STROKE PREDICTION

Ananya Nagpal, Yatharth Verma, Shivanshu Shekhar, Saurabh Chavan, Bontha Rishi

B22EE008, B22CS064, B22EE062, B22EE061, B22A1014

Abstract

This study provides a comprehensive overview of the process involved in constructing a stroke prediction classifier. The classifier is designed to estimate the likelihood of an individual experiencing a stroke, while also identifying the important elements that significantly contribute to the occurrence of a stroke. Stroke prediction can be performed by taking into account numerous factors such age, presence of heart disease, smoking status, and others. The dataset is analyzed to determine the likelihood of a person experiencing a stroke. The dataset's characteristics are then utilized in five distinct machine learning models to forecast strokes, and their effectiveness is subsequently compared. The objective is to examine the variables that influence the likelihood of experiencing a stroke. The data can subsequently be utilized to develop a predictive system for identifying stroke occurrences and the primary factors contributing to stroke.

Contents

1 Introduction

Stroke is the leading cause of mortality and morbidity worldwide, as stated by the World Health Organisation. A stroke is a cerebrovascular accident characterized by the presence of a blood clot or hemorrhage in the brain, resulting in potentially irreversible damage. It causes damage to the brain similar to how a heart attack causes damage to the heart. The causes of death from stroke are primarily due to the presence of co-morbidities and the occurrence of complications. Timely identification of diverse warning symptoms of a stroke can aid in mitigating the intensity of the stroke. Common predictors of mortality from stroke in those over the age of 65 include a history of previous stroke, atrial fibrillation, and hypertension. Furthermore, timely identification and proper treatment are necessary to avert additional harm to the damaged region of the brain and potential consequences in other bodily areas. The prevalence of stroke is expected to escalate, leading to a continual rise in stroke and heart disease mortality rates.

Stroke occurrence can be predicted by analyzing input characteristics such as gender, age, presence of different diseases, and smoking habits. Our project aims to utilize machine learning methods to accurately forecast the occurrence of a stroke by analyzing a dataset that contains numerous risk variables.

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_job", "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 a test size of 0.3.

1.1 citing paper

Receiver Operating Characteristic (ROC)

SMOTE Technique

Feature Selection

Encoding Techniques

SVM

Random Forest

1.2 Figures

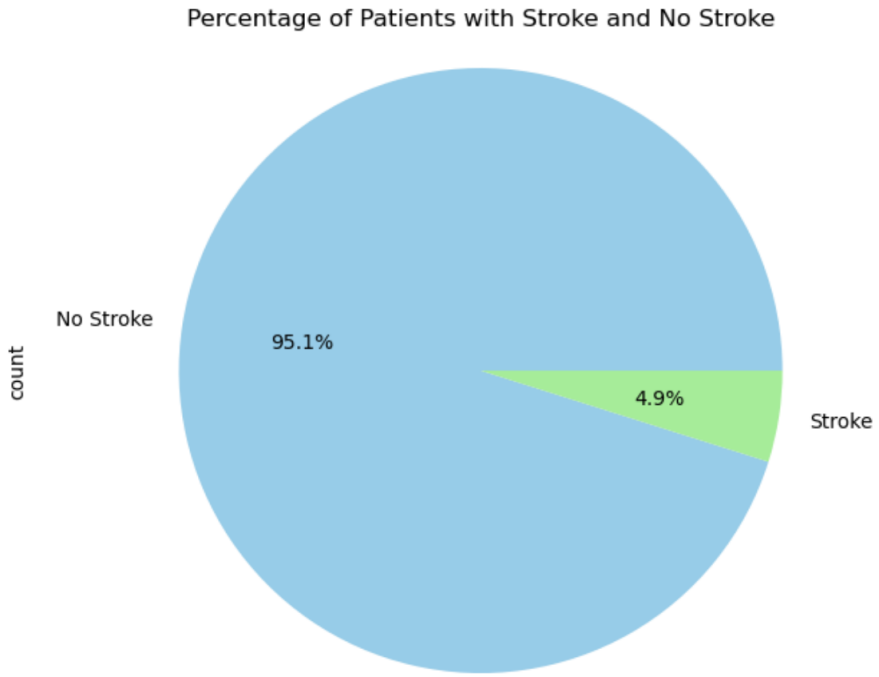

Figure 1: The dataset is highly imbalanced which will have a negative impact on training ML models.

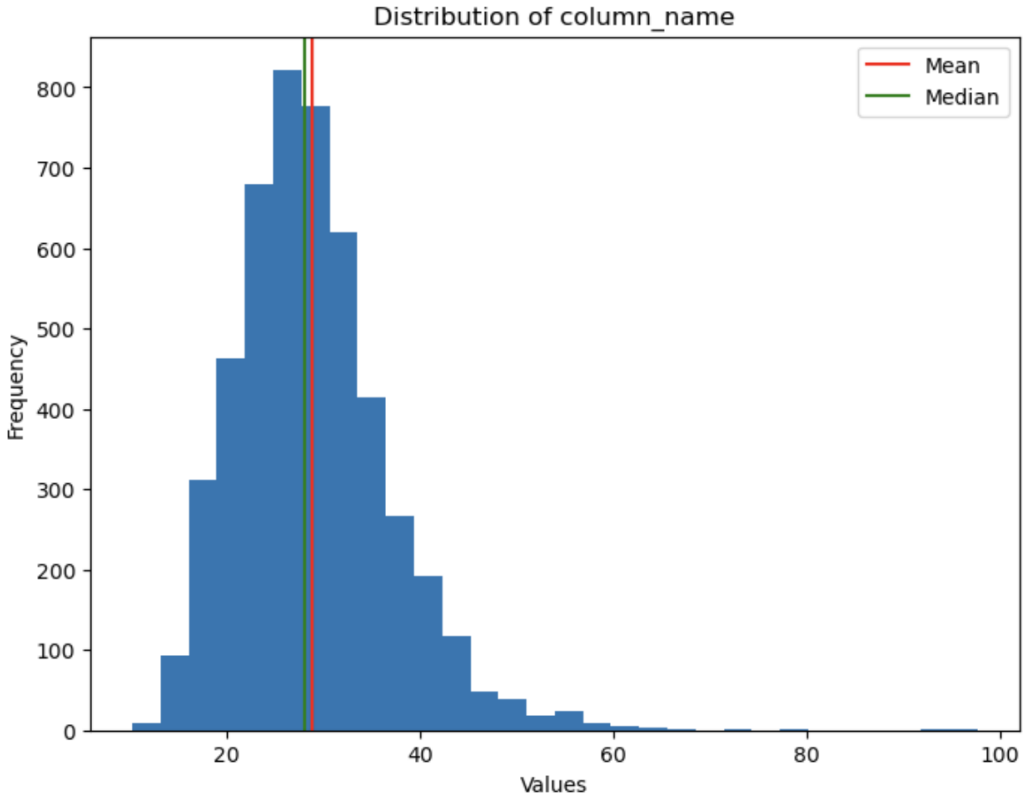

Figure 2: Dataset only had null values in the bmi column, As we can see that the distribution is rightly skewed therefore we replaced all the null values by the median of the bmi column.

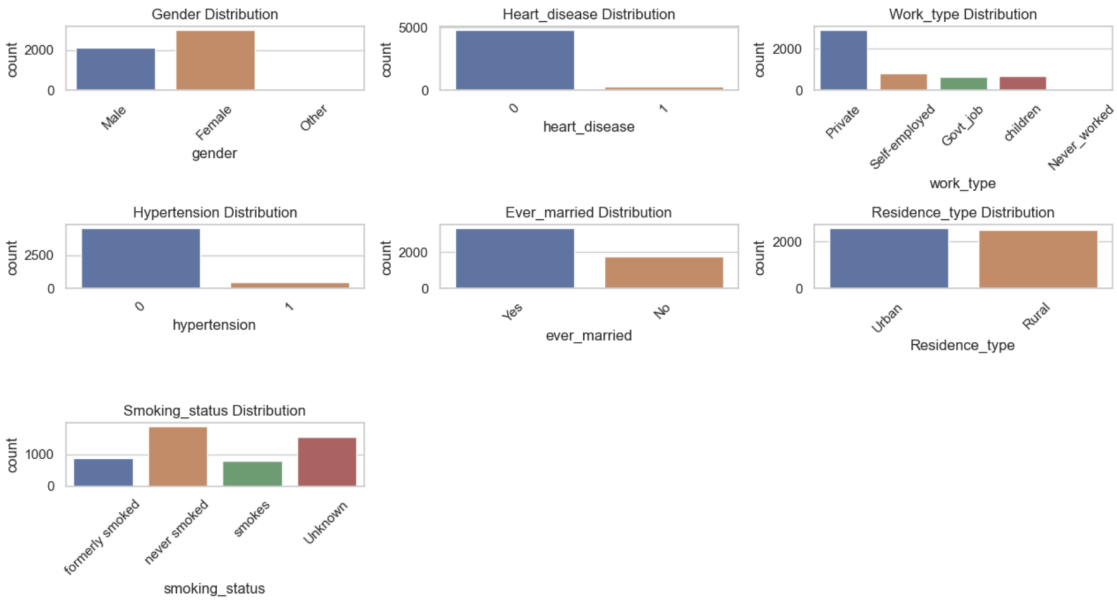

Figure 3: Univariate exploratory data analysis for the categorical variables.

```
age                                float64
hypertension                         int64
heart_disease                        int64
avg_glucose_level                  float64
bmi                                float64
stroke                               int64
age_group                            int64
BMI_class                            int64
glucose_class                        int64
work_type_Govt_job                 float64
work_type_Never_worked             float64
work_type_Private                  float64
work_type_Self-employed            float64
smoking_status_Unknown             float64
smoking_status_formerly smoked     float64
smoking_status_never smoked        float64
smoking_status_smokes              float64
gender_Female                      float64
gender_Male                        float64
ever_married_No                    float64
ever_married_Yes                   float64
Residence_type_Rural               float64
Residence_type_Urban               float64
```

Figure 4: List of features after encoding categorical data as converting them into numerical form is necessary for algorithms to process the data effectively

2 Approaches Tried

2.1 Overview

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

- K Nearest Neighbour
- Decision Tree
- Random Forest
- Support Vector Machine
- Naive Bayes Classification
- Perceptron

We also made use of Feature Selection where Top 10 features were selected based upon their weightage in predicting our output and SMOTE techniques which is Synthetic Minority Over-sampling Technique (SMOTE) is a technique for oversampling imbalanced data by generating synthetic samples for the minority class was used as the dataset was highly imbalanced.

2.2 ML Models for stroke prediction

1. **KNN:** It works well in classification problems by predicting based upon labels of nearest data points. GridSearchCV was used to find the best value of k which came out to be 1.
2. **Decision Tree:** 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. Decision Tree classifier with min sample split 2 was implemented.
3. **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 20, min sample split 2, and min sample leaf 1 was implemented.

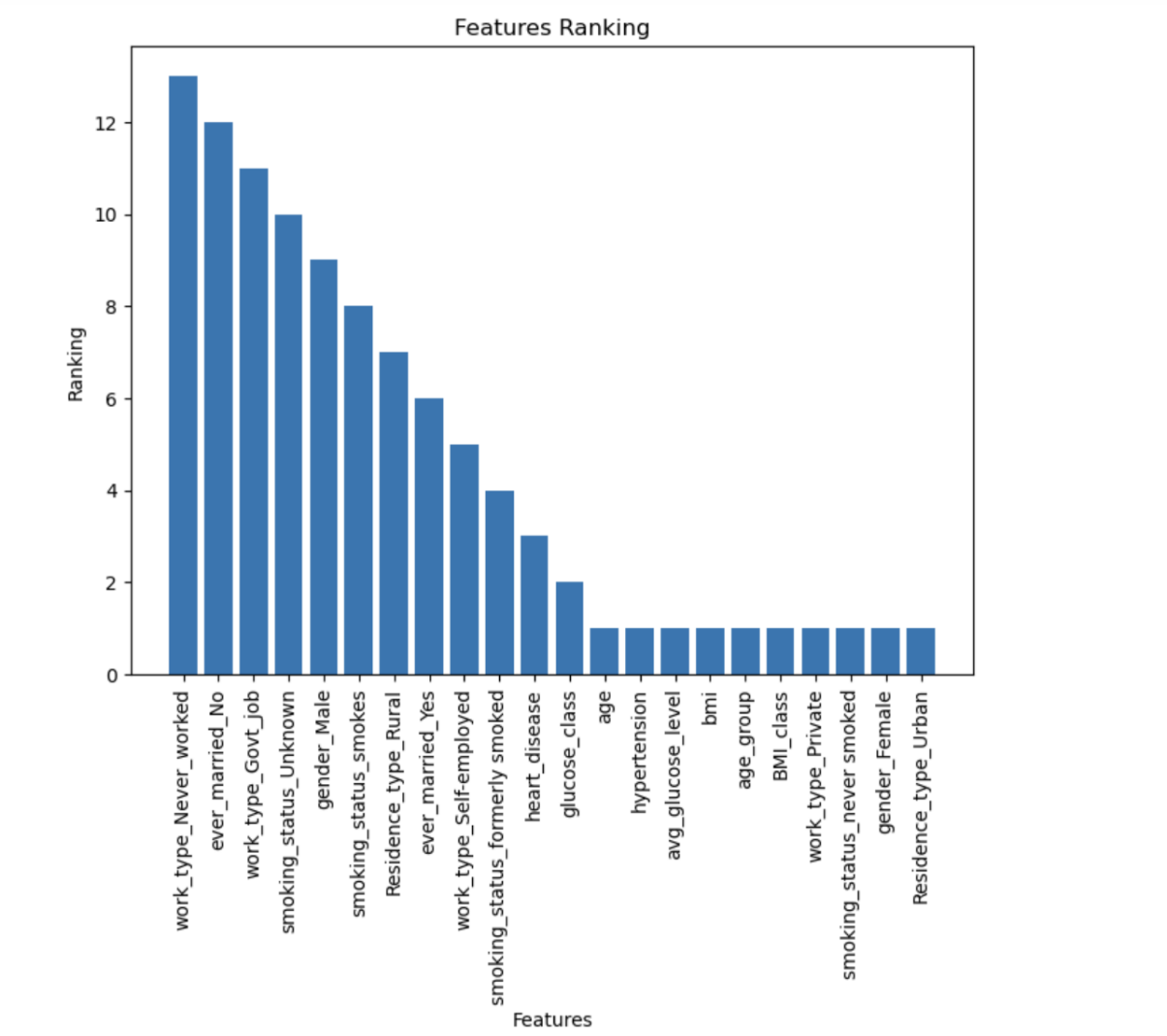

Figure 5: Feature selection was implemented and top 10 features were selected. Feature selection increases the prediction power of the algorithms by selecting the most critical variables and eliminating the redundant and irrelevant ones

```
Original set distribution:
 stroke
0    3403
1     174
Name: count, dtype: int64

Resampled set distribution:
 stroke
0    3403
1    3403
Name: count, dtype: int64
```

Figure 6: Before and after scenarios of labels after implementing Synthetic Minority Over-sampling Technique (SMOTE) which is a technique for oversampling imbalanced data by generating synthetic samples for the minority class was used as the dataset was highly imbalanced

```
Random Forest Test metrics
Confusion Matrix:

[[1396   62]
 [  59   16]]


Classification Report:

              precision    recall  f1-score   support

           0       0.96      0.96      0.96      1458
           1       0.21      0.21      0.21        75

    accuracy                           0.92      1533
   macro avg       0.58      0.59      0.58      1533
weighted avg       0.92      0.92      0.92      1533
```

Figure 7: Results of the models were analyzed with the help of confusion matrix and classification report. Here is the result of the Random Forest classifier.

4. **SVM:** It works by finding the optimal hyperplane that separates data points into different classes. GridSearchCV was used in the model to find the best parameters which directed to use $C = 100$, $Gamma = auto$, and $kernel = rbf$.

5. **Naive Bayes Classifier:** These are a collection of classification algorithms based on Bayes Theorem. For this, we used Bernoulli naive bayes classifier because it is a binary classification problem.

6. **Perceptron:** Perceptron is one of the most basic models of deep learning used for classification. This was used just to observe its performance in comparison with other machine learning models.

Results of the models were analyzed with the help of confusion matrix and classification report.

3 Experiments and Results

3.1 Exploring the dataset and pre-processing

1. The dataset is highly imbalanced which will have a negative impact on training ML models.

2. The dataset was checked for null values. It was found that the dataset only had null values in the BMI column, which were removed by replacing them with the median of the column as the distribution was rightly skewed.

3. It is always better to reduce the number of categories within the column. Therefore we changed the 'Other' value in the gender column to 'Female', changed the values of people under 18 with 'Unknown' smoking status to 'never smoked' assuming that they do not smoke and also changed the "children" under work_type to "Never_worked" assuming that children don't work.

4. We performed univariate analysis for continuous variables and these were the observations:

   (a) We have more females in the dataset. Also, there's a single patient whose gender is "Other". Since female is the mode of the gender feature, the patient with 'Other' will be recategorized to female. This way, we'll have just 2 categories in the column.

   (b) Most of the patients in the data are healthy in terms of heart disease.

   (c) More than 50% of the patients work in the private sector.

   (d) With the assumption that children can't work/never worked, we changed the instances of "children" category.

   (e) 90% of the patients are not hypertensive.

   (f) We have more patients who have married at one stage in their life than those who haven't.

   (g) We have almost an equal number of patients living in the Rural and Urban areas.

5. We also performed multivariate analysis and these were the observations:

   (a) More patients that are older than 40 years seem to have strokes with a small number of patients less than 40 years.

   (b) The males in the data tend to have strokes at an age over 40, while women tend to have strokes from around their 30s.

   (c) More patients from the private sector have strokes, followed by the self-employed, and government workers respectively.

   (d) The underweight patients are the least likely class to have strokes, followed by the healthy weight class.

6. We encoded the categorical data:

   (a) Label Encoding was used for the ordinal features so we can preserve the order of the categories.

   (b) OneHot Encoding was used for other nominal features since there is no inherent order in the categories.

3.2 Results

	KNN	Decision Tree	Random forest	Naive bayes	SVM	Perceptron
Accuracy	0.89	0.89	0.92	0.65	0.87	0.78
F1 Score	0.14	0.17	0.21	0.20	0.13	0.23

Figure 8: Tabular representation of accuracy and F1 score

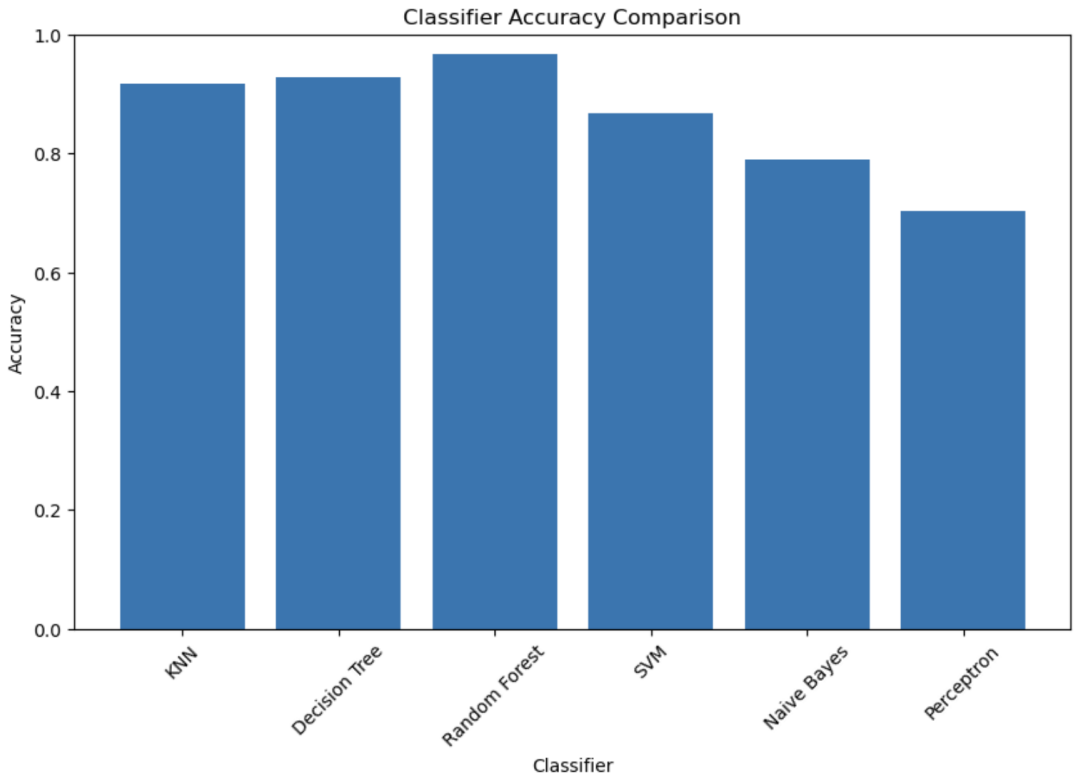

Figure 9: Performance comparison on the basis of accuracy

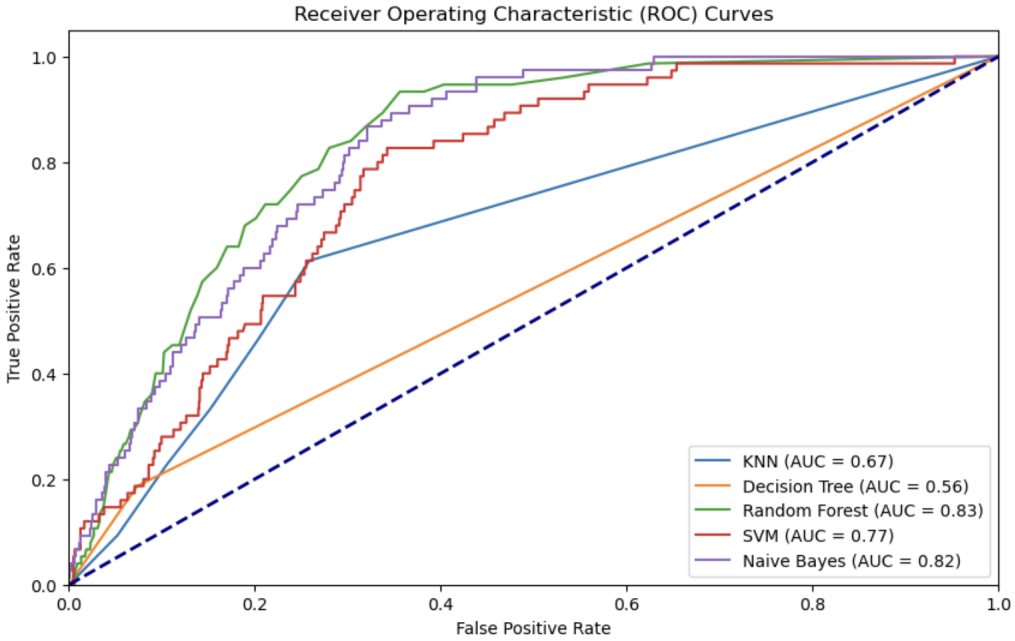

Figure 10: ROC curve

3.3 Analysis

On the basis of performance, Random forest performed best for the given dataset. Further analysis was done to get more insights.

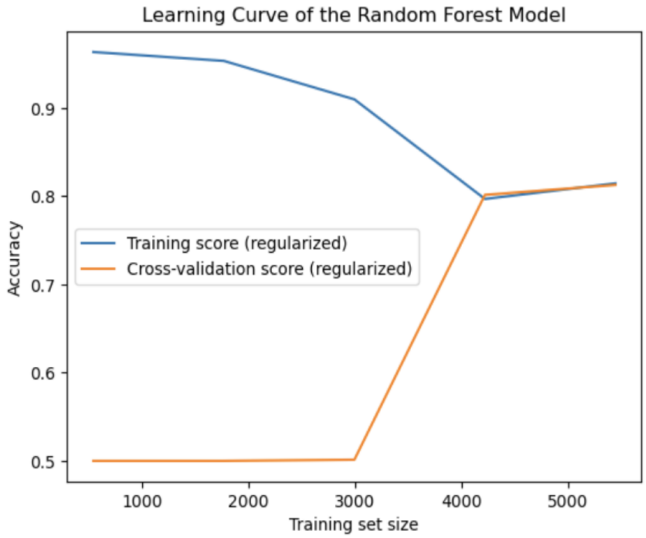

Figure 11: Learning curve for Random Forest

