Stroke prediction

YATHARTH VERMA, ANANYA NAGPAL, SHIVANSHU SHEKHAR, SAURABH CHAVAN, BONTHA RISHI
B22CS064, B22EE008, B22EE062, B22EE0061, B22AI014

Abstract

This study provides a comprehensive overview of the process involved in constructing a stroke prediction classifier. The classifier is designed to estimate the likelihood of an individual experiencing a stroke, while also identifying the important elements that significantly contribute to the occurrence of a stroke. Stroke prediction can be performed by taking into account numerous factors such age, presence of heart disease, smoking status, and others. The dataset is analyzed to determine the likelihood of a person experiencing a stroke. The dataset's characteristics are then utilized in five distinct machine learning models to forecast strokes, and their effectiveness is subsequently compared. The objective is to examine the variables that influence the likelihood of experiencing a stroke. The data can subsequently be utilized to develop a predictive system for identifying stroke occurrences and the primary factors contributing to stroke.

I. Introduction

Stroke is the leading cause of mortality and morbidity worldwide, as stated by the World Health Organisation. A stroke is a cerebrovascular accident characterized by the presence of a blood clot or hemorrhage in the brain, resulting in potentially irreversible damage. It causes damage to the brain similar to how a heart attack causes damage to the heart. The causes of death from stroke are primarily due to the presence of co-morbidities and the occurrence of complications. Timely identification of diverse warning symptoms of a stroke can aid in mitigating the intensity of the stroke. Common predictors of mortality from stroke in those over the age of 65 include a history of previous stroke, atrial fibrillation, and hypertension. Furthermore, timely identification and proper treatment are necessary to avert additional harm to the damaged region of the brain and potential consequences in other bodily areas. The prevalence of stroke is expected to escalate, leading to a continual rise in stroke and heart disease mortality rates. Stroke occurrence can be predicted by analyzing input characteristics such as gender, age, presence of different diseases, and smoking habits. Our project aims to utilize machine learning methods to accurately forecast the occurrence of a stroke by analyzing a dataset that contains numerous risk variables.

A. Dataset

The file healthcare-dataset-stroke-data.csv is used as the dataset. The train dataset contains 5110 rows with 12 columns containing :

- id: unique identifier
- gender: "Male", "Female" or "Other"
- age: age of the 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
- ever_married: "No" or "Yes"
- work_type: "children", "Govt_job", "Never_worked", "Private" or "Self-employed"
- Residence_type: "Rural" or "Urban"
- avg_glucose_level: Average glucose level in blood
- bmi: body mass index
- smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
- stroke: 1 if the patient had a stroke or 0 if not

