Stroke Prediction Using Machine Learning and XAI

Rashik Buksh Rafsan, Hasib Ar Rafiul Fahim

Department of Computer Science and Engineering
East West University, Dhaka, Bangladesh
Email: {2018-3-60-088, 2019-1-60-036}@std.ewubd.edu

I. Introduction

When the blood flow to a portion of the brain is interrupted, a dangerous medical condition known as a stroke can result, which is life-threatening. Strokes are a medical emergency that require prompt care. There are 2 main causes of strokes: 85% of cases are ischemic, in which the blood flow is interrupted by a blood clot, and another is Hemorrhagic, in which a brain blood artery that is weak bursts. [1]

A. Background

In 2019, there were 101.5 million cases of stroke worldwide, with ischemic stroke accounting for 77.2 million cases, intracerebral hemorrhage for 20.7 million cases, and subarachnoid hemorrhage for 8.4 million. [2]

B. Aim

This project identifies an accurate model for predicting Stroke Outcome with Machine Learning based on previous data. Our target is to find out which factors cause stroke. Also, we will find out which algorithm is best for these kind of dataset like, a decision is dependent on multiple factors like Stroke Prediction.

III. Dataset

We have analyzed the Kaggle's dataset for Stroke Prediction [3]. The datasets consist of 5110 number of Instances 12 number of Attribute data published by "**fedesoriano**" Data Scientist at Kaggle. But we have taken a total of 360 data for the project. This data set includes 271 instances of one class and 89 instances of another class. The instances are described by 12 attributes, some of which are linear, and some are nominal. The data type is a string type with any numerical value. It consisted of various sections named id, age, gender, avg_glucose_level, smoking_status etc.

Sample Dataset is given below:

id	gender	age	hypertension	heart_disease
9046	Male	67	0	1
51676	Female	61	0	0

ever_married	work_type	Residence_type
Yes	Private	Urban
Yes	Self-employed	Rural

avg_glucose_level	bmi	smoking_status
228.69	36.6	Formerly smoked
202.21	N/A	never smoked

stroke
1
1

IV. METHODOLOGY

A. Data Collection Method

The method of collecting data was one of the most necessary parts of the project. For this project, we needed data on Stroke prediction dataset which mostly will need to be related to the hypertension, age, smoking type, etc. Due to limitation in time and resources, we were unable to visit hospitals and collect the data on stroke patients' data ourselves. Instead, we used the 'Stroke Prediction Dataset' from Kaggle for this project.

B. Data Preprocessing

After collecting the data, we needed to perform some preprocessing task before we could proceed to use any classifiers. Data preprocessing tasks included adding attribute names, replacing missing values with median values, and encoding

feature values. Moreover, hyper parameter tuning for two classifiers have been used to find the set of optimal hyper parameters.

C. Classifiers Selection

The Stroke Dataset used in this project had 360 instances with 12 attributes. Since this is a classification task, classifiers used for classification were selected. A total of three classifiers have been used which include Random Forest Classifier, Decision Tree Classifier and Support Vector Machine Classifier.

D. Selecting the Best Model

The models created using the four classifiers will be used to compare with each other to find the best model. For each classifier, the confusion matrix will be made, and a classification report will be generated. The classification reports will be used to find different evaluation metrics like Accuracy, Sensitivity, Specificity, F1-score, Confusion Matrix, ROC-AUC of the three models.

V. Results

The Accuracy, Sensitivity, Specificity, F1-score, Confusion Matrix, ROC-AUC of the 3 models is presented below,

Model	Classification Metrics	Score
Random Forest	Sensitivity	0.71
	Specificity	0.875
	f1-score	0.72
	Confusion Matrix	[[42 6] [7 17]]
	Accuracy	82%
	Balanced Accuracy	0.7925
Decision Tree Classifier	Sensitivity	0.58
	Specificity	0.812
	f1-score	0.60
	Confusion Matrix	[[39 9] [10 14]]
	Accuracy	73%
	Balanced Accuracy	0.696
Support Vector Machine Classifier	Sensitivity	0.21
	Specificity	0.896
	f1-score	0.29
	Confusion Matrix	[[43 5] [19 5]]
	Accuracy	67%
	Balanced Accuracy	0.553

Among the 3 classifier models we got the best results from Random Forest classifier because of higher sensitivity, f1-score and accuracy which are respectively 0.71, 0.72 and 82%. Also, the Balanced Accuracy is higher for Random Forest which is 0.7925.

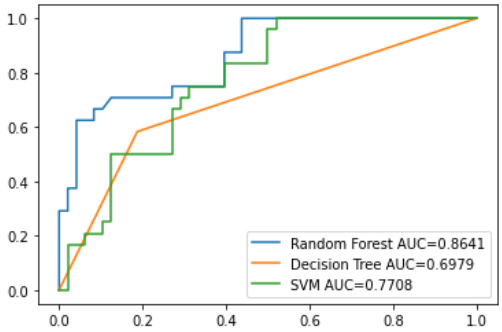
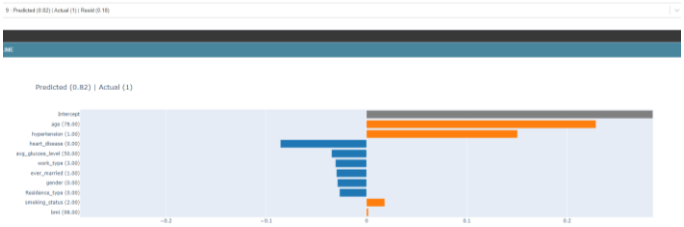


Fig-1: ROC Curve of 3 models

From the above curve we can see that Random Forest covers the highest area and the AUC score is 0.8641. From this evaluation we can interpret that random forest performs better among all the classifier models for our dataset.



We have used Explainable AI known as XAI which is a machine-learning systems that can justify their actions, identity, advantages, and disadvantages and provide an understanding of how they will act in the future. Using this system in this project was to identify what are the main reasons of stroke. After analyzing through LIME, we got these factors which are responsible for Stroke accordingly: Age, Hypertension, Smoking Status, Work Type etc.

VI. Conclusion

Machine learning techniques are largely being used in the domain of medical research. Predicting diseases in medical field is a big challenge and sometimes it can be a hard for humans to interpret. For this reason, to significantly improve detection accuracy for Stroke, we implemented 3 machine learning classifiers. From this study we found Random Forest Classifier gives the best result and we have used XAI for getting the factors which are responsible for Stroke.

REFERENCES

[1] Stroke. (2022). Retrieved 5 September 2022, from https://www.nhs.uk/conditions/stroke/#:~:text=A%20stroke%20is%20a%20serious,damage%20is%20likely%20to%20happen

[2] 2021 Heart Disease & Stroke Statistical Update Fact Sheet Global Burden of Disease. (n.d.). Professional Heart Daily. https://professional.heart.org/-/media/PHD-Files-2/Science-News/2/2021-Heart-and-Stroke-Stat-Update/2021_Stat_Update_factsheet_Global_Burden_of_Disease.pdf

[3] Stroke prediction dataset. (n.d.). Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset