****

**American International University Bangladesh**

**Stroke Prediction Using Machine Learning**

**Python Final Project**

Section : B

Course : Programming In Python

Semester : Spring 21-22

Date : 17-04-2022

Course Teacher : Akinul Islam Jony

|  |  |
| --- | --- |
| **Name** | **Id** |
| TAJWAR, MD. ABTAHI | 19-40281-1 |
| KHONNDOKAR AHAMED ALFA SUNNY | 19-40251-1 |
| WATHIN MARMA | 19-40379-1 |
| JAHNNABI MAZUMDER | 19-40294-1 |

**Stroke Prediction Using Machine Learning**

1. **Objective**

Stroke is a medical condition in which the blood vessels in the brain rupture, causing brain damage. Symptoms may appear if the brain's flow of blood and other nutrients is disrupted. Stroke is the leading cause of death and disability worldwide, according to the World Health Organization (WHO). Early awareness of the numerous stroke warning symptoms can assist to lessen the severity of the stroke.

In this project, we will try to accomplish following objectives

* Predict if a person has any stroke occurrence
* Discover which machine learning models works best for stroke prediction

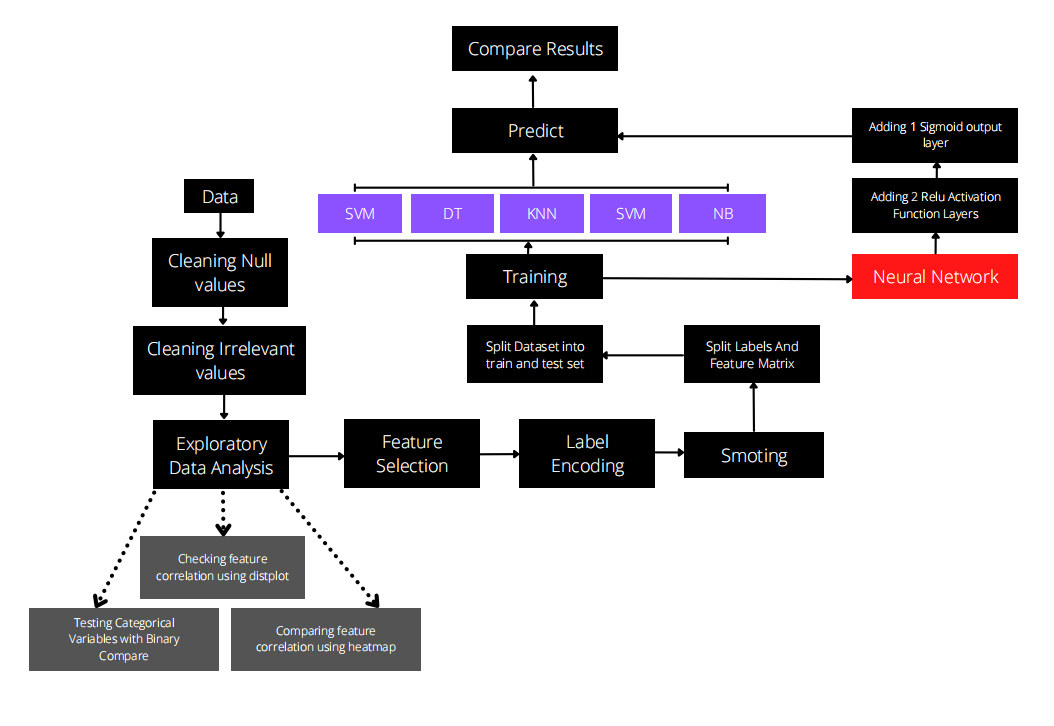
1. **Solution Design**
   1. **Data Source**

