Stroke prediction

Palak Singh, Sakshi Todi
B20EE086, B20EE088

Abstract

This report gives an overview of the process of building a classifier for stroke prediction, which predicts the probability of a person having a stroke along with the key factors which play a major role in causing a stroke. Stroke Prediction can be done considering various features such as age, heart disease, smoking status, etc. Analyzing the dataset to get insights about the probability of an individual to suffer from a stroke and the features of the dataset are applied to the five different machine learning (ML) models which are used to predict stroke, and their performance is compared. The goal is to investigate the factors that determine the chances of stroke. The data can then be used to create a system that predicts stroke and the major reasons which cause stroke.

I. Introduction

According to the World Health Organization, stroke is the greatest cause of death and disability globally. Stroke is a blood clot or bleed in the brain, which can cause permanent damage. It injures the brain like a heart attack which injures the heart. Causes of Death from stroke is as a result of co-morbidities and complications. Early recognition of various such warning signs of a stroke can help reduce the severity of the stroke. The most common predictors of death from stroke for those aged more than 65 years of age included previous stroke, atrial fibrillation and hypertension. Also, Early detection and appropriate management are required to prevent further damage to the affected area of the brain and other complications in other parts of the body. The burden of stroke will likely worsen with stroke and heart disease related deaths projected to increase continuously every year. We can predict whether a patient is likely to get a stroke based on the input parameters like gender, age, various diseases, and smoking status. The goal of our project is to apply principles of machine learning on existing dataset to effectively predict the stroke based on various risk factors.

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_jov", "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

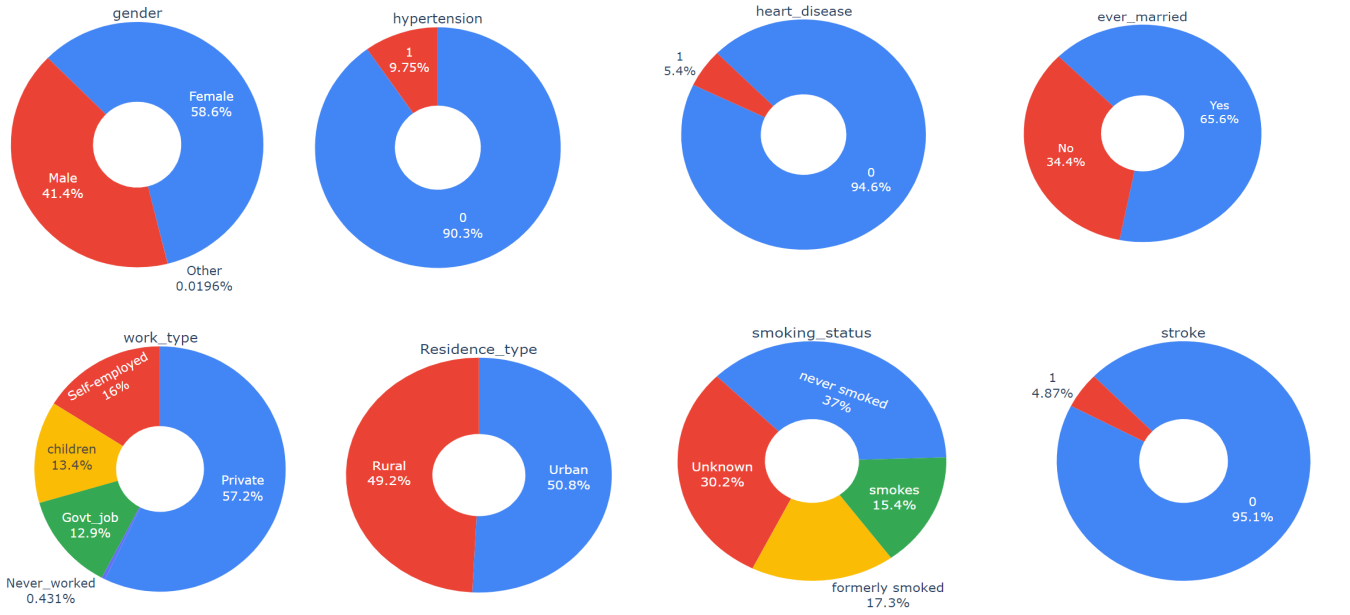
There are various classification algorithms present out of which we shall implement the following

