CSL 2050
PATTERN RECOGNITION AND MACHINE LEARNING

MAJOR PROJECT REPORT

SHASHWAT ROY - B21CS071

PRANAV PANT - B21CS088

YOGESH JANGIR - B 21CS083

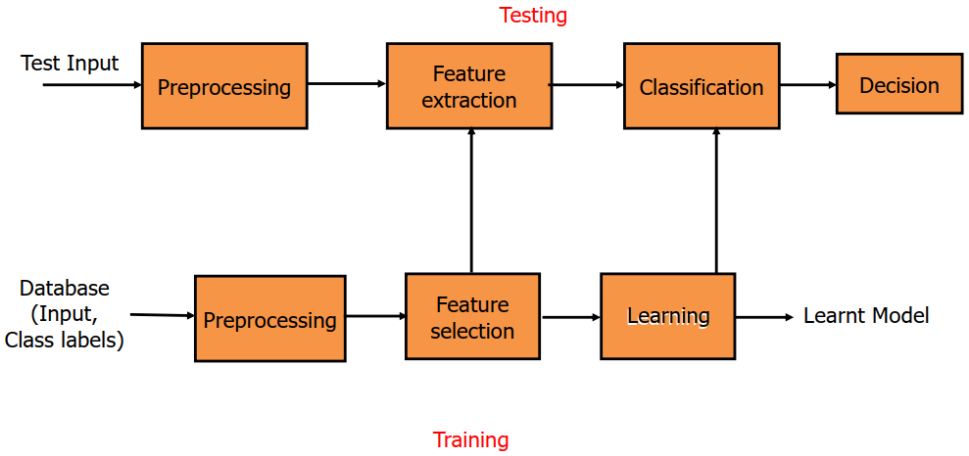
Dataset - Brain Stroke Dataset

Introduction

Brain stroke is a leading cause of disability and mortality worldwide. It occurs when the blood flow to the brain is interrupted, leading to cell death and potentially permanent brain damage.

Early detection and diagnosis of stroke are critical to prevent long-term disability and improve patient outcomes. Machine learning (ML) has shown great potential in the prediction of stroke risk, and several ML models have been developed for this purpose.

In this report, we present a pipeline for brain stroke prediction using machine learning algorithms.



Machine learning pipeline

We first discuss the dataset used for model training and testing, followed by the preprocessing steps, feature selection, and model development. We also evaluate the performance of our pipeline using various evaluation metrics and compare it with other state-of-the-art models. This pipeline has the potential to provide an accurate and efficient tool for stroke prediction, which can aid healthcare professionals in making timely and informed decisions for better patient care.

Importing data

The dataset we used is in the given link:

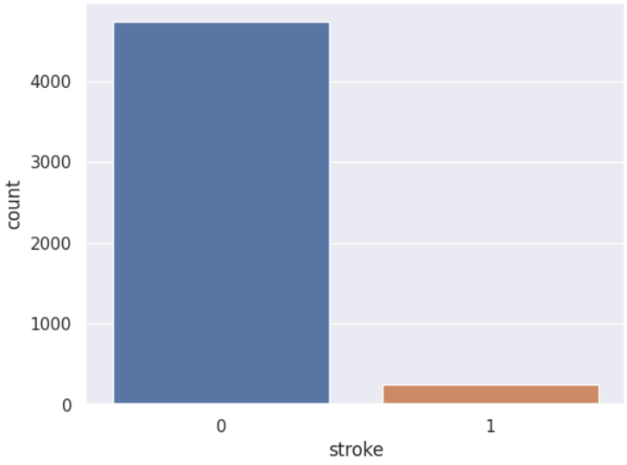
https://drive.google.com/file/d/1yruyN0GYpe0SiyyE57rXxANk1Mz0ckhX/view?usp=sharing

Checking the columns and we can say that both Categorical and numerical features are present.

- Categorical Features: gender, ever_married, work_type,Residence_type, smoking_status.
- Binary Numerical Features: hypertension,heart_disease, stroke
- Continuous Numerical Features: age, avg_glucose_level, bmi.

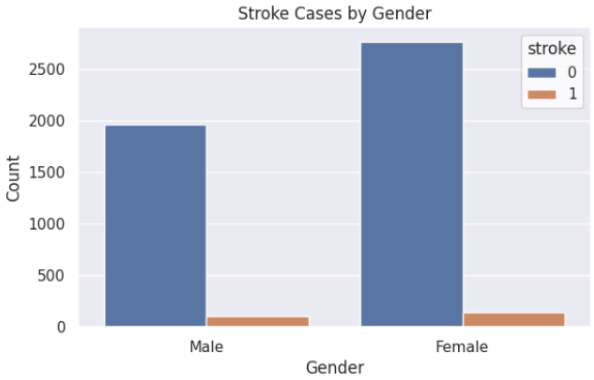
Visualizations :

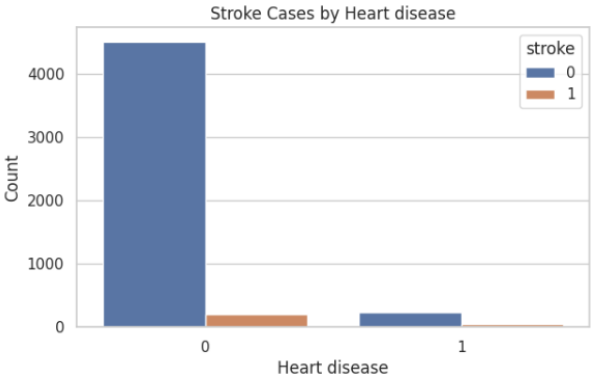
1. Visualized the features by plotting different kinds of plots including count plots, barplots and distplots for different kinds of features.
2. From the plots, we get the following inferences:



We have only 248 data points of positive class i.e, only 4.98% of data points belong to positive class. Hence, ,this is a highly unbalanced data distribution and we would have to do oversampling to obtain good results on the dataset.

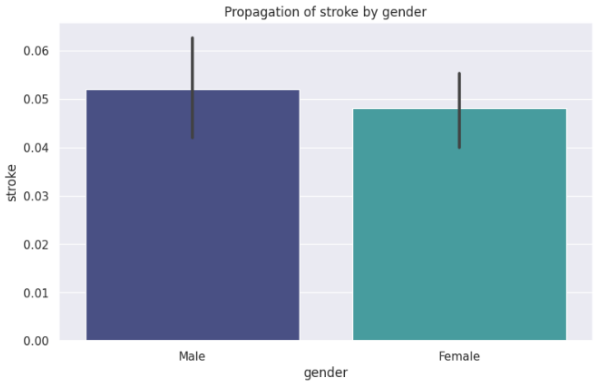
Note: Only few plots are plotted in the report. For all the plots, please refer to the collab file.

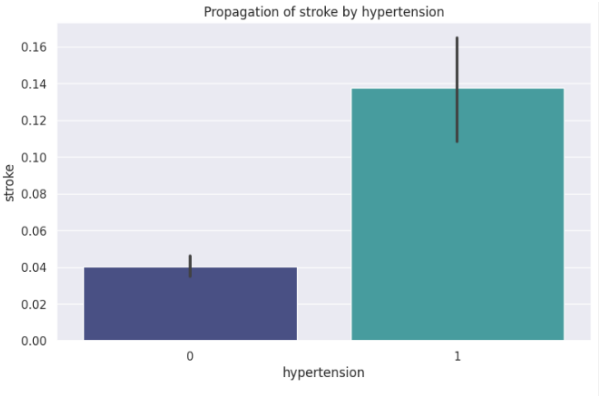




From the count plots, we can see all features with their counts of people who suffered from stroke and who did not suffer from stroke.

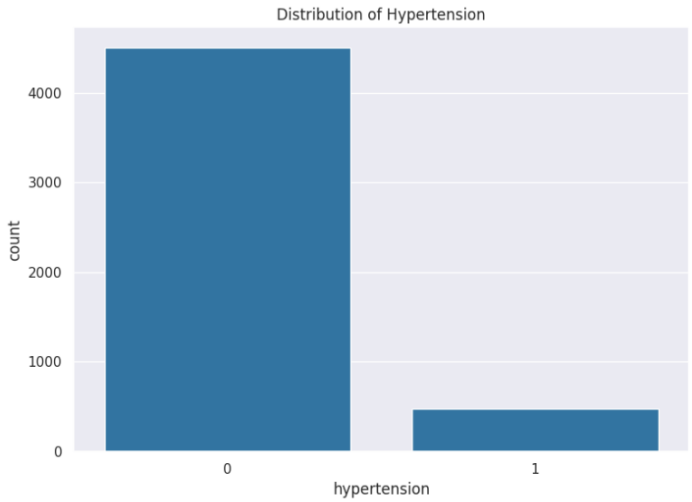
We can see that it is not correlated with any specific feature value of a column.

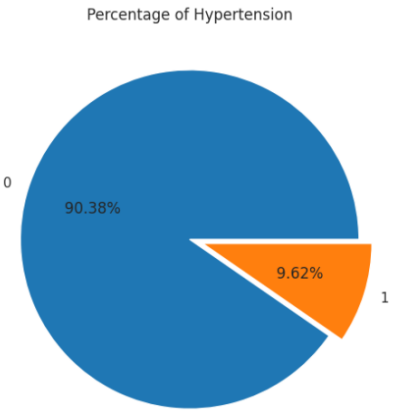




From the barplots, we have the following inferences:

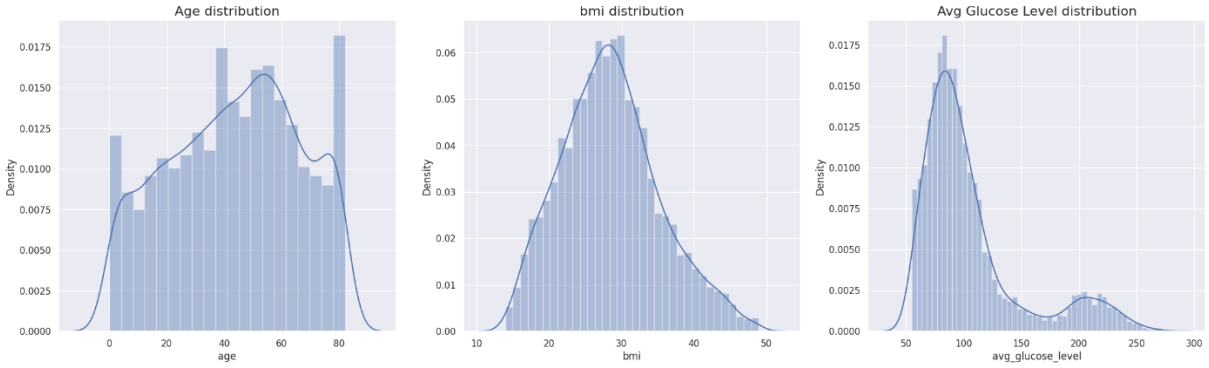
1. Both genders have almost similar chances of suffering from stroke. It is not skewed to any specific gender. Same goes for Residence (urban-rural) and work type.
2. People with hypertension, heart disease tend to have a higher chance of a brain stroke as they have more cases of stroke.





The plots showed the data-distribution of each feature set.

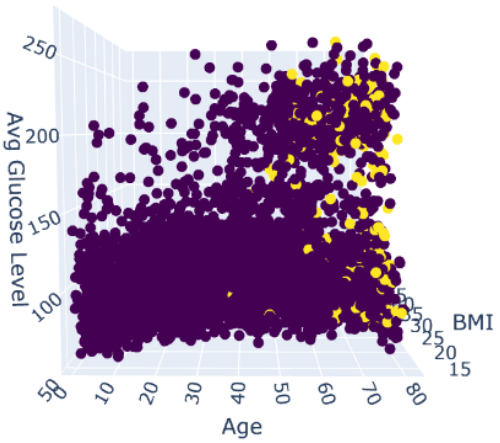
Interestingly, we have very less no. of people suffering from diseases like hypertension and heart disease and in the dataset. Also,we have more no. of females than males.



Plotting for continuous features, we have the following inferences:

1.The age feature is broadly spread, this means we have many people of each age group.

2. BMI has a gaussian curve with most people having bmi around 25-30.This means most of the people whose data was taken in the dataset were probably overweight. (as bmi>25 classifies as overweight).



Plotted 3-D plot between Age, BMI and Avg. Glucose level.

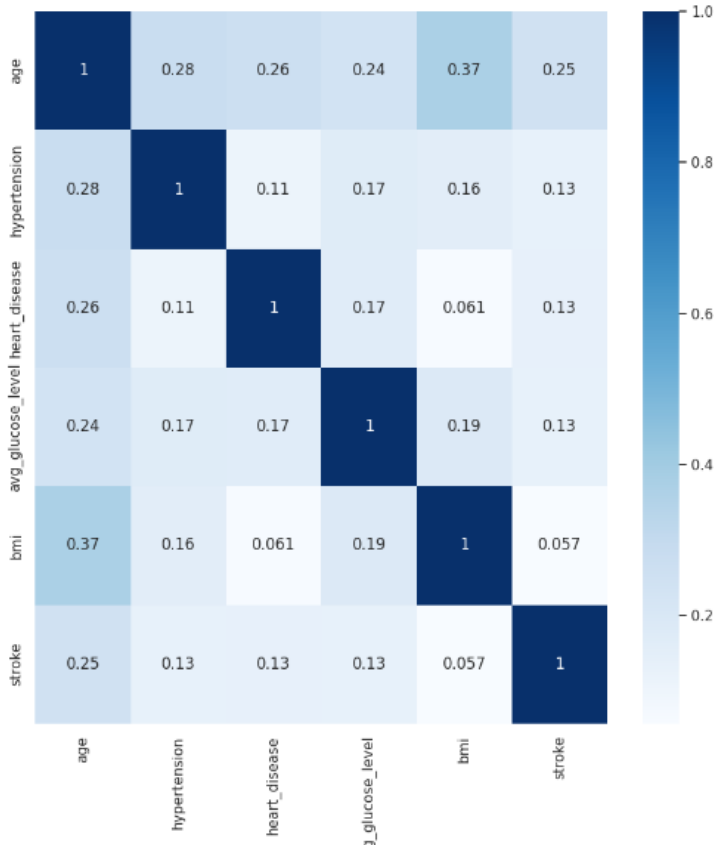
From the plot, we have the following inferences:

1.Old age people (age>60) have more chances of a brain stroke.

2. BMI does not seem to have a correlation with brain stroke, as people from all ranges of BMI have seem to suffer from brain stroke.

3. Same goes for Glucose level as it also does not have a straight relation with stroke.

Plotting the correlation matrix:



From the correlation matrix, we can see that:

1. Age has the highest correlation with stroke. As also seen from other kinds of plots, people of age>60 have a higher chance of suffering from stroke.
2. All other features do not seem to have a correlation with stroke.
3. Also, BMI and age have some correlation, meaning people of old age have a greater bmi.

Preprocessing

Fitted label encoder on the columns with object data type to convert them into numerical values and split the data into train and test sets with stratified splits. Stratified splits maintain the proportion of classes in both training and testing datasets. This is especially useful since we are using an imbalanced dataset.

