Project Proposal: *Stroke Prediction*

Abstract:

Stroke is a serious disease that affects millions of people all around the world. It is a cardiovascular disease that blocks oxygen from arteries leading to and from the brain. [1]. The goal of this project is to predict whether the patient will have a stroke based on a couple of features. I worked on an imbalanced public data on Kaggle website. To achieve this goal, I used feature categorization and a couple of models to evaluate and find the best accuracy after balancing the data.

Problems:

Predict whether I the patient will have stroke or not?
Does the gender impact on strokes?
Does the Hypertension impact on strokes?
Does the marriage impact on strokes?
Does the smoking impact on strokes?
Does the age impact on strokes?
What effects more BMI or the glucose level on strokes?

Data:

In this project, I will use a dataset from Kaggle [4]. It contains 11 features and one target, which is whether the patient has a stroke or not. The total data in the dataset is 5109. After balancing the data there was 9719	
ID	Unique identifier
Gender	Male, female, other
Age	The 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

Tools:

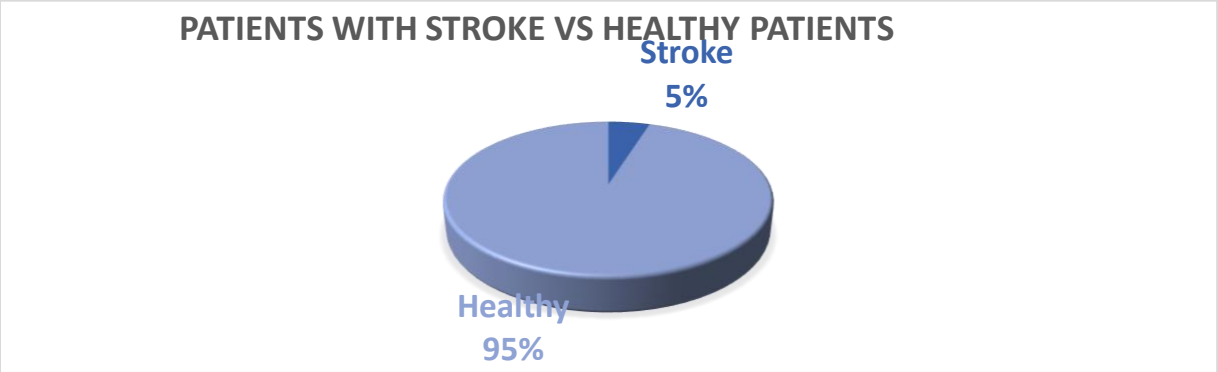
In this project I will use these tools to help me to achieve the requirement of this project
Scikit
NumPy
Xgboost
Pandas
Imblearn
Seaborn
Matplotlib
Bokeh
SciPy
Keras

Project Proposal: *Stroke Prediction*

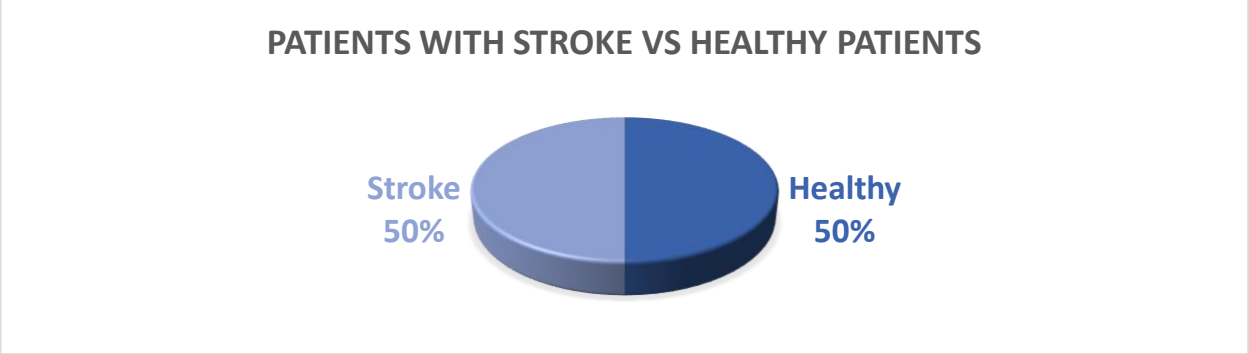
Algorithms

Feature engineering by applying categorical variable by using dummies from pandas library.

