Stroke Prediction using Machine Learning Techniques: A predictive analysis on Imbalanced Dataset

Monirul Islam Mahmud

Department of ECE

North South University

Dhaka, Bangladesh

mahmudislam2025@gmail.com

Hafeza Akter

Department of ECE

North South University

Dhaka, Bangladesh

hafeza.akter@northsouth.edu

*Abstract*— Stroke is a serious condition that can make it impossible for some body parts to move. A stroke patient needs to receive treatment right away because delaying care could seriously harm their health. The goal of this paper is to develop a model based on several machine learning techniques that can predict stroke risk given some data. The combined public and private dataset are used to train the model. To balance the data, we must use the oversampling and SMOTE techniques. The best accuracy of 96% among the 7 machine learning algorithms we used is provided by Random forest. Additionally, we created a web application that will take user input and give us the prediction.

Keywords—Stroke Prediction, Health, Machine Learning, Imbalanced Dataset, Web App.

# Introduction

Stroke is a medical emergency that occurs when the blood flow to a portion of the brain is blocked or diminished, brain tissue cannot receive oxygen and nutrients, which results in an ischemic stroke. A stroke is a potentially fatal disorder that emerges when there is insufficient blood flow to a certain area of the brain. The most frequent causes of this is a blocked artery or brain bleeding. The World Stroke Organization estimates that 13 million people worldwide get a stroke each year, with 5.5 million of them dying as a result [1]. Stroke does not only affect the elder peoples only, but also the leading causes of death and disability in the world, stroke impacts all part of any patient's life, including their family, social circle, and profession [2]. The two main global causes of death and disability are ischemic heart disease and stroke. The American Heart Association states that because stroke has a high mortality rate, it is regarded as a serious health problem [3]. Figure 1 shows the latest global health estimates by cause for ischemic stroke from the year 2000-2016 [4].

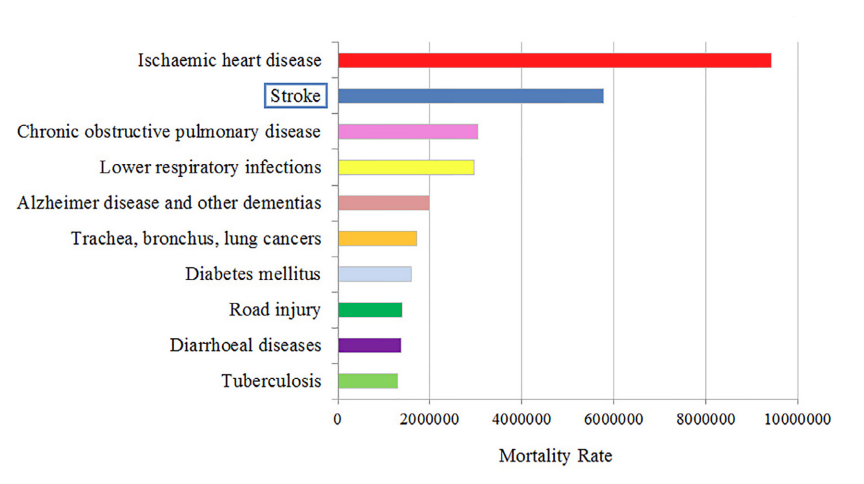


Figure 1: Figure ranks estimated mortality by cause, recorded in 2016

### For a stroke to be effectively treated, early detection is essential and ML can be very helpful in this process. Machine Learning (ML) is an advanced technology that enables healthcare professionals to make clinical judgments and predictions in order accomplish that. There have been numerous research on the use of ML for improving stroke diagnosis during the last few decades [4]. Existing research only performs insufficient amount of exploratory data analysis (EDA), and the majority of them use public dataset. However, we will implement this research using a mixture of public and private dataset while conducting enough data analysis to imbalanced dataset. Our unique contrbutions to this proposed research work are given below:

* Merged dataset of public and private stroke data is used in this research work.
* 7 different Machine Learning techniques including Logistic Regression, SVM, Decision Tree, Random Forest, K-nearest Neighbors, Gradient Boosting, Naïve Bayes algorithms are used in this research work to reach the best result.
* A complete web app is developed based on this research work which gives us the stroke prediction with real time inputs.

# Related Works

Existing literature were reviewed to gain essential information about several models that were used in Stroke Prediction. Some significant works are given below.

## Existing Dataset

Ferdib-Al-Islam at el. [5] used public data from Kaggle and private data from medical hospital in Bangladesh. The dataset contained 5110 instances of 12 columns. Atul Kumar Uttam at [6] used Kaggle data where 11 separate variables are included in the dataset with 5110 record out of which 4861 records of no stroke and 249 records of stroke are present. R.K Kavitha at el. [7] used 5110 records with 10 features, where 3577 (70%) was used as training data set and 1533 (30%) was used for testing the model. In the total data set, female was more than the male. Again, public data from Kaggle is used by Chetan Sharma at el. [8] also. Govindarajan at el. [9] used 507 patient data collected from Sugam Multispecialty hospital, Kumbakonam. G. Fang et al. [10] proposed machine learning algorithms to detect stroke by using IST dataset that contained 19435 patient data of 467 hospitals. Elias at el. [11] used Kaggle dataset, where they focused on participants who are over 18 years old. The number of participants was 3254, and 10 attributes.

## Existing Data Analysis

The strategies (under sampling, hyperparameter optimization on training, and validation on the testing fold) were used, and the validation results were combined over the rounds to provide a performance on the full imbalanced dataset in the paper [12]. Null values handle, normalization technique, Random Oversampling on imbalanced dataset, Hyperparameter tuning and label encoding are used in this paper [13]. Mean data used in missing values, label encoding, SMOTE, combination of Under-sampling and Oversampling, Normalization and standardization technique are used in [14].

Normalization, label encoding, missing value handle are done in [15]. In [5], Simple Imputer of Scikit learn, min-max scaling, label encoding is done. Under sampling and over sampling can be seen in [6]. In this research [7], feature selection techniques such as Principal Component Analysis (PCA), Chi Square, PCA combined with Chi Square, and Chi Square with outlier removal are used. Lastly, this research study [11] used the SMOTE approach to deal with the imbalanced participant distribution between the stroke and non-stroke classes.

## Machine Learning as Solution

Logistic Regression (LR), Random Forest (RF), XGBoost, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) were used by C. Kokkotis at el. [12]. The LR classifier achieved the best scores in accuracy. Support Vector Machine has the highest accuracy of 99.99% in [13]. We can see that Random Forest is the best-performing algorithm in terms of accuracy with a score of 90.36% in [14]. The proposed study in [15] has a 97% accuracy rate, with the weighted voting classifier outperforming the base classifiers where 10 different classifiers, including Logistics Regression, Stochastic Gradient Descent, Decision Tree Classifier, AdaBoost Classifier, Gaussian Classifier, Quadratic Discriminant Analysis, Multi Layer Perceptron Classifier, K Neighbors Classifier, Gradient Boosting Classifier, and XGBoost Classifier for predicting the stroke. The dataset has used to predict the result using a popular oversampling technique called SMOTE with different machine learning classifiers (Logistic Regression, Random Forest, and XG Boost) where the random forest model gave the 99.07% accuracy [5]. Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT) classifiers are applied in [16] and Random forest gave the maximum accuracy of 95.30%. Random forest algorithm obtains an accuracy of 98.94% in [8].

# Methodology

This section explains the proposed system’s methodology as well as how it was successfully implemented. This part is divided into 4 subsets: (1) Project Workflow (2) Dataset Details (3) Data Preprocessing (4) Machine Learning Model (5) Result Analysis & Evaluation (6) Web App development.