The dataset has been split into train and test with test size of 0.3

II. METHODOLOGY

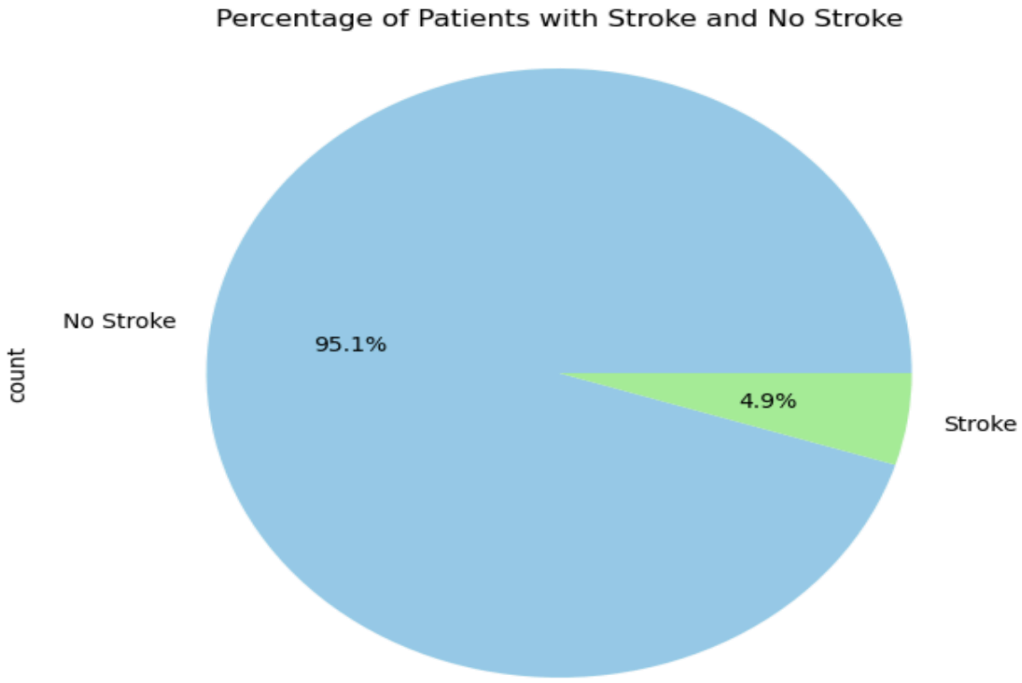
A. Overview

There are various classification algorithms present out of which we shall implement the following

- K Nearest Neighbour
- Decision Tree
- Random Forest
- Support Vector Machine
- Naive Bayes Classification
- Perceptron

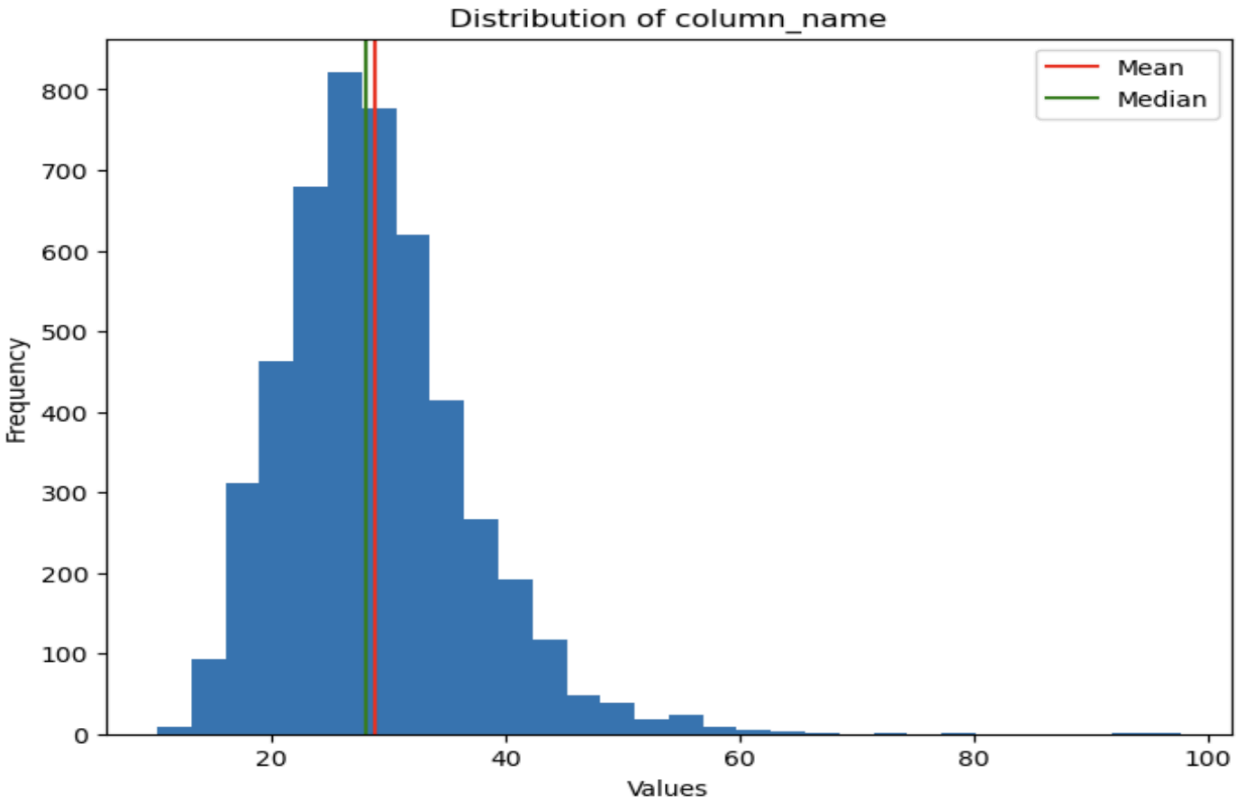
We also made use of Feature Selection and SMOTE techniques

B. Exploring the dataset and pre-processing



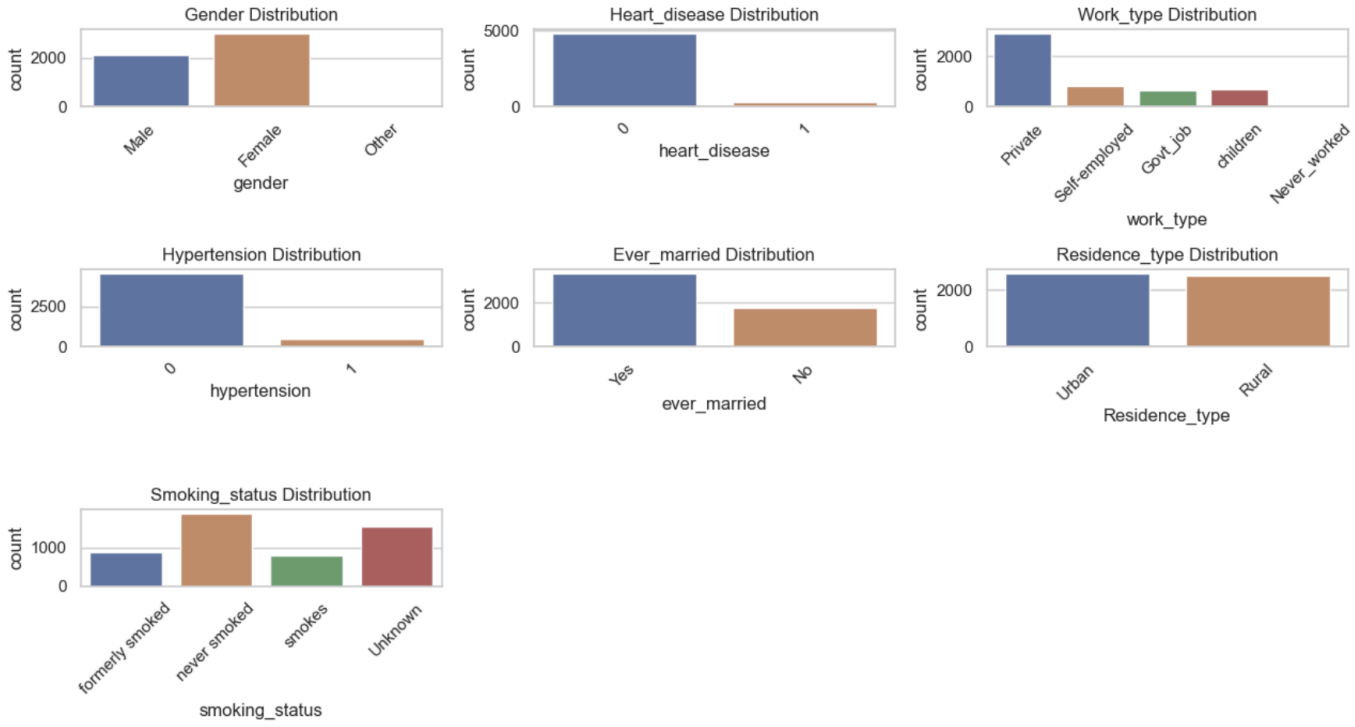
1. The dataset is highly imbalanced which will have a negative impact on training ML models.

2. The dataset was checked for the null values, the dataset only had null values in the bmi column which were removed by replacing them with the median of the column as the distribution was rightly skewed.



3. It is always better to reduce the number of categories within the column. Therefore we changed the 'Other' value in the gender column to ‘Female’, changed the values of people under 18 with 'Unknown' smoking status to 'never smoked' assuming that they do not smoke and also changed the "children" under work_type to "Never_worked" assuming that children don't work.

