Performance Analysis of Various Machine Learning Approaches in Stroke Prediction

Md. Shafiul Azam
Dept. of CSE
Pabna University of Science and Technology, Bangladesh

Md. Habibullah
Dept. of CSE
Pabna University of Science and Technology, Bangladesh

Humayan Kabir Rana
Dept. of CSE
Green University of Bangladesh, Bangladesh

ABSTRACT

Stroke is one of the most life threatening diseases. Now a day the difficulty of Stroke is a global health concern. Once a stroke disease occurs, it is not only matter of huge medical care and permanent disability but also can eventually lead to death. The most of the strokes can be prevent if we can identify or predict the occurrence of stroke in its early stage. In this situation, machine learning can be a hope. It plays a vital role in the prediction of diseases in health care industry. In this paper, the various machine learning approaches like Logistic Regression (LR), Random Forest (RF), Decision Tree (DT) are employed to predict the risk of stroke whether a patient will be affected by stroke or not. The main purpose of this research is to highlight the employing of machine learning algorithms in prediction of stroke risk and analysis the performance of these algorithms. This research also analyzed the significant features of datasets to predict the stroke risk.

Keywords

Stroke; Machine Learning; Decision Tree; Logistic Regression; Random Forest

1. INTRODUCTION

Stroke is a leading cause of disability and mortality throughout the world. According to recent survey of WHO (World Health Organization) 17.9 million people die each year globally because of Cardiovascular Diseases (CVD) and Four out of five CVD deaths are due to heart attacks and strokes and also it is increasing rapidly. A stroke occurs when blood vessels in the brain rupture or become occluded [1]. Typical symptoms include muscular weakness, loss of sensation, problems with vision, and impaired speech. Depending on the location and severity of neuronal damage, additional symptoms, including loss of consciousness, may occur [1].

There are two main types of stroke: hemorrhagic and thrombotic (also known as ischemic). A hemorrhagic stroke may be caused by a ruptured cerebral blood vessel, a ruptured intracranial aneurysm, or an arterio-venous malformation leading to an intracerebral hemorrhage in or near the brain [1].

A thrombotic stroke results from the occlusion of one or more cerebral blood vessels. A thrombus may form directly on a diseased small vessel, or a large-vessel atherosclerotic plaque may embolize and block a smaller cerebral artery [1,2].

The prediction of stroke risk can contribute to its prevention and early treatment. Numerous medical studies and data analyses have been conducted to identify effective predictors of stroke.

Potentially modifiable risk factors for stroke include hypertension, cardiac disease, atrial fibrillation, and lifestyle factors. Due to the lack of resources in the medical field, the prediction of stroke occasionally may be a problem. Utilization of suitable technological support in this regard can prove to be highly beneficial to the patients [3]. This issue can be resolved by machine learning approaches using potential datasets and motivate us to do this research.

As the stroke risk prediction can be hope in stroke prevention, the various the machine learning approaches are employed here to predict the stroke of a person and to analyze the performance of each approaches and discovers the significant risk factors. Machine learning algorithms are expected to effectively handle many more features.

In this paper we consider only some relatively some high-performance machine learning algorithms such as Logistic Regression, Random Forest, and Decision Tree. It is noted that some logical techniques are used for preprocessing the data to make the database balanced to gain high performance.

This paper presents details the performance of various machine learning approaches for stroke risk prediction and enlists the significant features. The algorithms that we have examined in our work are Decision tree (DT), Logistic regression (LR) and Random Forest (RF) with and without using smoking status attribute.

2. MATERIALS AND METHODS

2.1 Description of the dataset

The dataset [4] we used in our work has in total 12 columns and 62001 rows. First 11 of those columns are the features that we will be using later in order to predict the final column ‘target(stroke)’ which will tell us if the patient is going to be affected by stroke or not. The 62001 rows represent data of 62001 patients that we found from dataset. The short description of features of our used dataset is given in the following table-1.

Table 1. Feature description of the dataset

Attributes	Description	Range of values
id	Identification No.	Arbitrary
gender	Gender of the person [1: Male, 0: Female]	0, 1
age	Age of the person in years(0.08-82)	0.08-82
hypertension	Hypertension (0-No, 1-Yes)	0,1
heart_disease	Heart Disease (0-No, 1-Yes)	0,1

ever_married	Marital Status (0-No, 1-Yes)	0,1
work_type	1-Children 2-Govt_Job 3-Never Worked 4-Private 5-Self_Employed	1,2,3,4,5
Residence_type	0-Rural, 1-Urban	0,1
avg_glucose_level	Average Glicose Level	55-291.0
bmi	Body Mass Index	10.1-97.6
smoking_status	0-No-smoke, 1-Smoke	0,1
stroke	Class Attbute (0-No StrokeRisk 1-SrokeRisk	0,1

2.2 Methodologies

This paper intends to adopt and examine the three machine learning algorithms (Logistic Regression (LR), Random Forest (RF) and Decision Tree (DT)) for the prediction of stroke risk with and without using smoking attribute. It also compares the performance among them. The system flow is given in the following figure-1.