## Project Work Flow

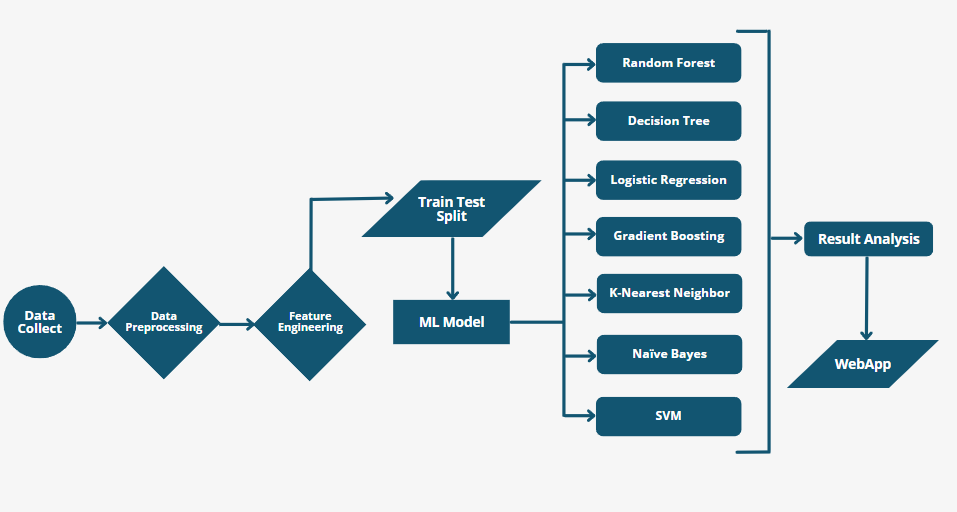


Figure 2: Project Workflow

## Dataset Details

First, this project gathered public data on stroke prediction from the Kaggle website. All libraries were imported into Google Collaboratory first, then the dataset. The dataset's shape was analyzed and it emerged that it had 12 features and 5110 records. The next 12 features are listed.

1. id: unique identifier
2. gender: "Male", "Female" or "Other"
3. age: age of the patient
4. hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5. heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6. ever\_married: "No" or "Yes"
7. work\_type: "children", "Govt\_job", "Never\_worked", "Private" or "Self-employed"
8. Residence\_type: "Rural" or "Urban"
9. avg\_glucose\_level: average glucose level in blood
10. bmi: body mass index
11. smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"\*
12. stroke: 1 if the patient had a stroke or 0 if not

## Data Preprocessing

The dataset's null values were analyzed afterwards. BMI has been found to have 201 null values. Mean values were used to fill up these null values. Following the analysis of the count plot, it became noticeable that there were extremely few people with strokes. We must use balancing techniques to balance the dataset.

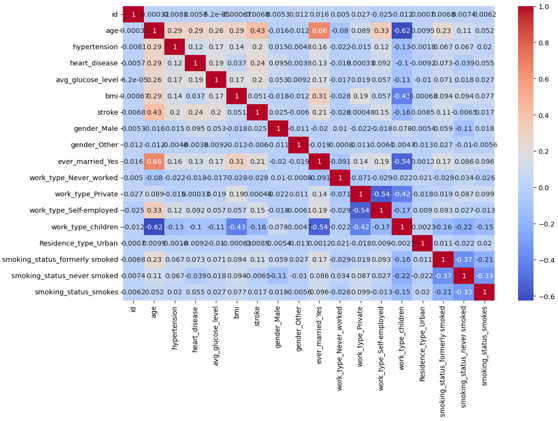


Figure 3: Correlation Matrix

This is the correlation matrix where we can find which features are correlated with the target value so that we can remove some unnecessary features. 'ID', an unnecessary column, was removed. We used 1 hot encoding to encode categorical values like gender, married, work type and smoking status into numerical features. We used Standard scaler to similarly scale all the numerical values, so that the model gives good performance. We have also used Normalization technique so that it falls between same range of 0 to 1. The label encoder was then used to encode all the categorical features. Categories each received a unique integer value in Label Encoding. Next, the dataset was divided into training and testing datasets, with training datasets using 80% of the data and testing datasets using 20% of the data. Finally, normalization methods like standard scalar were used to enhance the model's performance and training stability. The correlation matrix was used to check every feature that correlated to stroke.

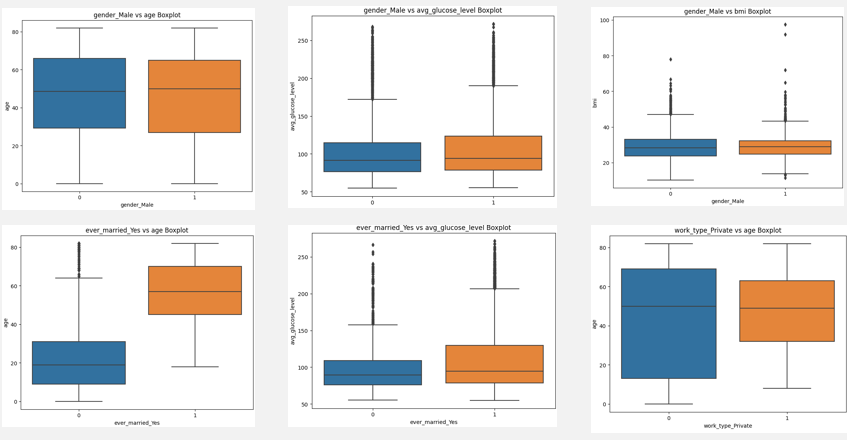


Figure 4: Box Plot for Categorical vs Numerical Features

This box plot shows us that from the age range 20-60, there is little chance of stroke. On the other hand, from the age 60 to upper range, there is high probability of having stroke.

We can see here bmi value higher than 30 have high chances of having stroke.

We have also implemented age vs. Gender, gender vs avg glucose level, married status vs age, work type vs age boxplot.

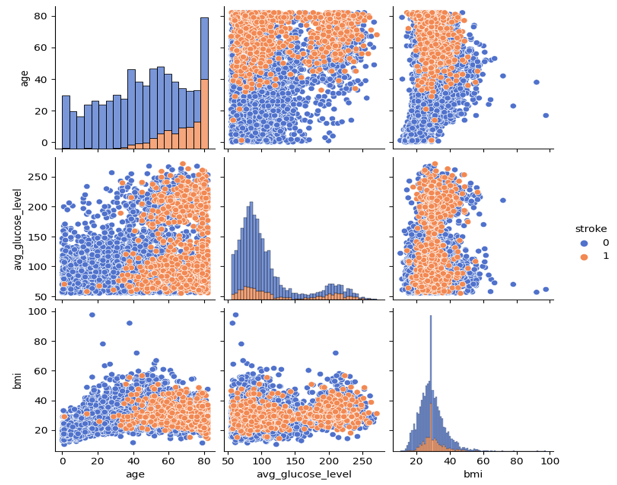


Figure 5: Pair plot for Numerical vs. Numerical Features

We can see from the pair plot that, Older people have chances of increasing avg glucose level and higher BMI, also we can see that average glucose level is higher in people with high BMI values. They also have higher chances of stroke.

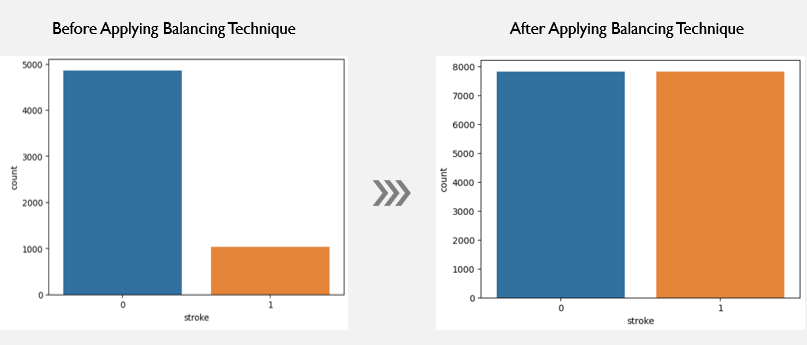


Figure 6: Before and after balancing dataset

We have used SMOTE and oversampling technique for balancing the dataset. SMOTE and oversample the minority class by generating synthetic instances to balance the representation. We also used Oversampling to prevent the model from being biased towards the majority class and classify instances from the minority class.

## Machine Learning Model

We have applied 7 machine learning algorithms, they are Random Forest, decision tree, K Nearest Neighbors, Naive Bayes, SVM, Gradient Boosting, Logistic Regression. For each method, cross validation and hyper parameter tuning were also performed. We have also evaluated model performance by calculating accuracy, precision, recall, F1 score.

1. Random Forest is a classifier that contains several decision trees with bagging on various subsets of the given dataset and improve the predictive accuracy of that dataset.
2. Decision Treeis a flow chart like tree classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
3. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
4. Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
5. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.
6. Gradient Boosting can use a wide range of base learners, such as decision trees, and linear models.
7. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

The training set of data is divided into several smaller sub-tree via Decision Tree approaches. On sets of labeled data, they are trained. It divides each subgroup recursively depending on the values of the following feature. This procedure is repeated until all the data is included in leaf nodes corresponding to the model's predictions.

## WebApp Develop

We have implemented a web app through Flask using API (Application Programming Interface). Flask works a middle worker between web app and machine learning code. In this web app a user will give inputs such as age, BMI, average Glucose level, hypertension, heart disease etc. and click on the submit button. Then the web app will connect with the machine learning code and predict if this patient has a chance of stroke or not based on those data.

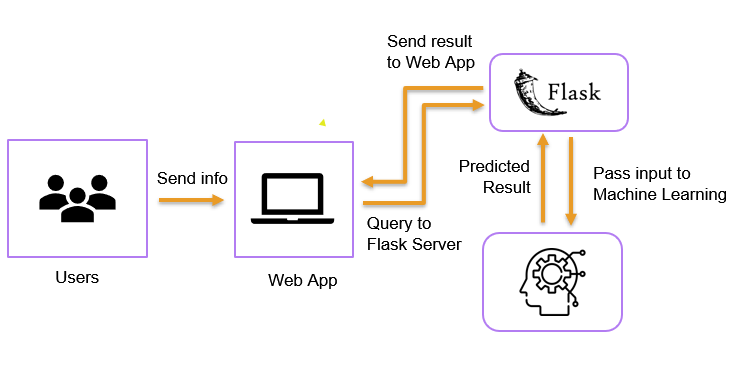


Figure 7: High Level Software Diagram

In this diagram, it is shown that at first user will send the information to the website. Then it will pass the information to flask and ML. After receiving the information, it will predict the result and return the information to the website so that user can see the predicted result.

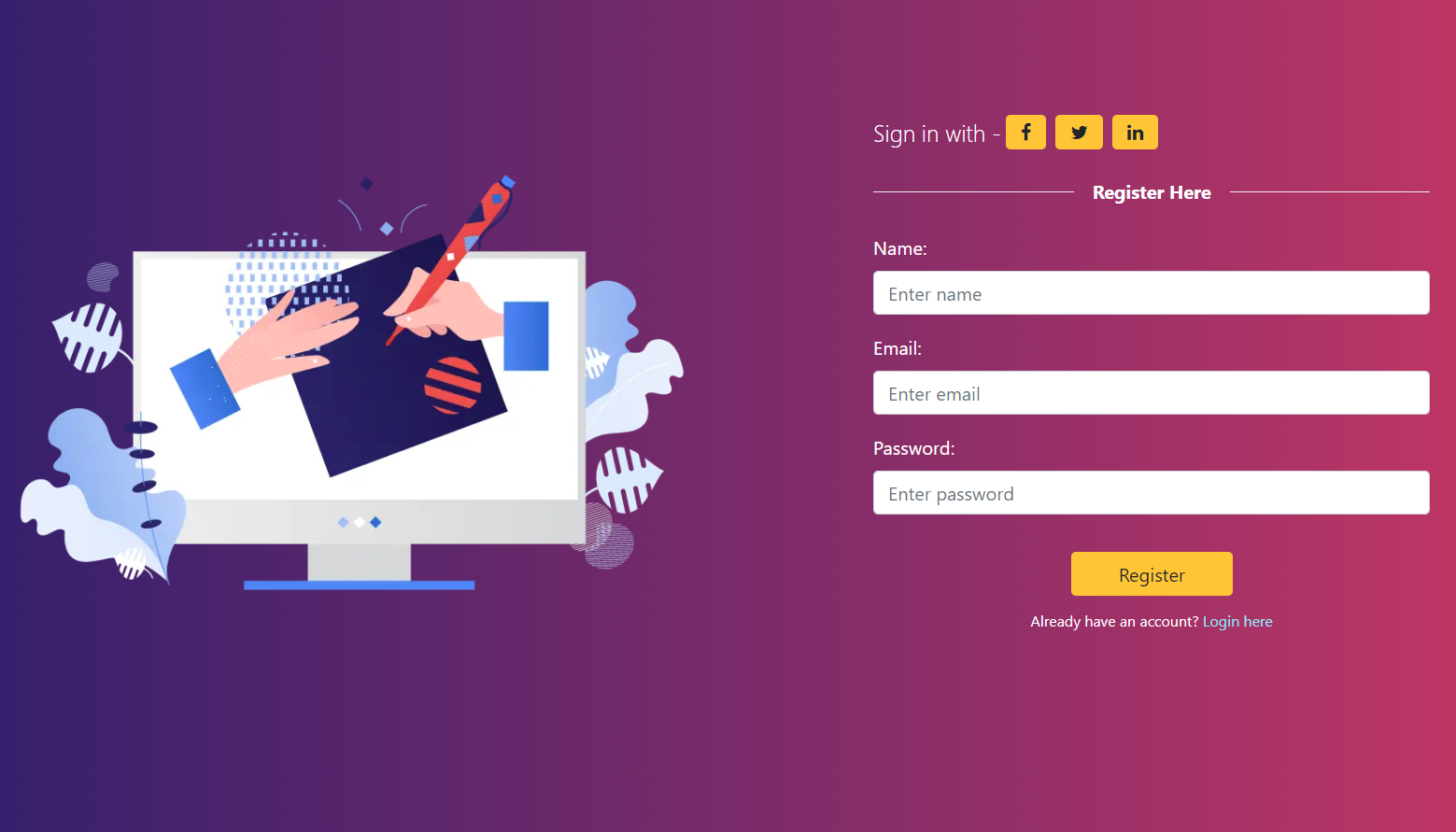


Figure 8: Register Page



Figure 9: Home page

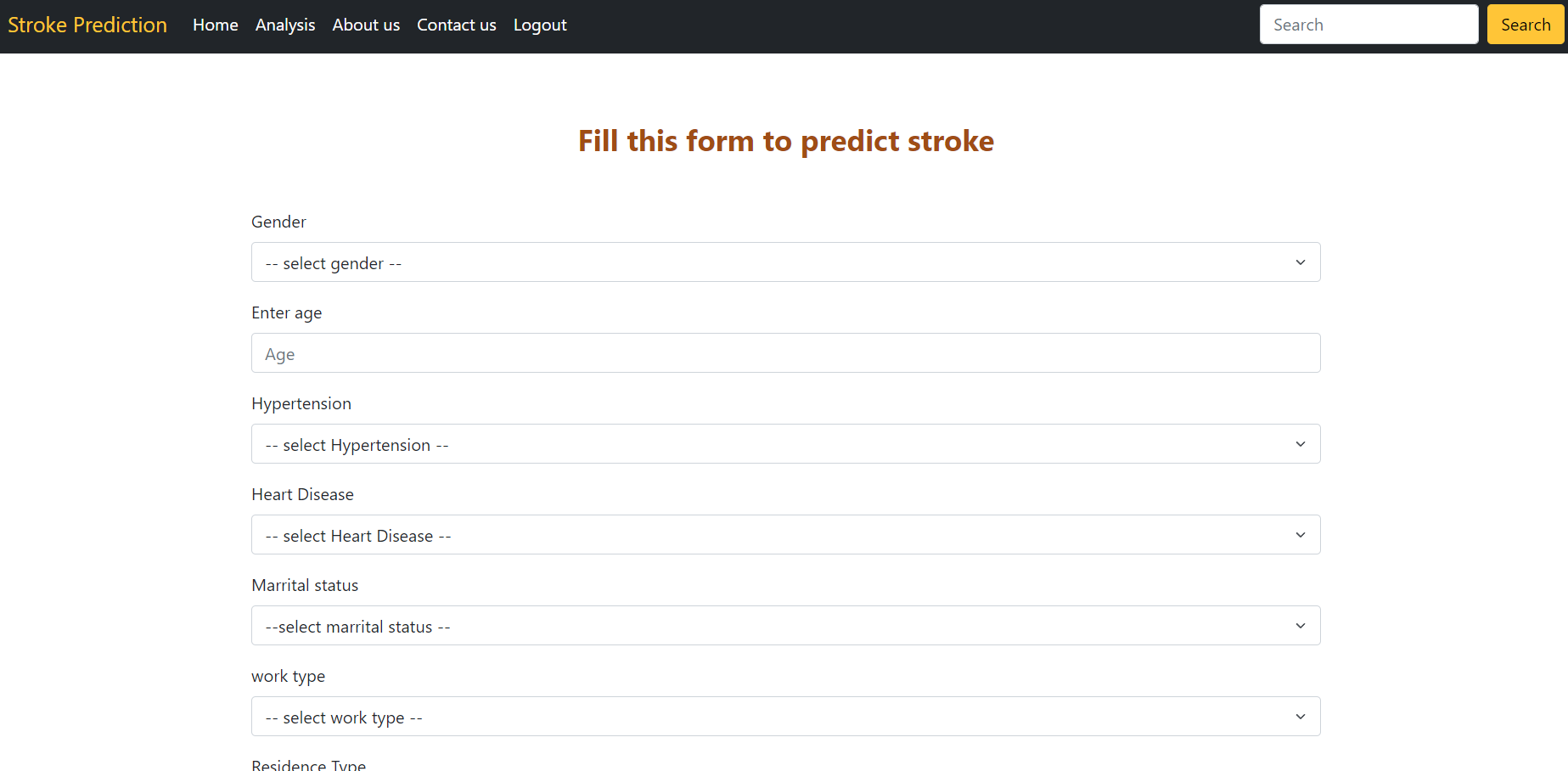


Figure 10: Prediction Page

# Result Analysis & Evaluation

In this section, the outcomes for prediction performance of the proposed ML models are summarized. The performance of the models is evaluated based on accuracy Precision and recall. A measurement of how successfully a machine learning algorithm predicts its outcomes is called its accuracy. Predicting the true positive measurement while dividing the total amount of positive predictions by the true positives is called precision. Recall is an approach of measuring the completeness of positive predictions.

* Accuracy =
* Precision =
* Recall =

We have applied different strategy like smote, under sampling and oversampling individually. But we didn't get expected result, so we applied both SMOTE and oversampling method, then we get highest accuracy of 96% in this project.

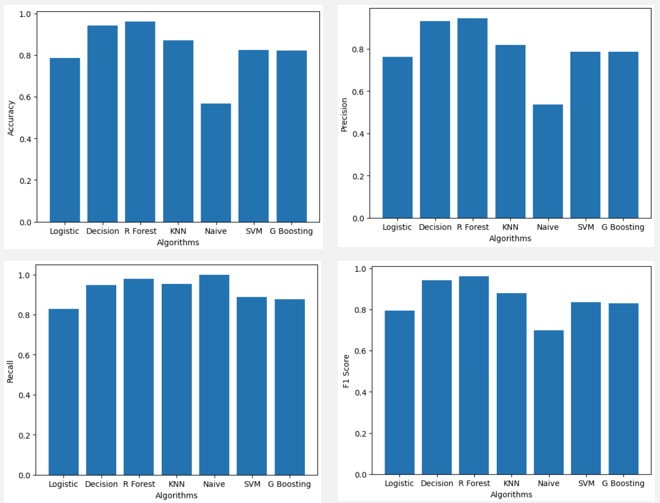


Figure 11: Model Evaluation

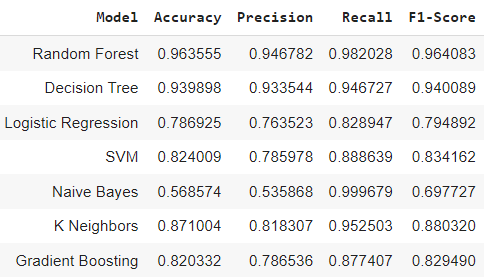


Figure 12: Performance Metrics

This is the diagram where we can see, Random forest is higher in accuracy, precision and f1 score. But Recall is higher in Naive Bayes algorithm.

The resulting ROC curve plot provides a visual representation of the performance of each classifier in terms of the true positive rate vs the false positive rate. Here we can see Random Forest and Decision Tree can perfectly separate the positive and negative classes, achieving 100% Area under the curve.

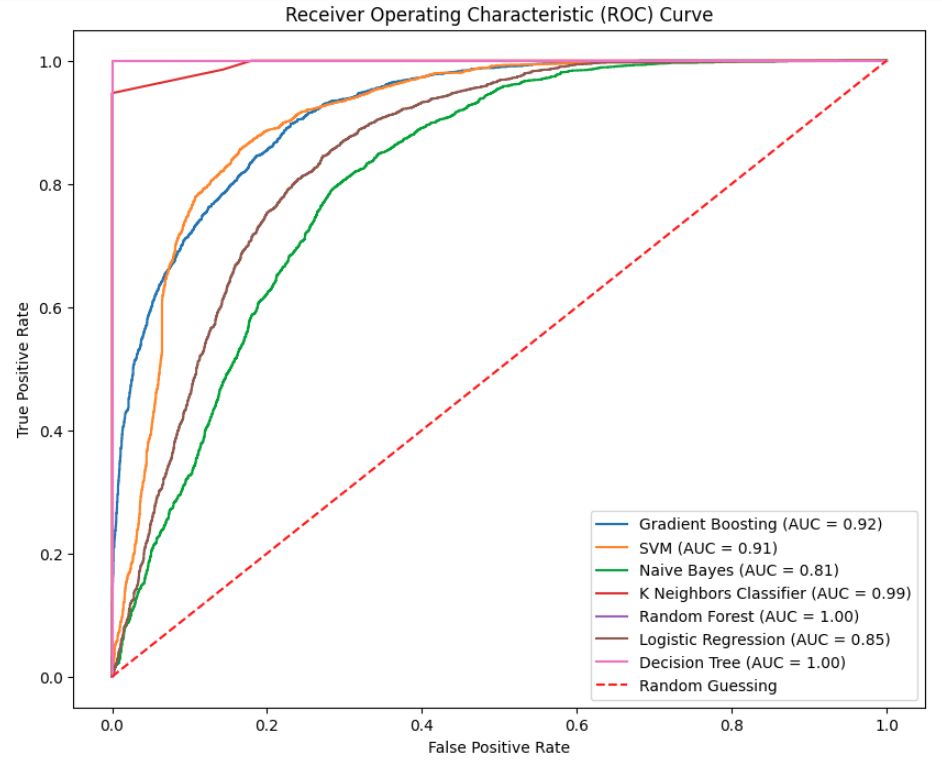


Figure 13: ROC Curve

We have calculated RMSE, MSE and MAE for all algorithms. Decision Tree and Random forest gives the lowest RMSE, MSE, MAE value of 0.04, 0.02 and 0.00 which means predicted values are closer to the actual values and the model makes fewer errors in its predictions.