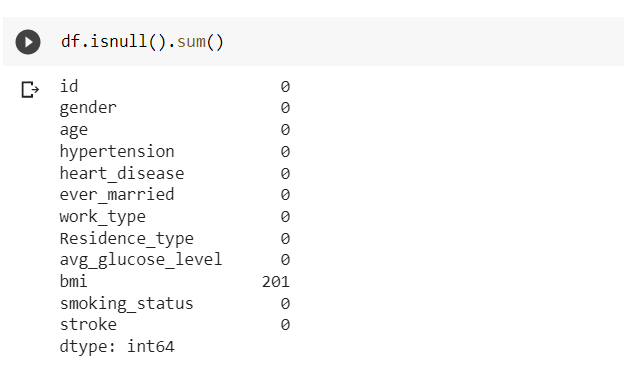
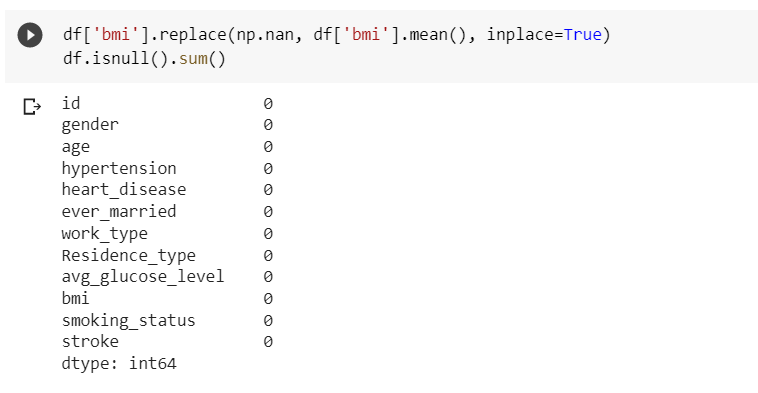
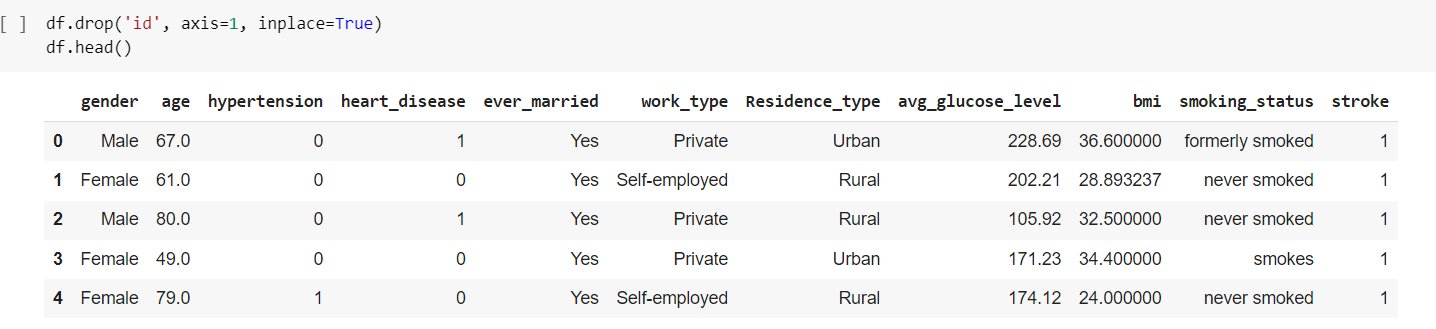
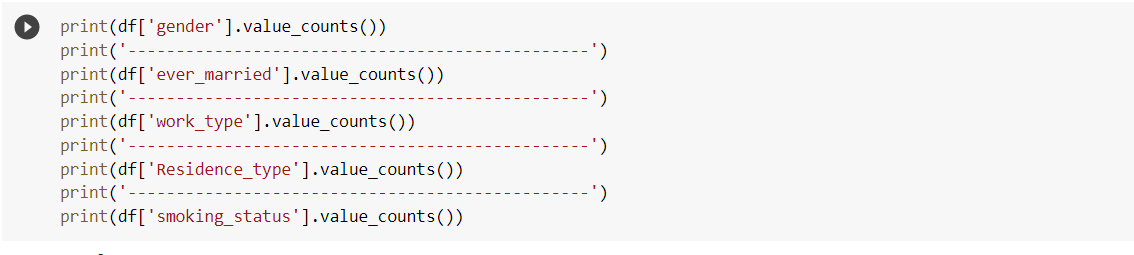
