Stroke Prediction with Classification Methods

Md. Tasluf Morshed Rhyme[1], Md. Assadujjaman Tilok[2], Md. Riazul Islam Jishan[3],

Md. Towhidul Islam Sayon[4]

Daffodil International University, Dhaka-1207, Bangladesh

[1,2,3,4]Department of Computer science and Engineering, Faculty of Science and Information Technology,

Email: tasluf15-12089@diu.edu.bd, assadujjaman15-12594@diu.edu.bd, riazul15-12772@diu.edu.bd

Keywords:

Heart disease, Stroke, Data mining, Data Set, Machine learning, Random forest, Support Vector Machine (SVM), Decision tree, Prediction, Accuracy

Abstract

Stroke is a life-threatening disease that has been ranked the third leading cause of death in states and in developing countries. It is a neurological disease that occurs when a brain cell dies as a result of oxygen and nutrient deficiency (Vieira et al., 2021). This paper contains a comparative analysis performed for the classification of the stroke dataset in order to predict Stroke cases with minimal attributes. The dataset contains 76 attributes including the class attribute, for 1025 patients collected from the USA, Canada & Switzerland, but in this paper, only a subset consists of 5111 observations and 11 columns attributes are used, and each attribute has a given set value. The algorithms used Decision tree (DT), SVM & random forest classifiers to show the performance of the selected classifications algorithms to identify the best predicting model to predict pre-stroke independence at 6 and 18 months earlier, considering sociodemographic and clinical characteristics, and then identifying differences between countries.

1. Introduction

Stroke is considered to be a burning issue not only in South Asia but also all over the world. Every year, around 16 million people are affected for the first time by stroke, and 5.7 million of them die (Sudha et al., 2012). Due to population aging, the number of individuals affected by a stroke is increasing over the years. It is generating a lot of controversies when the boom began from 1.1 to 13.7 million between 2000 to 2016 (Sudha et al., 2012). In the 21st century while the world is developing day after day in that very time researchers estimated that by 2030, around 77 million strokes will occur worldwide (Sudha et al., 2012). The main objectives of this research are: i) Use data mining techniques to predict people who are at a high risk of developing stroke. ii) Find the patient who has a higher chance to develop stroke near future (Olesen et al., 2011). Therefore, this paper showed that using different classification algorithms for the stroke patient dataset which gives very promising results in term of the classification accuracy for the SVN, Decision Tree, Random forest which are 94%, 91%, and 94% respectively. Different clustering

methods were used and compared to classify the best prediction model, and we concluded the benefit of having a reliable feature selection method for Stoke prediction by using a minimum number of attributes instead of having to consider all available ones.

---

2. Literature Review

A good number of research papers have already been written on this topic. Different researchers try to predict the stroke with different algorithms and different attributes.

According to .Sudha, conducted a study in a Patient dataset that is collected from healthcare institutes. They use different kinds of algorithms like decision trees, naive Bayes, and neural networks to find whether the patient is suffering from stroke disease or not. From their study, they have found neural networks having the best performance compared with the other two algorithms (Sudha et al., 2012).

According to Ohoud Almadani, they conducted research on a data set that is obtained from the Ministry of National Guards Health Affairs hospitals. They have used J48 (C4.5), JRip, and Neural Network algorithm in Weka software for their prediction. From their research, they have found patients with the following medical conditions, such as heart diseases, immunity diseases, diabetes militias, kidney diseases, hyperlipidemia, epilepsy, or blood disorders have a higher probability to develop stroke. The accuracy of this approach is approximately 95% for the C4.5-J48 algorithm (Almadani & Alshammari, 2018).

According to Devansh Shah, published a report on an existing dataset from the Cleveland database of UCI repository of heart disease patients. From 303 instances and 76 attributes, they have considered 14 attributes for testing which are related to heart disease. Using Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm they have found K-nearest neighbors (K=7) have the best accuracy about 90.789% (Shah et al., 2020).

According to Md. Murad Hossain, performed a study on the Kaggle Heart disease UCI dataset. In that dataset contains around 899 observations with 14 Attributes. They have also used different kinds of algorithms like KNN, Naïve Bayes, Random forest, Logistic regression, Support vector machine, J48, and Decision tree. To implement this algorithm they have used Weka software and find 83.6485% accuracy from the Naïve Bayes algorithm in their study.(Hossain et al., 2021)

According to Minhaz Uddin Emon, published a report predicting stroke diseases. They have collected the information from the medical clinic of Bangladesh. This research is conducted on 5110 people considering around 11 attributes such as age, heart disease, BMI, etc. Applying around ten different

classifiers, they have achieved an accuracy of 97%, where the weighted voting classifier performs better than the base classifiers (Emon et al., 2020).

According to Gangavarapu Sailasya performed a study on the same dataset that we use in our research. They have also used Logistic Regression, Decision Tree Classification, Random Forest Classification, K-Nearest Neighbors, Support Vector Machine, and Naïve Bayes Classification to train five different models. From there finding the best accuracy of approximately 82% from the Naïve Bayes algorithm (Sailasya & Kumari, 2021).

---

3. Methodology

A. Dataset: The data consists of 5111 observations and 11 columns. Here in this study, we want to classify whether a person has a risk of stroke or not. We want to predict this stroke depending on the various attributes. In the stroke column 0 means no risk for stroke and 1 means he or she has a risk of stroke. Other independent variables are gender, age, hypertension, heart disease, ever married, work type, Residence type, avg glucose level, BMI, and smoking status. Here age, hypertension, heart disease, avg glucose level, BMI are numeric variables.

	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

Table 1: Sample dataset

B. Descriptive Statistics:

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110.000000	5110.000000	5110.000000	5110.000000	5110.000000	4909.000000	5110.000000
mean	36517.829354	43.226614	0.097456	0.054012	106.147677	28.893237	0.048728
std	21161.721625	22.612647	0.296607	0.226063	45.283560	7.854067	0.215320
min	67.000000	0.080000	0.000000	0.000000	55.120000	10.300000	0.000000
25%	17741.250000	25.000000	0.000000	0.000000	77.245000	23.500000	0.000000
50%	36932.000000	45.000000	0.000000	0.000000	91.885000	28.100000	0.000000
75%	54682.000000	61.000000	0.000000	0.000000	114.090000	33.100000	0.000000
max	72940.000000	82.000000	1.000000	1.000000	271.740000	97.600000	1.000000

Table 2: Describe numerical variables

In the above table, describe all numeric variables from our dataset. It shows the count, max, min, and all others information of numeric variables. The columns of numeric variables are id, age, hypertension, heart disease, avg glucose level, BMI and the last one is stroke. Although the id column only describes the id of a patient. We will remove this from our dataset in the normalization part.

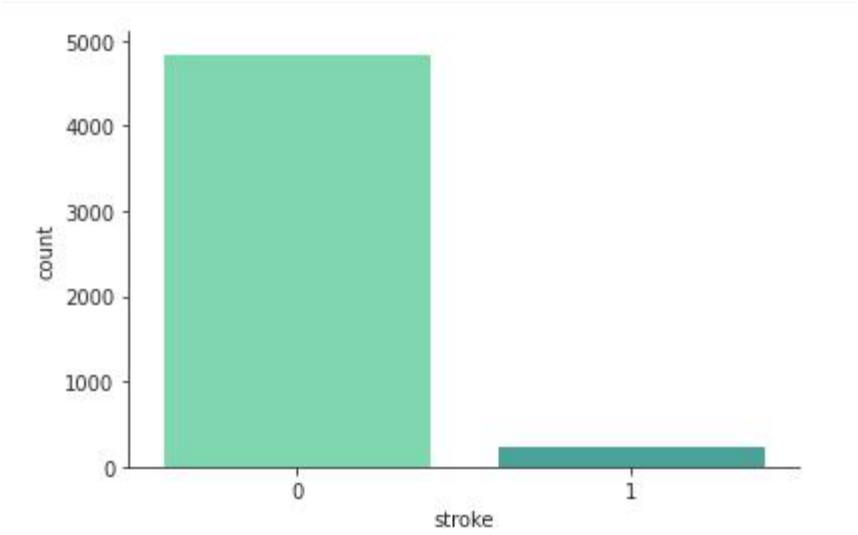

Figure 1: Distribution of 0's and 1's in the target variable

From the above figure, we can the Distribution of the target variable. Here around 4800 people have no risk for stroke which is denoting by 0 and around 300 have the risk for stroke which is denoting by 1