# Conclusion

We can see from the performance of Machine Learning algorithm is that machine learning approaches can help us in the field of health and disease.

We have used different machine learning algorithms such as Random forest, Decision Tree, SVM, K-Nearest Neighbors, Logistic Regression, Naïve Bayes and Gradient Boosting. From the analysis of result, we have found that Random forest gives the best accuracy of 96% which is highest among all other algorithms. Random Forest is less sensitive to noisy and outlier data points compared to other algorithms. Precision, recall and f1 score is also higher in Random forest. Naive Bayes gives the lowest accuracy of 56% but it gives the highest recall. We have performed lots of Data analysis in this project, we also use feature scaling techniques to scale the features. Also, we have used SMOTE and Oversampling techniques to balance the dataset as our dataset was imbalanced.

In the future, our priority will be adding more stroke dataset from hospitals. We will try to gain more accuracy, precision, recall and f1 score of machine learning algorithms in future. We will try to add more feature in our web app such as patients medical report scanning, giving medicine suggestion based on the prediction result etc. Lastly, we will try to implement web extension with this model instead of the web app.

# References

|  |  |
| --- | --- |
| [1] | "World Stroke Organization," [Online]. Available: https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learn-about-stroke. |
| [2] | T. Elloker and A. Rhoda, "The relationship between social support and participation in stroke: A systematic review.," *Afr. J. Disabil,* pp. 1-9, 2018. |
| [3] | V. S. C. C. C. A. A. C. S. C. S. C. M. D. D. R. Benjamin EJ, "Heart disease and stroke statistics-2018update: a report from the american heart association.," *Circulation,* vol. e67, p. 137(12), 2018. |
| [4] | E. F. J. C. Manisha Sanjay Sirsat, "Machine Learning for Brain Stroke: A Review," *Journal of Stroke and Cerebrovascular Diseases,* vol. 29, 2020. |
| [5] | M. G. Ferdib-Al-Islam, "An Enhanced Stroke Prediction Scheme Using SMOTE and Machine Learning Techniques," *nternational Conference on Computing Communication and Networking Technologies (ICCCNT),* 2021. |
| [6] | A. K. Uttam, "Analysis of Uneven Stroke Prediction Dataset using Machine Learning," *International Conference on Intelligent Computing and Control Systems (ICICCS),* 2022. |
| [7] | D. R. K. D. W. J. M. S. Sujithra, "Applying Machine Learning Techniques for Stroke Prediction in Patients," *International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA) ,* 2021. |
| [8] | S. S. M. K. A. S. Chetan Sharma, "Early Stroke Prediction Using Machine Learning," *International Conference on Decision Aid Sciences and Applications (DASA),* 2022. |
| [9] | R. K. S. A. H. G. R. P. P. J. &. R. M. Priya Govindarajan, "Classification of stroke disease using machine learning algorithms," *Neural Computing and Applications,* vol. 32, 2020. |
| [10] | W. L. L. W. Gang Fang, "A machine learning approach to select features important to stroke prognosis," *Computational Biology and Chemistry,* vol. 88, 2020. |
| [11] | E. D. a. M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," *Sensors,* 2022. |
| [12] | G. G. E. G. S. M. T. T. D. T. K. V. a. N. A. Christos Kokkotis, "Pipeline for Stroke Prediction on Imbalanced Data," *Diagnostics,* 2022. |
| [13] | ,. K. M. M. U. S. T. R. S. K. D. Nitish Biswas, "A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach," *Elsevier,* 2022. |
| [14] | S. G. J. S. A. A. M. M. U. M. M. KRISHNA MRIDHA, "Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study with a Web Application for Early Intervention," *IEEE Access,* pp. 1-1, 2023. |
| [15] | M. S. K. T. I. M. M. M. R. S. A. M. a. M. S. K. Minhaz Uddin Emon, "Performance Analysis of Machine Learning Approaches in Stroke Prediction," *IEEE Xplore,* 2020. |
| [16] | A. R. S. S. R. S. J. B. a. U. S. B. Akter, "A Machine Learning Approach to Detect the Brain Stroke Disease," *4th International Conference on Smart Systems and Inventive Technology (ICSSIT),* 2022. |