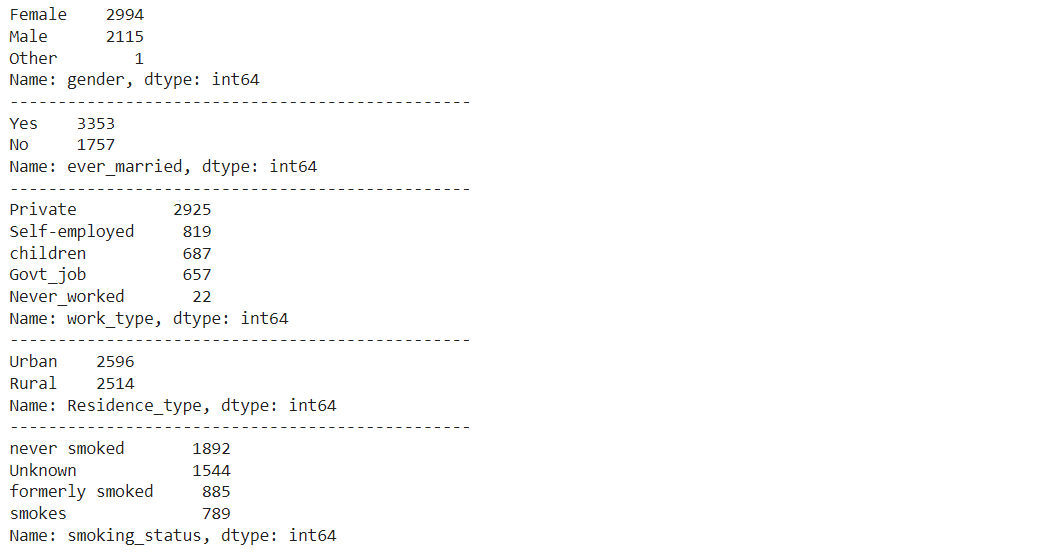
Source: [Stroke Prediction Dataset | Kaggle](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset)