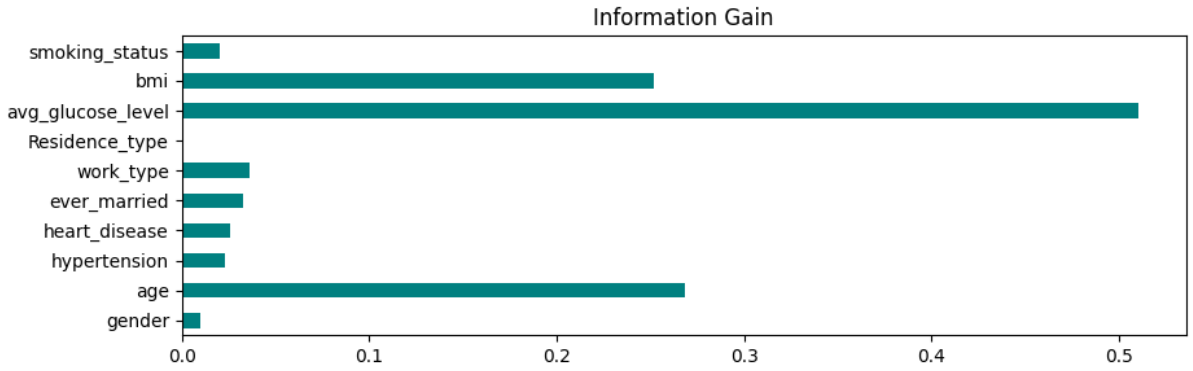
Data Transformations

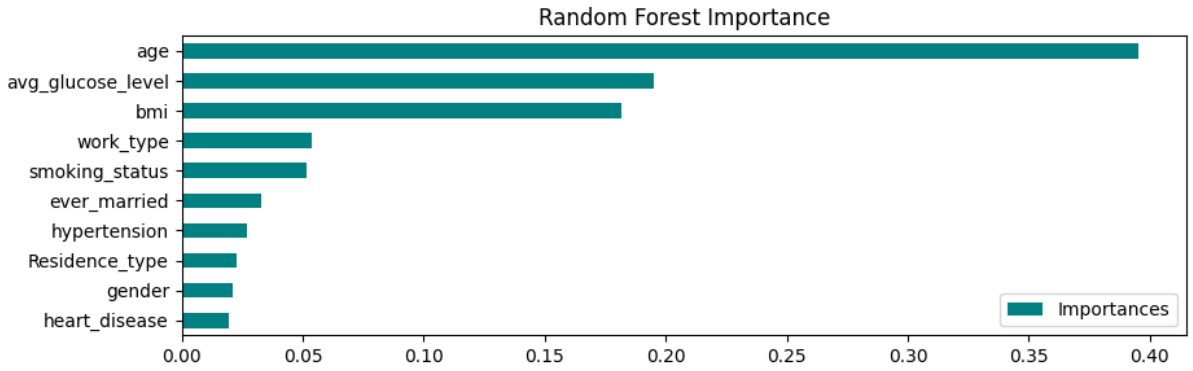
Oversampling using SMOTE

The positive class (stroke positive) constitutes only 4.9% of the data set. Hence, we apply SMOTE for oversampling. The resulting data set has 3312 data points for each stroke class.

Feature Selection

As is apparent from the correlation heatmap, not all features significantly affect the probability of suffering a stroke which is why we implemented 2 feature selection methods:



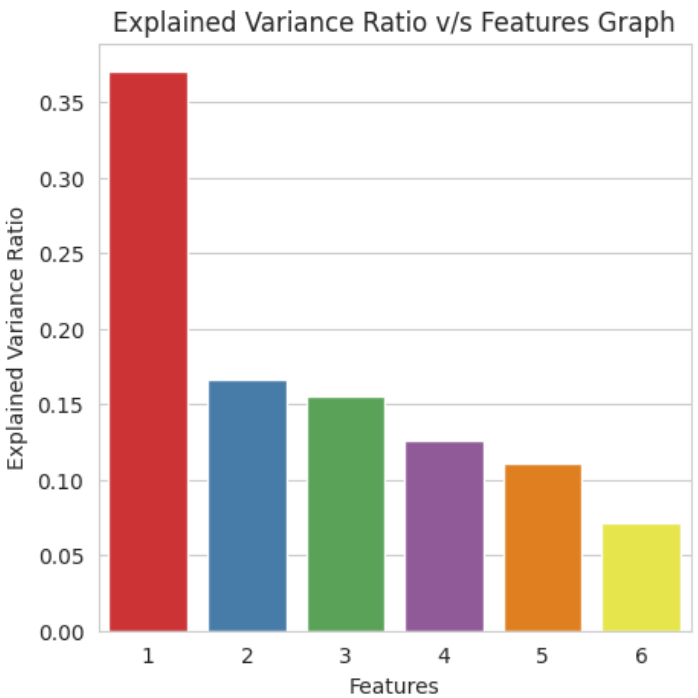


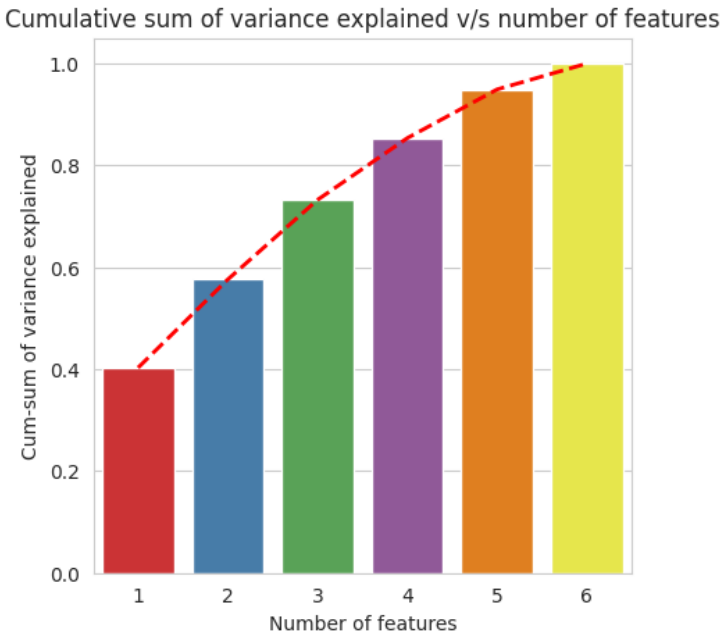
Using the above two methods, we select the following features for training:

[age,avg_glucose_level,bmi,work_type,smoking_status,ever_married]

PCA Transformation

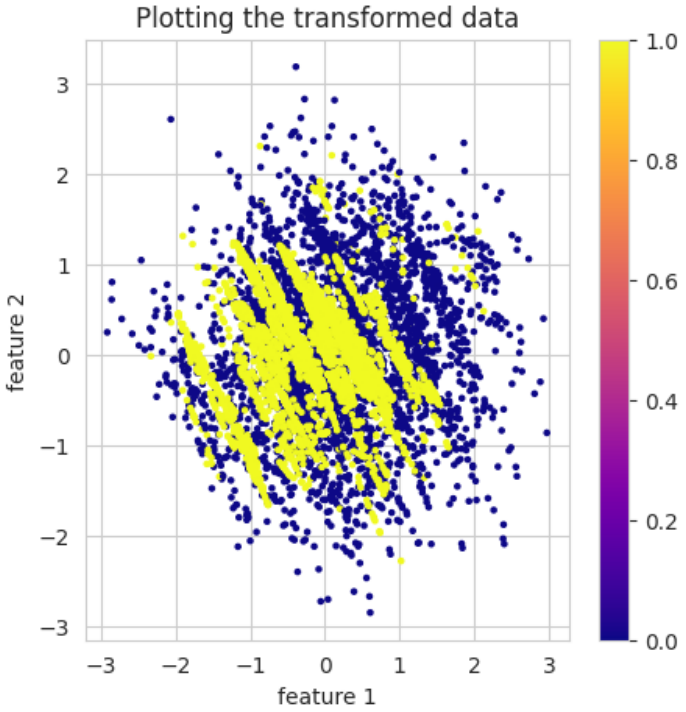
Applied PCA to see the variance distribution in the dataset-





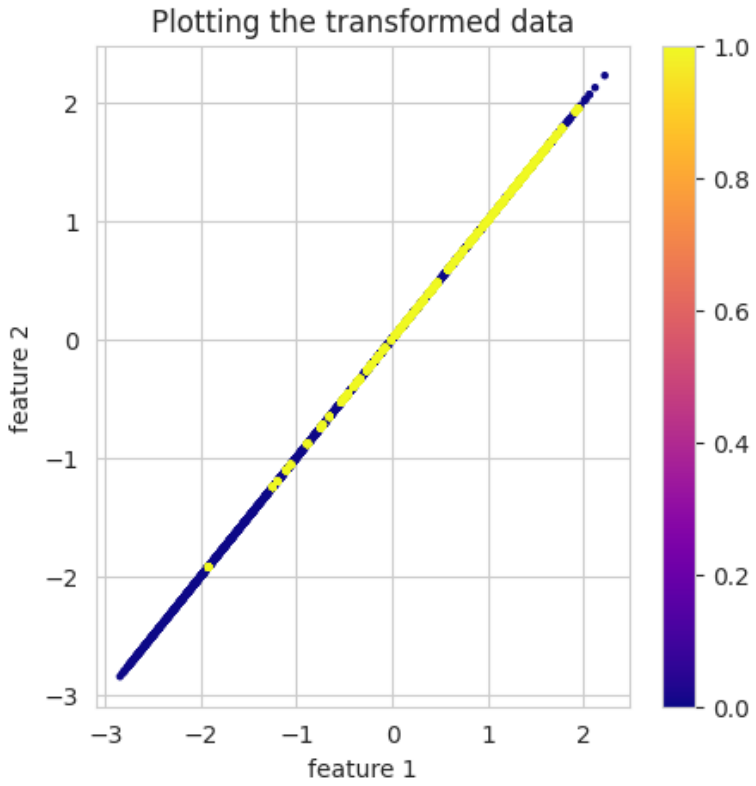
Plotted cumulative variance graph to find optimal no. of features.