C. Exploratory Data Analysis:

1. Residence type

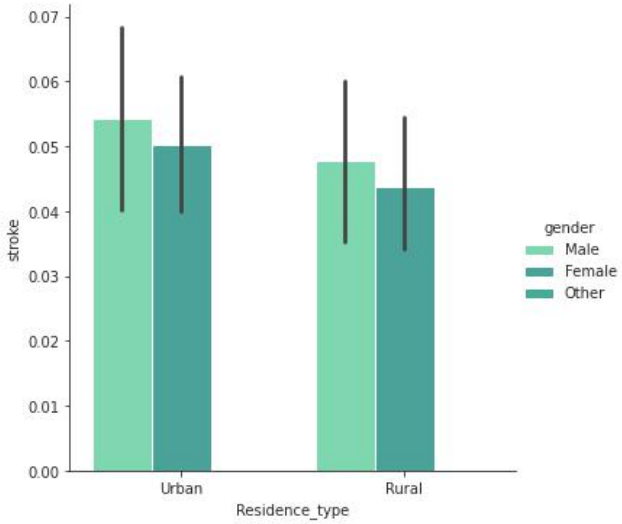

Figure 2: Distribution of Urban and Rural

From the above figure, we can see that in both Urban and Rural areas, males have more risk of stroke than females. And we also see that, have more Urban people than Rural in the dataset. So, males have a greater risk of stroke than females.

2. Ever married

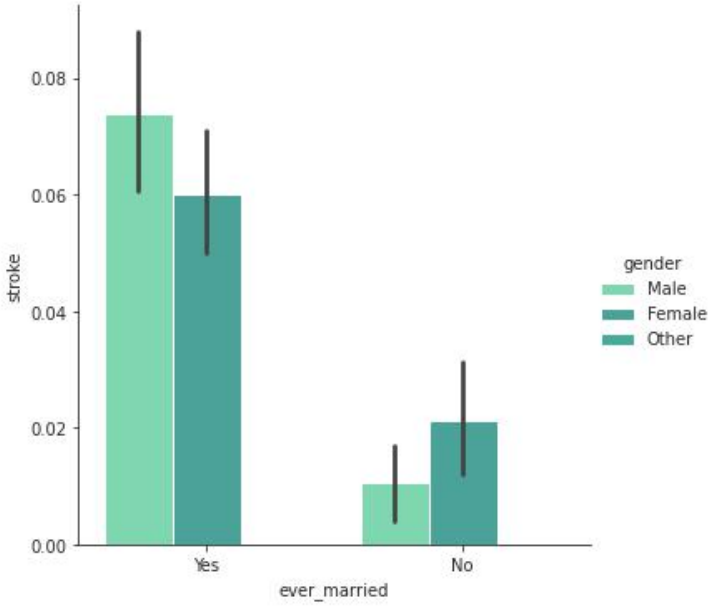

Figure 2: Distribution of Ever married

In here we see those people who have married have a high risk for stroke and people who don't marry have a fewer risk of stroke. But the interesting part is Male have a high risk of stroke after marriage and females have a high risk when they don't get married.

3. Hypertension

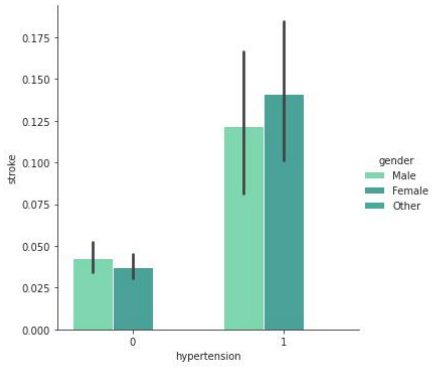

Figure 3: Distribution of hypertension

Here we can see those people who have hypertension have a risk for stroke. And In the male and female distribution, we can say that, if females have hypertension then they have a high risk for stroke than males.

4. Heart disease

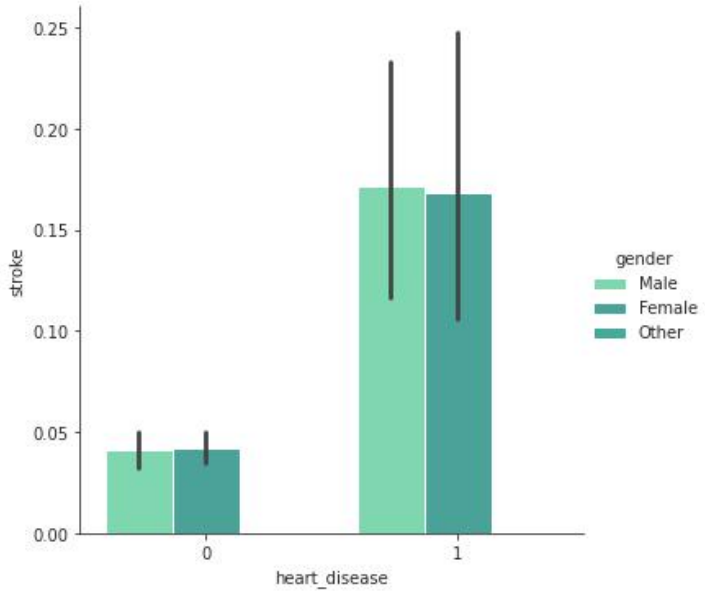

Figure 4: Distribution of heart disease

The above plot show, male and female both of the case if they have heart disease then there will be high risk for stroke. Here the plot describes that the higher the heart disease and the higher the risk of stroke.

5. Age

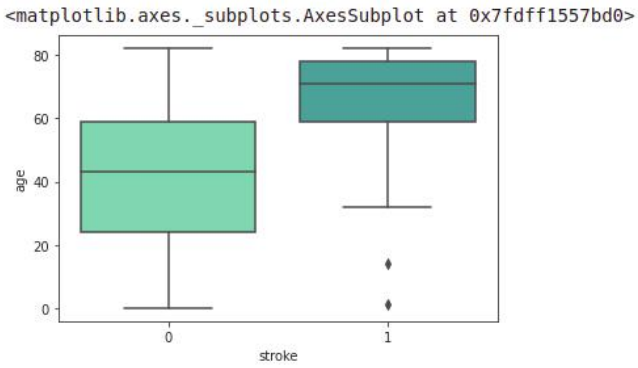

Figure 5: Age vs Stroke

Here the plot describes the relationship between age and stroke. We can see that. People between 60-80 age have a high risk for stroke. It means older citizens have more risk for stroke than younger.

b) MULTICOLLINEARITY CHECK:

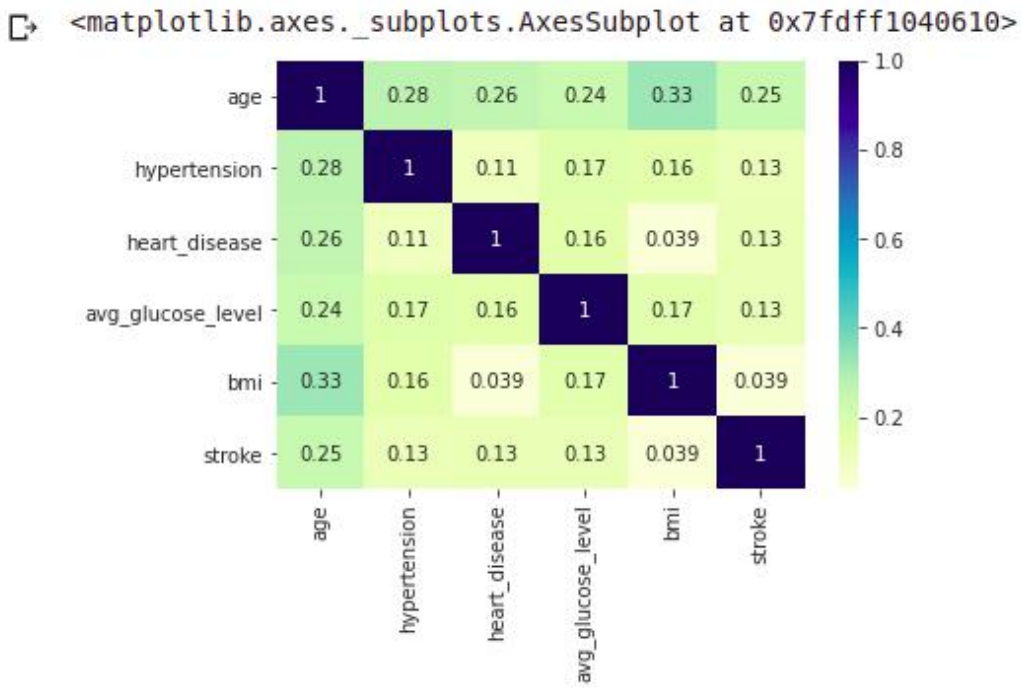

Figure 6: Correlation between attributes

From the above heatmap, we can see the correlation between the different attributes of our dataset. Here stroke and BMI have a good correlation of about 0.039 it means this stroke variable highly depends on the BMI variable. We also see BMI and heart disease have also high correlation from the plot.

D. Pre- Processing:

There are some missing values in the BMI and Smoking status column of our dataset. As BMI is a numeric column so we use mean to replace all the missing values. And smoking status is a nominal column so we use mean to replace all the missing values.

	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	36.600000	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	28.893237	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.500000	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.400000	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.000000	never smoked	1

Table 3: After Handling missing values

E. Modelling

1. Random Forest:

Random forest is a supervised machine learning algorithm that is constructed from decision tree algorithms ("Introduction to Random Forest in Machine Learning", 2021). Random forest is a flexible and easy machine-learning algorithm to use that produces a great result most of the time(Sisodia & Sisodia, 2018). A random forest algorithm consists of many decision trees ("Introduction to Random Forest in Machine Learning", 2021). The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating ("Introduction to Random Forest in Machine Learning", 2021). It is one of the most used algorithms; because of its simplicity and diversity, it can be used for both classification and regression methods (Sisodia & Sisodia, 2018).

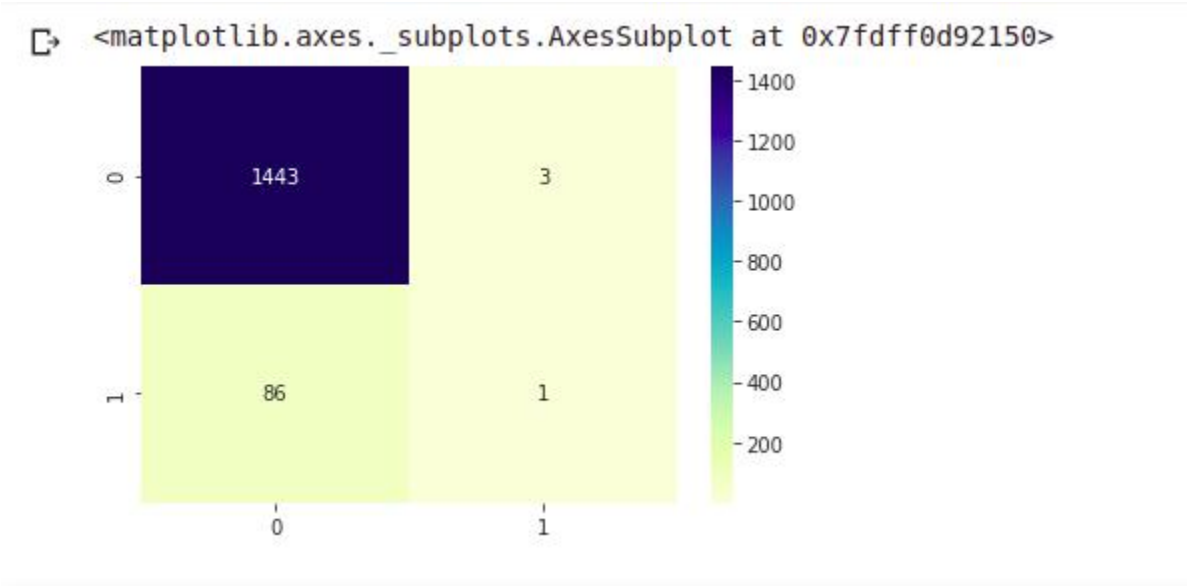

Figure 6: Decision tree Confusion Matrix

Here, we split our dataset into 30% percentages. So, it will randomly pick around 1500 instances from the dataset for the test set and the rest of the instances will be in the train set.

	precision	recall	f1-score	support
0	0.94	1.00	0.97	1446
1	0.25	0.01	0.02	87
accuracy			0.94	1533
macro avg	0.60	0.50	0.50	1533
weighted avg	0.90	0.94	0.92	1533

A precision of 0.94 and recall is 1 from the decision tree. Finally, the accuracy of this algorithm is:

Accuracy = (TP + TN) / (TP + TN + FP + FN) * 100
= (1443+1)/ (1443+1+3+86) * 100
= 94.19%

2. Decision Tree Classifier:

A Decision Tree is a supervised machine learning algorithm used to solve classification problems by using entropy and information gain. The main objective of using a Decision Tree in this research work is the prediction of the target class using decision rules.

It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification ("Introduction to Random Forest in Machine Learning", 2021). In every stage, the Decision tree chooses each node by evaluating the highest information gain among all the attributes. The evaluated performance of the Decision Tree technique using Confusion Matrix is as follows (Hossain et al., 2021).

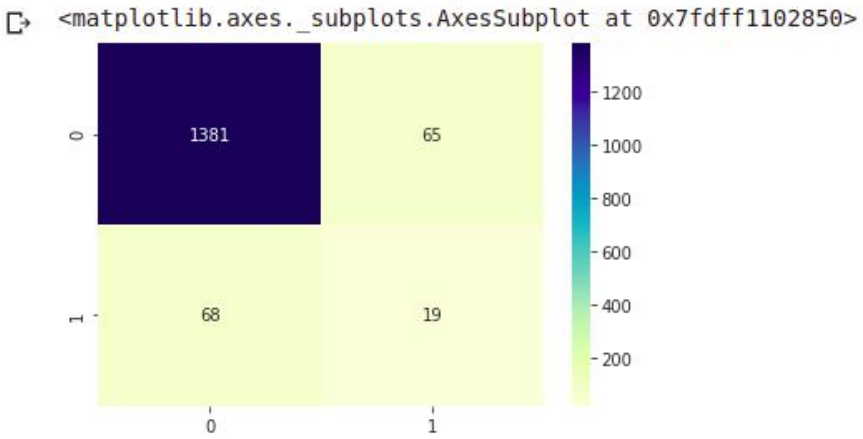

Figure 7: Decision tree Confusion Matrix

Here, we split our dataset into 30% percentages. So, 1533 instances are in the test set and the rest of the instances are in the train set.

	precision	recall	f1-score	support
0	0.95	0.96	0.95	1446
1	0.23	0.22	0.22	87
accuracy			0.91	1533
macro avg	0.59	0.59	0.59	1533
weighted avg	0.91	0.91	0.91	1533

A precision of 0.95 and recall is 0.96 from the decision tree. Finally, the accuracy of this algorithm is:

Accuracy = (TP + TN) / (TP + TN + FP + FN) * 100
= (1381+19)/ (1381+19+65+68) * 100
= 91.32%

3. Support Vector Machine (SVM):

A support vector machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression processes (Sisodia & Sisodia, 2018). It can solve linear as well as non-linear problems and work effectively for many practical problems (Sisodia & Sisodia, 2018). The idea of a support vector machine (SVM) is simple: The algorithm creates a line or a hyperactive plane, which separates the data into two classes (Sisodia & Sisodia, 2018). Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both
classification and regression analysis (Sisodia & Sisodia, 2018). It is one of the most popular and widely used machine learning techniques (Sisodia & Sisodia, 2018). This algorithm is also known as binary approach algorithm because it is used for binary classification like present or absence, either normal or abnormal, impactful or none impactful (Sisodia & Sisodia, 2018).

In this study, it is used for the prediction of a heart disease that is Stroke. SVM uses kernel tricks for performing non-linear classification (Sisodia & Sisodia, 2018). A kernel is used to transform low-dimensional space into high-dimensional space. There are three types of kernel such as linear, poly, and radial (Sisodia & Sisodia, 2018). In our method, we use a poly kernel. Before using this algorithm, I have normalized my data. Because SVM works well in normalized data.

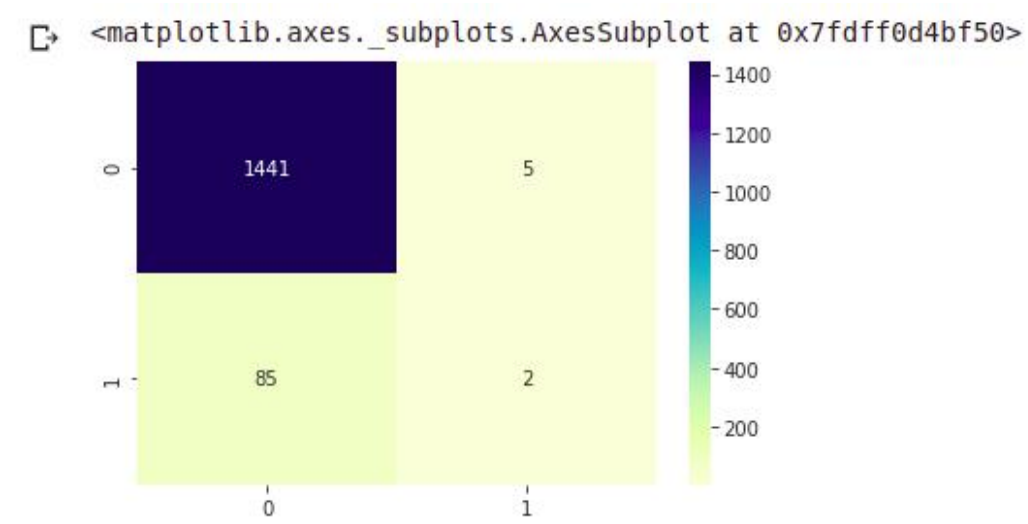

Figure 8: SVM Confusion Matrix

Accuracy = (TP + TN) / (TP + TN + FP + FN) * 100
= (1441+2)/ (1441+2+85+5) * 100
= 94.13%

	precision	recall	f1-score	support
0	0.94	1.00	0.97	1446
1	0.29	0.02	0.04	87
accuracy			0.94	1533
macro avg	0.62	0.51	0.51	1533
weighted avg	0.91	0.94	0.92	1533

Here the precision is 0.94 and recall is 1 from the SVM algorithm.

---

4. Result

Here this table represents the performance of three algorithms that we use in this paper. From the table, we can see that Random forest has more Correctly Classified than any other. So can say

that for our dataset random forest will be the best option. We can classify up to 94% by a random forest algorithm.

No. of Instance	Algorithm Name	Correctly Classified	Incorrectly Classified
1533	Random Forest	1444	89
1533	Decision Tree	1400	133
1533	SVM	1443	90

Table: Algorithm Performance

Performances of all classifiers based on various measures are plotted via a graph in Figure-2

