AHSANULLAH UNIVERSITY OF SCIENCE & TECHNOLOGY

Department of Computer Science and Engineering

CSE4108: Artificial Intelligence Lab

Fall 2020

PROJECT REPORT

Brain Stroke Prediction

Submitted To

Md. Siam Ansary
Department of Aust, CSE

Submitted By

Sanjida Akter Ishita 170204089

Contents

1 Introduction

"Stroke" occurs when a blood vessel in the brain ruptures and bleeds, or when there's a blockage in the blood supply to the brain. The rupture or blockage prevents blood and oxygen from reaching the brain's tissues. Without oxygen, brain cells and tissue become damaged and begin to die within minutes.

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This project is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status.

The dataset contains 250 real world observations and 11 different attributes

2 Attributes

- Gender: 'Male' or 'Female'
- Age: age of patient
- Hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension.
- Heart Disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- Marital Status: 'No' or 'Yes'
- Work Type: 'Private' or 'Self-employed' or 'Govt_job' or 'Never_worked' or 'children'
- Residence Type: 'Rural' or 'Urban'
- Avg Glucose Level: average glucose level in blood
- BMI: body mass index
- Smoking Status: 'never smoked' or 'formerly smoked' or 'Unknown' or 'smokes'
- Stroke: 1 if the patient had a stroke or 0 if not

3 Models

3.1 SVM(Support Vector Machine) Classifier

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text.

SVM Algorithm in Machine Learning Support Vector Machine or SVM algorithm is a simple yet powerful Supervised Machine Learning algorithm that can be used for building both regression and classification models. SVM algorithm can perform really well with both linearly separable and non-linearly separable datasets.

3.2 Decision Tree Classifier

A Decision Tree algorithm is one of the most popular machine learning algorithms. It uses a tree like structure and their possible combinations to solve a particular problem. It belongs to the class of supervised learning algorithms where it can be used for both classification and regression purposes.

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

4 Main Functionalities

4.1 Import sklearn libraries

Some sklearn libraries are imported in code which is related to project.

4.2 Openning the dataset

Firstly I saved the dataset csv file in my drive. Then I executed the lines which contain import drive from the google colab. After mounting the drive I copied the path from my drive of dataset file and wrote the command that read the csv file. That is how the dataset file opened.

4.3 Handling missing value

In this part missing values are handled in dataset. Here main function is replace missing values of BMI with it's mean value, remove (drop) data associated with missing values in variable 'smoking_status' with 'never smoked' 'formerly smoked' 'Unknown' 'smokes'.

4.4 Exchanging class to Numerical values

Here normally converting the class to digits for further calculation.

4.5 Histogram features

Here it shows some plots that discover, and show, the underlying frequency distribution (shape) of a set of continuous data.

4.6 Data Processing

Collection, manipulation, and processing collected data for the required use is known as data processing. There are some part of it,

4.6.1 Balance target(Stroke) class

In this part it shows the percentage of target class in dataset. Simply saying, how many patients are having stroke or not among 250 patients.

4.6.2 Handling Imbalanced Class

Mainly scaling the data for further calculation. Here used one of oversampling technique called Synthetic Minority Oversampling Technique (SMOTE), by synthesising new samples from the minority class to have the same number of samples as the majority class (illustrated in figure below). Oversampling technique is chosen because we do not want to loose significant amount of information as if we use undersampling technique.

4.6.3 Data Splitting

4.6.4 Data Normalization

4.7 Model Trainning and Evaluation

To model trainning and evaluation here trained two types of model to see the result for which model will have more accuracy for the data set. Here also we found accuracy, f1, precision and recall. This will help to make a comparison between two models and also found performance matric.

4.8 Performance comparison between 2 models(SVM and DT)

4.8.1 Result Table

Classifiers	**Accuracy**	**Recall**	**Precision**	**F1 Score)**
SVM(Support Vector Machine) Classifier	0.86	1.00	0.78	0.88
Decision Tree Classifier	0.79	0.86	0.75	0.80

Table 1: Result Table

4.8.2 Result Chart

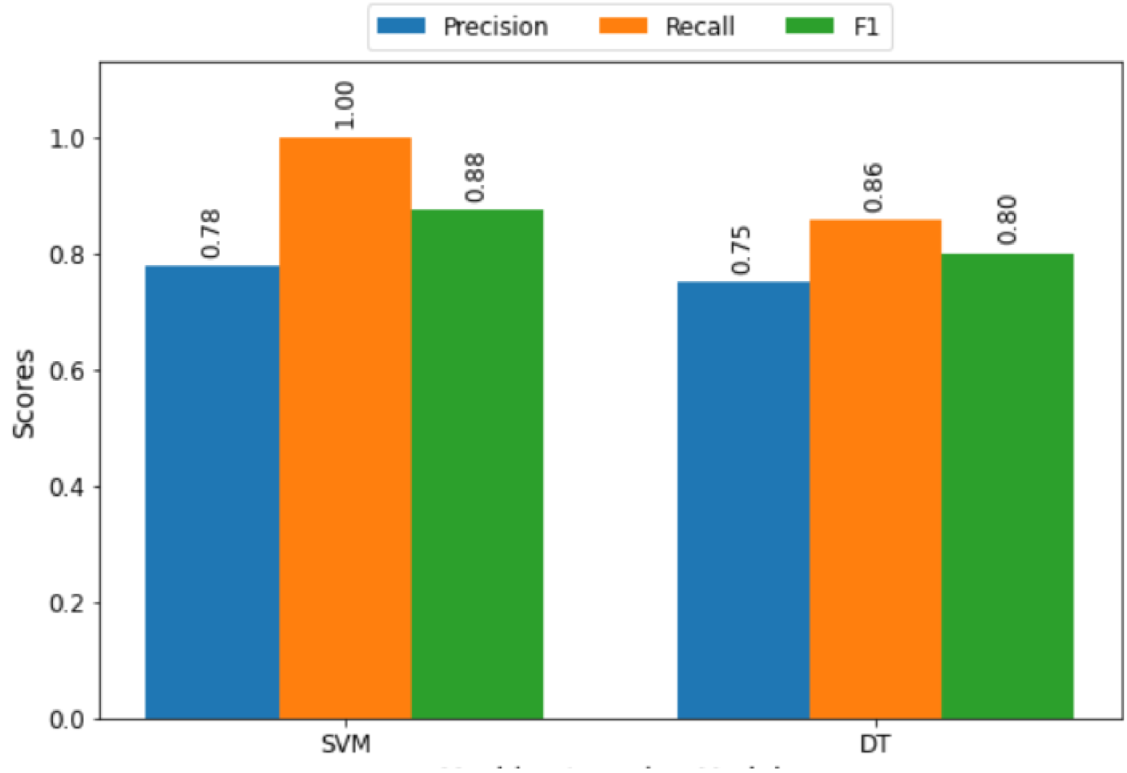

Figure 1: Machine Learning Model

From the result table it is seen that SVM has gained the highest accuracy, which is 86%.

5 Conclusion

After the model analysis, the accuracy of SVM classfier is 86% which is highest.The precision is 78%, it predicts that a patient has stroke, it is correct around 78% of the time. The recall is 100%. It gives measure that 100% of the time the model is able to identify the relevant data.

For the Decision Tree classification, the accuracy is 79%. The precision is 75%, it predicts that a patient has stroke or not, it is correct around 75% of the time. The recall is 86%. It gives measure that 86% of the time the model is able to identify the relevant data.

Lastly,SVM(Support Vector Machine) Classification works good for the dataset.