Team Sage

Brain Stroke Prediction

Jerrick Gerald

Renganathan Laxmanan

Sai Dontukurti

Sowmya Maddali

Why brain stroke?

- In the US, stroke is one of the main causes of mortality and disability. A stroke may happen to anybody, regardless of age or background.
- Approximately 795,000 individuals in the US suffer strokes each year, and 137,000 of them face death, according to the National Institutes of Health (NIH), US Department of Health and Human Services.
- According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

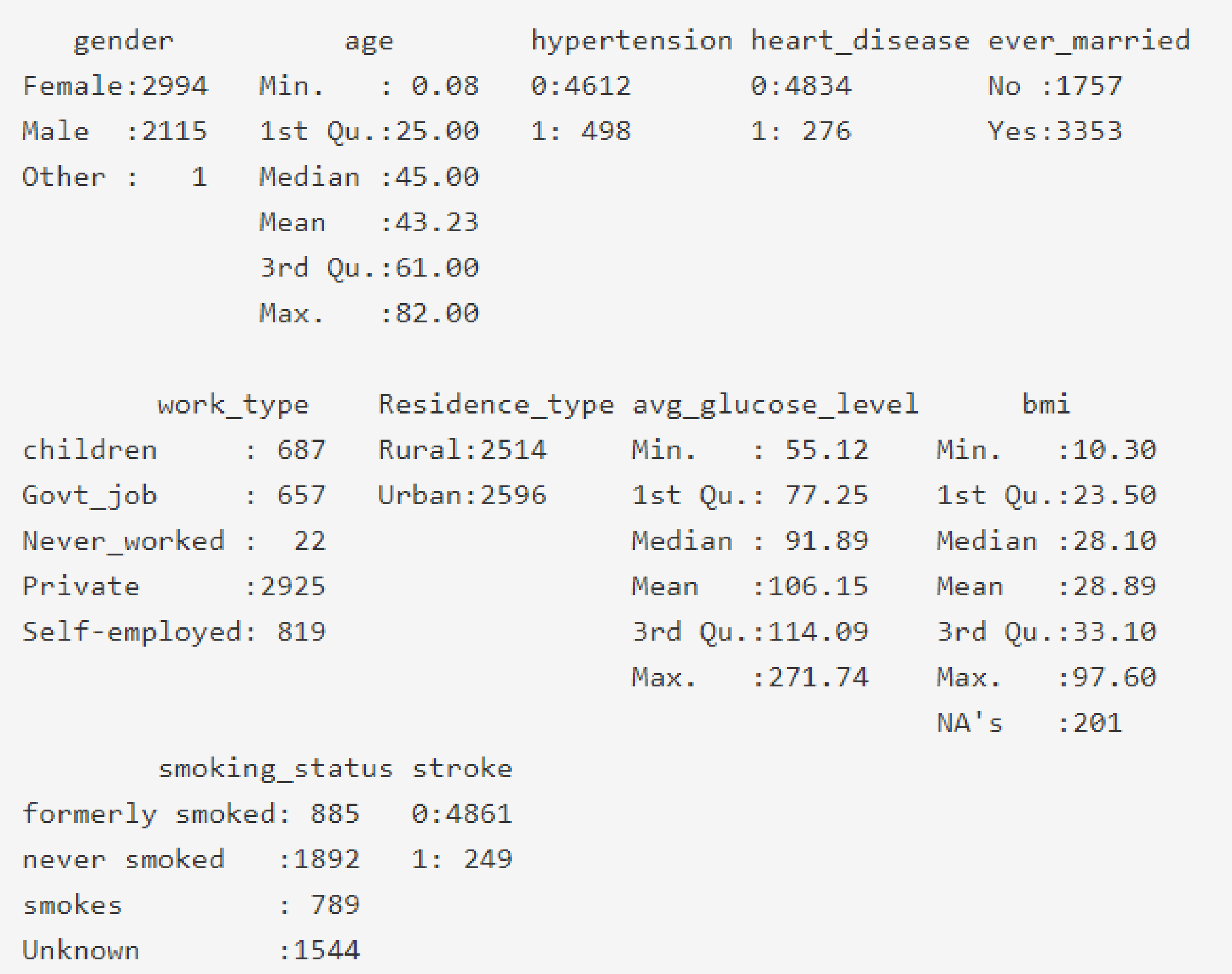
- Ref: Eunice Kennedy Shriver National Institute of Child Health and Human Development

About the dataset

- The dataset has been taken from Kaggle under the name Stroke Prediction dataset
- It is used to predict whether a patient is likely to get a stroke based on the input parameters like gender, age, BMI, hypertension, average glucose level, marital status, work type, residence type, various diseases, and smoking status.
- Each row in the dataset provides relevant information about the patient.

