BRAC
UNIVERSITY

Inspiring Excellence

CSE422- Artificial Intelligence-Machine Learning Lab Project

Project Name- STROKE PREDICTION

Section-10

Date: August 30, 2022.

S.N	**Name**	**Student ID**
1	NAZMUL HASSAN OYON	20101528
2	SANJANA BINTE ALAM	20301455
3	SAMIHA JUBAIDA ALAM	20301458
4	RIFHA HOSSAIN MUNAJA	20301466

Table of Contents

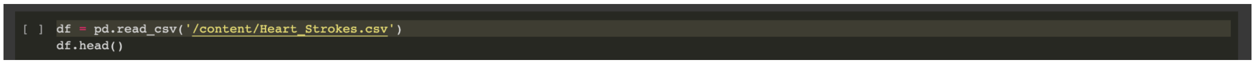
INTRODUCTION:

One of the disorders that poses the greatest threat to life for people over 65 is stroke. It affects the brain similarly to a "heart attack," which damages the heart. It is the third greatest cause of death in both developed and developing nations. Whenever a stroke illness develops, it can potentially cause death as well as expensive medical care and permanent disability. A stroke claims the lives of one person every four minutes, although up to 80% of strokes can be checked if we could recognize or anticipate them early on. A person can be saved if there is a predictor that can predict if the person is going to have a stroke or not which will be categorized by different characteristics. So that can be a good invention for all.

The goal of this project is to develop a model that can determine if a person will have a stroke or not. It does this by using machine learning. For this procedure, we made use of a dataset that contains information about a person's BMI, work_type, age, residence_type, ang_glucose level,gender, smoking status, etc and other information . We use the data gathered from the various features which included in the dataset to evaluate whether or not this person is likely to have a stroke.

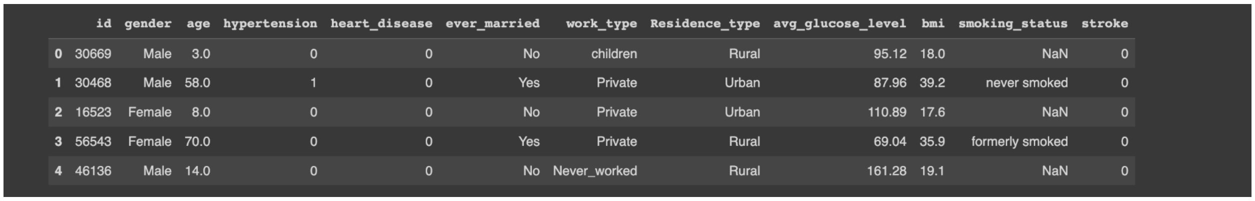
Methodology:

Dataset: We choose our dataset from an online dataset source known as Kaggle and choosed dataset that has 11 clinical features for predicting stroke events. In the project we have opened our dataset as df.



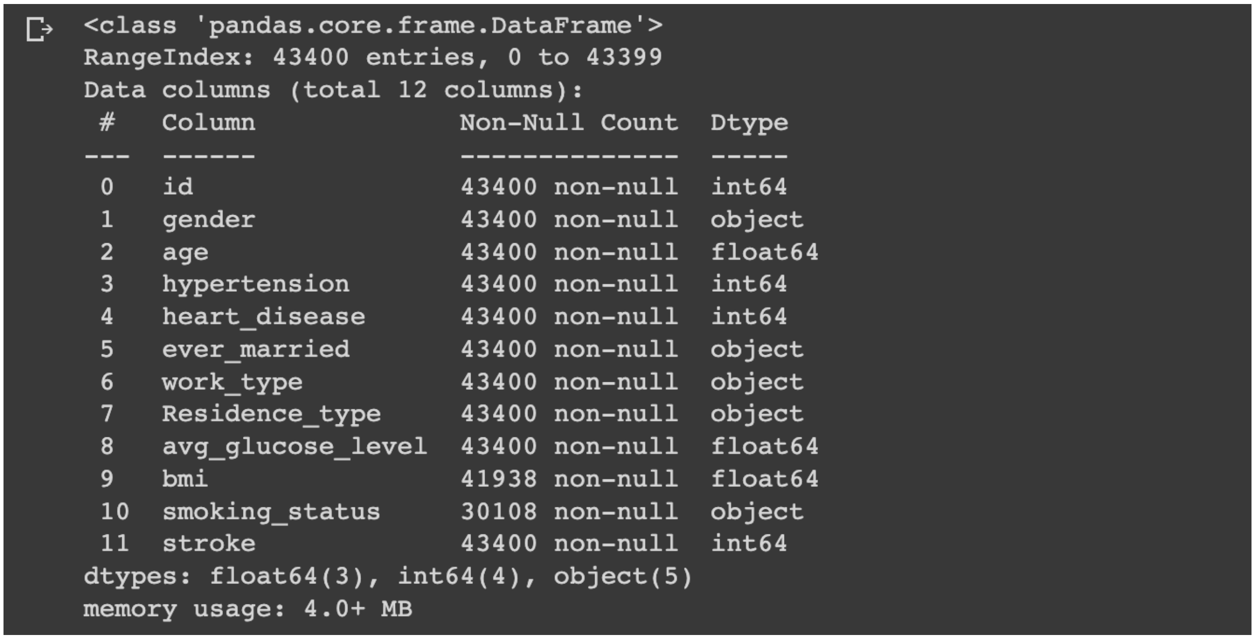
```
[ ]  df = pd.read_csv('/content/Heart_Strokes.csv')
     df.head()
```

Dataset details: The dataset we have used there has 43400 rows and 12 columns. And, these 12 column are id, gender, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke.



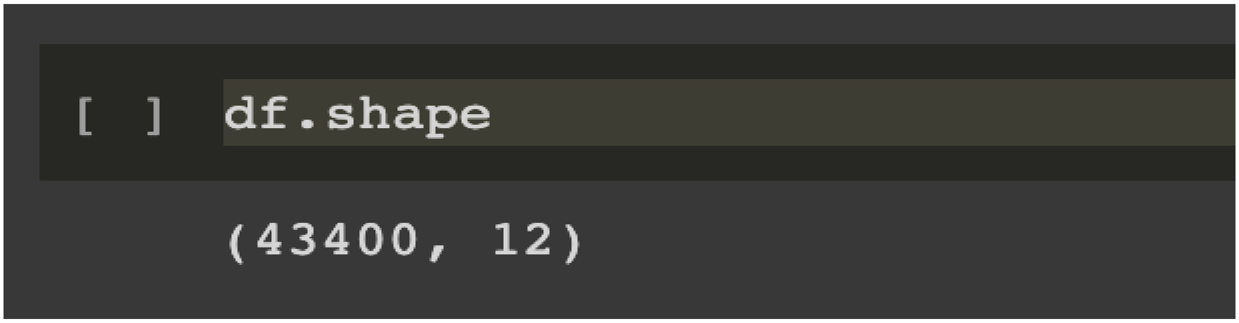
	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	30669	Male	3.0	0	0	No	children	Rural	95.12	18.0	NaN	0
1	30468	Male	58.0	1	0	Yes	Private	Urban	87.96	39.2	never smoked	0
2	16523	Female	8.0	0	0	No	Private	Urban	110.89	17.6	NaN	0
3	56543	Female	70.0	0	0	Yes	Private	Rural	69.04	35.9	formerly smoked	0
4	46136	Male	14.0	0	0	No	Never_worked	Rural	161.28	19.1	NaN	0

There are three types of data type of columns are int64, object, float64.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43400 entries, 0 to 43399
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 43400 non-null  int64  
 1   gender             43400 non-null  object 
 2   age                43400 non-null  float64
 3   hypertension       43400 non-null  int64  
 4   heart_disease      43400 non-null  int64  
 5   ever_married       43400 non-null  object 
 6   work_type          43400 non-null  object 
 7   Residence_type     43400 non-null  object 
 8   avg_glucose_level  43400 non-null  float64
 9   bmi                41938 non-null  float64
 10  smoking_status     30108 non-null  object 
 11  stroke             43400 non-null  int64  
dtypes: float64(3), int64(4), object(5)
memory usage: 4.0+ MB
```

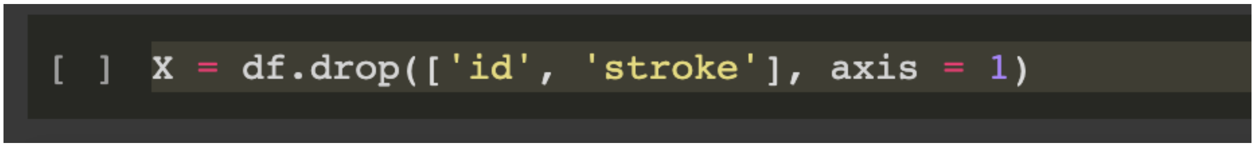
In the total dataset we have 43400 row and 12 column over all, here are the shape of dataset are given below:



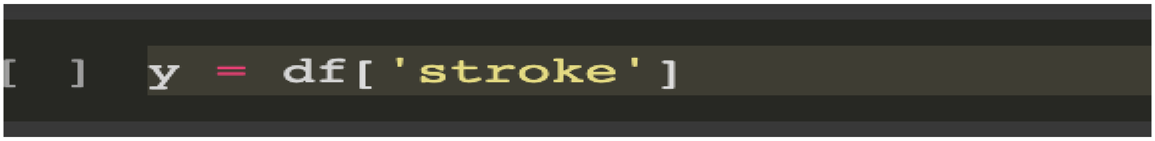
```
[ ]  df.shape