4. We performed univariate analysis for continuous variables and these were the observations:
   1.We have more female in the dataset. Also there's a single patient whose gender is "Other".Since female is the mode of the gender feature, the patient with 'Other' will be re-categorised to female. This way, we'll have just 2 categories in the column.
   2.Most of the patients in the data are healthy in terms of heart disease.
   3.More than 50% of the patients work in the private sector.
   4.With the assumption that children can't work/never worked, we changed the instances of "children" category.
   5.90% of the patients are not hypertensive.
   6.We have more patients who have married at one stage in their life than those who haven't.
   7.We have almost equal number of patients living in the Rural and Urban areas.



5. We also performed multivariate analysis and these were the observations:
   1.More patients that are older than 40 years seem to have strokes with little number of patients less than 40 years.
   2.The males in the data tend to have strokes at age over 40, while women tend to have strokes from age around 30s.
   3.More patients from the private sector have strokes, followed by the self employed, and govt workers respectively.
   4.The underweight patients are the least class that has strokes, followed by the healthy weight class.

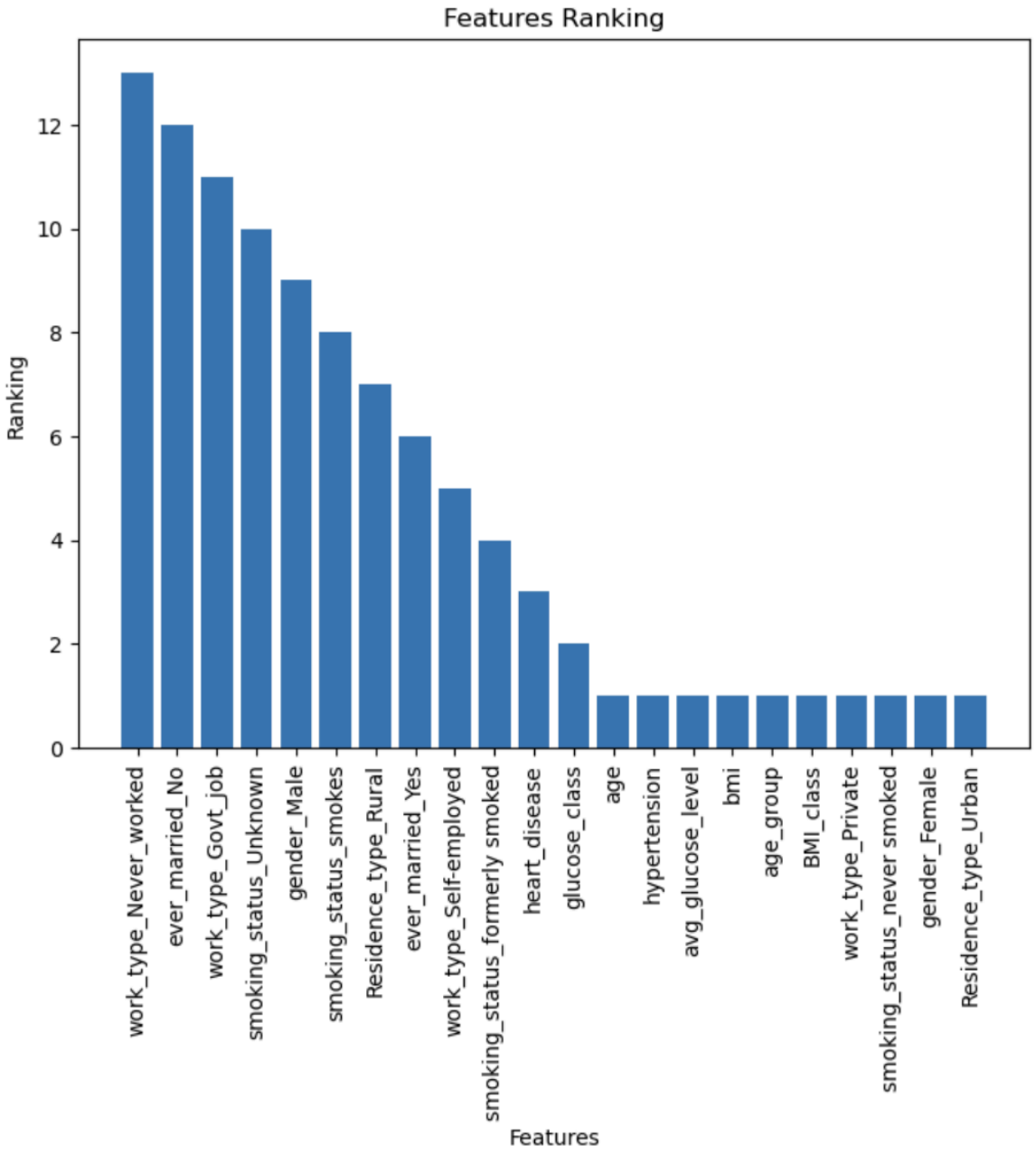
6. We encoded the categorical data :
   1.Label Encoding was used for the ordinal features so we can preserve the order of the categories
   2.OneHot Encoding was used for other nominal features since there is no inherent order in the categories.

