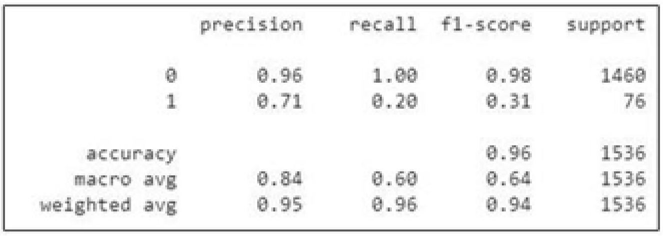


Figure 6. Matrix and Confusion Matrix Display

In Figure 4. Confusion Matrix results from python execution are one of the tools that display and compare the actual value or the actual value with the predicted model value. There are numbers 1 and 0 in the picture above, and number 1 explains that someone is indicated for a stroke, and number 0 explains that someone is not shown for a stroke. The results obtained from the Confusion Matrix process are as follows: True Positive (TP) is 1454, True Negative (TN) is 15, False Positive (FP) is 6, and False Negative (FN) is 61. From the evaluation results confusion matrix above, proceed to the binary classification calculation to assess the model's performance built using the tested XGBoost (Xtreme Gradient Boosting) algorithm. It can be seen that the best results from experiments carried out with several tested parameters. The 5th experiment resulted in the best accuracy of existing trials with a value of 96%. The following is the overall result of the 5th test can be seen in Figure 5 Classification Report.



```
              precision    recall  f1-score   support

           0       0.96      1.00      0.98      1460
           1       0.71      0.20      0.31        76

    accuracy                           0.96      1536
   macro avg       0.84      0.60      0.64      1536
weighted avg       0.95      0.96      0.94      1536
```

Figure 7. Classification Report

From Figure 5 above, it can be seen that the classification results have values of accuracy, recall, precision, etc. The accuracy value shows that the correct ratio of patients who predicted stroke and no stroke from all patients in the dataset is 96%. Then recall knows that patients predicted stroke compared to all patients who stroke, with a value of 20%, and precision shows The ratio of patients with true strokes in the dataset of all patients with predicted strokes is 71%. From the recall results obtained, only 20%, of course, is a low result. These results are produced because the total number of patients who have had a stroke is around 76 of the total data that has been split or divided into the previous dataset, so the results obtained are low. After all, the sum between True Positive and False Positive in the recalled formula produces very high results, therefore, the division of the sum with the True Positive value of the recall value gets low results.

Table 5. Resaerch Comparation

Author	Method	Accuracy
Hager Ahmed, et. (2019) [13]	Random Forrest	90%.
Nugroho, (2020) [15]	K-Nearest Neighbor	82.75%
Colak, et, al. (2015) [14]	Support Vector machine (SVM)	92%
Ramdhan, et,. al (2020	PSO + GA+C4.5	92.02%
Proposed	Xtreme Gradient Boosting Algorithm	96%

A comparison of research conducted previously with this research is described in Tabel 5. The topics raised in previous studies with this research are the same, namely the topic of stroke, as for the differences between the previous study and the research conducted, namely when applying the machine method. Learning processing classification methods include SVM, Decision three, etc. And there are also differences in the data preprocessing technique. Previous research was to apply deletion to data that had missing values. There was a lot of data loss from the deletion; some studies had not implemented preprocessing without overcoming the missing value. This, of course, will impact the results of the classification being biased.

In this study, the author will refine the research that is being carried out based on the shortcomings of previous research and conduct testing by adjusting the method used in the form of setting parameters, etc. The performance of the accuracy of the model used in this stroke classification process for the better, namely research. Regarding stroke prediction using the Xtreme Gradient Boosting method.

4. CONCLUSION

The accuracy results obtained get better results than previous studies using the same dataset pattern. The dataset is the Stroke Prediction Dataset. It consists of four stages in this study in the Stroke Prediction Dataset: preprocessing data, Split data, classification by XGBClassifier, and classifier evaluation. The changes made in this study from previous research include the data preprocessing process. Previous research carried out the deletion of data that would lose 202 data in data processing, and there were also previous studies that had not deleted data or other things that could overcome the empty values contained in the data. BMI attribute, with the classification results, will be biased. The preprocessing technique carried out by this study makes the empty values contained in the BMI attribute with an average BMI of stroke and an average of no stroke, there is no data loss, and also the classification results because the value on the BMI attribute has been resolved. The second change is related to training and testing data distribution. In this study, there was an increase of 10% of test data to 30% of test data compared to previous studies, which only used 20% of test data. The results showed that the classification using XGBoost (Xtreme Gradient Boosting) achieved the best accuracy of 96%, which was a better result than previous studies.