Explained variance for n_features=5 is 94.91%



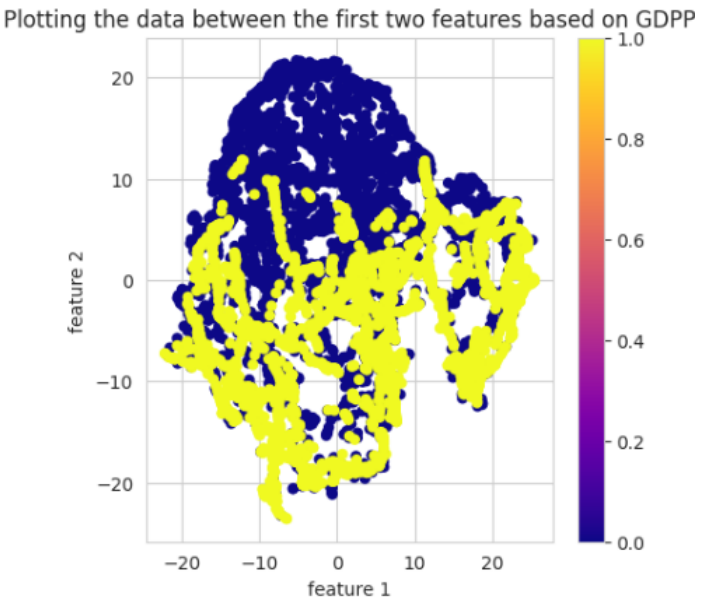
LDA Transformation

Since there are only 2 classes, LDA transforms the data set into a 1 dimensional data set.



t-SNE Transformation

We set n_components=3 for the transformation.



Model Training & Testing

We trained and tested the model through the following datasets:-

DTC

The best results are obtained for the LDA dataset.

```
              precision    recall  f1-score   support

           0       0.97      0.70      0.81      1421
           1       0.08      0.53      0.14        74

    accuracy                           0.69      1495
   macro avg       0.52      0.61      0.48      1495
weighted avg       0.92      0.69      0.78      1495


[[995 426]
 [ 35  39]]
```

Bagging Classifier

The best results are obtained for the t-sne dataset.

```
              precision    recall  f1-score   support

           0       0.98      0.74      0.84      1421
           1       0.13      0.76      0.22        74

    accuracy                           0.74      1495
   macro avg       0.56      0.75      0.53      1495
weighted avg       0.94      0.74      0.81      1495


[[1046  375]
 [  18   56]]
```

Random Forest Classifier

The best results are obtained for the untransformed dataset.

```
              precision    recall  f1-score   support

           0       0.97      0.78      0.87      1421
           1       0.13      0.61      0.21        74

    accuracy                           0.77      1495
   macro avg       0.55      0.69      0.54      1495
weighted avg       0.93      0.77      0.84      1495


[[1111  310]
 [  29   45]]
```

Logistic Regression

We get the best results for training on the untransformed dataset:

```
                precision    recall  f1-score   support

             0       0.99      0.69      0.81      1421
             1       0.13      0.85      0.22        74

      accuracy                           0.70      1495
     macro avg       0.56      0.77      0.52      1495
  weighted avg       0.95      0.70      0.78      1495


  [[981 440]
   [ 11  63]]
```

XGB Classifier

The best performance is obtained with the t-sne transformed data.

```
              precision    recall  f1-score   support

           0       0.96      0.78      0.86      1421
           1       0.09      0.45      0.16        74

    accuracy                           0.76      1495
   macro avg       0.53      0.61      0.51      1495
weighted avg       0.92      0.76      0.83      1495

[[1104  317]
 [  41   33]]
```

Naive Bayes Classifier

The best performance is obtained with the t-sne transformed data.

```
              precision    recall  f1-score   support

           0       0.99      0.61      0.76      1421
           1       0.11      0.93      0.20        74

    accuracy                           0.63      1495
   macro avg       0.55      0.77      0.48      1495
weighted avg       0.95      0.63      0.73      1495


[[868 553]
 [  5  69]]
```

Voting Classification with Bagging

We obtain the best f1 score,recall when using XGBClassifier and Logistic Regression as the ensemble and training on the untransformed dataset.

```
              precision    recall  f1-score   support

           0       0.98      0.74      0.84      1421
           1       0.13      0.76      0.22        74

    accuracy                           0.74      1495
   macro avg       0.56      0.75      0.53      1495
weighted avg       0.94      0.74      0.81      1495


[[1046  375]
 [  18   56]]
```

When using SVM with sigmoid kernel,XGBClassifier and Logistic Regression as an ensemble, we obtain better accuracy.

