STROKE PREDICTION

DEPARTMENT OF CSE(AI & ML)
TEAM - 9

20WH1A6643 - T.Sreeya
20WH1A6644 - S.Chethana
20WH1A6645 - K.Sudiksha
20WH1A6646 - B. Sreevidya
20WH1A6647 - P. Shiromani

April 1, 2023

PROBLEM STATEMENT

- Stroke is a major cause of disability and mortality worldwide. Early identification of individuals at high risk of stroke can help prevent or mitigate its devastating consequences.
- The problem is to develop an accurate and reliable predictive model for stroke based on a combination of demographic, clinical, and lifestyle factors.
- The ML model should be able to identify individuals at high risk of stroke.

PYTHON PACKAGES AND LIBRARIES

- Numpy
- Pandas
- Matplotlib
- seaborn
- sklearn
- flask

Algorithms

- Support Vector Machine
- Random Forest Classifier
- Logistic Regression
- Naive Bayes
- Xgboost

Support Vector Machine

- Support Vector Machine (SVM) is a machine learning algorithm used for classification and regression problems. It works by finding a hyperplane that best separates the data into different classes.
- The hyperplane is chosen in such a way that it maximizes the margin, which is the distance between the hyperplane and the nearest points from each class.
- In other words, SVM finds the decision boundary that maximizes the margin between the different classes. The goal is to find the hyperplane that generalizes well to new data, which means it is able to correctly classify new data points that were not used during training.

Random Forest Classifier

- Random Forest Regression is a machine learning algorithm that uses multiple decision trees to make predictions for regression problems. It combines the results of each decision tree to produce a more accurate prediction.
- It is called "random forest" because each decision tree is built using a random subset of the features and a random subset of the training data.
- This helps to reduce overfitting and improve the accuracy of the model.

Logistic Regression

- Logistic Regression is a machine learning algorithm used for binary classification problems, where the goal is to predict the probability of a binary outcome (e.g., Yes/No, True/False).
- It works by fitting a logistic function to the input data, which maps the input variables to a probability between 0 and 1.
- The logistic function is used to model the relationship between the input variables and the output variable, by estimating the coefficients that best fit the data.

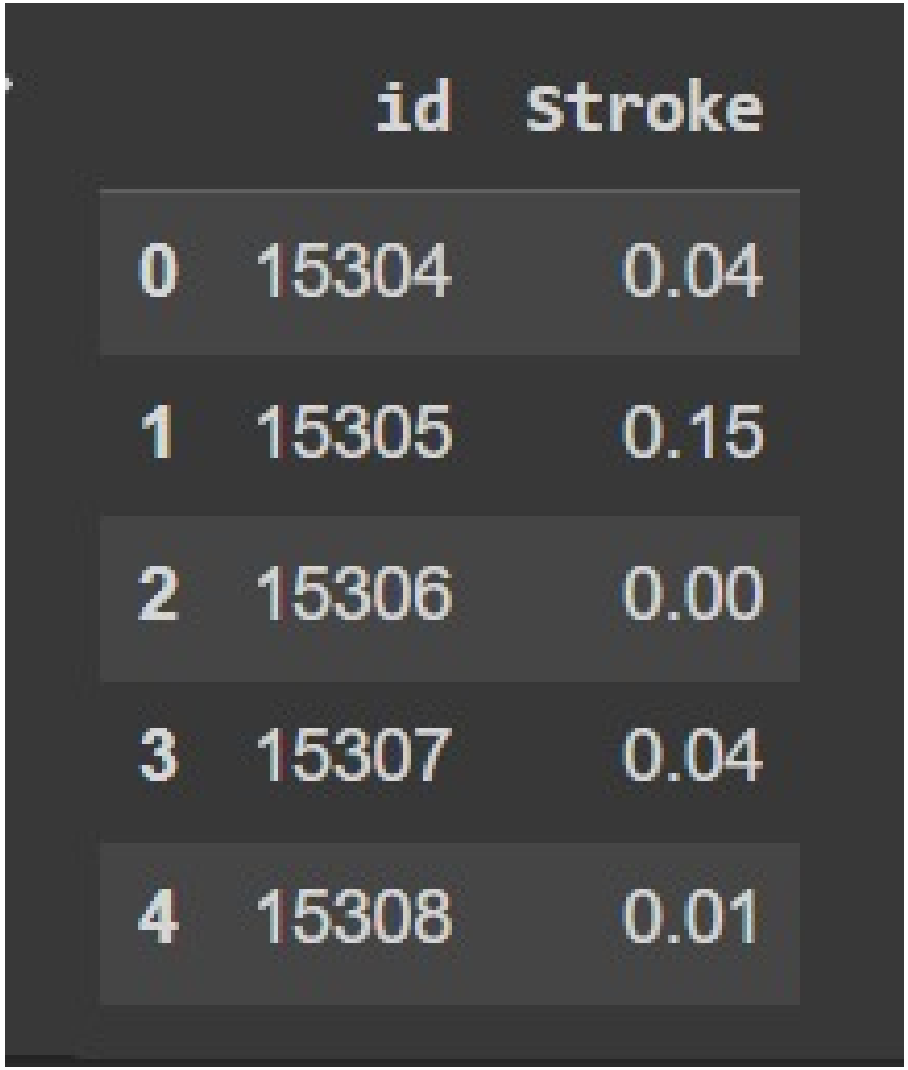
Naive Bayes

- Naive Bayes is a popular algorithm in machine learning for classification tasks. It is based on the Bayes theorem, which describes the probability of an event based on prior knowledge or evidence.
- In the context of machine learning, Naive Bayes is often used for text classification tasks, such as spam detection or sentiment analysis.
- The "naive" part of the name comes from the assumption that the features used in the classification task are independent of each other, which may not be entirely true in practice.

XGBoost (eXtreme Gradient Boosting)

- XGBoost (eXtreme Gradient Boosting) is a machine learning algorithm that is used for both regression and classification problems.
- It is a type of ensemble model that combines the predictions of multiple decision trees to improve the accuracy of the model.
- XGBoost works by iteratively adding decision trees to the model, with each new tree attempting to correct the errors made by the previous trees.