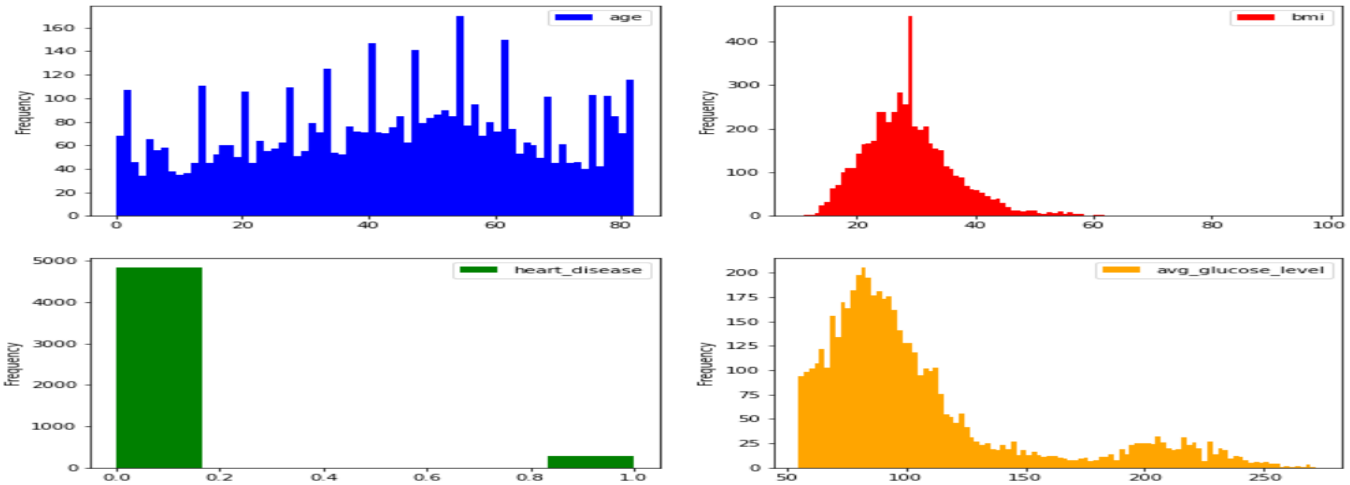
- Random Forest Classification
- LightGBM classification
- Decision Tree Classification
- XgBoost Classification
- Naive Bayes Classification

We also make use of SFS for feature selection.

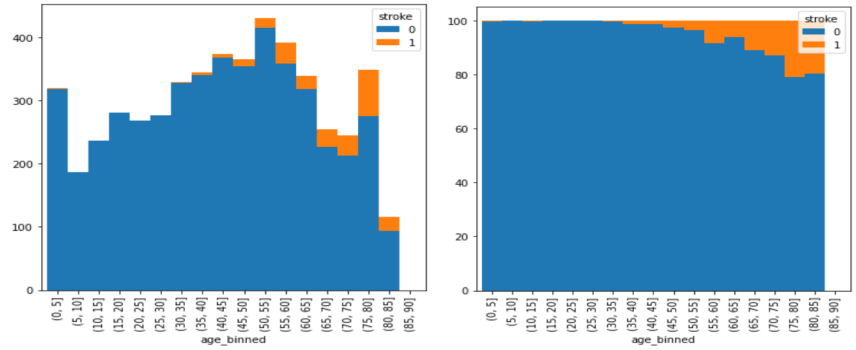
B. *Exploring the dataset and pre-processing*

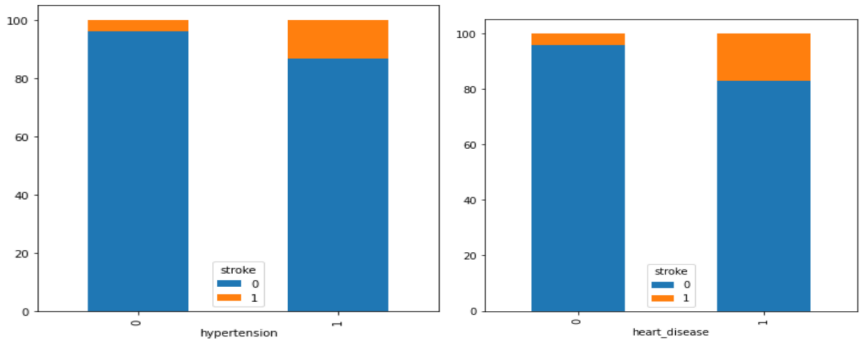
The descriptive analysis of data was done, the database was checked for null values, the null values were replaced by mean values of the columns. Normalization was done except for the columns with categorical data. Next, we explore the categorical data and apply label encoding on it. Visualization of distribution of data in every attribute is done below.





From the above plots we can infer the following:

- We have good distribution for age
- Average glucose distribution is reasonable because the normal average of blood in sugar is less than 140 , that may be not good this feature will not be helpful to know if diabetes have correlation between diabetes and strokes
- Most people have a BMI between 25-35 which indicates that higher bmi does not increase stroke risk.
- 

- From the above plot we can see that Older patients are more likely to suffer a stroke than a younger patient.
- Work and marital_status are related with age.
- 

- From the above plot, we see that Higher proportion of patients who suffered from hypertension or heart disease experienced a stroke.
- Gender and Residence_type does not affect the occurence of stroke.

We observe that our dataset is imbalanced. So to balance the data, the SMOTE method was used. It populated our data with records similar to our minor class.

C. Implementation of classification algorithms

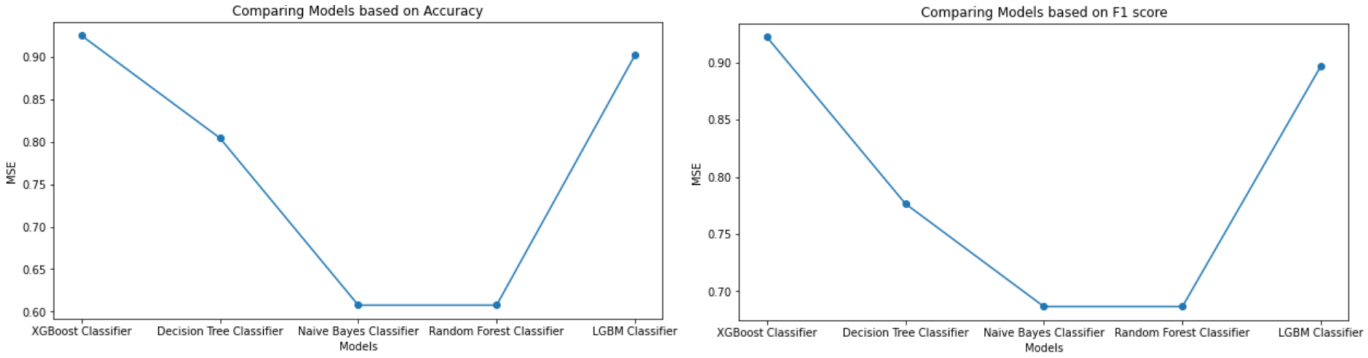
1. *LGBM Classifier:* It is a gradient boosting framework that uses tree based learning algorithms. It is designed to be distributed and efficient.
   - RandomizedSearchCV was used to find the best parameters for the mode and then they were used.l
2. *XGBoost Classifier:* It is a powerful approach for building supervised classification models.The objective function contains loss function and a regularization term. It tells about the difference between actual values and predicted values, i.e how far the model results are from the real values.
   - RandomizedSearchCV was used to find the best parameters for the model and then they were used.
   - Total 25 fittings of the data was done to find the best parameters.
3. *Decision Tree Classifier*: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..
4. *Naive Bayes Classifier:*These are a collection of classification algorithms based on Bayes' Theorem.For this classification problem we used Bernoulli naive bayes classifier because it is a binary classification problem.
   - Bernoulli Naive Bayes was used and it was fine tuned using RandomSearchCV.
5. *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 22 and min sample split 4 was implemented.

III. PREDICTIONS

	LGBM Classifier	XGBoost Classifier	Decision Tree Classifier	Naive Bayes Classifier	Random Forest Classifier
Accuracy	87.67	90.47	80.54	60.75	60.75
F1 score	86.68	89.81	77.90	68.66	68.66