Balance the data by using SMOTE from imblearn library

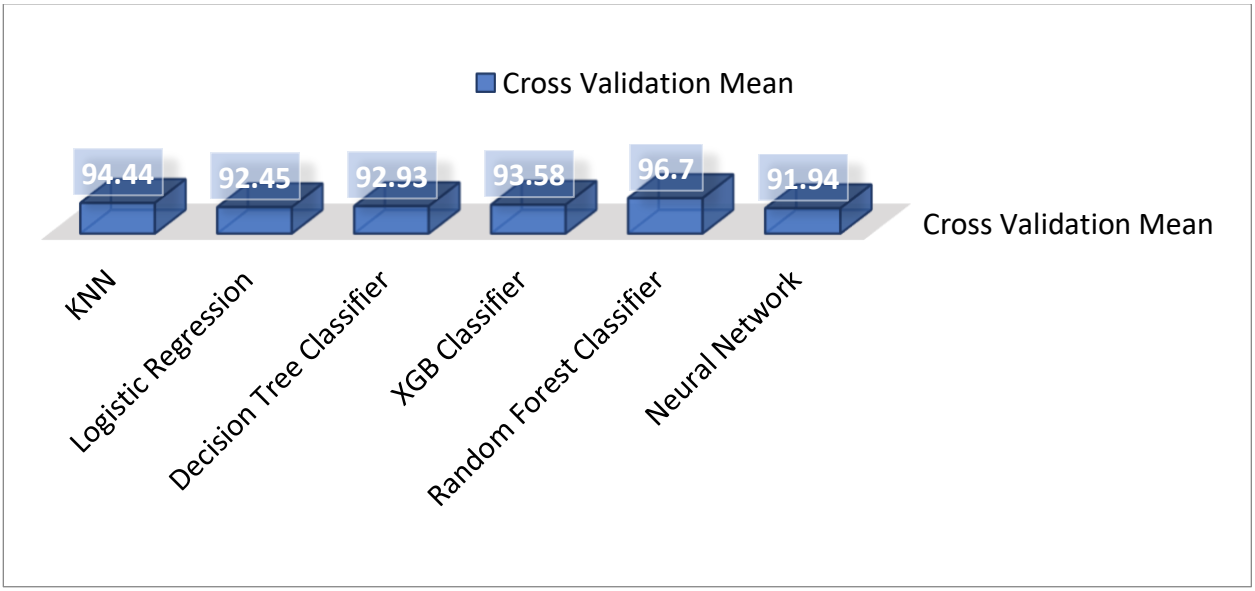


So clearly the data is imbalanced so I will balance it using SMOTE minority sampling.

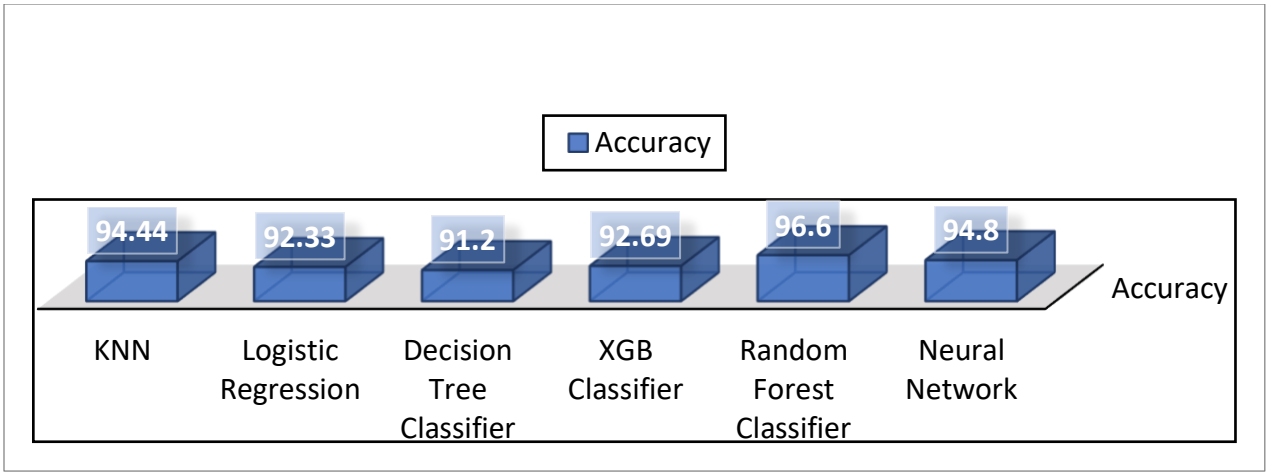


The data after balanced.

Cross Validation with 10-fold CV



Project Proposal: *Stroke Prediction*



With this dataset after doing the EDA and categorization and dealing with imbalanced data by using the minority sampling, after tuning the hyperparameter and by doing the cross validation I figured out that the best model is the random forest classification with accuracy = 96.60%.

Communication:

I did a PowerPoint presentation to present this project.

Reference:

[1]: https://www.cdc.gov/stroke/types_of_stroke.htm#hemorrhagic

Project link: https://github.com/eyad718/T5-project/blob/main/project_T5.ipynb