Gender	**Age**	**Heart_disease**	**BMI**	**Hypertension**
Avg Glucose Level	**Ever_married**	**Work_Type**	**Residence_Type**	**Stroke - Target**

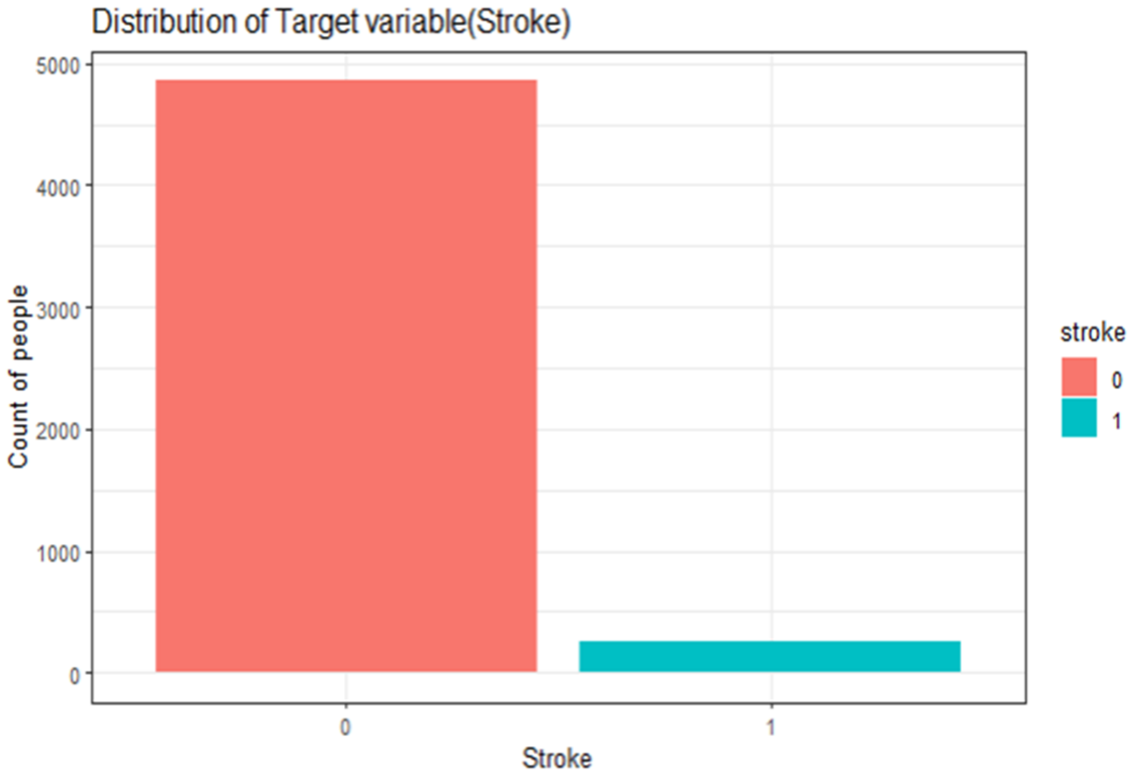
Summary of the dataset



```
   gender          age        hypertension heart_disease ever_married
Female:2994   Min.   : 0.08   0:4612       0:4834        No :1757
Male  :2115   1st Qu.:25.00   1: 498       1: 276        Yes:3353
Other :   1   Median :45.00
              Mean   :43.23
              3rd Qu.:61.00
              Max.   :82.00

        work_type    Residence_type avg_glucose_level      bmi
children     : 687   Rural:2514     Min.   : 55.12    Min.   :10.30
Govt_job     : 657   Urban:2596     1st Qu.: 77.25    1st Qu.:23.50
Never_worked :  22                  Median : 91.89    Median :28.10
Private      :2925                  Mean   :106.15    Mean   :28.89
Self-employed: 819                  3rd Qu.:114.09    3rd Qu.:33.10
                                    Max.   :271.74    Max.   :97.60
                                                      NA's   :201
        smoking_status stroke
formerly smoked: 885   0:4861
never smoked   :1892   1: 249
smokes         : 789
Unknown        :1544
```

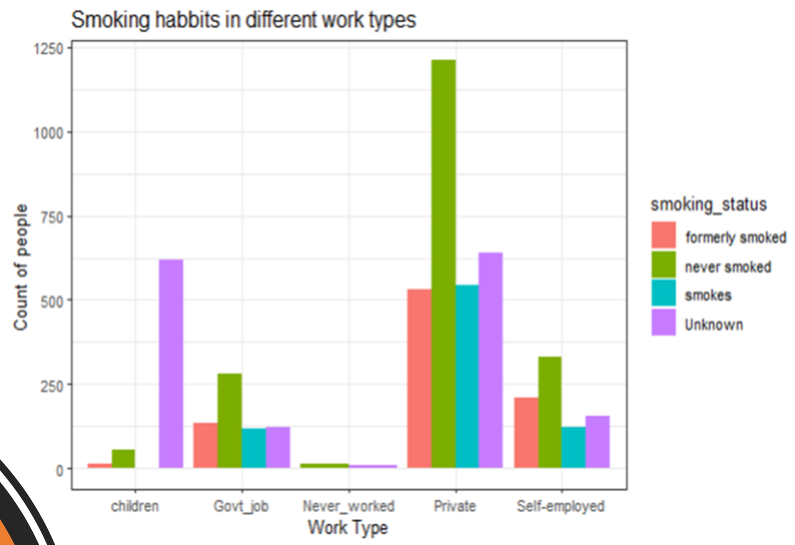
Distribution of the target variable

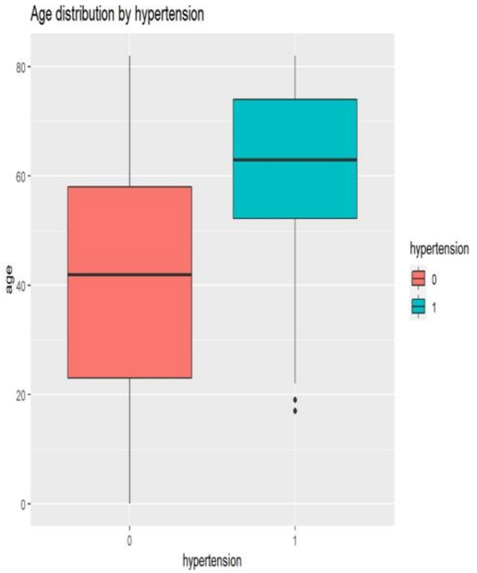


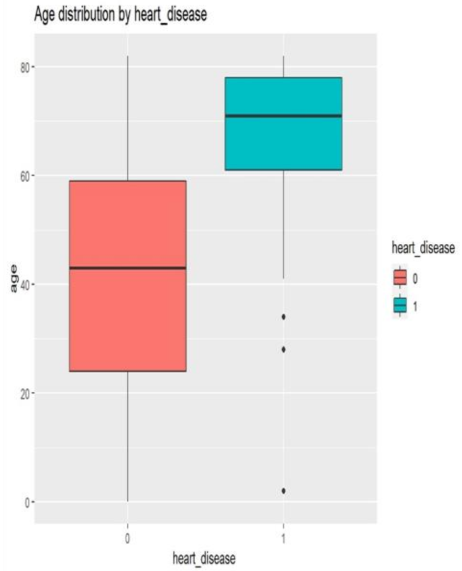
Data Pre-Processing

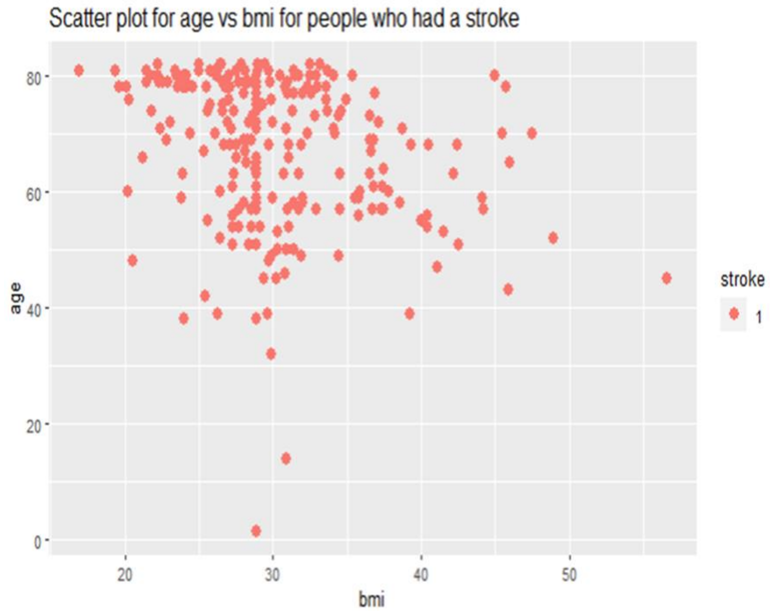
- Converting the columns into their respective data types.
- NA removal from BMI.
- Subsetting of dataset purely for Data Analysis.

