Stroke Prediction

Machine Learning-I

- Shikha Sharma
- Aron Rock
- Sanjana Godolkar

Introduction

Stroke is a life-threatening condition that disrupts blood flow to the brain, leading to a high risk of disability and death
Our project aims to utilize machine learning algorithms to develop models that predict the risk of having a stroke.

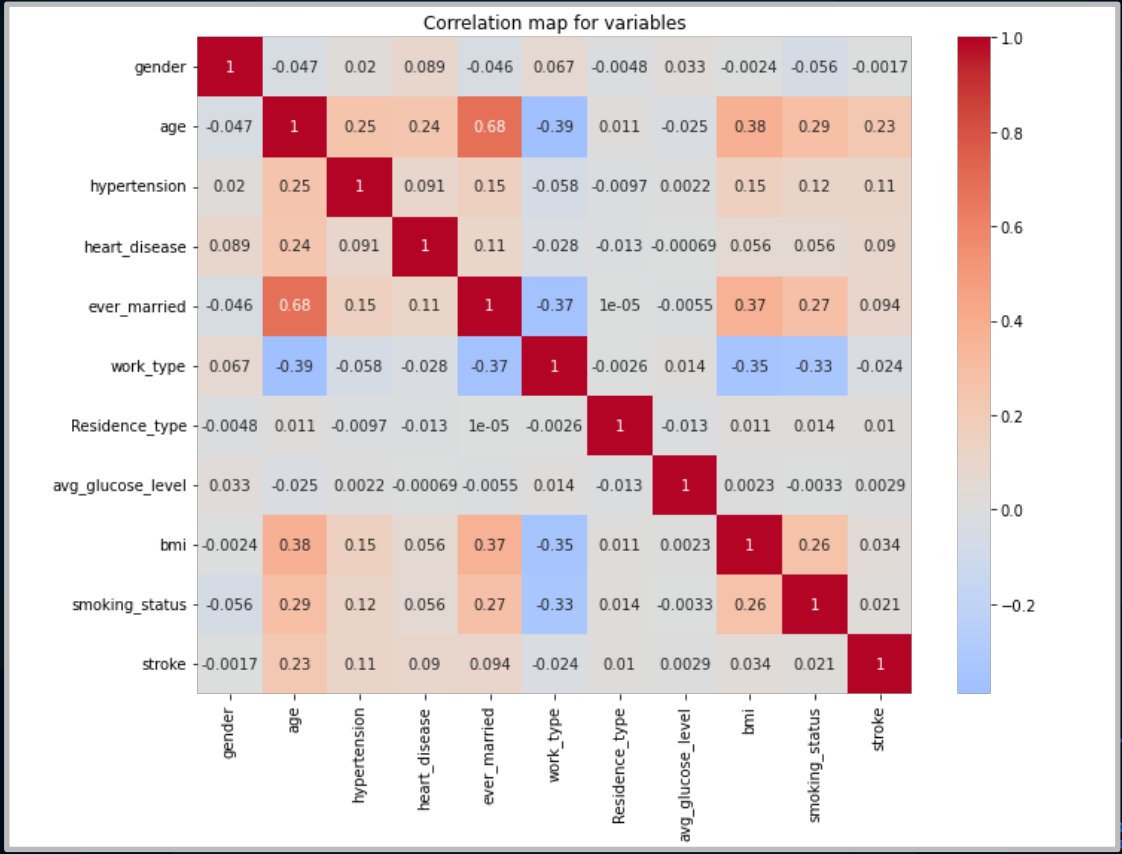
Dataset Description

- Source: Kaggle ("healthcare-dataset-stroke-data")
- 5,110 observations with 12 attributes
- Attributes: ID, Gender, Age, Hypertension, Heart Disease, Ever Married, Work Type, Residence Type, Avg. Glucose Level, BMI, Smoking Status, Stroke
- Imbalanced dataset:
  - Stroke cases: 249
  - Non-stroke cases: 4861

Data Preprocessing

- Cleaning and preparing data for model development
- Excluding 'id' column
- Handling missing values: Filling null values in 'BMI' column with mean
- Label encoding for categorical variables
- Balancing the dataset using the SMOTE technique