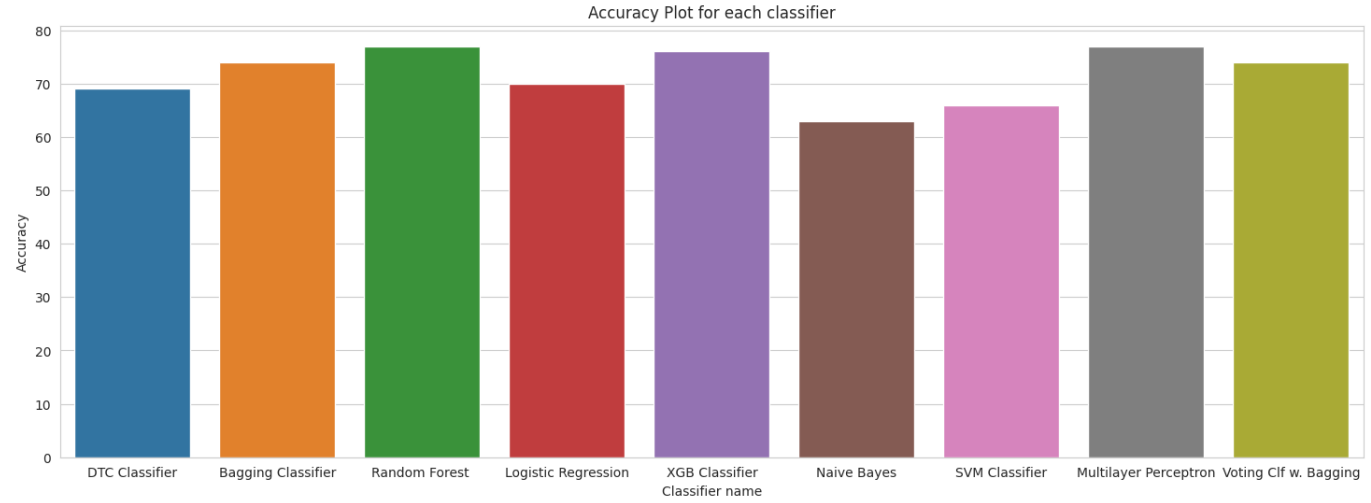
```
              precision    recall  f1-score   support

           0       0.96      0.84      0.90      1421
           1       0.11      0.38      0.17        74

    accuracy                           0.82      1495
   macro avg       0.54      0.61      0.53      1495
weighted avg       0.92      0.82      0.86      1495


[[1192  229]
 [  46   28]]
```