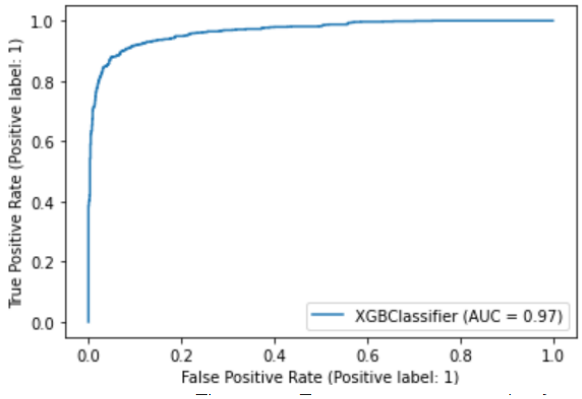
IV. EVALUATION OF MODELS

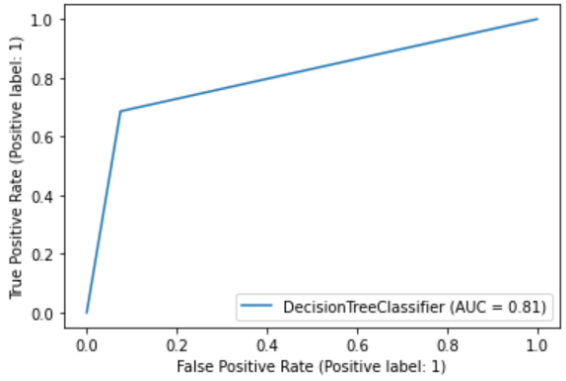
Graphs for comparing the accuracy and F1 score for different models is plotted below.

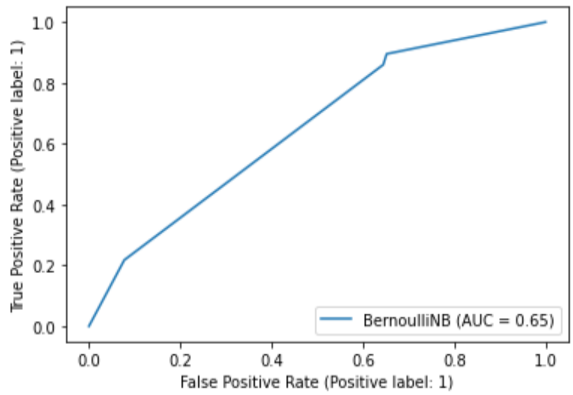


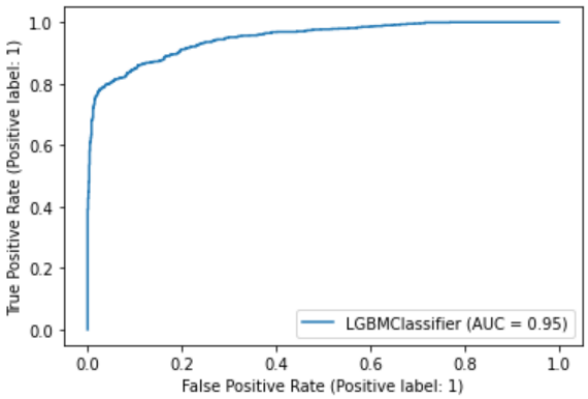
It was found from the predictions doen and the graphs plotted that, XGBoost Classifier was the one that performed the best in the prediction of stroke.

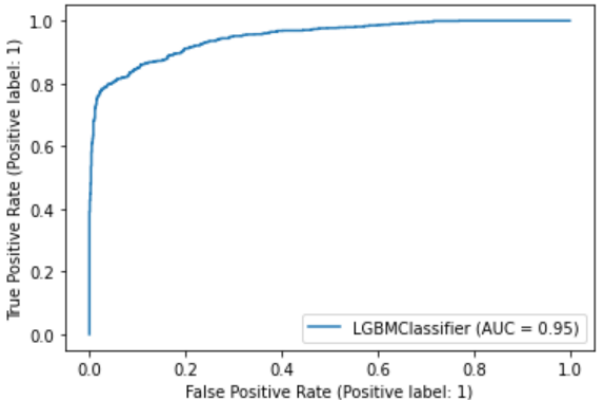
Fig: ROC plots for Random Forest Classification, LightGBM, DTC, XgBoost Classification, Naive Bayes Classification











V. RESULTS AND ANALYSIS

The table shows the comparison between various models. From the table, we can see that Random Forest and Naives Bayes Classifier did not perform well in comparison to the other three. Comparing the other three, the maximum accuracy score was found to be 90.47% for the XGBoost model. Therefore, the XGBoost model is preferred because it gives the max accuracy and F1 score.

VI. CONTRIBUTIONS

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

- Sakshi Todi (B20EE088): Bernoulli Naive Bayes, XGBoost and LightGBM, Report
- Palak Singh (B20EE088): Data pre-processing and exploratory analysis, Random Forest, Decision Tree, Report

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://machinelearningmastery.com/develop-first-xgboost-model-python-scikit-learn/

➔ https://analyticsindiamag.com/understanding-the-auc-roc-curve-in-machine-learning-classification/#:~:text=ROC%20curve%2C%20also%20known%20as,sensitivity%20of%20the%20classifier%20model