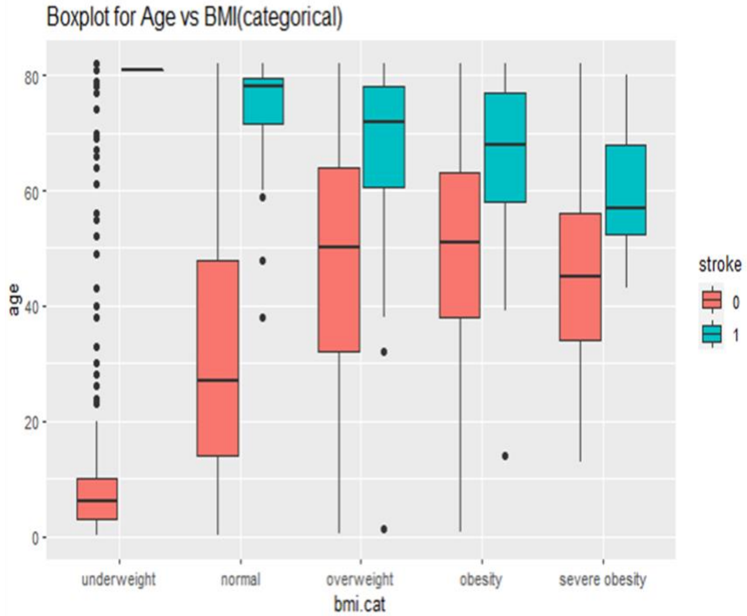
Basic EDA


Smoking habbits in different work types
Count of people
1250
1000
750
500
250
0
children
Govt_job
Never_worked
Private
Self-employed
Work Type
smoking_status
formerly smoked
never smoked
smokes
Unknown

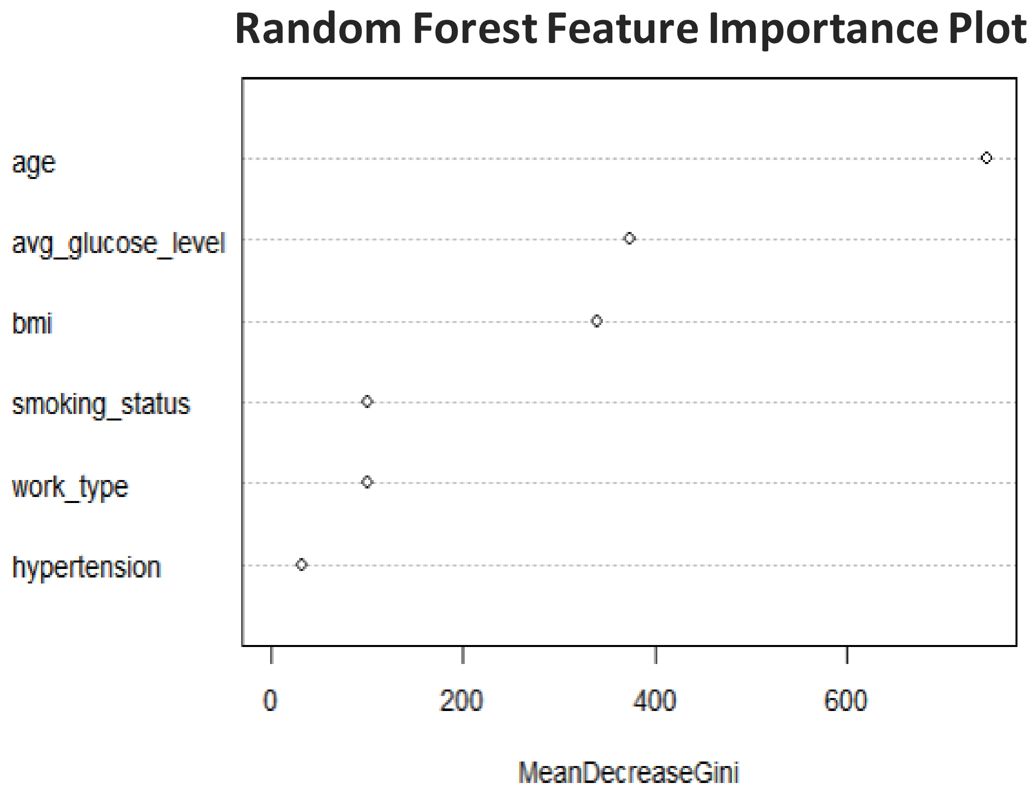

Age distribution by hypertension
age
80
60
40
20
0
0
1
hypertension
hypertension
0
1


Age distribution by heart_disease
age
80
60
40
20
0
0
1
heart_disease
heart_disease
0
1


Scatter plot for age vs bmi for people who had a stroke
age
80
60
40
20
0
20
30
40
50
bmi
stroke
1


Boxplot for Age vs BMI(categorical)
age
80
60
40
20
0
underweight
normal
overweight
obesity
severe obesity
bmi.cat
stroke
0
1

Feature Selection



SMART: How to address the data imbalance issue in the dataset?