Heat Map


Correlation map for variables
gender 1 -0.047 0.02 0.089 -0.046 0.067 -0.0048 0.033 -0.0024 -0.056 -0.0017
age -0.047 1 0.25 0.24 0.68 -0.39 0.011 -0.025 0.38 0.29 0.23
hypertension 0.02 0.25 1 0.091 0.15 -0.058 -0.0097 0.0022 0.15 0.12 0.11
heart_disease 0.089 0.24 0.091 1 0.11 -0.028 -0.013 -0.00069 0.056 0.056 0.09
ever_married -0.046 0.68 0.15 0.11 1 -0.37 1e-05 -0.0055 0.37 0.27 0.094
work_type 0.067 -0.39 -0.058 -0.028 -0.37 1 -0.0026 0.014 -0.35 -0.33 -0.024
Residence_type -0.0048 0.011 -0.0097 -0.013 1e-05 -0.0026 1 -0.013 0.011 0.014 0.01
avg_glucose_level 0.033 -0.025 0.0022 -0.00069 -0.0055 0.014 -0.013 1 0.0023 -0.0033 0.0029
bmi -0.0024 0.38 0.15 0.056 0.37 -0.35 0.011 0.0023 1 0.26 0.034
smoking_status -0.056 0.29 0.12 0.056 0.27 -0.33 0.014 -0.0033 0.26 1 0.021
stroke -0.0017 0.23 0.11 0.09 0.094 -0.024 0.01 0.0029 0.034 0.021 1
gender
age
hypertension
heart_disease
ever_married
work_type
Residence_type
avg_glucose_level
bmi
smoking_status
stroke
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

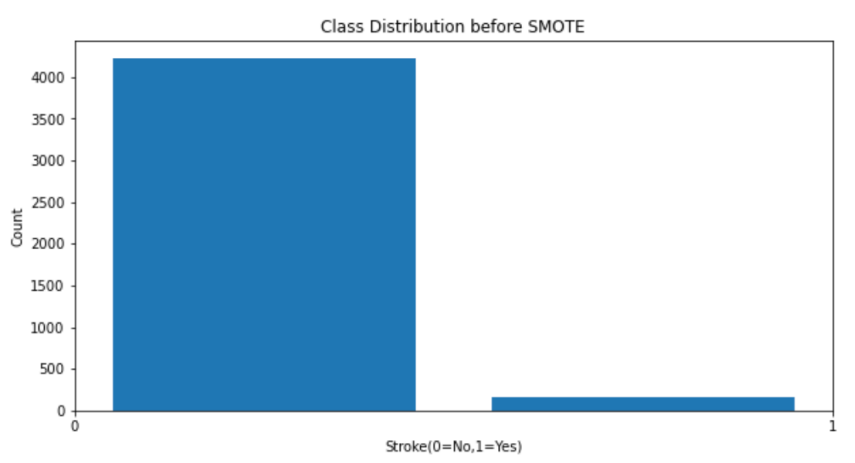
Feature Reduction

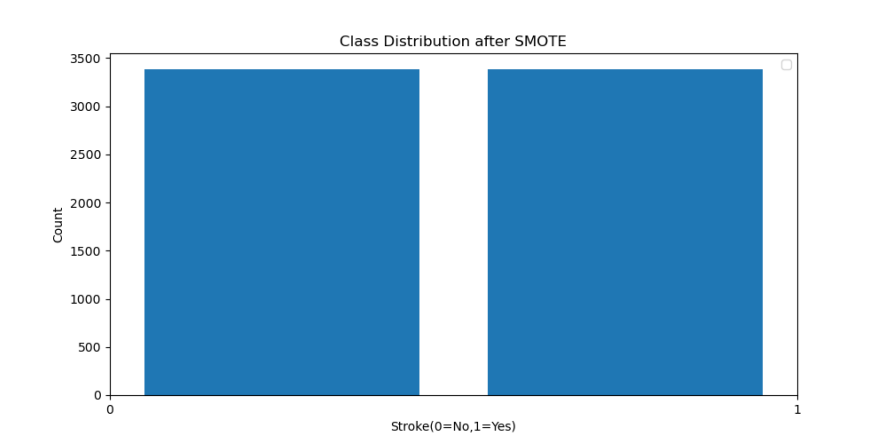
- Dropped the following features:
  ‘gender’ , ’ever_married’ , ‘work_type’ , ‘Residence_type’ , ‘avg_glucose_level’ , ‘bmi’ , ‘smoking_status’
- Keeping the following features:
  ‘age’ , ‘hypertension’, ‘heart_disease’

SMOTE

- SMOTE (Synthetic Minority Over-sampling Technique) is an oversampling technique used in machine learning to handle imbalanced datasets.
- It creates synthetic examples of the minority class by selecting the nearest neighbors and changing them to generate new samples.
- SMOTE helps to balance the class distribution, allowing for better performance of machine learning models and reducing the risk of overfitting.
- SMOTE can be applied to various classification algorithms, such as logistic regression, decision trees, and support vector machines.
-

Before and After SMOTE






```
Class counts before SMOTE oversampling:
Class 0 : 3383
Class 1 : 129
```

```
Class counts after SMOTE oversampling:
Class 0 : 3383
Class 1 : 3383
```

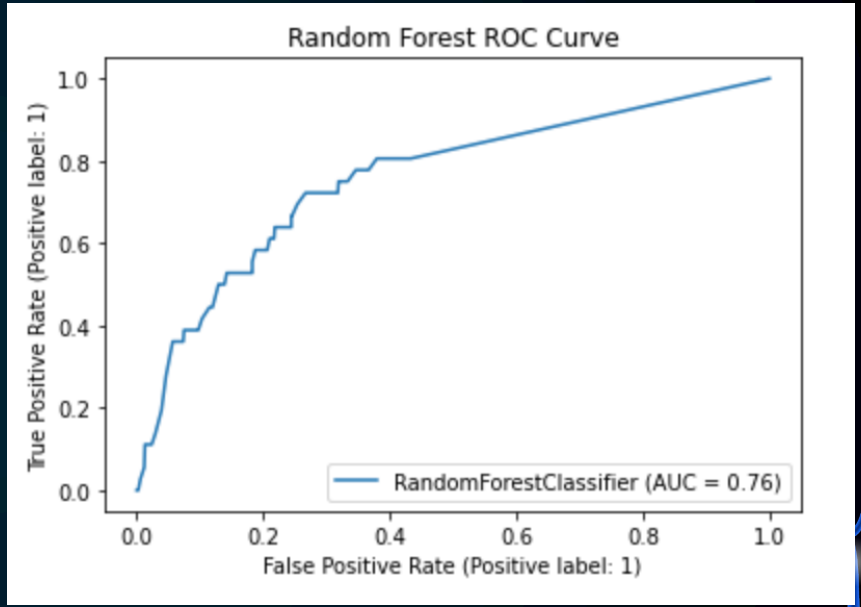
Models

Random Forest

random-forest Classification report

	precision	recall	f1-score	support
0	0.97	0.92	0.94	843
1	0.17	0.39	0.23	36
accuracy			0.90	879
macro avg	0.57	0.65	0.59	879
weighted avg	0.94	0.90	0.91	879

Accuracy-Random-forest
: 0.8953356086461889



XGBoost

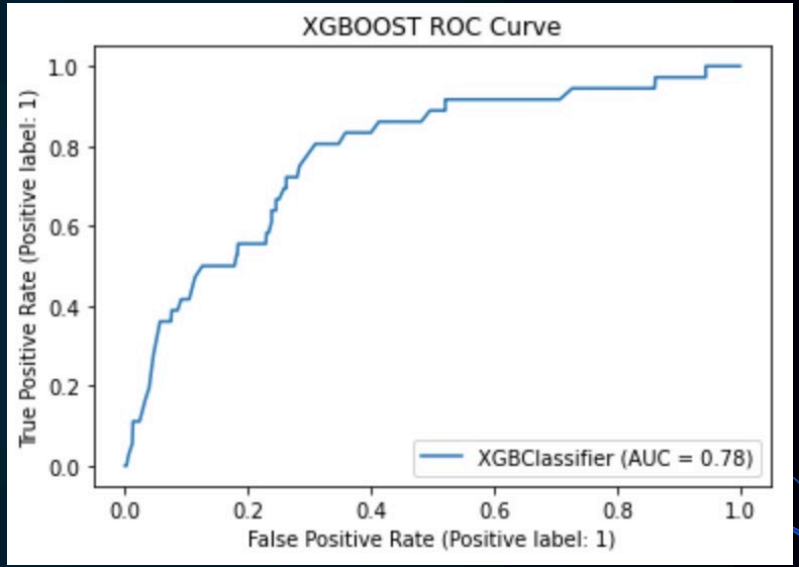
XGBOOST confusion matrix-

```
 [[766  77]
 [ 21  15]]
```