C. Implementation of classification algorithms

a) Before training ml models, we prepared our dataset by the following procedure:

1. The dataset was split into a 70:30 ratio.

2. Feature selection : Top 10 features were selected based upon their weightage in predicting our output.



3. Data scaling : We scaled the input columns of the selected features.

4. SMOTE : Synthetic Minority Over-sampling Technique (SMOTE) is a technique for oversampling imbalanced data by generating synthetic samples for the minority class was used as the dataset was highly imbalanced.

This technique helps in getting better results as it removes biasedness during the training process of ML models.

Below is the before and after scenario of labels after we applied SMOTE technique.

```
Original set distribution:
 stroke
0    3403
1     174
Name: count, dtype: int64

Resampled set distribution:
 stroke
0    3403
1    3403
Name: count, dtype: int64
```

b) Training various machine learning models for stroke prediction.

1. KNN*:* It works well in classification problems by predicting based upon labels of nearest data points. GridSearchCV was used to find best value of k which came out be 1

2. Decision Tree :It creates the classification model by building a decision tree.Each node in the tree specifies a test on an attribute, each branch descending from that node corresponds to one of the possible values for that attribute. GridSearchCV was used in the model to find the best parameters for fine tuning of the model.Decision Tree classifier with min sample split 2 was implemented.

3. Random Forest Classifier*:*It is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Random forest classifier with max depth 20 and min sample split 2 and min sample leaf 1 was implemented.

4. SVM : It works by finding the optimal hyperplane that separates data points into different classes. GridSearchCV was used in the model to find the best parameter which directed to use 'C'='100', 'Gamma'='auto' and 'kernel'='rbf'

5. Naive Bayes Classifier*:* These are a collection of classification algorithms based on Bayes Theorem.For this We used Bernoulli naive bayes classifier because it is a binary classification problem.

6. Perceptron: Perceptron is one of the most basic models of deep learning used for classification. This was used Just to observe its performance in comparison with other machine learning models

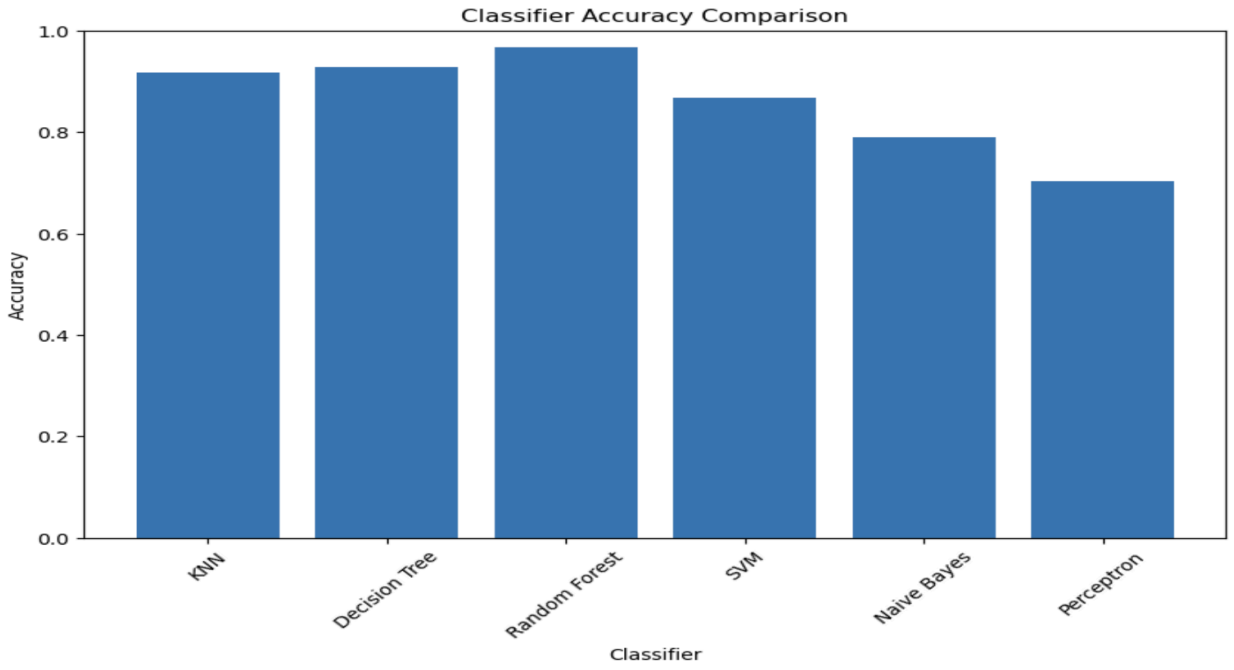
c) Results of the models were analyzed with the help of confusion matrix and classification report. Below is the result of the Random Forest classifier.