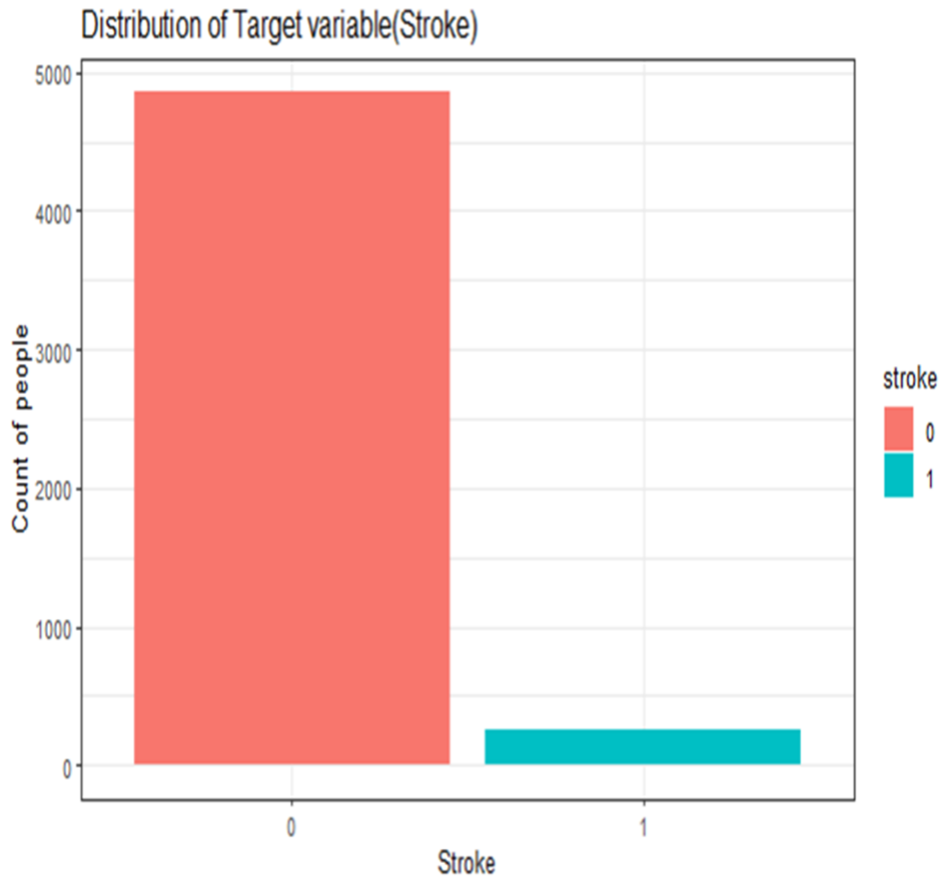
Data distribution is **96% of no stroke** and **4% of stroke**.

Balancing Techniques:

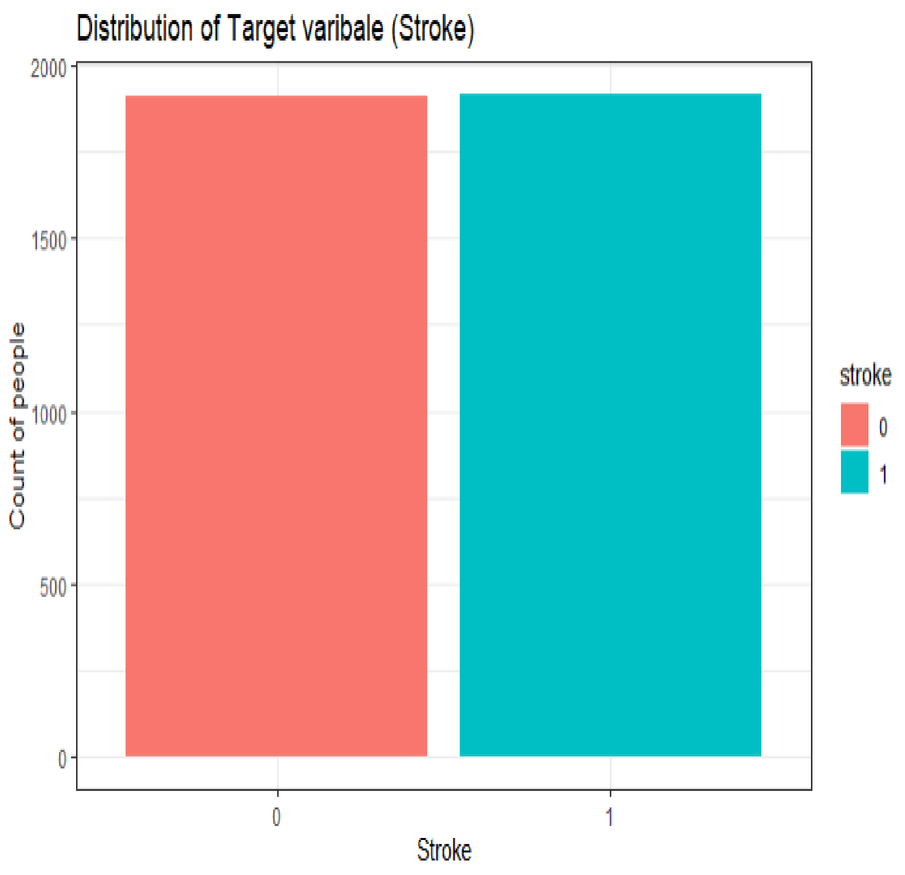
- **ROSE** – Random Over Sampling Examples
- **BOTH** – Under and Over Sampling

Data Balancing

BEFORE BALANCING



AFTER BALANCING



Model Building

Model Techniques:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier

SMART: What evaluation metrics can be used to find the best model and why?

Evaluation Metrics:

Metrics to be considered:

- **TN – True Negative** : Patient has no stroke and model classified as having no stroke.
- **TP – True Positive** : Patient has a stroke and the model classifies it as a stroke.
- **FP – False Positive** : A Patient who does not have stroke is classified as a patient having a stroke.
- **FN – False Negative** : A Patient who has a stroke but is classified as no stroke.

Recall(IMPORTANT), F-1 score, and Accuracy is considered for metrics.

Why Recall ?

- **Recall** is given more preference than precision because of two drawbacks,
  1. **Case 1**: The model predicts the patient has no stroke, but the has stroke (FN).
  2. **Case 2**: The model predicts the patient has stroke, but the patient has no stroke(FP).

Modeling on Balanced Dataset

ROSE Technique

MODELS	RECALL	ACCURACY
LOGISTIC REGRESSION	0.80	0.74
DECISION TREE	0.796	0.734
DECISION TREE-TUNED	0.796	0.736
RANDOM FOREST	0.52	0.78

Modeling on Balanced Data with Feature Selection

ROSE Technique

MODELS	RECALL	ACCURACY
LOGISTIC REGRESSION	0.87	0.74
DECISION TREE	0.79	0.73
DECISION TREE - TUNED	0.79	0.73
RANDOM FOREST	0.64	0.75

Modeling on Balanced Data

BOTH (Under And Over) Sampling

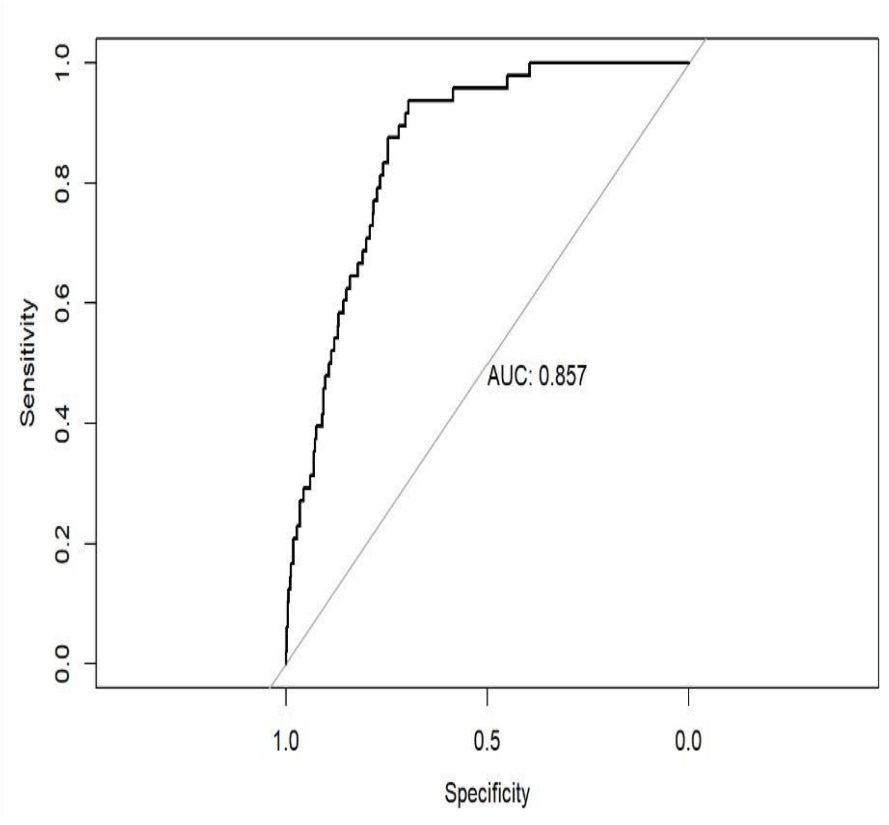
MODELS	RECALL	ACCURACY
LOGISTIC REGRESSION	**0.84**	0.74
DECISION TREE	0.796	0.734
DECISION TREE- TUNED	0.796	0.736
RANDOM FOREST	**0.52**	0.78

Modeling on Balanced Data with Feature Selection

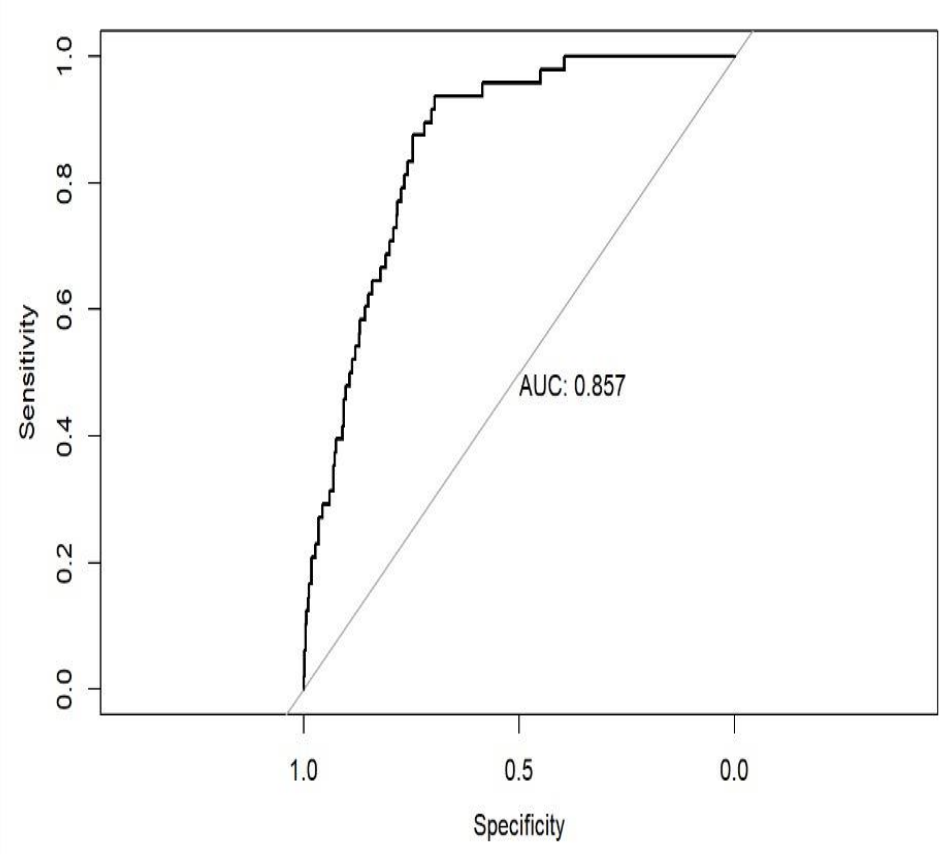
BOTH (Under And Over) Technique

MODELS	RECALL	ACCURACY
LOGISTIC REGRESSION	0.87	0.73
DECISION TREE	0.84	0.70
DECISION TREE- TUNED	0.68	0.75
RANDOM FOREST	0.31	0.90