XGBOOST Classification report

	precision	recall	f1-score	support
0	0.97	0.91	0.94	843
1	0.16	0.42	0.23	36
accuracy			0.89	879
macro avg	0.57	0.66	0.59	879
weighted avg	0.94	0.89	0.91	879

Accuracy-XGBOOST
: 0.888509670079636

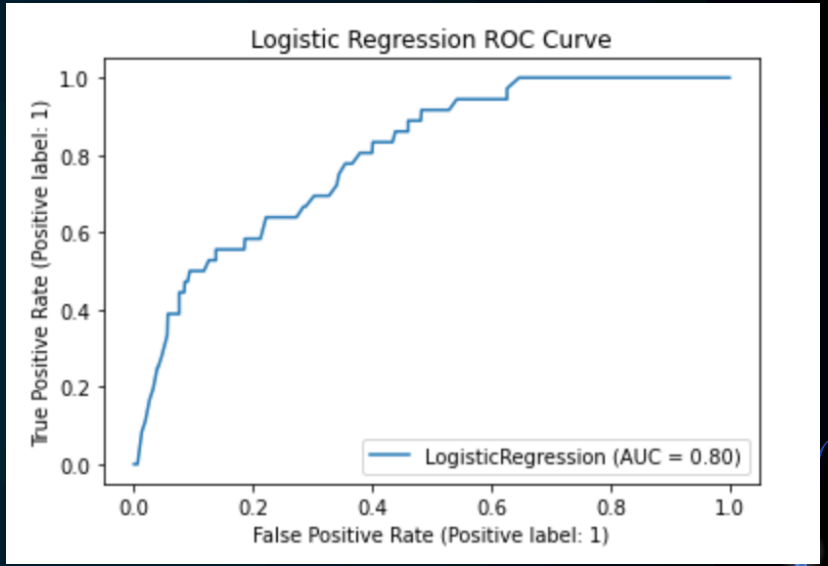


Logistic Regression

Logistic Regression Accuracy: 0.7690557451649602

```
[[653 190]
 [ 13  23]]
```

	precision	recall	f1-score	support
0	0.98	0.77	0.87	843
1	0.11	0.64	0.18	36
accuracy			0.77	879
macro avg	0.54	0.71	0.53	879
weighted avg	0.94	0.77	0.84	879