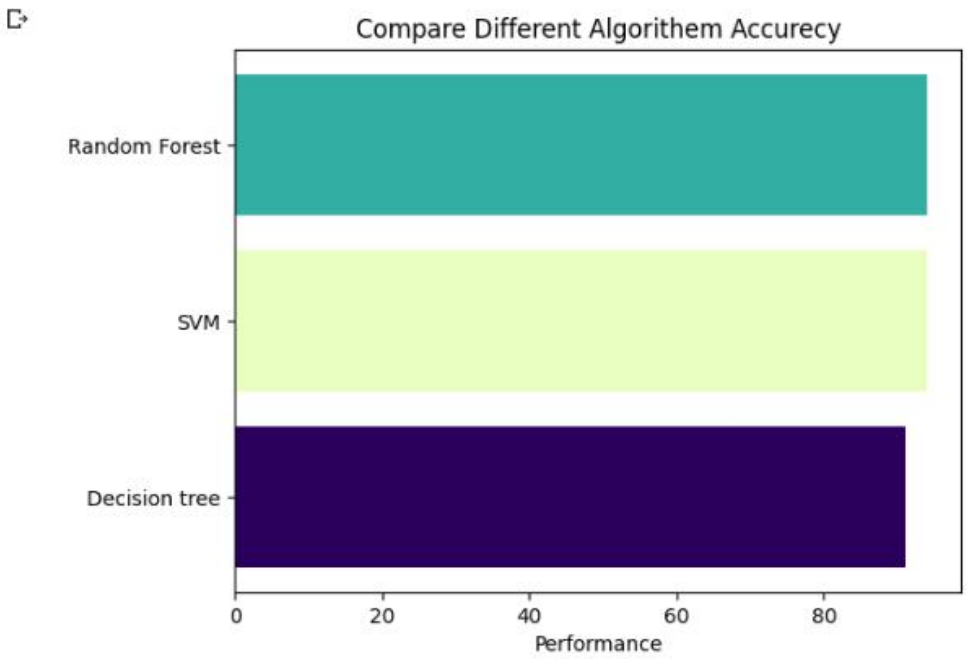

Figure 9: Compare Different Algorithm Accuracy

5. Conclusion

In this paper, we used three classifiers to find out the performance of stroke Occurrence of a person. The proposed Random forest and Svg (Support Vector Machine) classifiers both score the same to predict stroke. They both considered gender, age, hypertension, heart disease, average glucose level, marital status, work type, residence type, BMI, smoking status to predict stroke. The performance evaluation reveals that Random forest and Svg both provided the same accuracy as 94% compared with the decision tree algorithm which scores 91% accuracy. As a result, random forest or SVG algorithm can be considered for the prediction of stroke. In that case, we choose the random forest algorithm to predict the stroke. We have evaluated the relationship between these diseases and the possibility of occurring stroke in a human individual. So, if we can maintain this disease from an early stage then it will help to reduce stroke in our life. In the future, we would like to include deep learning and neural networks in our project. Also, we want to work with brain CT scan and MRI Together with an existing model that will boost the accuracy and performance of our project.

Reference

Hossain, M., Khurshid, S., Fatema, K., Hasan, M., & Hossain, M. (2021). Analysis and Prediction of Heart Disease Using Machine Learning and Data Mining Techniques. *Canadian Journal Of Medicine*. https://doi.org/10.33844/cjm.2021.60500

Sisodia, D., & Sisodia, D. (2018). Prediction of Diabetes using Classification Algorithms. *Procedia Computer Science*, *132*, 1578-1585. https://doi.org/10.1016/j.procs.2018.05.122

Introduction to Random Forest in Machine Learning. Engineering Education (EngEd) Program | Section. (2021). Retrieved 30 August 2021, from https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/.

Almadani, O., & Alshammari, R. (2018). Prediction of Stroke using Data Mining Classification Techniques. *International Journal Of Advanced Computer Science And Applications*, *9*(1). https://doi.org/10.14569/ijacsa.2018.090163

Olesen, J., Lip, G., Hansen, M., Hansen, P., Tolstrup, J., & Lindhardsen, J. et al. (2011). Validation of risk stratification schemes for predicting stroke and thromboembolism in patients with atrial fibrillation: nationwide cohort study. *BMJ*, *342*(jan31 1), d124-d124. https://doi.org/10.1136/bmj.d124

Sudha, A., Gayathri, P., & Jaisankar, N. (2012). Effective Analysis and Predictive Model of Stroke Disease using Classification Methods. *International Journal Of Computer Applications*, *43*(14), 26-31. https://doi.org/10.5120/6172-8599

Vieira, A., Soares, P., & Nunes, C. (2021). Predicting Independence 6 and 18 Months after Ischemic Stroke Considering Differences in 12 Countries: A Secondary Analysis of the IST-3 Trial. *Stroke Research And Treatment*, *2021*, 1-13. https://doi.org/10.1155/2021/5627868

Shah, D., Patel, S., & Bharti, S. (2020). Heart Disease Prediction using Machine Learning Techniques. *SN Computer Science*, *1*(6). https://doi.org/10.1007/s42979-020-00365-y

Emon, M., Keya, M., Meghla, T., Rahman, M., Mamun, M., & Kaiser, M. (2020). Performance Analysis of Machine Learning Approaches in Stroke Prediction. *2020 4Th International Conference On Electronics, Communication And Aerospace Technology (ICECA)*. https://doi.org/10.1109/iceca49313.2020.9297525

Sailasya, G., & Kumari, G. (2021). Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. International Journal Of Advanced Computer Science And Applications, 12(6). https://doi.org/10.14569/ijacsa.2021.0120662