ROC CURVE OF THE BEST MODEL WITH FEATURE SELECTION

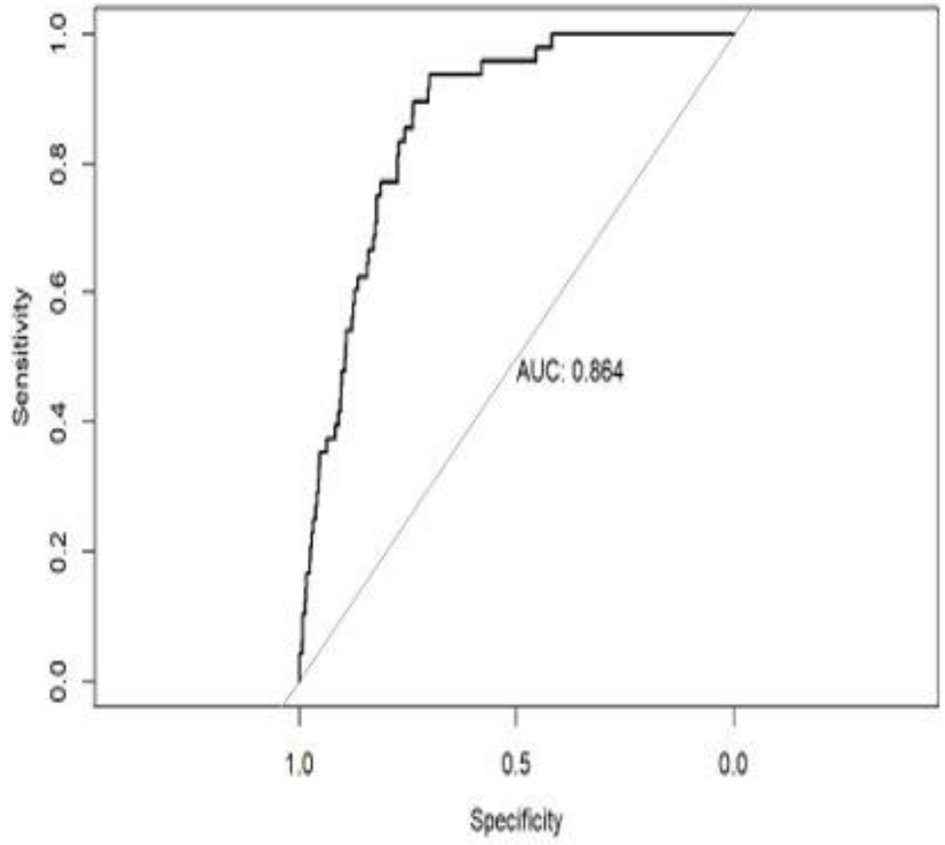


Logistic with feature selection- ROSE

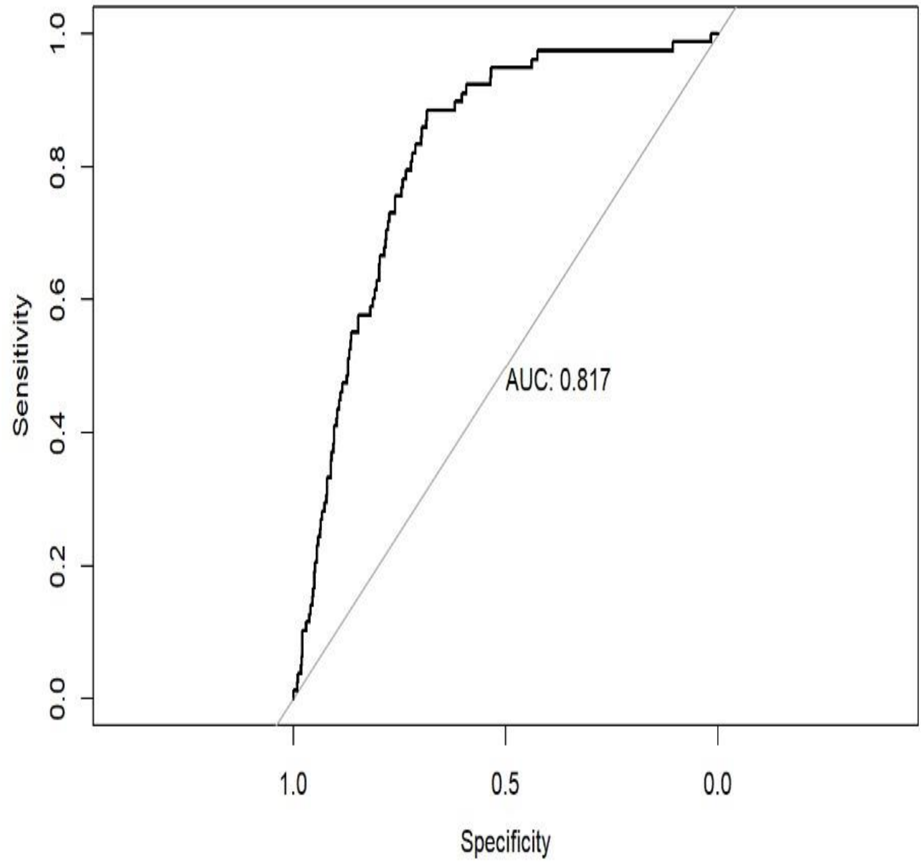


Logistic Regression- BOTH (Under & Over sampling)

ROC CURVE OF THE BEST MODEL WITHOUT FEATURE SELECTION



Logistic Regression - ROSE



Logistic Regression – Under & Over (BOTH) sampling

SMART: What is the best machine learning model that predict the likelihood of a stroke?

MODELS	RECALL	AUC Score	ACCURACY
Logistic Regression with feature selection using BOTH technique (Under and Over Sampling)	0.87	0.85	0.73

CONCLUSION

- The main factors that we used to predict the likelihood of a stroke are age, average glucose level, BMI, smoking status, work type, and hypertension.

- The data imbalance is addressed by using **BOTH (Under and Oversampling** Technique) & **ROSE** – Random Over sampling Examples.

- **Recall** is mainly used as an evaluation metric to find the best model.

- **Logistic Regression with feature selection using BOTH** (Under and Over sampling) technique performs the best.

THANK YOU