Conclusion: The model is overfitting on the train data and the cross-validation score increases as the training size increases which is a good sign that the model is generalizing well to new data. The training score decreasing as the training size increases is expected because of the regularization, and the increasing cross-validation score indicates in its ability to generalize to new data.

4 Summary

This project involves the construction of a classifier using machine learning models to estimate the likelihood of an individual experiencing a stroke. The dataset used contains essential factors such as age, gender, presence of diseases, and lifestyle habits like smoking status. The dataset underwent preprocessing steps such as handling null values, encoding categorical data, and balancing imbalanced data using the SMOTE technique to ensure accurate predictions.

Various classification algorithms including K Nearest Neighbour, Decision Tree, Random Forest, Support Vector Machine, Naive Bayes, and Perceptron were implemented and evaluated for stroke prediction. Among these, the Random Forest algorithm performed the best with an accuracy of 92percent and an F1 score of 0.21. The project's objective is to develop a predictive system that can identify stroke occurrences and determine the primary factors contributing to strokes.

A Contribution of each member

1. Ananya Nagpal: Prepared dataset for training. Implemented KNN. Coordinated in making web page and report.
2. Yatharth Verma: Performed EDA. Implemented Random Forest and SVM. Coordinated in making web page and report.
3. Shivanshu Shekhar: Performed data processing. Implemented Perceptron. Made spotlight video.
4. Saurabh Chavan: Implemented Decision Tree. Made presentation and spotlight video.
5. Bontha Rishi: Implemented Naive Bayes. Made presentation.