First we searched datasets on google and kaggle.com. Finally we found a prediction dataset from kaggle.com containing following features

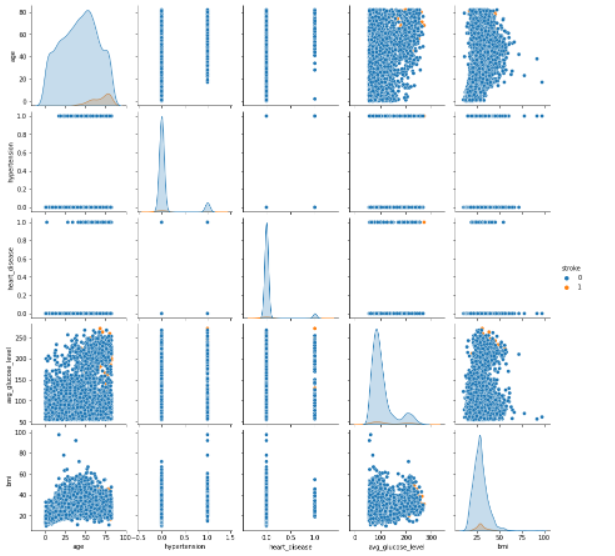
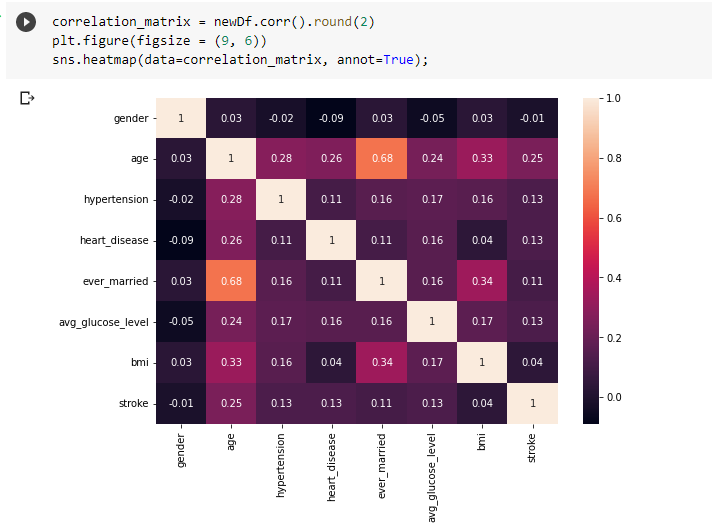
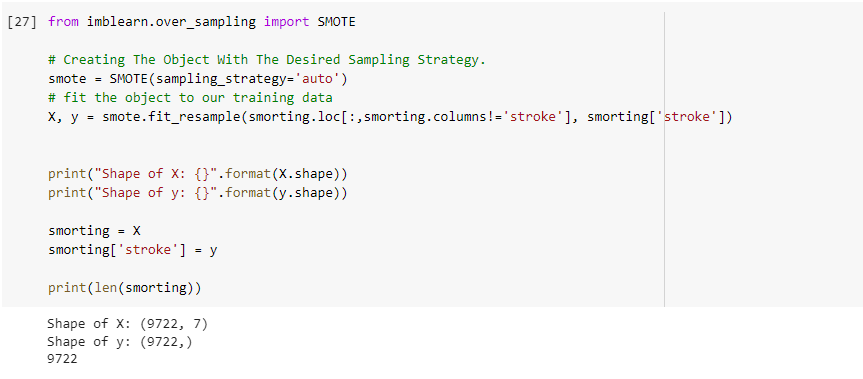
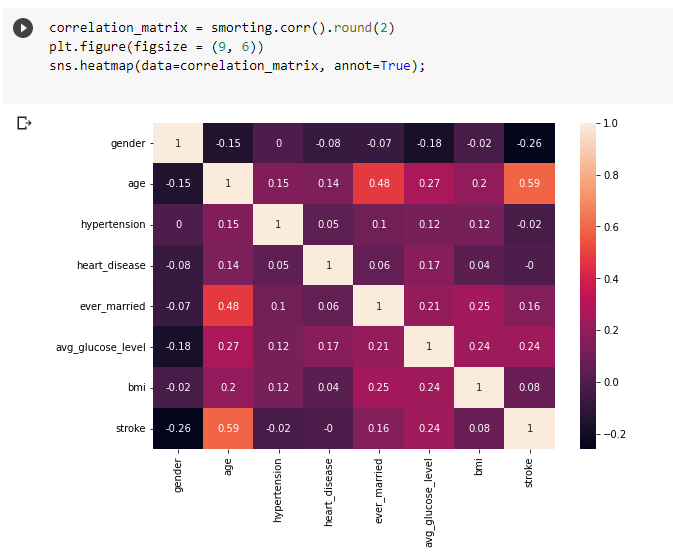
* id: Unique Id
* gender: Gender of the person
* age: Age of the person
* hypertension: Hypertension binary feature
* heart\_disease: Heart disease binary feature
* ever\_married: Has the patient ever been married?
* work\_type: Work type of the patient
* Residence\_type: Residence type of the patient
* avg\_glucose\_level: Average glucose level in blood
* bmi: Body Mass Index
* smoking\_status: Patient smokes or not
* stroke: Patient has stroke or not **[target]**

**2.2 Flowchart**

****

1. **Data Preprocessing**
   1. **Null Check  
      **
   2. **Replacing NaN values**Replaced 201 bmi nan values with average value, as there only 3.93% nan value exists  
      
   3. **Dropping Unnecessary Columns**As the Id column does not have any impact on the whole process, this column was dropped.  
      
   4. **Checking for unwanted values  
        
      **
2. **Exploratory Data Analysis**
   1. **Comparing relation with non numeric data**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

* 1. **Checking inner relation of features  
       
     **
  2. **Checking inner relation after applying SMOT  
       
     **

1. **Feature Selection**After the exploratory data analysis, we were able to pick the right features for better results. Following principles are used for feature selection

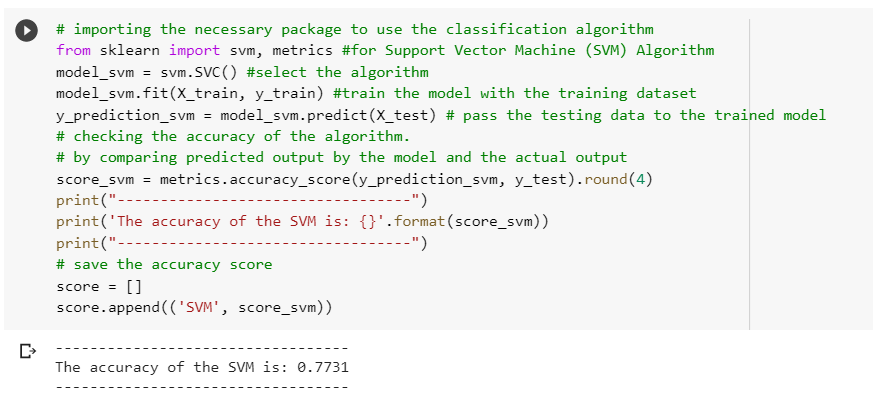
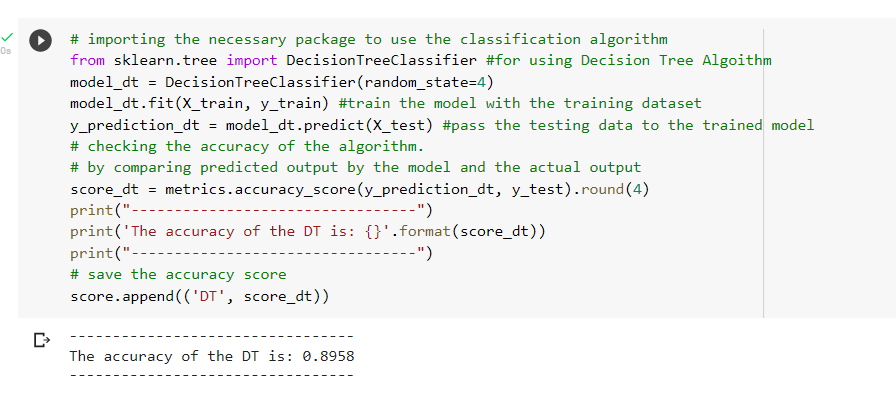
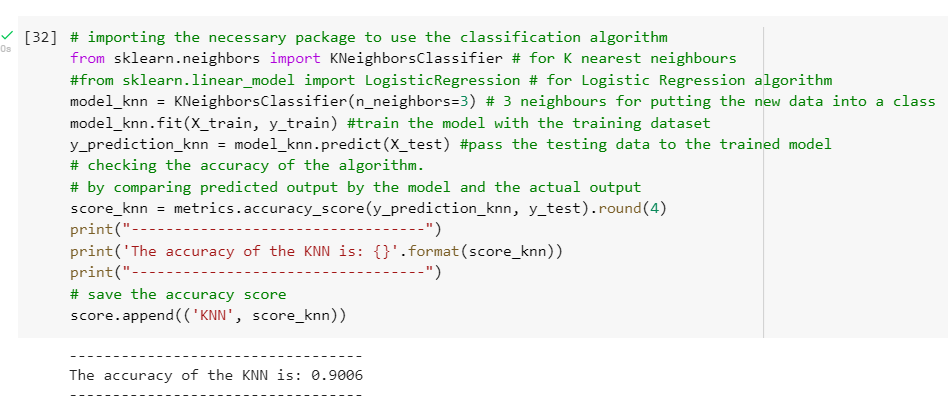
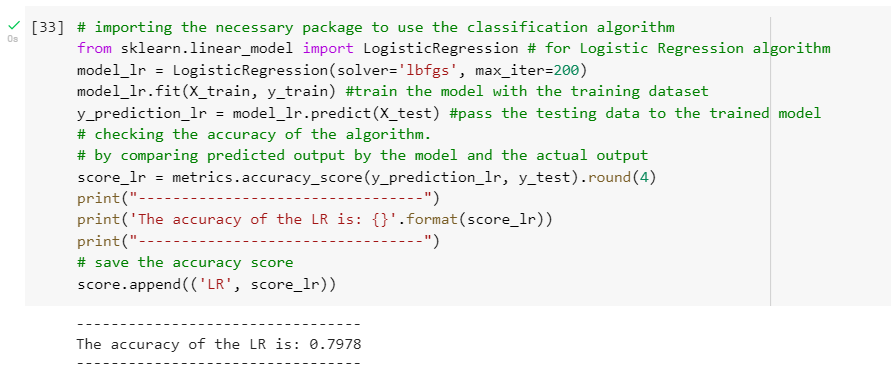
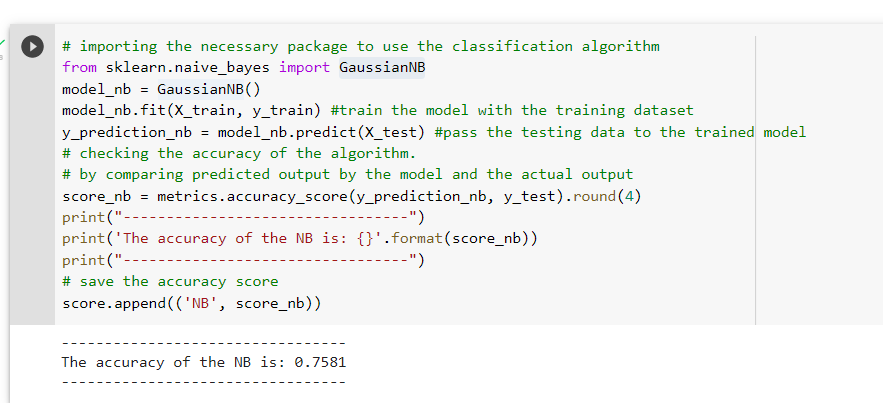
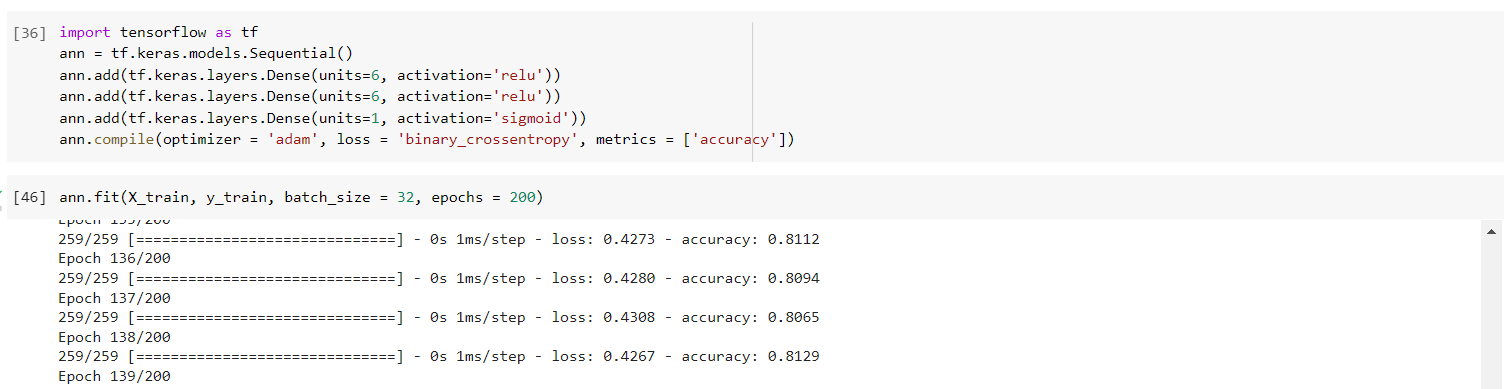
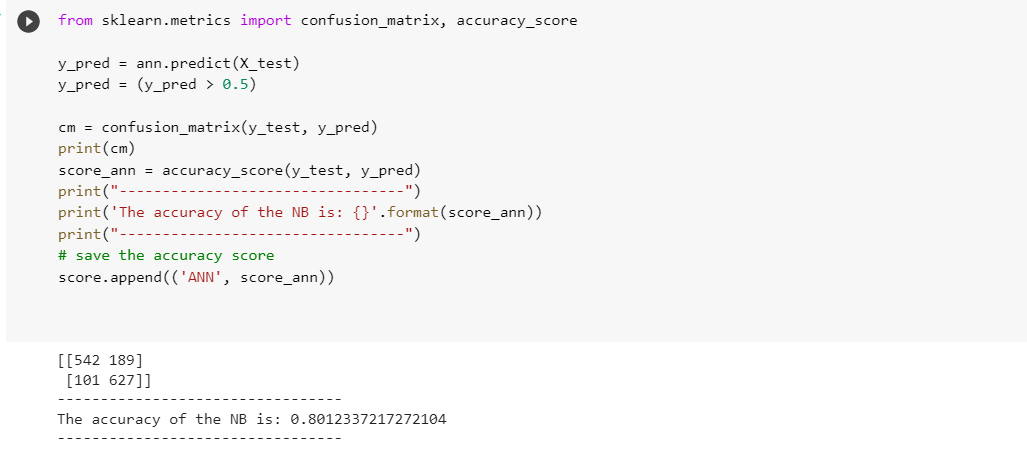
* Checked relation with target variable, picked if there is higher relation
* Checked inner relations of features. Features with higher inner relation were dropped for better results.

**Selection Features:**

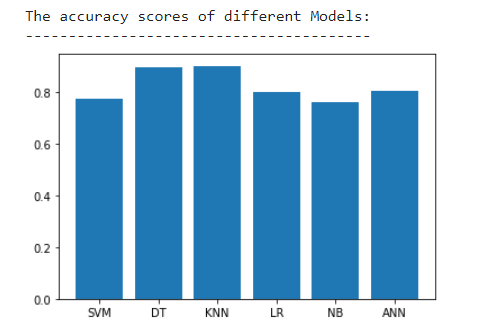
* gender
* age
* hypertension
* heart\_disease
* ever\_married
* avg\_glucose\_level
* bmi

**Rejected Features**

* smoking\_status
* Residence\_type
* work\_type

1. **Model Creation**
   1. **Support Vector Machine  
      **
   2. **Decision Tree Classifier  
      **
   3. **KNeighbors Classifier  
      **
   4. **Logistic Regression  
      **
   5. **Gaussian Naive Bayes  
      **
   6. **Artificial Neural Network  
        
      **
2. **Discussion**

In this project we have compared different models to predict stroke. And from the experiment we observed fluctuation of performance.



For this experiment, we have used classifier models. For the experiment, KNearest Neighbors Classifier provides the best accuracy which is 90%. Decision tree performs close to KNN and it has 89.58% accuracy. From the graph, it can be seen that the worst performing model is Gaussian Naive Bayes which produced 75.81% correct results.

We searched through various websites for example: google.com, kaggle.com, ICS archive etc. We found our desired dataset in kaggle.com. We faced problems while feature selection, which is the features of the dataset, didn't have much correlation. Also there was a potential column ‘smoking\_status’ which had to be dropped because of its large amount of unknown data.