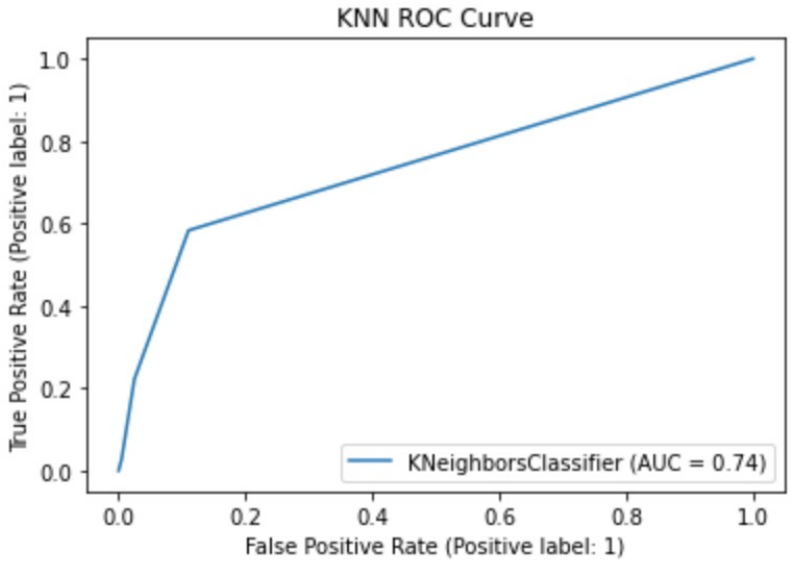


K Nearest Neighbor

K-Nearest Neighbors Accuracy: 0.9556313993174061

```
[[839   4]
 [ 35   1]]
```

	precision	recall	f1-score	support
0	0.96	1.00	0.98	843
1	0.20	0.03	0.05	36
accuracy			0.96	879
macro avg	0.58	0.51	0.51	879
weighted avg	0.93	0.96	0.94	879

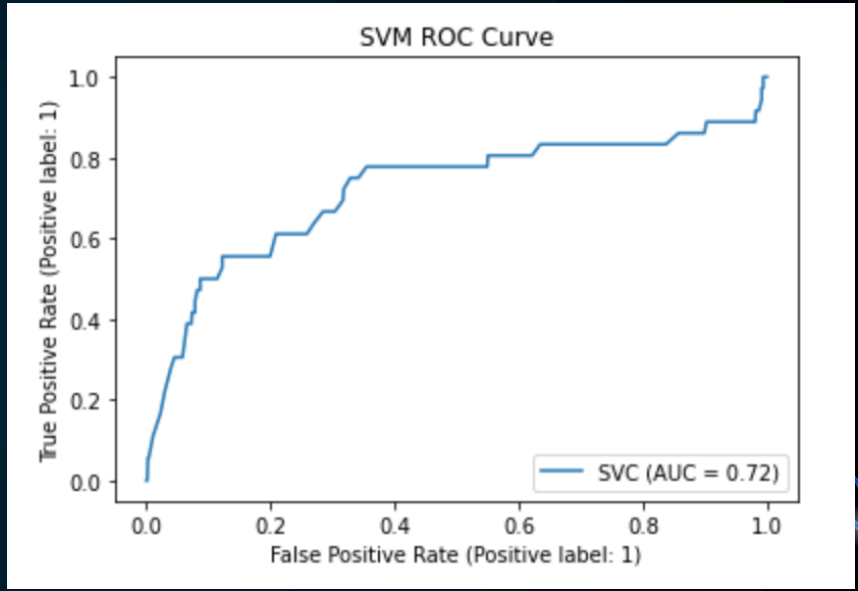


Support Vector Machine

Support Vector Machine Accuracy: 0.7565415244596132

[[643 200]
 [14 22]]

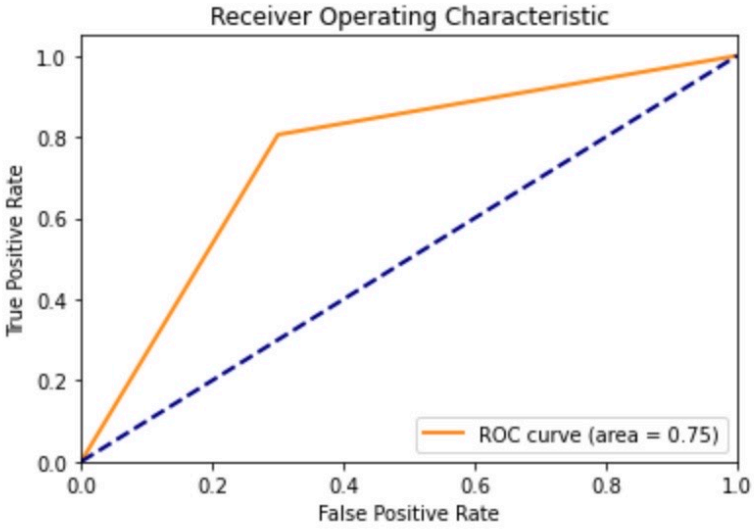
	precision	recall	f1-score	support
0	0.98	0.76	0.86	843
1	0.10	0.61	0.17	36
accuracy			0.76	879
macro avg	0.54	0.69	0.51	879
weighted avg	0.94	0.76	0.83	879