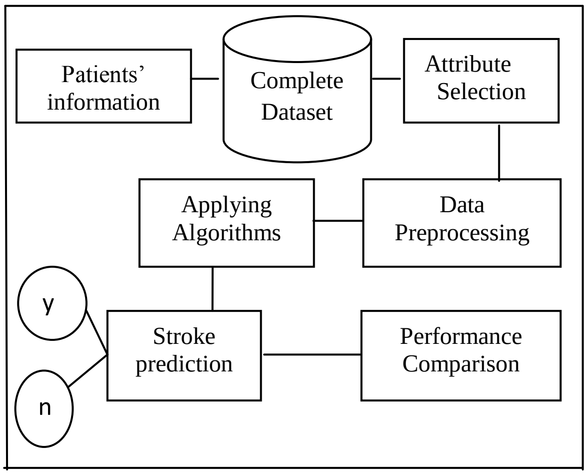


Figure 1: System of Stroke Prediction

2.2.1 Decision Tree

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too [5,6]. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node. Important Terminology related to Decision Trees are:

Root Node: It represents the entire population or sample and this further gets divided into two or more homogeneous sets.

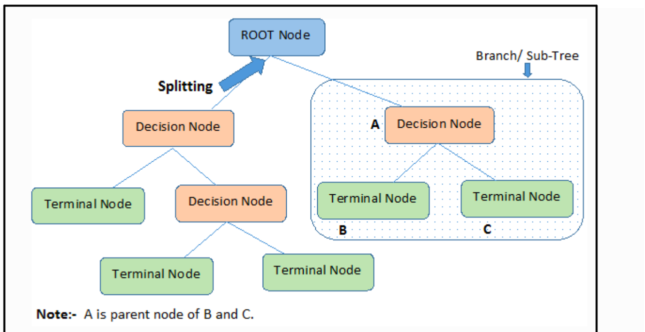
Splitting: It is a process of dividing a node into two or more sub-nodes.

Decision Node: When a sub-node splits into further sub-nodes, then it is called the decision node.

Leaf / Terminal Node: Nodes do not split is called Leaf or Terminal node. Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.

Branch / Sub-Tree: A subsection of the entire tree is called branch or sub-tree.

Parent and Child Node: A node, which is divided into sub-nodes, is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.



Note:- A is parent node of B and C.

Figure 2: A picture for decision tree

Decision trees classify the examples by sorting them down the tree from the root to some leaf/terminal node, with the leaf/terminal node providing the classification of the example. Each node in the tree acts as a test case for some attribute, and each edge descending from the node corresponds to the possible answers to the test case. This process is recursive in nature and is repeated for every subtree rooted at the new node.

2.2.2 Logistic Regression

Logistic Regression is a supervised learning algorithm that trains the model by taking input variables(x) and a target variable(y). In Logistic Regression the output or target variable is a categorical variable, unlike Linear Regression, and is thus a binary classification algorithm that categorizes a data point to one of the classes of the data. The general equation of Logistic Regression is-

$$log(p(X)/(1-p(X))) = \beta_0 + \beta_1 X \ldots\ldots\ldots (1)$$

Where, p(X) is the dependent variable, X is the independent variable, β_0 is the intercept and β_1 is the slope co-efficient [7,8]. Logistic Regression measures the relationship between the dependent variable, the output, and the independent variables, the input, by estimating probabilities using its underlying logistic function. It uses L2 penalty for regularization. The resultant probabilities are then converted to binary values 0 or 1 by the logistic function, also known as the sigmoid function. The sigmoid function takes any real-valued number and maps it into a value between the ranges 0-1 excluding the limits themselves. Afterwards, a threshold classifier transforms the result to a binary value.

2.2.3 Random Forest

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result [9,10,11].

The Working Steps of Random Forest Algorithm

Step 1 − First, start with the selection of random samples from a given dataset.

Step 2 − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.

Step 3 − In this step, voting will be performed for every predicted result.

Step 4 − At last, select the most voted prediction result as the final prediction result.