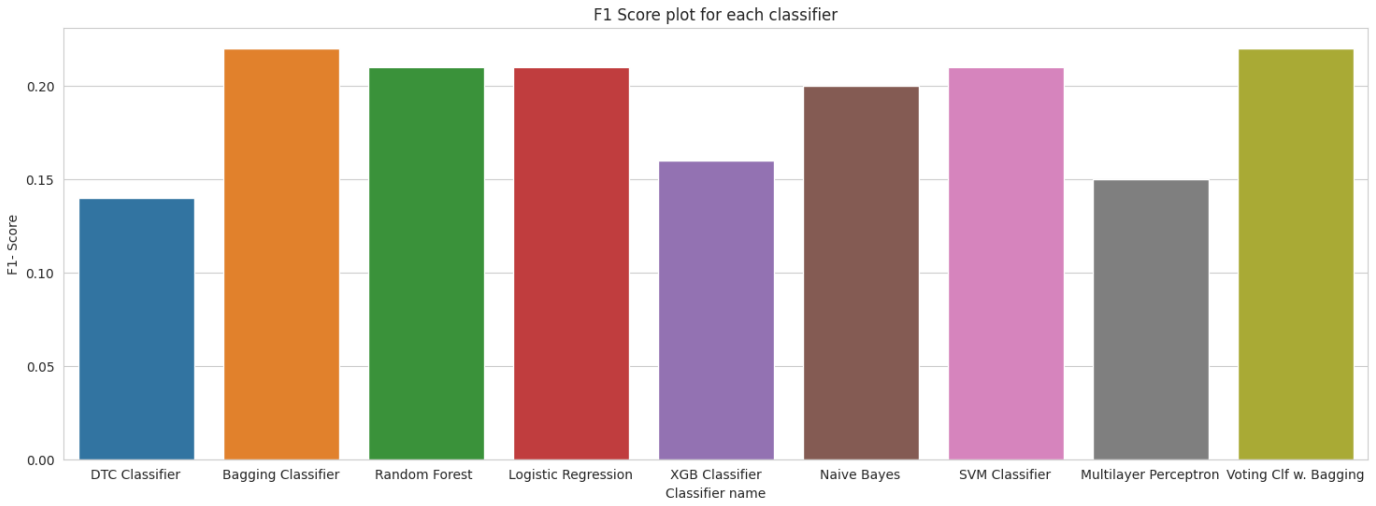
Multi Layer Perceptron

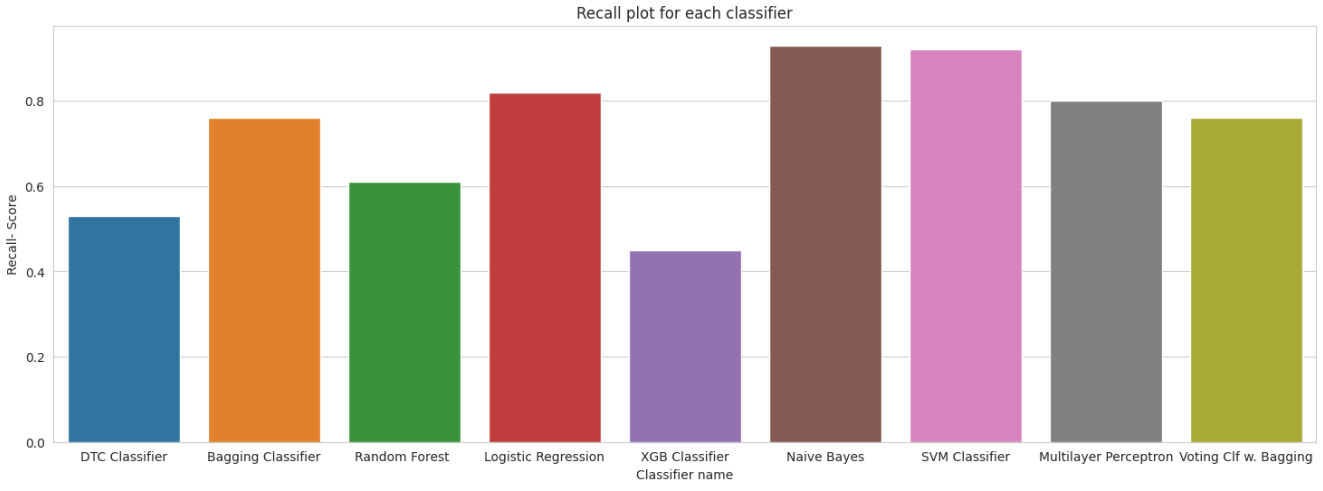
The best performance is obtained with the t-sne transformed data using tanh and ReLU activation functions and adam optimizer.

```
47/47 [==============================] - 0s 2ms/step
              precision    recall  f1-score   support

           0       0.98      0.64      0.77      1421
           1       0.10      0.80      0.18        74

    accuracy                           0.65      1495
   macro avg       0.54      0.72      0.48      1495
weighted avg       0.94      0.65      0.74      1495

[[906 515]
 [ 15  59]]
```

Performance Comparison







Inferences & Conclusions

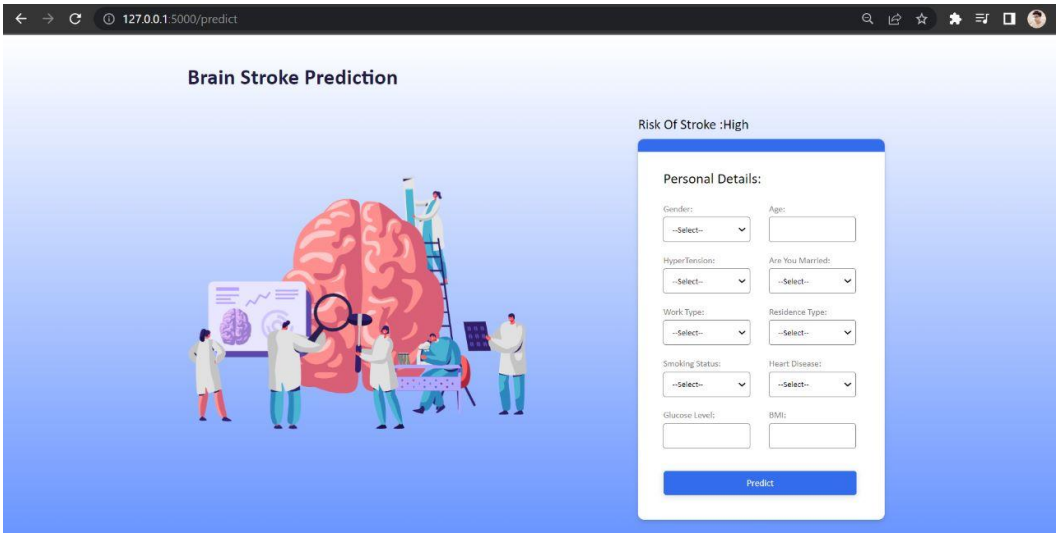
The dataset was highly imbalanced with a very small proportion of positive class. As a result, models struggled to perfectly capture the characteristics of the positive class. Moreover, when tuning the hyperparameters, it was noticed that there was a trade-off between recall of the positive class and the model accuracy after a certain point. Due to the sheer size of the majority class, the precision of all models were subpar for the positive class.

Brain stroke is a life-threatening medical condition. Its detection is essential in saving countless lives. Priority must be given to minimizing the number of false negative cases as misclassifying is hazardous to the potential stroke victim's health while the issue of false positives is not as nearly hazardous to the patients who are falsely classified as potential stroke candidates.

Although almost all of the models gave comparable results across different parameters, the bagging classifier can be considered to have performed the best out of all of them.

Creating Website and Hosting

We created a website through HTML, CSS and Flask in which the user can enter the details of the features. These values of features are then fetched to the backend written in flask language, which is then passed as a test datapoint to the model. The model then predicts the brain stroke as 0 or 1 which is then displayed on the site.



Note: The website is hosted locally.

Contributions

Shashwat Roy (B21CS071)

Responsible for data transformations, training and parameter tuning of models, analyzing performances of the same.

Contributed in report-writing.

Pranav Pant (B21CS088)

Responsible for data analysis, preprocessing , training and parameter tuning of models.

Contributed in report-writing and website.

Yogesh Jangir (B21CS083)

Responsible for data visualization , training and parameter tuning of models.

Contributed in report-writing and website coding.