Multi-Layer Perceptron

	precision	recall	f1-score	support
0	0.99	0.70	0.82	843
1	0.10	0.81	0.18	36
accuracy			0.70	879
macro avg	0.55	0.75	0.50	879
weighted avg	0.95	0.70	0.79	879

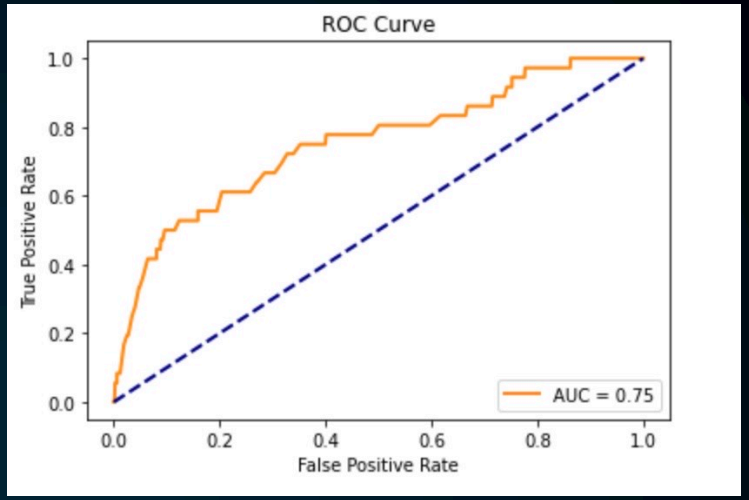
Accuracy: 0.704
Precision: 0.103
Recall: 0.806
F1-score: 0.182

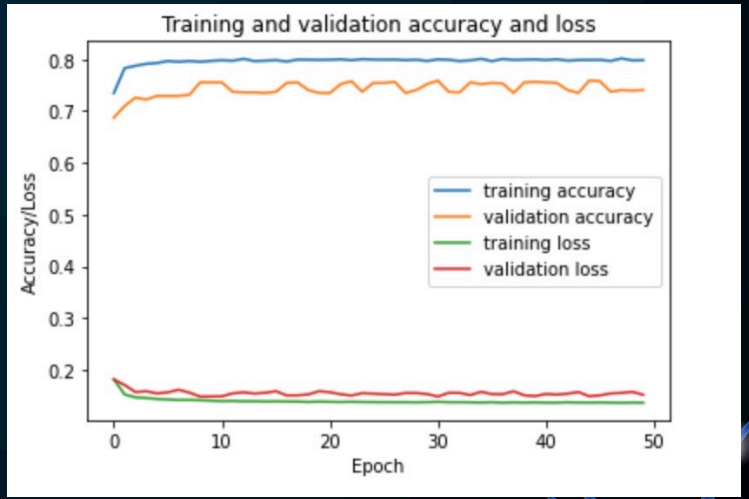


Keras Model

	precision	recall	f1-score	support
0	0.98	0.74	0.84	843
1	0.10	0.64	0.17	36
accuracy			0.74	879
macro avg	0.54	0.69	0.51	879
weighted avg	0.94	0.74	0.82	879

Accuracy: 0.737
Precision: 0.095
Recall: 0.639
F1-score: 0.166





Model Comparison

Model Name	Precision	F1 Score	ROC	Accuracy	Recall
Random Forest	0.17	0.23	0.76	0.89	0.39
XGboost	0.16	0.23	0.78	0.88	0.42
Logistic Regression	0.11	0.18	0.80	0.77	0.64
KNN	0.20	0.05	0.74	0.95	0.03
SVM	0.10	0.17	0.72	0.75	0.61
MLP	0.10	0.18	0.75	0.70	0.81
Keras	0.09	0.16	0.75	0.74	0.61

Conclusion

- Based off all the models we have built, we choose MLP to be the best model
- This is because MLP has the highest recall score
- Since our problem statement deals with stroke prediction, recall is our evaluating metric.


Questions?

Thank you