     (43400, 12)
```

Pre-processing technique:

In any machine learning project, after reading the dataset we have to pre-process our data for using the machine learning algorithms. Since, we have these 12 columns in our dataset (id, gender, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke). If we look carefully here id can not be a feature for heart stroke prediction. Because there is no relation with id for predicting the heart stroke. So, we will drop the id from our dataset and also we will put our dataset into variable X and y. In the X we will store every column except the target column. And, we will store the target column as a stroke in the y variable.

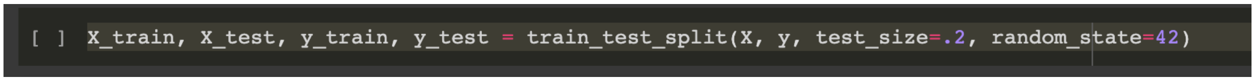


```
[ ] X = df.drop(['id', 'stroke'], axis = 1)
```



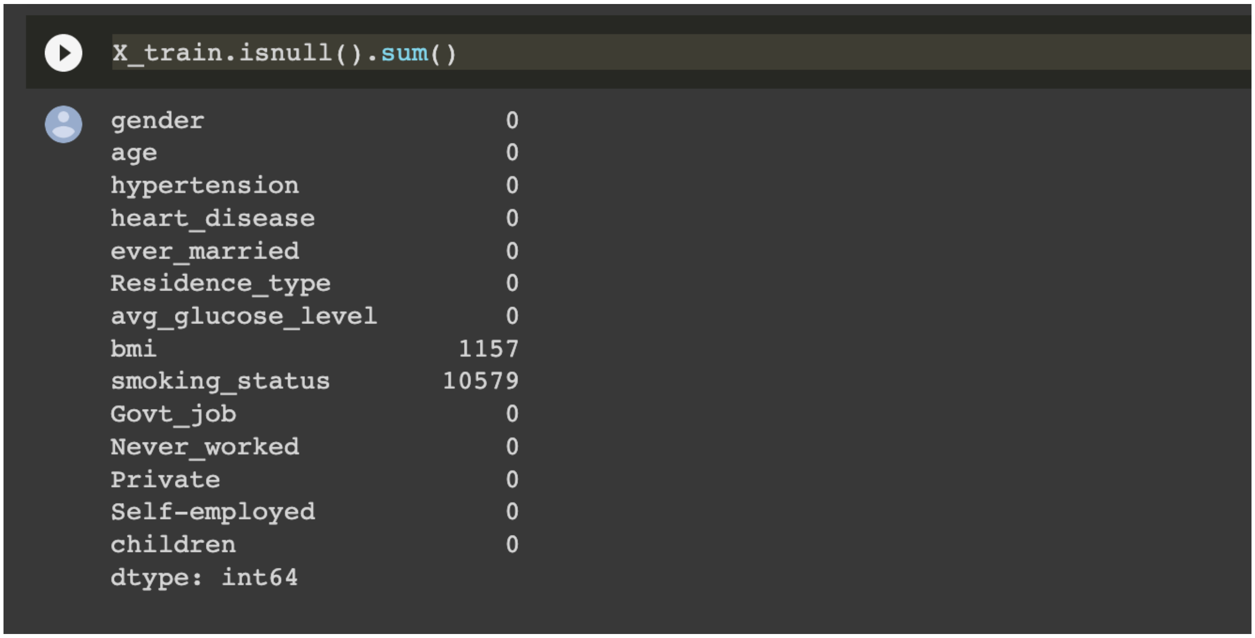
```
[ ] y = df['stroke']
```

After these, we will test 20% of our data and the rest of the data will be on the training part. For X variable training and test part will be as X_train, X_test and again for y variable it will be y_train, y_test.