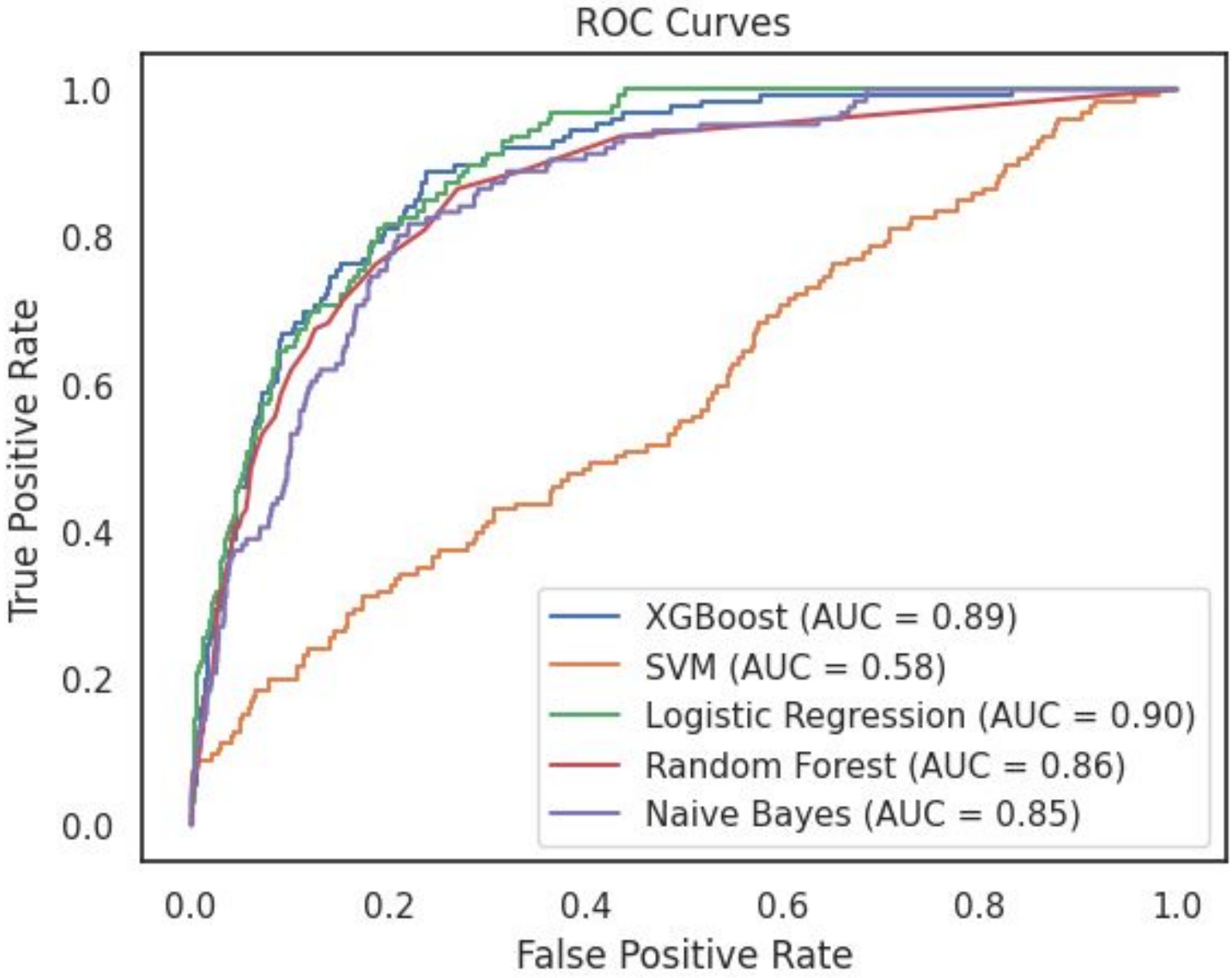
OUTPUT



	id	Stroke
0	15304	0.04
1	15305	0.15
2	15306	0.00
3	15307	0.04
4	15308	0.01

Figure: Output

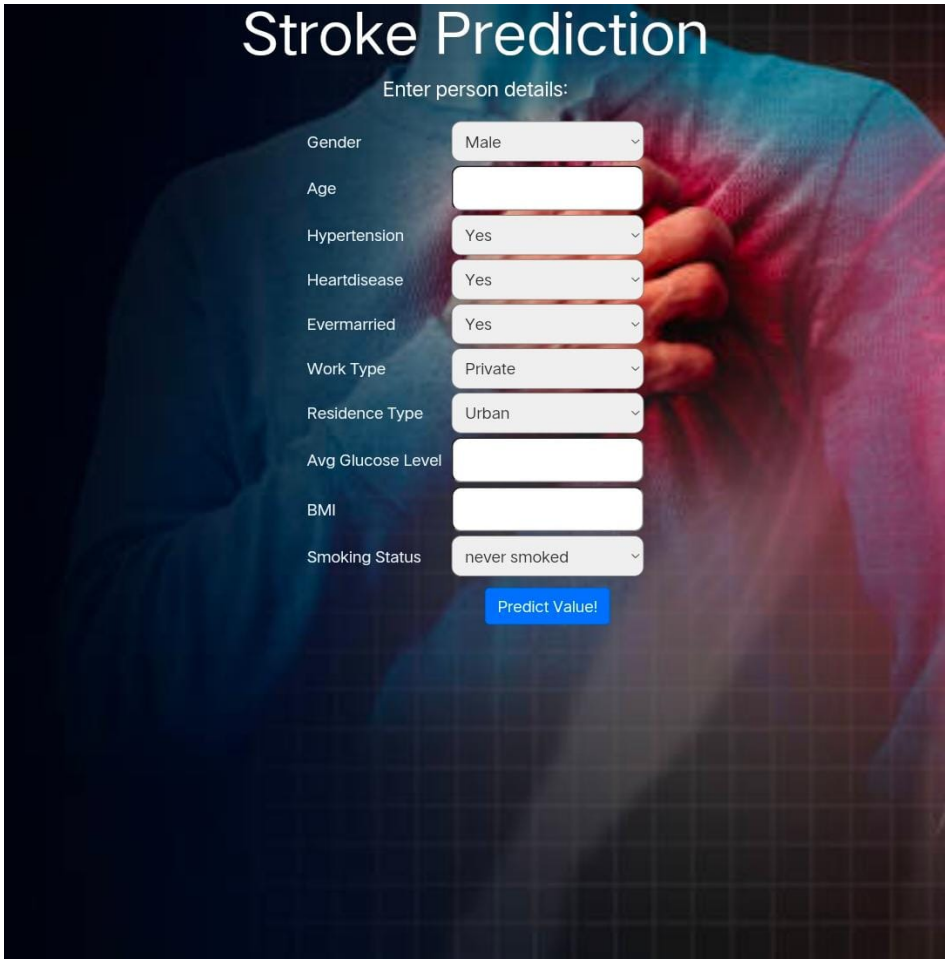
Comparison Graph



Comparison Table

Model Name	AUC
XGBoost	0.89
Support Vector Machine	0.58
Logistic Regression	0.90
Random Forest	0.86
Naive Bayes	0.85

DEMO



Website link: http://20wh1a6646.pythonanywhere.com/

THANK YOU !