Suggestions from further research are to be able to overcome the imbalance of the dataset class. The dataset used is a stroke prediction dataset from Kaggle, the class of the dataset has a class imbalance, the class of someone who has a stroke is 249 and does not have a stroke is 4860. Class imbalance such as this dataset can affect athe model during classification. The model can only determine the majority class. Most likely, the predicted minority class will be predicted as the majority class. With the imbalance in this dataset, it is necessary to apply a method to overcome this.

ACKNOWLEDGEMENTS

The Acknowledgments section is optional. Research sources can be included in this section.

REFERENCES

[1] M. C. Medeiros, G. F. R. Vasconcelos, and Á. Veiga, “ce pt us cr t,” vol. 0015, 2019.

[2] B. Hakim, “Analisa Sentimen Data Text Preprocessing pada Data Mining dengan Menggunakan Machine Learning Data Text Pre-Processing Sentiment Analysis in Data Mining Using Machine Learning School of Computer Science and Technology , Harbin Institute of Technology,” vol. 4, no. 2, pp. 16–22, 2021.

[3] V. Chandani, “Komparasi Algoritma Klasifikasi Machine Learning dan Feature Selection pada Analisis Sentimen Review Film,” vol. 1, no. 1, pp. 56–60, 2015.

[4] I. Lishania, R. Goejantoro, and Y. N. Nasution, “Perbandingan Klasifikasi Metode Naive Bayes dan Metode Decision Tree Algoritma (J48) pada Pasien Penderita Penyakit Stroke di RSUD Abdul Wahab Sjahranie Samarinda,” *Jurnal Eksponensial*, vol. 10, no. 2, pp. 135–142, 2019.

[5] H. Kamel, B. B. Navi, N. S. Parikh, A. E. Merkler, P. M. Okin, R. B. Devereux, J. W. Weinsaft, J. Kim, J. W. Cheung, L. K. Kim, B. Casadei, C. Iadecola, M. R. Sabuncu, A. Gupta, and I. Díaz, “Machine Learning Prediction of Stroke Mechanism in Embolic Strokes of Undetermined Source,” *Stroke*, no. September, pp. 203–210, 2020.

[6] H. Ahmed, S. F. Abd-El Ghany, E. M. Youn, N. F. Omran, and A. A. Ali, “Stroke Prediction Using Distributed Machine Learning Based on Apache Spark,” *International Journal of Advanced Science and Technology*, vol. 28, no. 15, pp. 89–97, 2019.

[7] C. Colak, E. Karaman, and M. G. Turtay, “Application of Knowledge Discovery Process on The Prediction of stroke,” *Computer Methods and Programs in Biomedicine*, vol. 119, no. 3, pp. 181–185, 2015.

[8] T. Liu, W. Fan, and C. Wu, “A Hybrid Machine Learning Approach to Cerebral Stroke Prediction Based on Imbalanced Medical Dataset,” *Artificial Intelligence in Medicine*, vol. 101, p. 101723, 2019.

[9] S. A. J. I. Nugroho, “Naskah Publikasi Perbandingan Metode Fuzzy K-Nearest Neighbor dan Neighbor Weighted K-Nearest Neighbor untuk Deteksi Penyakit Stroke,” 2020.

[10] R. S. Rohman, R. A. Saputra, and D. A. Firmansaha, “Komparasi Algoritma C4.5 Berbasis PSO dan GA untuk Diagnosa Penyakit Stroke,” *CESS (Journal of Computer Engineering, System and Science)*, vol. 5, no. 1, p. 155, 2020.

[11] H. J. Van Os, L. A. Ramos, A. Hilbert, M. Van Leeuwen, M. A. Van Walderveen, N. D. Kruyt, D. W. Dippel, E. W. Steyerberg, I. C. Van Der Schaaf, H. F. Lingsma, W. J. Schonewille, C. B. Majoie, S. D. Olabarriaga, K. H. Zwinderman, E. Venema, H. A. Marquering, and M. J. Wermer, “Predicting Outcome of Endovascular Treatment for Acute Ischemic Stroke: Potential Value of Machine Learning Algorithms,” *Frontiers in Neurology*, vol. 9, no. SEP, pp. 1–8, 2018.

[12] A. M. Khalimi, “Dataset adalah Data untuk Data Mining,” 2020.

[13] Fedesoriano, “archive,” 2021. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?resource=download

[14] H. Ma’rifah, A. P. Wibawa, and M. I. Akbar, “Klasifikasi Artikel Ilmiah dengan Berbagai Skenario Preprocessing,” *Sains, Aplikasi, Komputasi dan Teknologi Informasi*, vol. 2, no. 2, p. 70, 2020.

[15] A. A. Rizal and S. Soraya, “Multi Time Steps Prediction dengan Recurrent Neural Network Long Short Term Memory,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 18, no. 1, pp. 115–124, 2018.