The following figure-3 will illustrate its working –

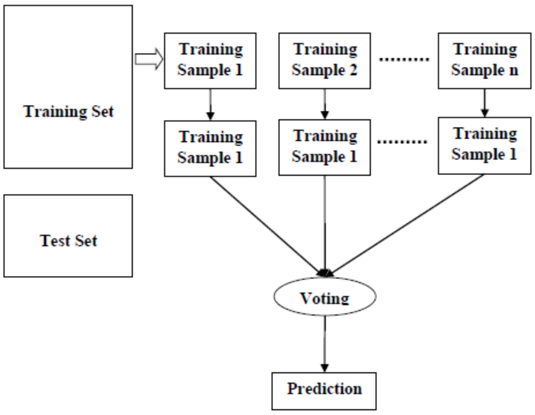


Figure 3: A picture for Random Forest Algorithm

3. EXPERIMENT SETTING & RESULTS

In this work the three most important machine learning algorithms include Decision Tree, Logistic Regression and Random Forest are employed and examined. The experiments were constructed on python. Applying those algorithms which were discussed in the above session, the prediction is done whether a patient will be affected by stroke or not. Finally the various performances evaluation methods named Accuracy, Precision, Recall and F1-score are used to measure the performance of those algorithms using confusion matrix in this work.

3.1 Confusion Matrix

Confusion Matrix is the easiest way to determine the performance of a classification model by comparing how many positive instances were correctly/incorrectly classified and how many negative instances were correctly/incorrectly classified. In a Confusion Matrix, the rows represent the actual labels and the columns represent the predicted labels shown in the following table-2.

Table 2. The Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

True Positives (TP):

True positives are the instance where both the predicted class and actual class is True (1), i.e., when patient actually has complications and is also classified by the model to have complications.

True Negatives (TN):

True negatives are the instances where both the predicted class and actual class is False (0), i.e., when a patient does not have complications and is also classified by the model as not having complications.

False Negatives (FN):

False negatives are the instances where the predicted class is False (0) but actual class is True (1), i.e., when a patient is classified by the model as not having complications even though in reality, they do.

False Positives (FP):

False positives are the instances where the predicted class is True (1) while actual class is False (0), i.e., when a patient is classified by the model as having complications even though in reality, they do not.

3.1.1 Accuracy (ACC)

Accuracy determines the number of correct predictions over the total number of predictions made by the model. Even though it is widely used, it is not a very good measure of performance especially when the dataset is imbalanced like in this case. The formula for Accuracy is:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

3.1.2 Precision

Precision is a measure of the proportion of patients that actually had complications among those classified to have complications by the system. The formula for Precision is:

$$Precision = \frac{TP}{TP + FP}$$

3.1.3 Recall/ Sensitivity

Recall or sensitivity is a measure of the proportion of patients that were predicted to have complications among those patients that actually had the complications. The formula is:

$$Recall = \frac{TP}{TP + FN}$$

3.1.4 F1 Score

F1 Score is the harmonic mean of the Recall and Precision that is used to test for Accuracy. The formula is:

$$F1\ 2 * \frac{Precision * Recall}{Precision + Recall}$$

3.2 Experimental Results

3.2.1 Significant Features with random Forest

The ranks of importance of the features are found from each algorithm and finally the accumulated results are shown in figure-4. It is seen that the Age, hypertension, heart disease, Residence type, Avg Glucose level, BMI and Smoking status come as significant variables here, but Gender, Marriage status and Work Status are some less significant features.

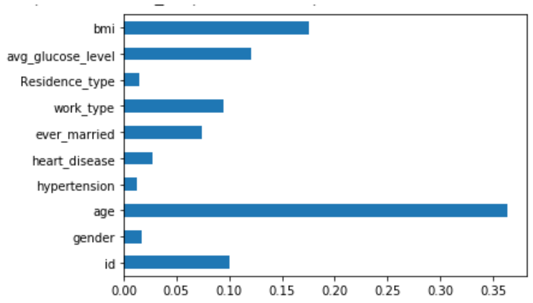


Figure 4: Feature Importance with random Forest

3.2.2 Performance comparison

The performance of each model is examined with our dataset and compared with and without using smoking status and shown in the table-3 below.

Table 3. Performance Analysis of various approaches

%	Accuracy	Precision	Recall	F1-score
DTCWSS	98.78	98	99	99
DTCWOSS	99.46	99	99	99
LRCWSS	71.21	70	76	73
LRCWOSS	81.34	78	87	82
RFCWSS	99.98	99	99	99

The following figure-5 gives an overview of performance of each approach.