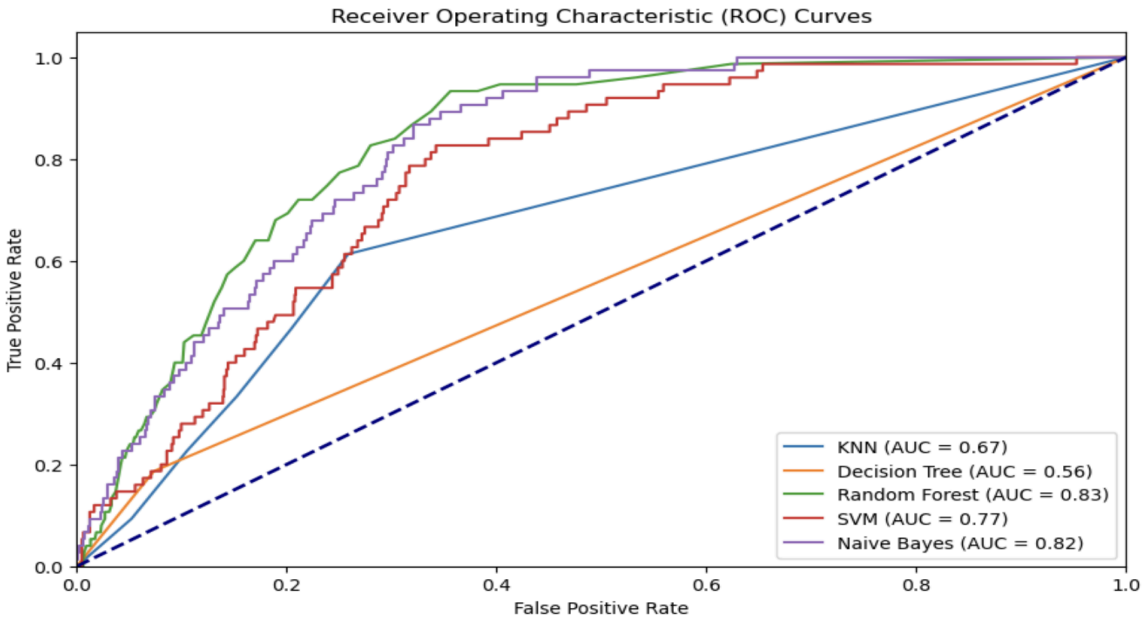
```
Random Forest Test metrics
Confusion Matrix:

[[1396   62]
 [  59   16]]


Classification Report:

              precision    recall  f1-score   support

           0       0.96      0.96      0.96      1458
           1       0.21      0.21      0.21        75

    accuracy                           0.92      1533
   macro avg       0.58      0.59      0.58      1533
weighted avg       0.92      0.92      0.92      1533
```