```
[ ] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=.2, random_state=42)
```

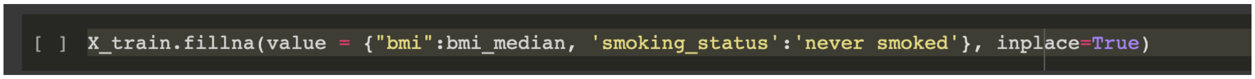
Generally , machine learning algorithms don't work on the dataset if there exists a ‘NaN’ value. So, in the dataset we have to ensure that no NaN value exist in the dataset.



```
X_train.isnull().sum()
```

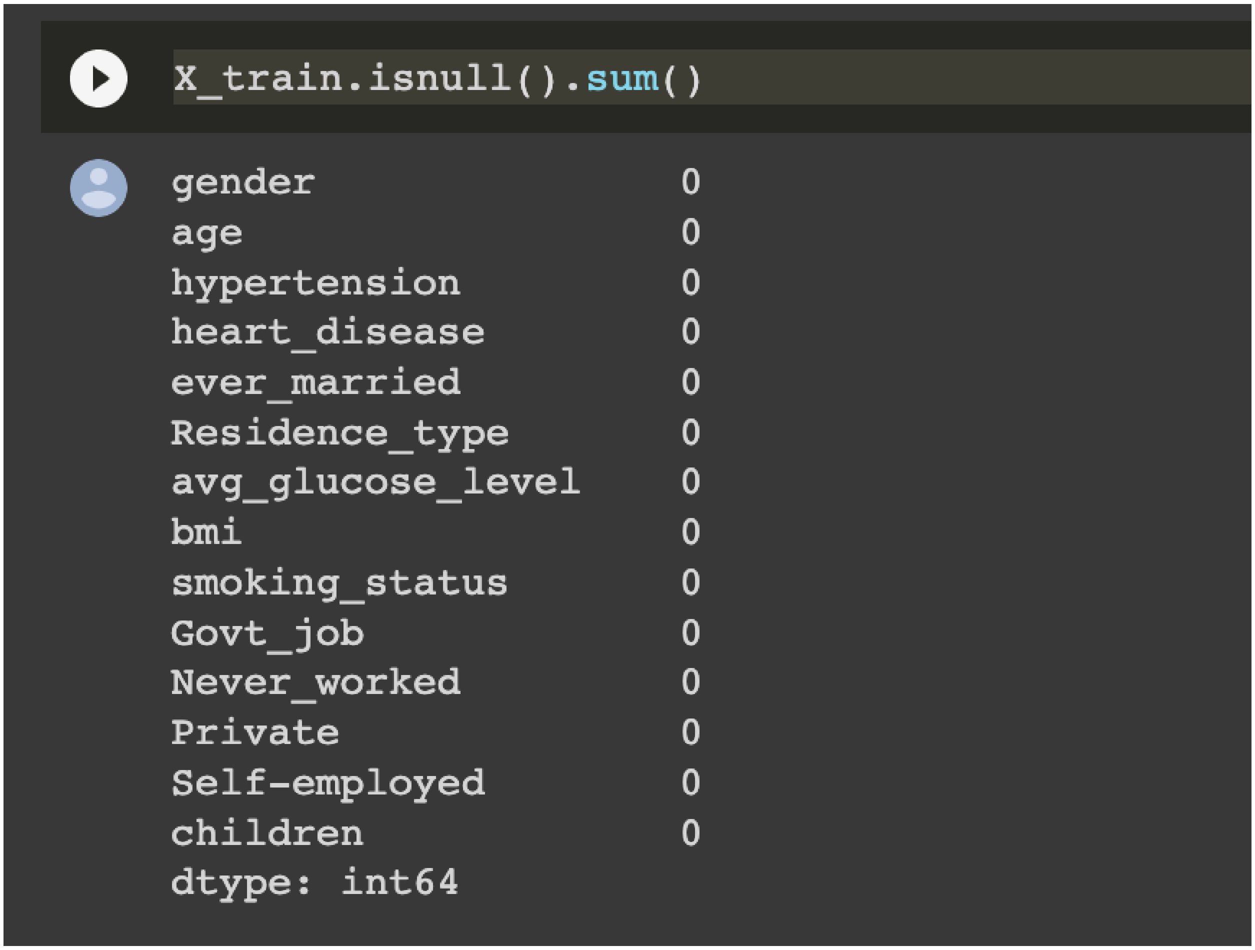
```
gender                   0
age                      0
hypertension             0
heart_disease            0
ever_married             0
Residence_type           0
avg_glucose_level        0
bmi                   1157
smoking_status       10579
Govt_job                 0
Never_worked             0
Private                  0
Self-employed            0
children                 0
dtype: int64
```

After implementing ‘isnull().sum()’ we can see that in bmi and smoking status, their null value exists. So, in order to work with null values, we will fill the median value of bmi in the corresponding null values. Again, from the dataset if we execute mode then the mode will be ‘never_smoked’. As this value is the mode, we will replace the null value with ‘never_smoked’ in the smoking_status feature.