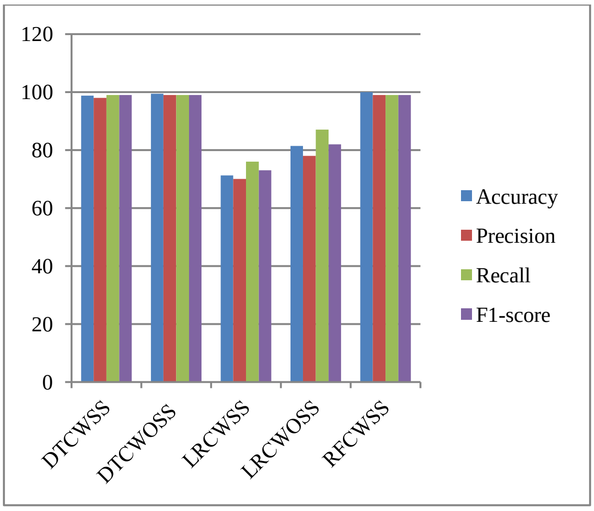


Figure 5: Performance Comparison of each Approach

By comparing all methods from the above figure-5 and table-3 it is seen that the performance of Random forest algorithm gives the better performance in our work.

4. CONCLUSION

In this paper we have tried to make the dataset balanced by using some preprocessing techniques. Secondly, the three machine learning approaches include Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT) are employed to analyze the significant features and to predict the stroke risk of a patient. Lastly, the comparison of performance of each approach is shown here. This work is considered as the basement for the healthcare system for stroke patients.

In future we will extend our works with various deep learning mechanisms using big data to predict the stroke risk and analyze the performance.

5. ABBREVIATION

The Abbreviations:
DTCWSS-Decision Tree Classifier with Smoking Status
DTCWOSS-Decision Tree Classifier without Smoking Status
LRCWSS-Logistic Regression Classifier with Smoking Status
LRCWOSS-Logistic Regression Classifier without Smoking Status
RFCWSS-Random Forest Classifier with Smoking Status

6. ACKNOWLEDGMENTS

We are grateful to the dept. of Computer Science and Engineering, Pabna University of Science and Technology for permitting us to use the Advanced Computing Lab (ACL) for our experiments.

7. REFERENCES

[1] Stroke. “Neurological, Psychiatric, and Developmental Disorders-NCBI Bookshelf” ncbi.nlm.nih.gov/books/NBK223479/

[2] NK Podder, HK Rana and et al., "A system biological approach to investigate the genetic profiling and comorbidities of type 2 diabetes", Gene Reports, pp. 100830, https://doi.org/10.1016/j.genrep.2020.100830, 2020.

[3] Priyanka N., Pushpa Ravi Kumar. "Usage of data mining techniques in predicting the heart diseases -Naïve Bayes & decision tree", 2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT), 2017

[4] Kaggle. “Healthcare stroke Patients in Python” kaggle.com/surajdidwania/healthcare-stroke-patients-in-python/data?fbclid=IwAR21bmwdw1jItWPIMRMEB-_CehjYuh5wH6IQOlVvOvOyBp0umH4X1d9Zk7g

[5] T. Lumley, R. A. Kronmal, M. Cushman, T. A. Manolio, and S. Goldstein. A stroke prediction score in the elderly: Validation and web-based application. J. Clin. Epidemiol., 55(2):129–136, February 2002.

[6] Mujtaba, M. A., Azam, M. S., &Rana, H. K., "Performance evaluation of various data mining classification techniques that correctly classify banking transaction as fraudulent", GUB Journal of Science and Engineering, vol. 4, no. 1, pp. 59-63, 2017.

[7] HosmerJr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. "Applied logistic regression". vol. 398. John Wiley & Sons, 2013.

[8] Couronné, Raphael, Philipp Probst, and Anne-Laure Boulesteix. "Random forest versus logistic regression: a

large-scale benchmark experiment." BMC bioinformatics 19.1 (2018): 270.

[9] Tutorialspoint. "Classification Algorithms - Random Forest"www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm

[10] Hossen, M. R., Azam, M. S., &Rana, H. K., "Performance evaluation of various DNA pattern matching algorithms using different genome datasets", Pabna University of Science and Technology Studies, vol. 3, no. 1, pp. 14-8, 2018.

[11] Chang, Chuan-Yu, Man-Ju Cheng, and Matthew Huei-Ming Ma. "Application of Machine Learning for Facial Stroke Detection." 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP). IEEE, 2018.