# Abstract

I tackled the problem of predicting strokes based on many health factors in this stroke prediction research. To obtain an understanding of the characteristics influencing the risk of stroke, the dataset was examined and purified. I preprocessed the data, and cross-validation was used to assess the models. SMOTE was used to address imbalanced class concerns. Different machine learning models were used in cross validation including Logistic regression, svm, knn and Ensemble ML models. After undergoing hyperparameter adjustment, the best-performing model was found to be XGBoost. On the test set, the final XGBoost model produced an accuracy that was good enough. This study shows how to predict strokes in a comprehensive way by classic machine learning techniques.

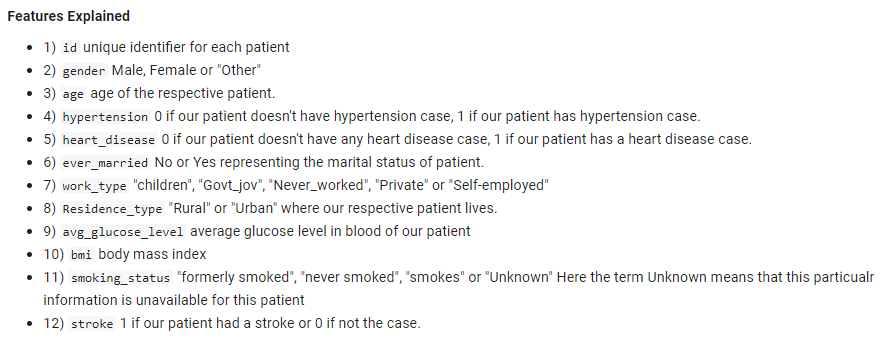
# Introduction

This work uses machine learning to provide early identification and preventive treatments for stroke, a global public health concern. The objective is to create accurate prediction models that consider a variety of variables, including age, average blood sugar, heart disease, and hypertension. The goal of the project is to improve stroke prediction by finding patterns and relationships in a heterogeneous dataset. The impact of strokes can be considerably decreased by early intervention based on these predictive models, which makes this effort essential for enhancing public health outcomes. Detailed feature engineering, data exploration, and machine learning-based model selection are all part of the process. The study's findings demonstrate how machine learning and medical research can be used to detect high-risk patients and improve healthcare.

# Approach

## Data Representation and Description (Data assessing)

A detailed breakdown of the key features in our dataset is provided in the following artifact:



The 5110-row dataset includes 11 input features and the "stroke" target feature, representing personal patient data. With the possible exception of the age variable's translation from float to integer during the data cleaning process, all feature data types are valid and have no bearing on the outcome. Patients range in age from newborns to adults 82 years of age, according to summary statistics. The BMI and average glucose both show significant variances, with standard deviations of 7.85 and 45 mg/dl, respectively. These results highlight the diversity of the dataset. A visual examination of feature distributions is intended for data analysis that comes after.

## Data Cleaning Steps

### Dealing with Missing Values

For machine learning models, handling missing values is essential because some algorithms, such as XGBoost, cannot function with them. In our data set bmi column has missing values. With a threshold of 5% (3.93% values missing in bmi), the decision to remove missing rows is made with the least amount of impact on data analysis and model believability.

### Checking for Duplicate Values

I use a two-step procedure to guarantee dataset integrity and unbiased model performance: removing exact duplicates and looking for possible replication in feature subsets. Notably, there are no duplicate values in our dataset.

### Changing Incorrect Data Types

I have converted the data type of age from float to integer for better performance in algorithm. It would be more beneficial if our data size was large.

### Removing Unnecessary Features

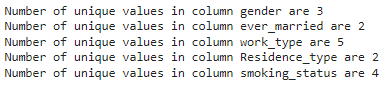
I removed the id column from our dataset as it was not serving any purpose either in data analysis or in ML models building. As it is understood that id column is simply for indexing, and it has nothing to do with predicting strokes cases.

### Organizing Object and Numerical Columns

I divide the object and numerical columns into lists to enable smooth operations during pre-processing. This makes it possible to apply particular operations that are appropriate for each type of data efficiently.

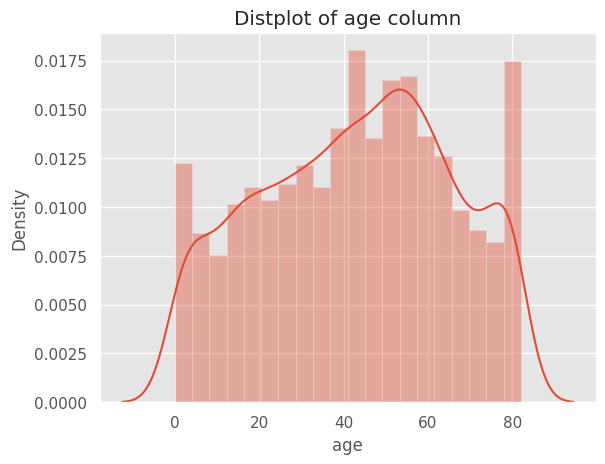
### Assessing Unique Values of Object Data Types:

Recognizing the distinct values included in categorical features is essential for evaluating the quality of the data as well as for upcoming feature engineering. I carefully examine the distinct values of object data types. It is given as artifact below:



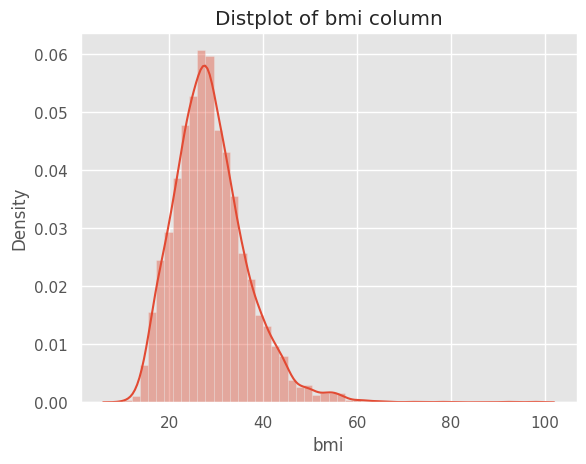
### Dealing with Outliers

The age distribution looks nearly normal with no outliers when seen using a boxplot and histogram. The right skewness of average glucose level and BMI indicates a large number of outliers, mostly in the higher value range. Since these outliers contain important information, it is imperative to keep them around, especially when dealing with obese and diabetic individuals whose elevated BMI readings and blood glucose levels are normal. It is therefore judged unnecessary to remove these data points from the data frame. Graphs are given below:



A graph of blood glucose level

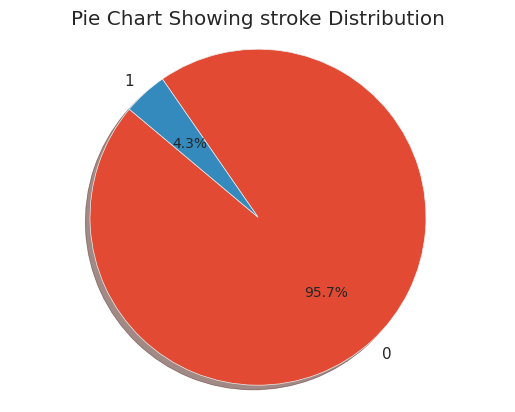
Description automatically generated



## Data Analysis & Visualization

### Univariate analysis

* There is a clear case of class imbalance with respect to target variable. Thats why will use SMOTE to handle the class imbalance.



* The Female and married are slightly more as compared to male and single.
* Most of the patients belong to the work type of Private as compared to other work categories.
* The rural and urban population is almost balanced.
* The individuals having heart disease and hypertension are predominately present in the data as compared to counter category.
* From smoking habits, the nonsmokers who never smokes are present more as compared to other available categories.

### Bivariate and Multivariate Analysis with respect to stroke feature

* It is visible from gender proportion data frame that male is slightly more (5.1%) at the risk of stroke as compared to female (4.7%)

A graph of a graph with text

Description automatically generated with medium confidence

* It is visible from hypertension proportion data frame that hypertensive patients are slightly more at the risk of stroke (13.2%) as compared to non-hypertensive (3.9%)

A graph of stroke by hypertension

Description automatically generated

* It is visible from heart disease proportion data frame that heart disease patients are slightly more at the risk of stroke (17%) as compared to non-heart disease patients (3%)

A graph with a bar and a number of orange squares

Description automatically generated with medium confidence

* It is visible from ever married proportion data frame that married patients are slightly more at the risk of stroke (6%) as compared to non-heart disease patients (1%)

A graph with orange squares

Description automatically generated

* It is visible from work type proportion data frame that self-employed individuals are slightly more at the risk of stroke (7%) as compared to other work type individuals. Another higher proportion include government jobs and private individuals (4.5%). The unusual impact is of non-workers which is quite contradictory to medical sciences evidence.

A graph of different colored bars

Description automatically generated

* From Residence type proportion data frame, it can be concluded that both rural and urban areas residencies have almost same impact on stroke.

A graph of a graph with orange squares

Description automatically generated with medium confidence

* It is visible from smoking status proportion data frame that smokers and formerly smoker patients are slightly more at risk of stroke (6%) as compared to nonsmoker patients (1%).

A graph of smoking status

Description automatically generated

* age clearly indicates that stroke patients are mostly aged persons.

A chart with a graph

Description automatically generated with medium confidence

* In avg glucose suggests plot I can see that stroke patients have mostly high levels of blood glucose suggesting diabetes leading to stroke.

A chart with a graph of different colored squares

Description automatically generated with medium confidence

* In bmi boxplot I can see that the inter quartile range of stroke patients is slightly higher suggesting obesity leads to stroke.

A graph of a graph showing different colored squares

Description automatically generated with medium confidence

## Data Preprocessing for Machine Learning

### Label Encoding

The dataset's categorical columns were labelled in order to make the machine learning techniques more compatible with non-numeric data. For such columns the Label Encoder was used to make sure these features were integrated into the model without any problems.

### Making Features and Target

After that, the dataset was divided into the target variable (y) and features (X). The feature set (X) consisted of the remaining features after the target variable, "stroke" was isolated.

### Splitting Data into Training and Test Set

As part of the model construction process, splitting the dataset into training and test sets was a crucial step. An 80-20 split ratio was used to divide the data, setting aside 20% for testing and 80% for training.

### Scaling

Standardization was used to guarantee consistency across the feature scales and stop any one characteristic from predominating over the others. Using StandardScaler, the features were scaled independently for the test and training sets.

### Handling Imbalanced Data Using SMOTE

Due to the target variable's imbalance only 4.3% of cases are labelled as 0. Conventional machine learning methods may show bias in favor of the majority class. The Synthetic Minority Over-Sampling Technique (SMOTE) was used to correct for this imbalance. By oversampling the minority class, this method produced a training set distribution of both groups that was more evenly distributed. For computational efficiency, the resampled data was then down sampled, producing a final training dataset (X train, y train) that was prepared for model training.

# Machine Learning Model Evaluation and Testing

A wide range of machine learning models were used in the search for an efficient stroke prediction model. The distinct advantages and traits that every model offers enhance the solution's overall resilience. Among the models utilized in this project are:

* 1. K-Nearest Neighbors
  2. Decision Tree Classifier
  3. Random Forest Classifier
  4. Gradient Boosting Classifier
  5. Support Vector Classifier
  6. Logistic Regression
  7. XGBClassifier
  8. Ada Boost
  9. Stochastic Gradient
  10. Bagging Classifier

## Evaluation Metrics

Three key evaluation metrics were selected to assess the performance of each model:

### Accuracy Score

A model's overall correctness is gauged by its accuracy, which is determined by dividing the number of properly predicted occurrences by the total number of examples in the dataset.

### Classification Report

A comprehensive evaluation of a classification model's performance, including measures such as support for each class, recall, F1 score, and precision, is provided via a classification report.

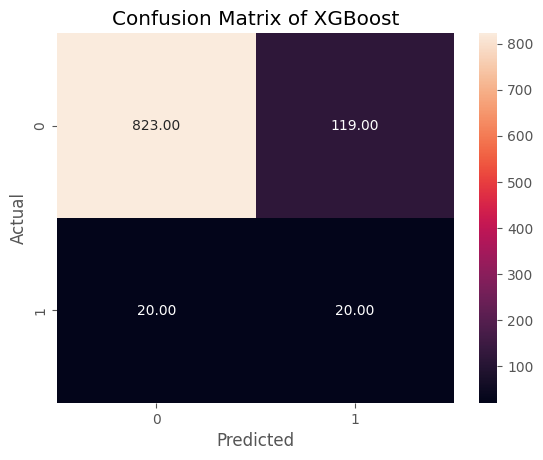
**1. Precision:** Proportion of correctly identified positive instances out of all instances predicted as positive by model.

**2. Recall:** Proportion of positive instances correctly identified by the model.

**3. F1 Score:** The harmonic means of precision and recall.

### c. Confusion Matrix

For every class, a thorough breakdown of true positives, false positives, true negatives, and false negatives is provided by the confusion matrix. To help with focused changes, this matrix provides insights into the areas of the model where misunderstanding and misclassification occur.



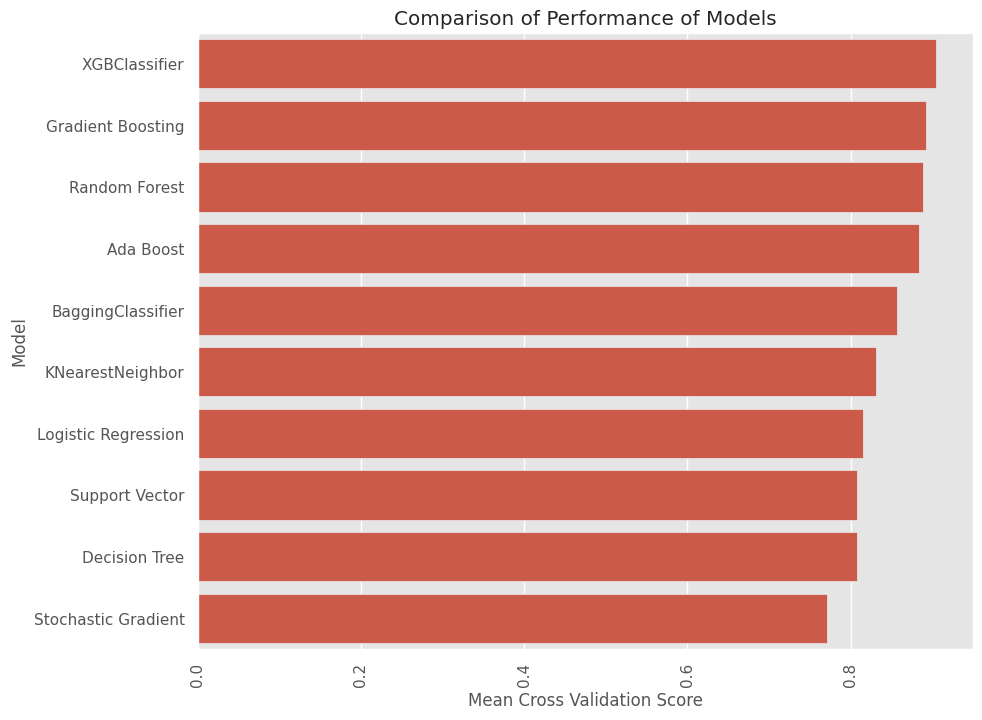
## Model Selection Process

Selecting the right machine learning model requires methodical thought. Following industry best practices, the dataset was subjected to a range of algorithms with their default hyperparameters. Five folds of cross-validation were used to evaluate each model's performance. A more thorough hyperparameter tweaking using grid search CV was subsequently applied to the model or models demonstrating the best performance.

Reference: Hands on Machine Learning" by Oreil'y.

## Machine Learning Models Evaluation

Each model's mean performance over several folds was determined by computing the cross-validation scores. The models are compared and visualized using a bar chart, which makes it easier to determine which model performs best. This examination led to the conclusion that the XGBoost model was the most promising option.



## Best Selected Model

Our best model is the XGBoost Model, then I perform its hyperparameter tuning using the Randomized search CV to see if its performance can be enhanced or not.

### Hyperparameters Tuning using Randomized Search CV

The Accuracy score after performing hyperparameters tuning using Randomized search CV is slightly better than our default XGBoost model. So, I selected this hyper tuned model for our final testing on test dataset. Initially it was 0.9042. After tuning the results become 0.9122.

### Testing Final Selected Model on Test Set (XGBoost) using best Hyperparameters

Following hyperparameter adjustment, the testing set was subjected to the XGBoost model. The performance of the system on data that had not been seen before was assessed using the accuracy score (0.8584), classification report, and confusion matrix.

# Conclusion

The Stroke cases detection project demonstrates how well XGBoost performs by utilizing thorough preparation procedures including feature selection and handling skewed data. Although many machine learning models were assessed, additional improvements can be made by adjusting hyperparameters and investigating cutting-edge methods. Subsequent advancements could encompass advanced algorithms for detecting anomalies, investigating deep learning, and cooperating to address stroke detection issues in real-time.

# References

* https://towardsdatascience.com/practical-implementation-of-outlier-detection-in-python-90680453b3ce
* https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
* https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/