```
[ ] X_train.fillna(value = {"bmi":bmi_median, 'smoking_status':'never smoked'}, inplace=True)
```

After the processing of the null values, we will further check if their null values exist or not.



```
X_train.isnull().sum()
```

```
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
Residence_type       0
avg_glucose_level    0
bmi                  0
smoking_status       0
Govt_job             0
Never_worked         0
Private              0
Self-employed        0
children             0
dtype: int64
```

Finally, we have completed all the necessary steps to implement our machine learning model and now our dataset is ready to perform.

Applied Model:

SVC MODEL :

The Linear Support Vector Classifier (SVC) method applies a linear kernel function to perform classification and it performs well with a large number of samples.A support vector classifier is a supervised machine learning algorithm that can be used for both classification and regression tasks.Support vector machines are a set of supervised learning methods used for classification, regression, and outliers detection.SVC model can be used to to detect cancerous cells based on millions of images or it can be used to predict future driving routes with a well-fitted regression model.

Decision Tree :

A decision tree is a supervised learning system in which each leaf node indicates the goal variable or the result and each internal node in the tree reflects the attribute in the dataset. The decision tree operates in the method described below:-

1. The top main node is where all of the samples are located.
2. Features can have a categorical nature.
3. After choosing the optimal splitting attribute, the model is built using that data.
4. The model is created using statistics, which is then applied to the test data.

Logistic Regression:

A statistical analysis method called logistic regression uses previous observations from a data set to predict a binary outcome, such as yes or no. By examining the correlation between one or more already present independent variables, a logistic regression model predicts a dependent data variable.This logistics regression algorithm is used to predict the probability of a target variable.Logistic regression can also play a role in data preparation activities by allowing data sets to be put into specifically predefined buckets during the extract, transform, load process in order to stage the information for analysis.

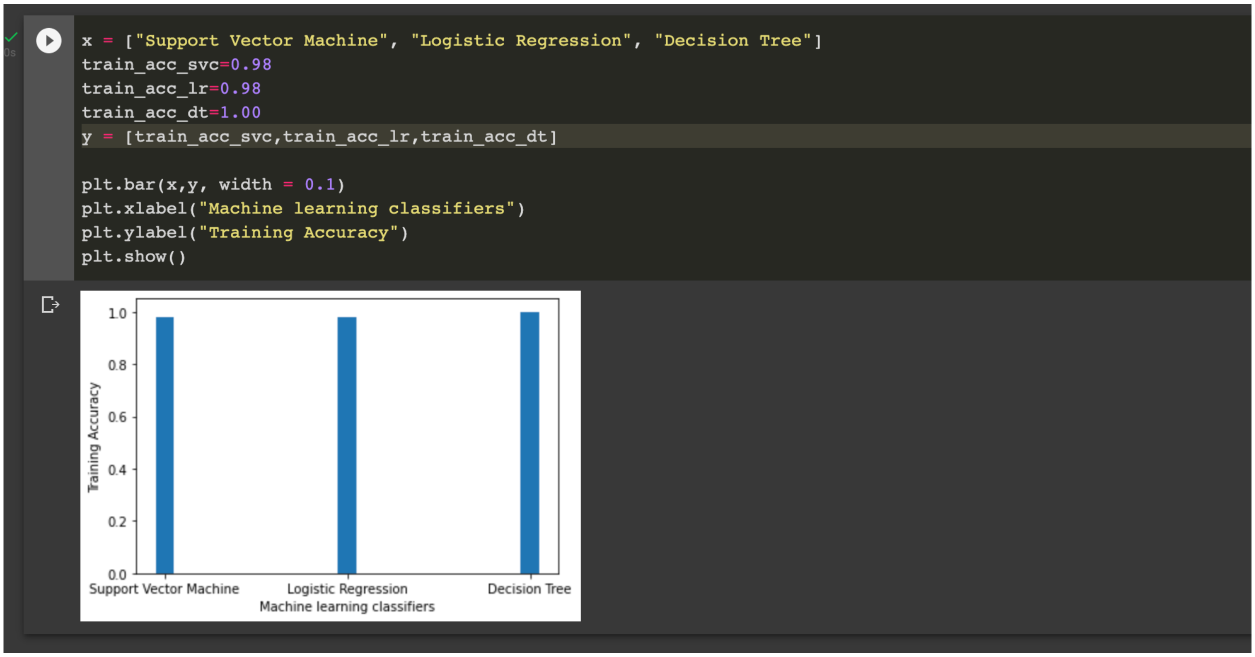
Results:

The table below shows the three machine learning model accuracy which is SVC, Decision Tree and Logistic Regression and confusion matrix models technique which applied to this model.

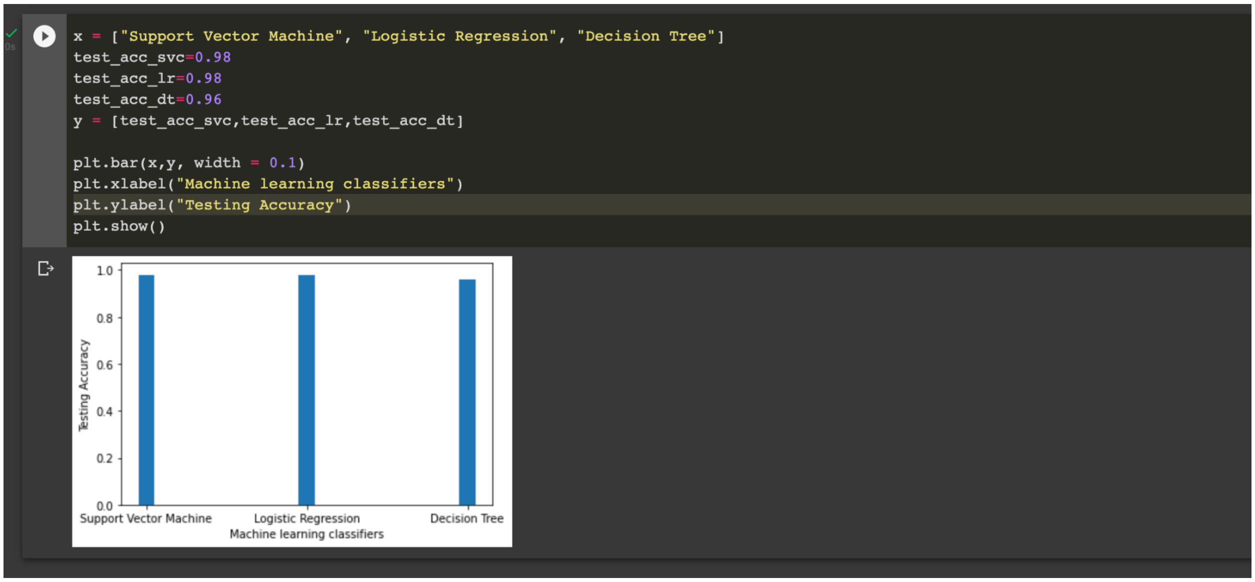
Models	Training Accuracy	Testing Accuracy
SVC	0.98	0.98
Decision Tree	1.00	0.96
Logistic Regression	0.98	0.98

Comparison:

Here we can see the three training accuracy of the model in graphical way-

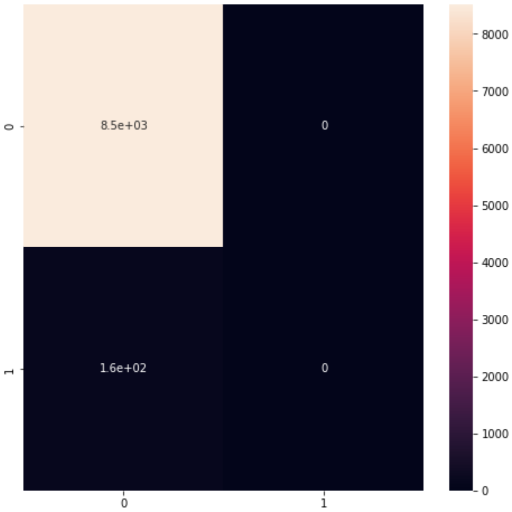


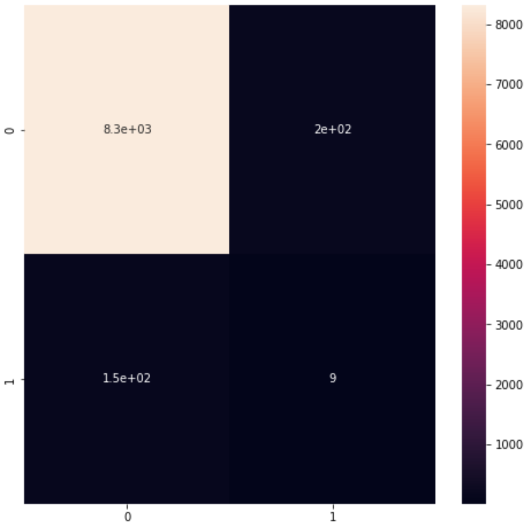
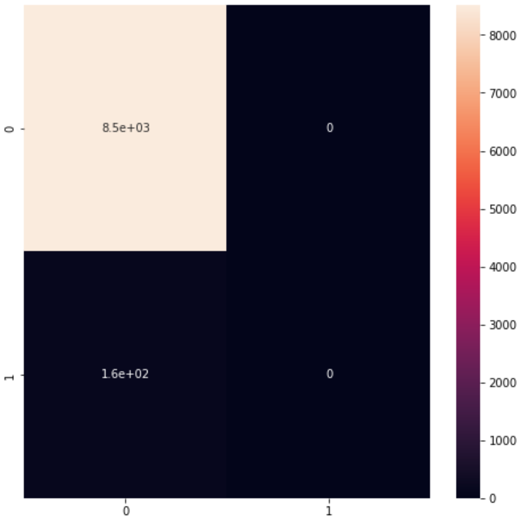
```
x = ["Support Vector Machine", "Logistic Regression", "Decision Tree"]
train_acc_svc=0.98
train_acc_lr=0.98
train_acc_dt=1.00
y = [train_acc_svc,train_acc_lr,train_acc_dt]