III. PREDICTIONS

	KNN	Decision Tree	Random forest	Naive bayes	SVM	Perceptron
Accuracy	0.89	0.89	0.92	0.65	0.87	0.78
F1 Score	0.14	0.17	0.21	0.20	0.13	0.23

IV. EVALUATION OF MODELS



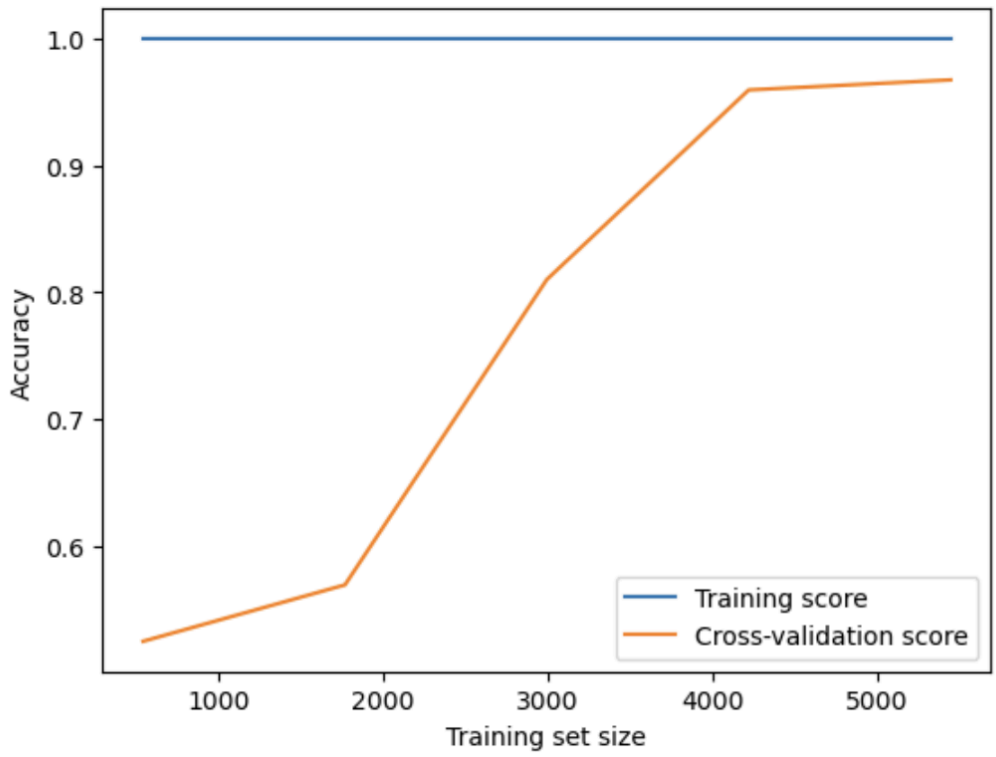


Graphs for comparing the performance for different models are plotted above.

It was found from the result that Random Forest was the one that performed the best in the prediction of stroke.

V. RESULTS AND ANALYSIS

The table shows the comparison between various models. From the table Random Forest performed well. Therefore, the Random Forest model is preferred because it gives the max accuracy and F1 score.



Conclusion:

The model is overfitting on the train data and the cross-validation score increases as the training size increases which is a good sign that the model is generalizing well to new data. The training score decreasing as the training size increases is expected because of the regularization, and the increasing cross-validation score indicates in its ability to generalize to new data.

VI. CONTRIBUTIONS

The learning and planning was done as a team.The individual contributions are as given

- Ananya Nagpal : KNN, Prepared dataset for modeling, Web Page, Report
- Yatharth Verma : EDA, Random Forest and SVM, Web Page, Report
- Shivanshu Shekhar : Data Processing , Decision Tree, Spotlight video
- Saurabh Chavan : Perceptron, Presentation, Spotlight Video
- Rishi : Naive Bayes, Presentation

VII. REFERENCES

- https://www.javatpoint.com/machine-learning-random-forest-algorithm#:~:text=As%20the%20name%20suggests%2C%20%22Random,tree%20and%20based%20on%20the
- https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
- https://analyticsindiamag.com/understanding-the-auc-roc-curve-in-machine-learning-classification/#:~:text=ROC%20curve%2C%20also%20known%20as,sensitivity%20of%20the%20classifier%20model