plt.bar(x,y, width = 0.1)
plt.xlabel("Machine learning classifiers")
plt.ylabel("Training Accuracy")
plt.show()
```



```
x = ["Support Vector Machine", "Logistic Regression", "Decision Tree"]
test_acc_svc=0.98
test_acc_lr=0.98
test_acc_dt=0.96
y = [test_acc_svc,test_acc_lr,test_acc_dt]

plt.bar(x,y, width = 0.1)
plt.xlabel("Machine learning classifiers")
plt.ylabel("Testing Accuracy")
plt.show()
```

<u>CONFUSION MATRIX:</u>

Models	Confusion Matrix
SVC	

Decision Tree	
Logistic Regression	

CONCLUSION:

In this model, we have shown how different characteristics methods can be used in three different models in machine learning classification.The results show that, of the three models, the logistic decision tree's training accuracy of the model is 1.00 and its testing accuracy of the model is 0.96, while the training accuracy of the other two models is 0.98 and 0.98, respectively. The SVC(linear model) accuracy is 0.98 for training accuracy and also 0.98 is for testing accuracy.

We can claim that our model can predict if a person will have a heart stroke or not when given data with particular attributes. The output of this machine model is quite precise. This model has the potential to predict the heart stroke for saving the life of human beings.

REFERENCE:

1. Minhaz Uddin Emon, M. S. (2020, December 28). *IEEE.* Retrieved from IEEE Xplore: https://ieeexplore.ieee.org/abstract/document/9297525/authors#authors
2. George Lawton, E. B. (n.d.). *TechTarget.* Retrieved from SearchBusinessAnalytics: https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression
3. Dataset-https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset