Section 1 Introduction

The data is of stroke dataset. Data is generated from critical trails and hospital data. There is missing data in the dataset. Dataset has a high imbalance for stroke. These are detailed below in input and output data
Data is upsampled before classification as target has imbalance distribution and Missing value is predicted

Section 2.1 :Input and Output Data

Input:

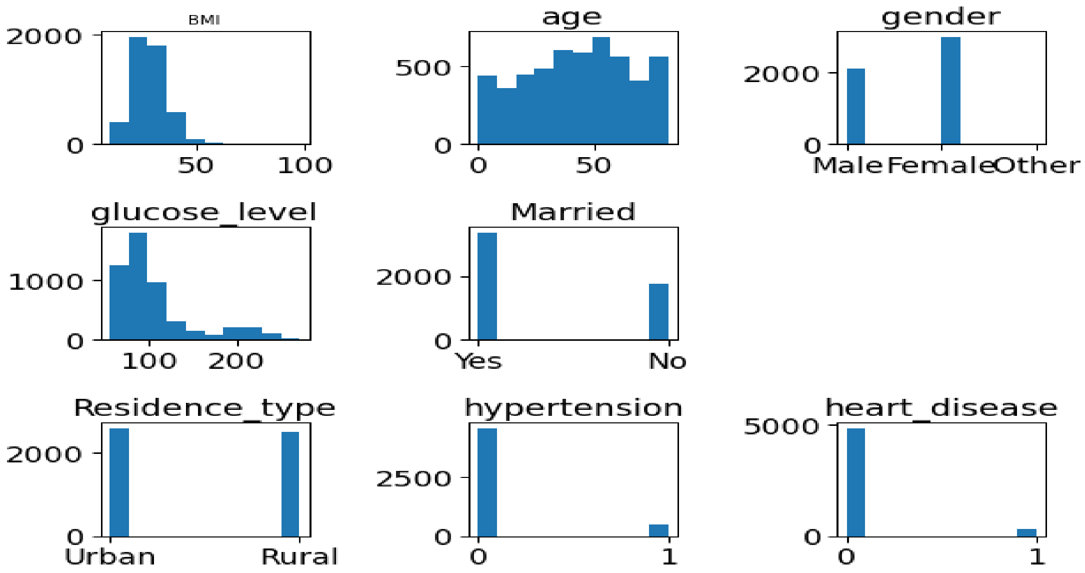
1. Gender={Male,Female}={0,1}
2. Age=0-80
3. hypertension={0,1}
4. Heart Disease={0,1}
5. avg_Glucose_level=55-271
6. bmi=10-100
7. Work_type={'Private':0,'Self-employed':1,'Govt_job':2,'children':-1,'Never_worked':-2}
8. Residence =Rural,Urban
9. Smoking_Status={Never Smoked,Formerly Smoked.Unknown,Smokes}
10. Ever Married ={Yes,No}

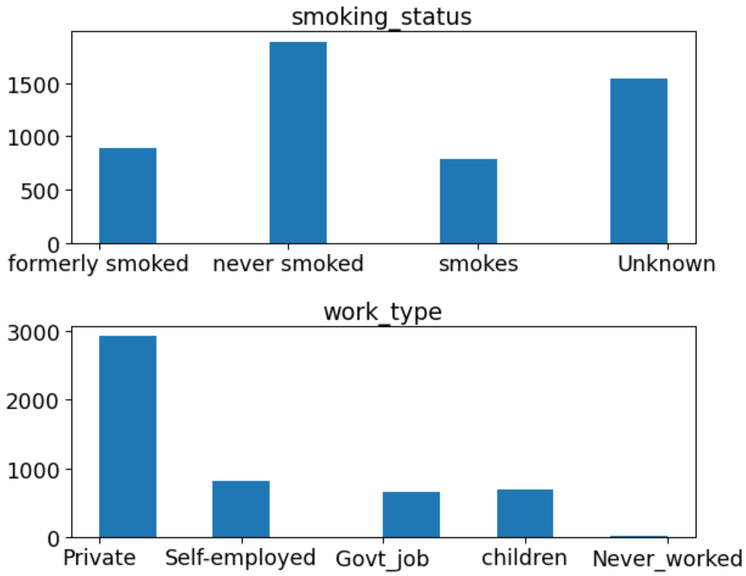
Output

Stroke: Binary 1-Yes 0-No

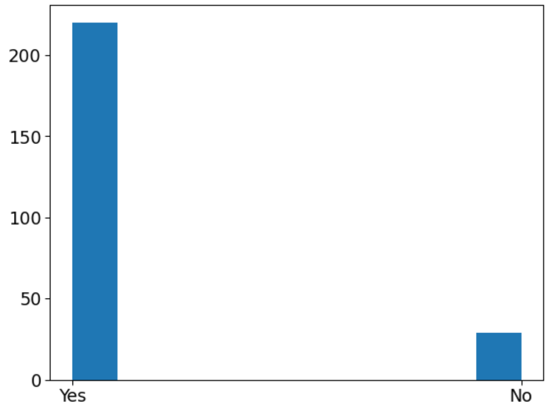
Section 2.2: Data Visualisation

1. Overall Population Distribution

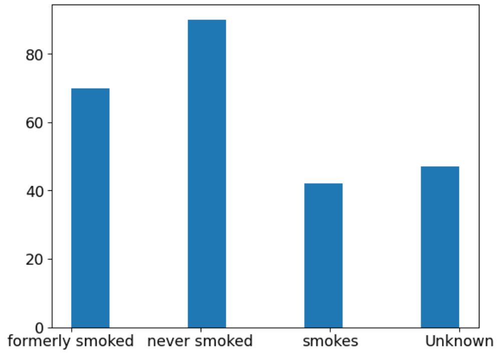

BMI
2000
0
50
100
age
500
0
0
50
gender
2000
0
Male
Female
Other
glucose_level
1000
0
100
200
Married
2000
0
Yes
No
Residence_type
2000
0
Urban
Rural
hypertension
2500
0
0
1
heart_disease
5000
0
0
1



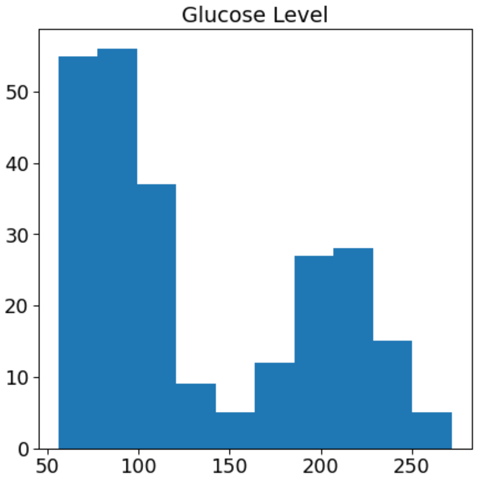
2. Stroke Population
    a. Marriage Status



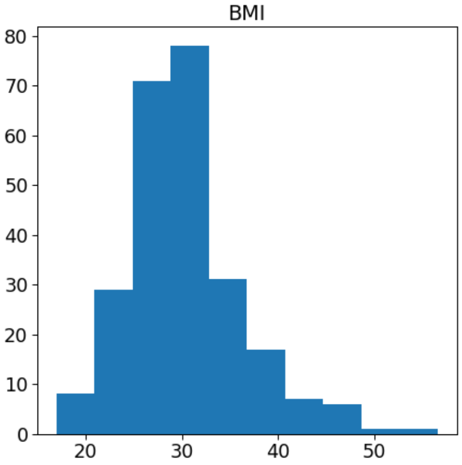
    b. Smoking Status



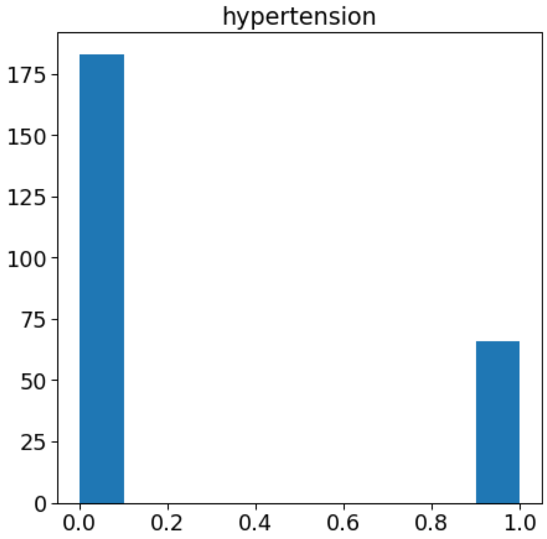
c. Average Glucose Level



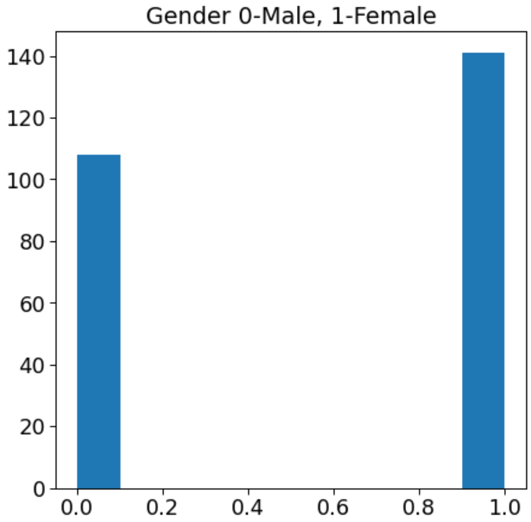
d. BMI



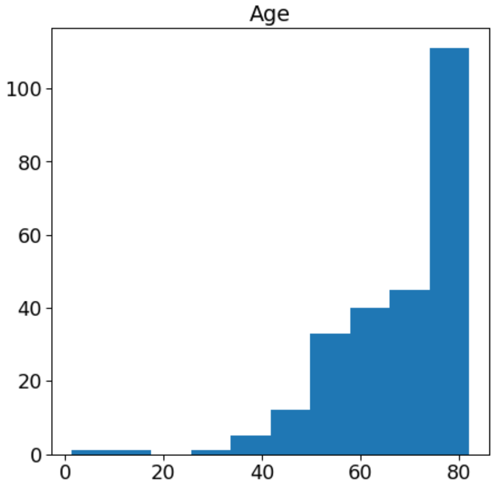
e. Hypertension



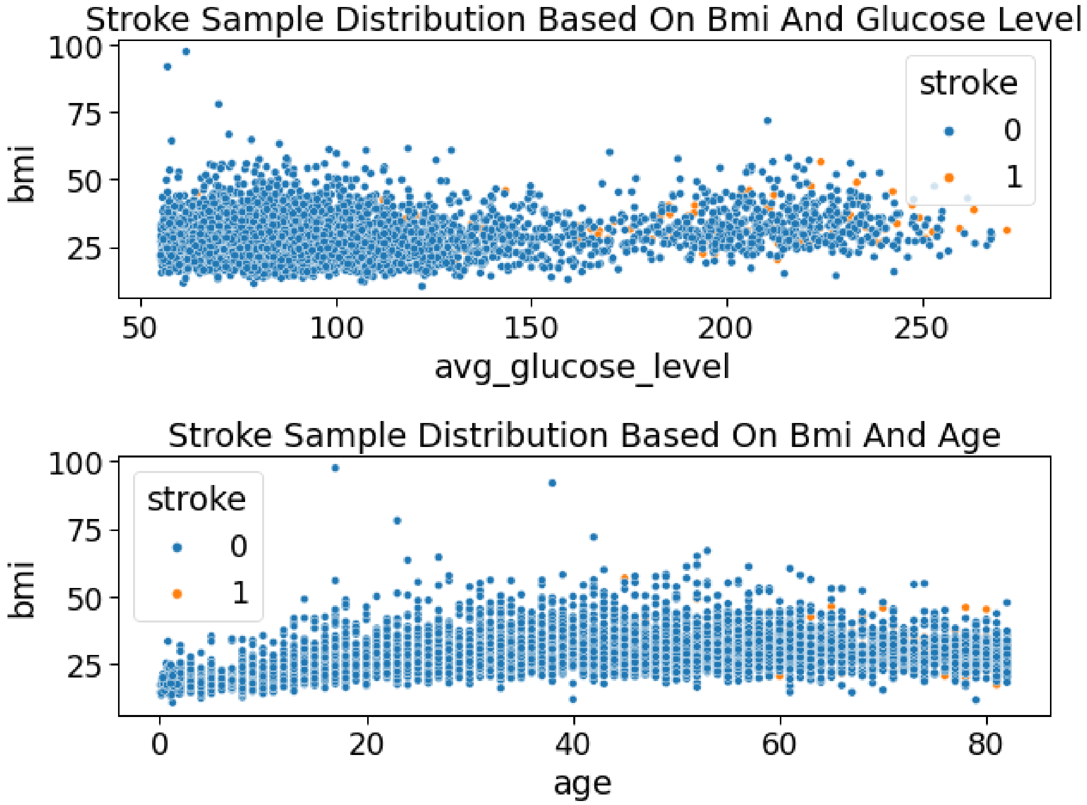
f. Gender



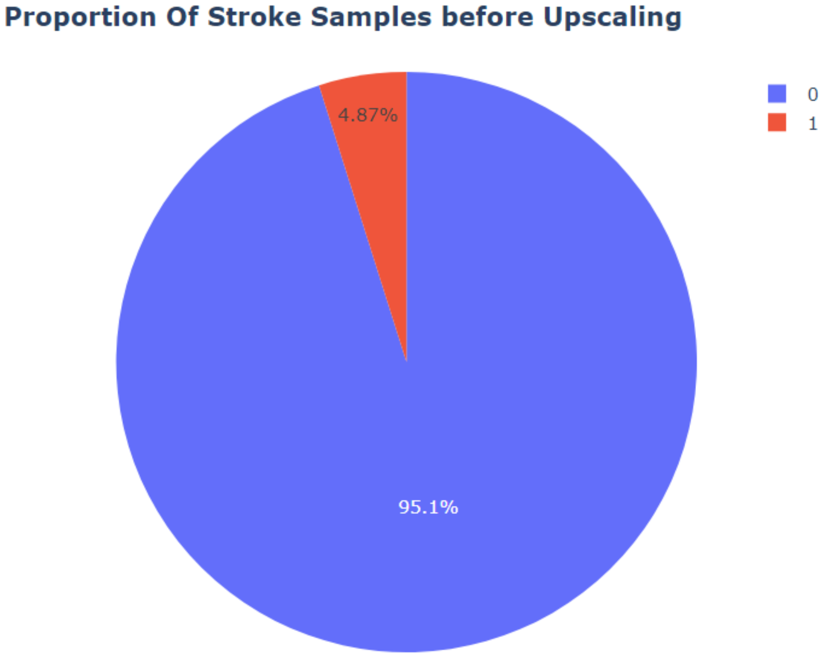
g. Age



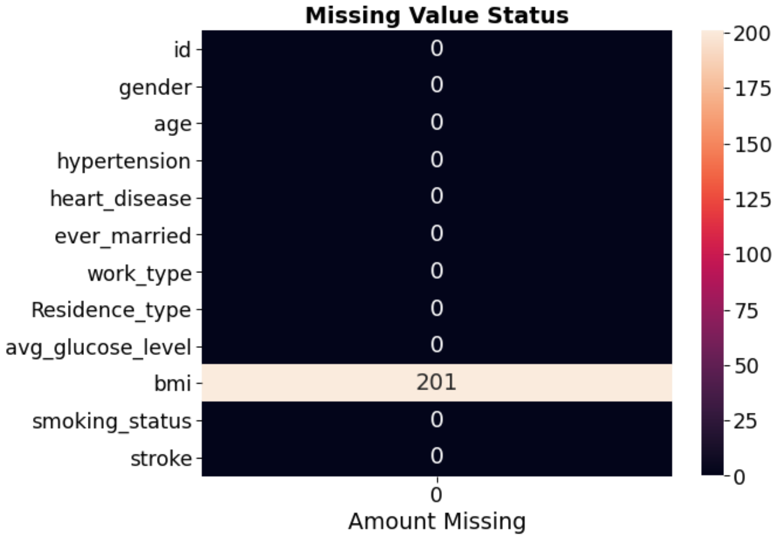
3. Stroke Sample Distribution Based on BMI Glucose Level and Age



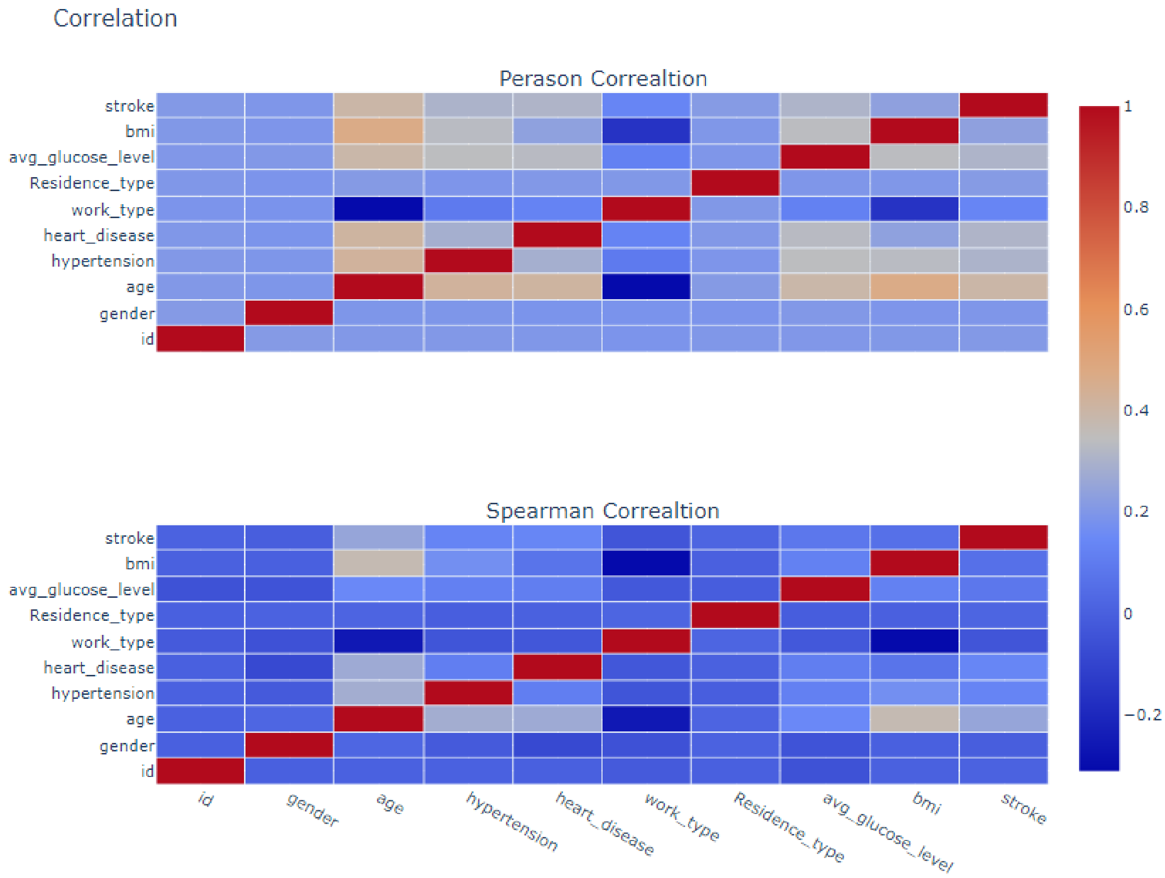
4. Stroke Distribution in Overall Population



5. Missing Values



6. Correlation



Section 2.3 Inference

1. Distribution of Marriage status and residence type doesn’t say anything about stroke population , as marriage status is the same as the underlying distribution and residence type is normally distributed in stroke population. This is also visible in correlation.
2. The distribution of stroke in the population is heavily skewed towards no stroke so Upsampling is needed to make the distribution uniform. This is done using SMOTE reresambling
3. There are **missing values in BMI,** but these contain stroke populations too, so cannot be dropped as very few stroke samples. These are predicted using **Random forest**.
4. Smoking Status also doesn't give any info on stroke population.
5. There is high correlation in age, bmi, avg_glucose_level and work type

Section 2.4 Upsampling

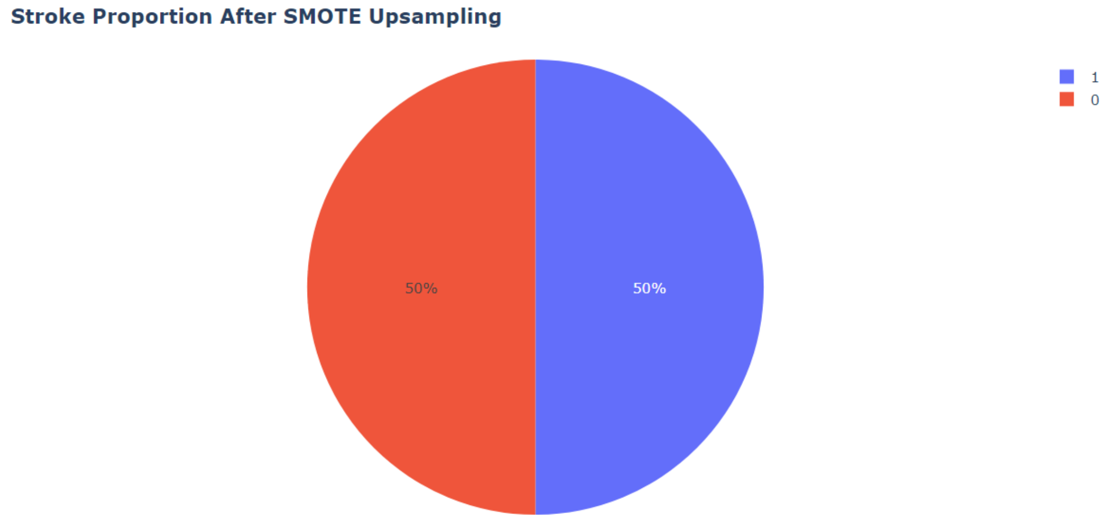
Since the stroke population is below 5%, data is imbalanced, we need to upsample data . This is done using SMOTE resample , else all will be classified as no stroke making model obsolete.

This works by synthesising new samples in minority class, by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line. The synthetic instances are generated as a convex combination of the two chosen instances a and b

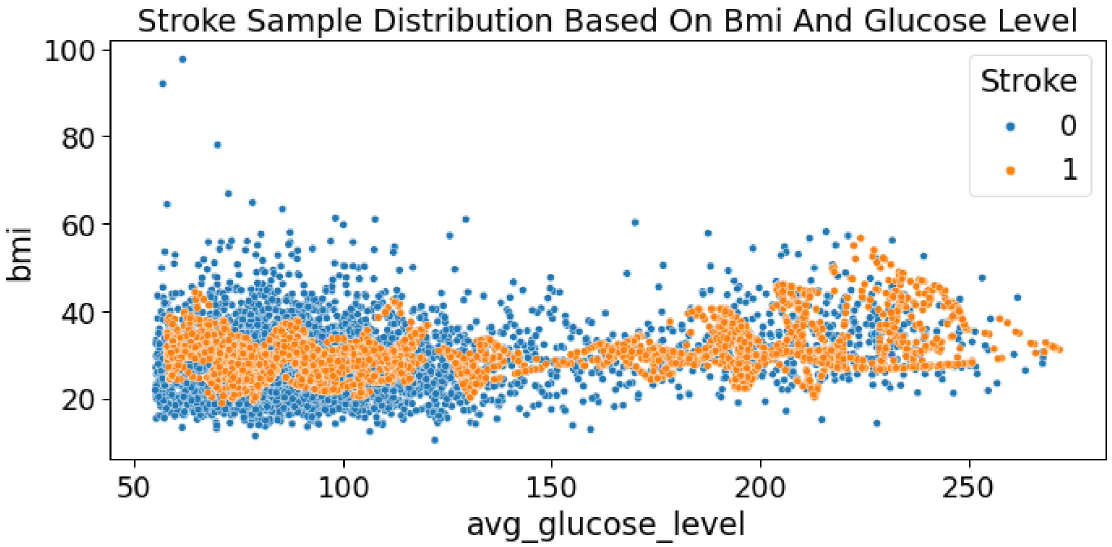
Reference:

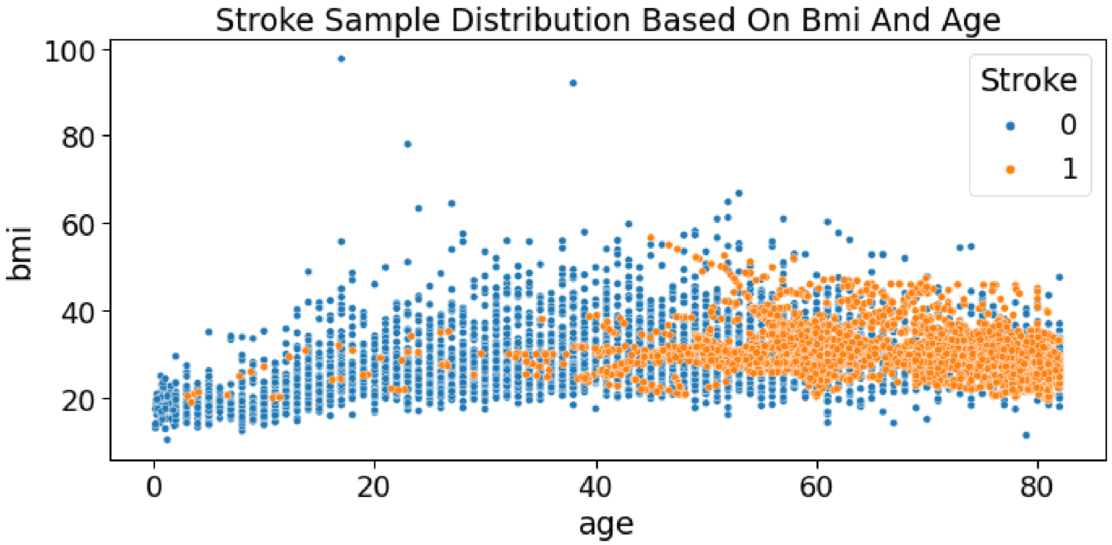
https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Stroke Population After Resampling:



Distribution After Upsampling





Section 3: Problem Statements

1. Classify whether the person had a stroke based on the health profile of the person
   Classification is done using following methods
   a. Decision Tree Classifier
   b. Random Forest Classifier
   c. KNN Classifier
   d. MLP Classifier
   e. Logistic Regression
   f. SVM Classifier
2. Do regression analysis to compute the BMI of the person based on following methods
   a. MLP Regressor
   b. Random Forest Regressor
   c. Decision Tree Regressor
   d. Ridge Regression
   e. Linear Regressor

Section 4.1: Classification

Classification is done using ML pipeline by sklearn. All models are from Sklearn
All data is scaled using standard scaler before training the model for classification Random state =42 is used for consistency of results

1. Decision Tree Classifier
   The function to measure the quality of a split = “gini”
   Best split at each node
   Nodes are expanded until all leaves are pure or until all leaves contain less than 2 samples.

2. Random Forest Classifier
   100 trees
   Gini used for each split
   nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
   max_features=sqrt(7)
   unlimited number of leaf nodes
   bootstrap samples are used when building trees

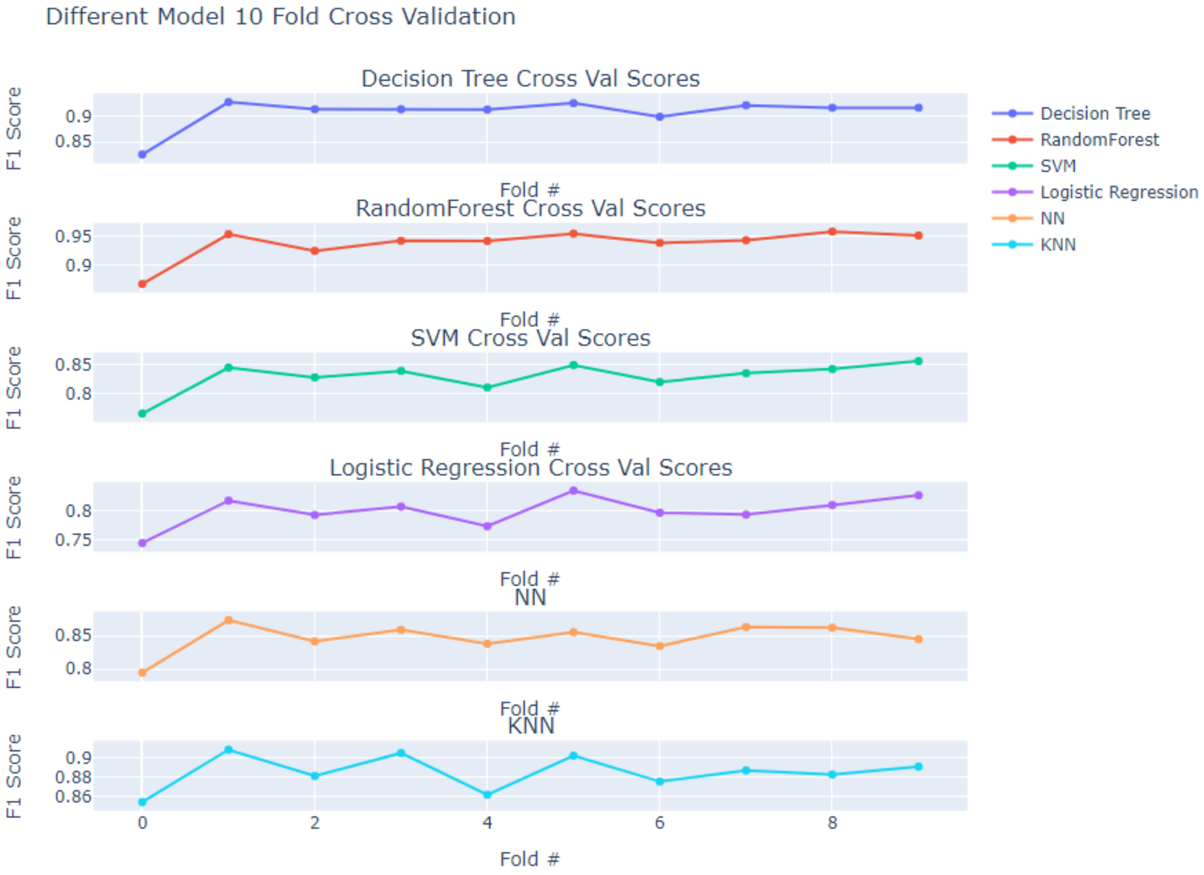
max_samples=full dataset

3. KNN Classifier
   n=5
   metric=euclidean
4. MLP Classifier
   Maximum iterations= 500
   ‘relu’ activation function
   stochastic gradient-based optimizer for iteration
   L2 penalty=1e-4
   batch_size=200
   learning_rate=1e-3

5. Logistic Regression
   L2 regularisation used 1e-4,
   Solver = lbfgs,Limited-memory Broyden–Fletcher–Goldfarb–Shanno. It approximates the second derivative matrix updates with gradient evaluations.
   max_iter=100
6. SVM Classifier
   L2 regularisation =1
   Radial basis function kernel deg=3
   1 iteration

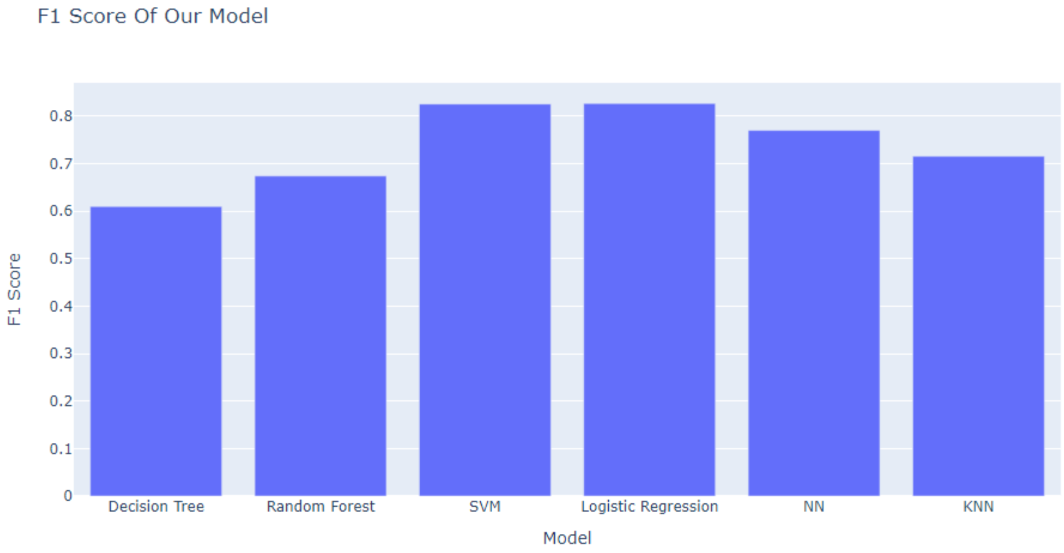
Section 4.2.1 Classification Performance Parameters

1. Ten fold cross validation using f1score is used and results given below

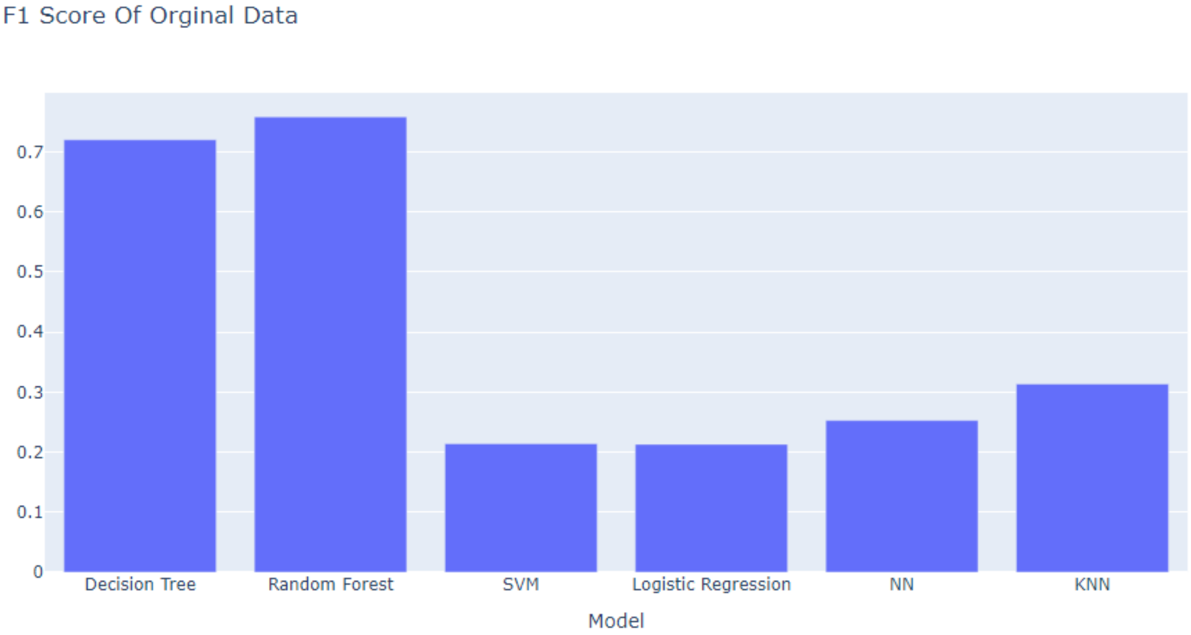


All the models are giving similar cross validation scores highest being Random forest and lowest being Logistic regression

2. F1 score
    a. for upscaled train data

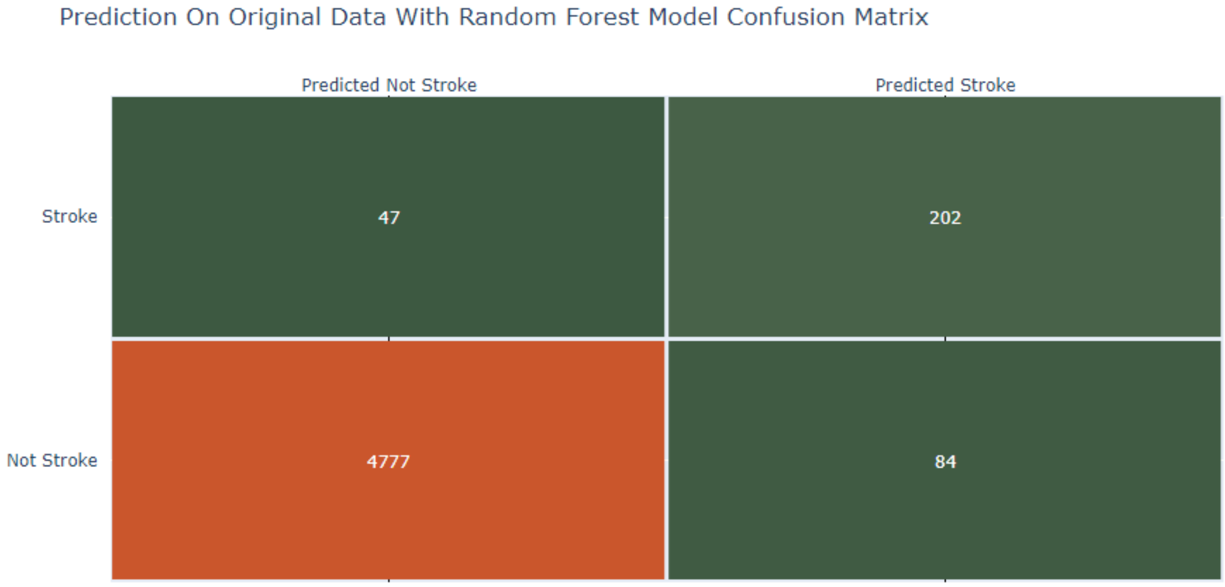


    b. For Original Data



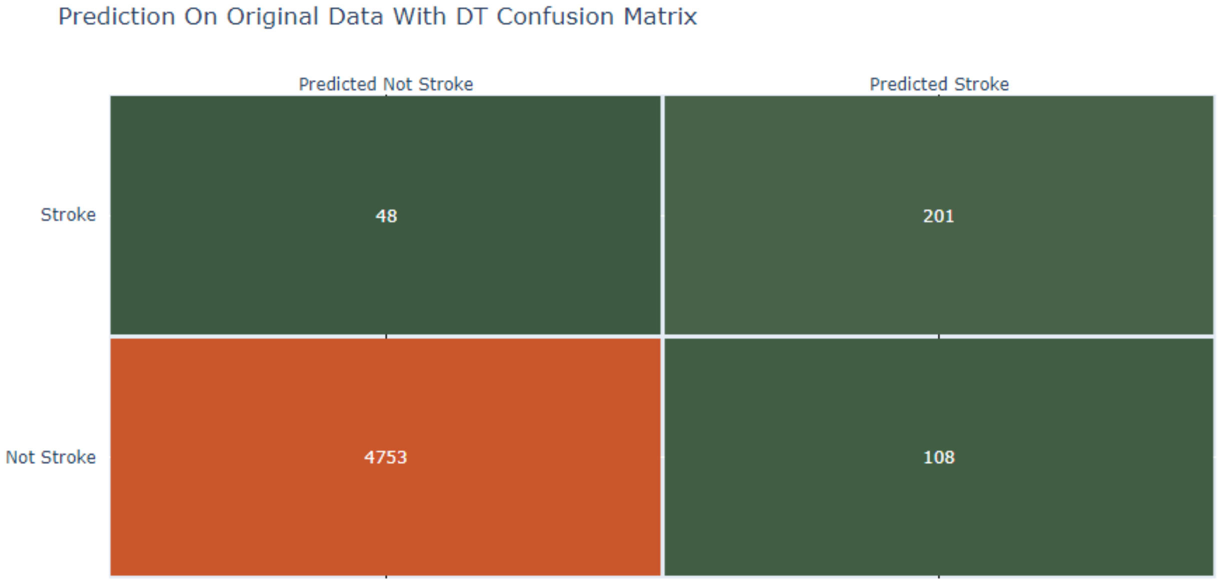
3. Confusion Matrix

a. Original Data



Prediction On Original Data With Random Forest Model Confusion Matrix

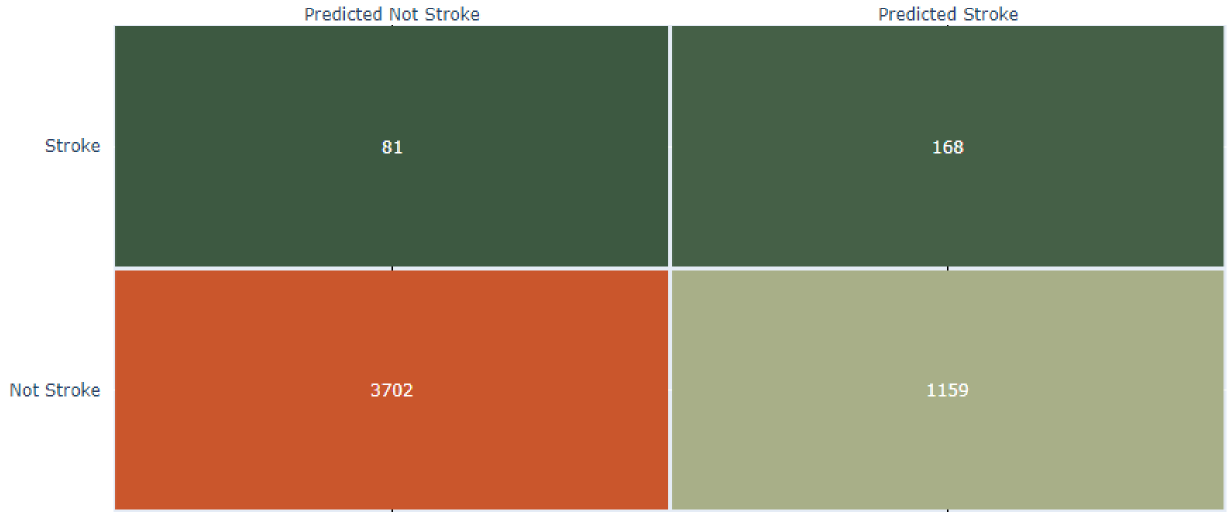
	Predicted Not Stroke	Predicted Stroke
Stroke	47	202
Not Stroke	4777	84



Prediction On Original Data With DT Confusion Matrix

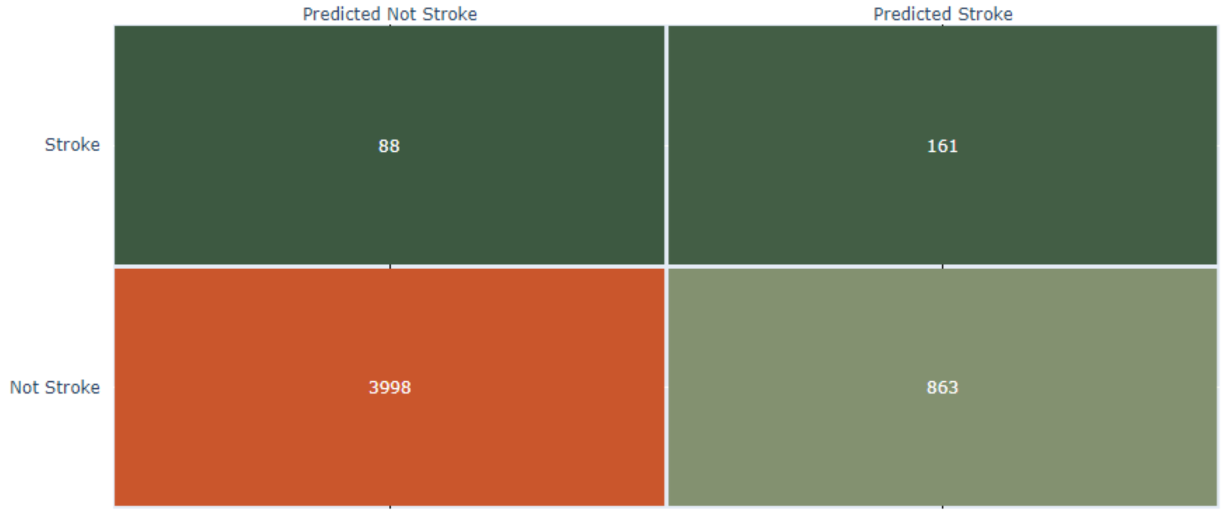
	Predicted Not Stroke	Predicted Stroke
Stroke	48	201
Not Stroke	4753	108

Prediction On Original Data With LR Confusion Matrix



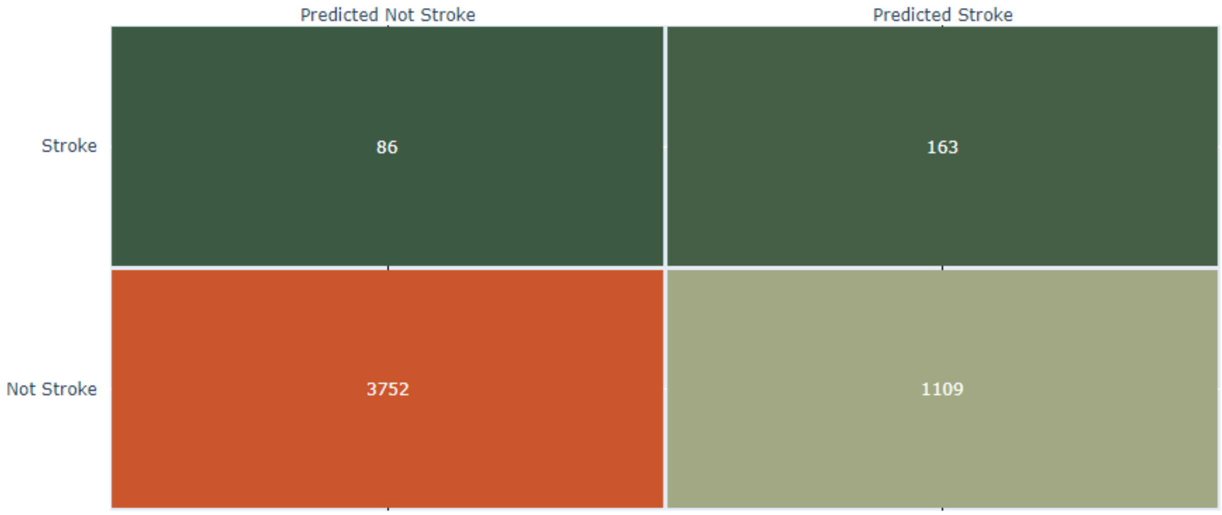
	Predicted Not Stroke	Predicted Stroke
Stroke	81	168
Not Stroke	3702	1159

Prediction On Original Data With NN Confusion Matrix



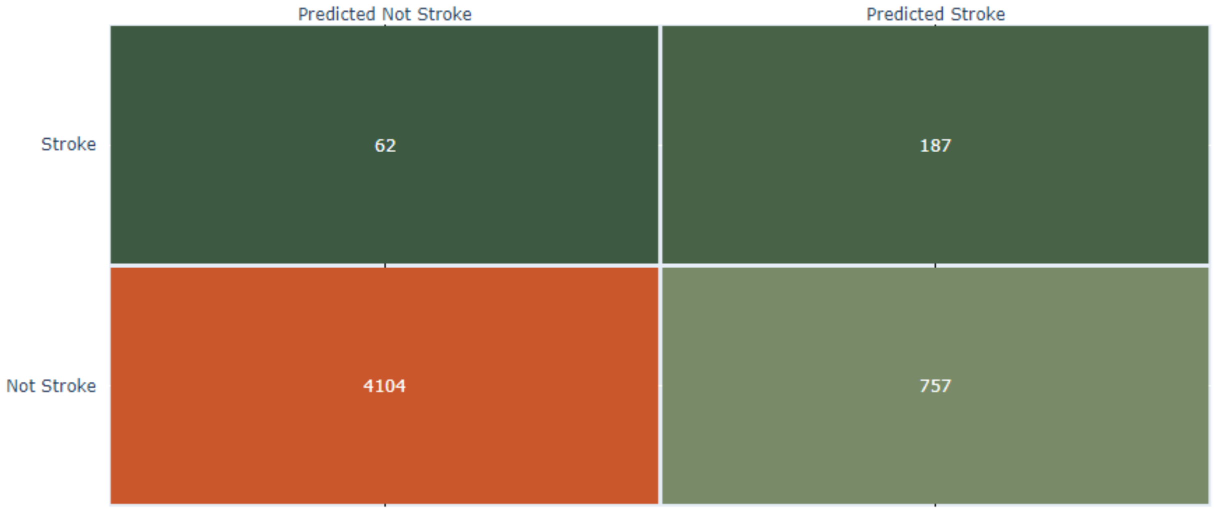
	Predicted Not Stroke	Predicted Stroke
Stroke	88	161
Not Stroke	3998	863

Prediction On Original Data With SVM Confusion Matrix



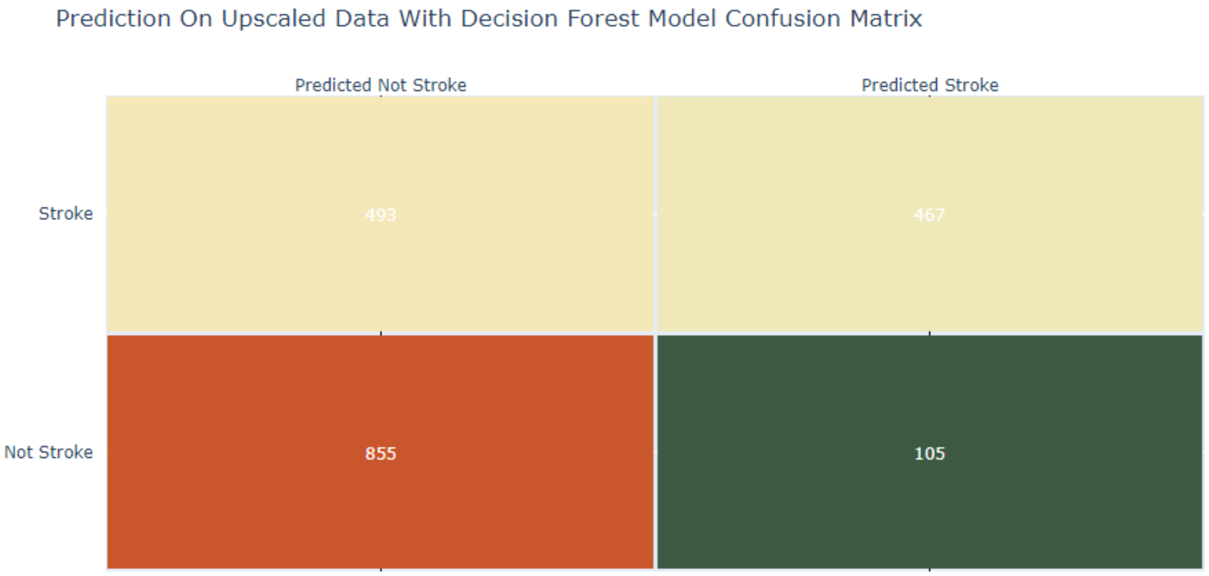
	Predicted Not Stroke	Predicted Stroke
Stroke	86	163
Not Stroke	3752	1109

Prediction On Original Data With KNN Confusion Matrix



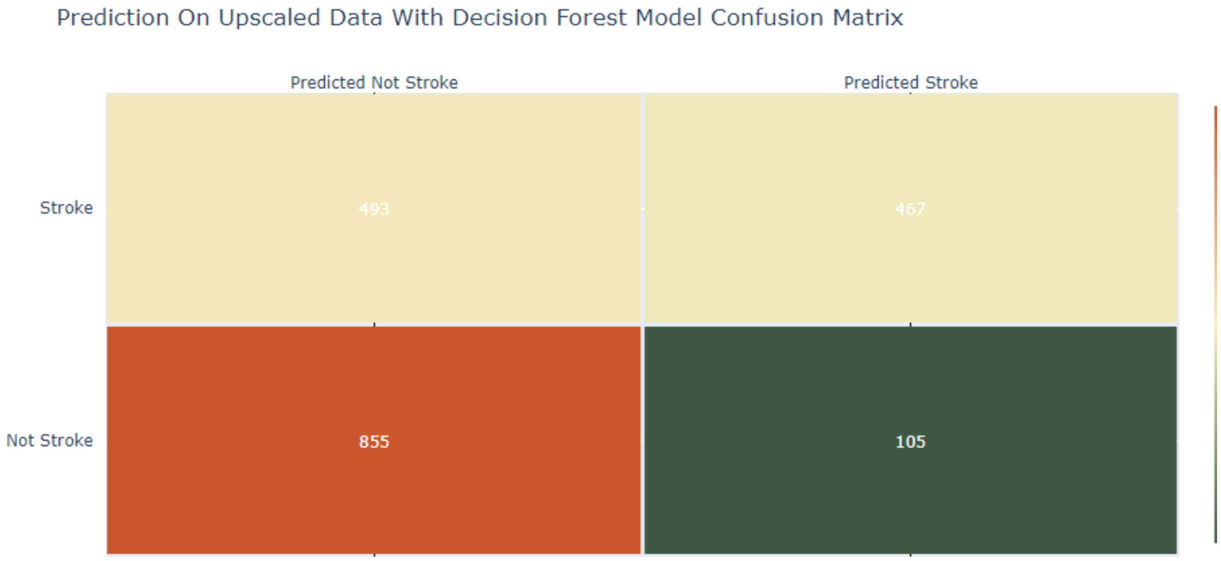
	Predicted Not Stroke	Predicted Stroke
Stroke	62	187
Not Stroke	4104	757

b. On Evaluation set in Upscaled data



Prediction On Upscaled Data With Decision Forest Model Confusion Matrix

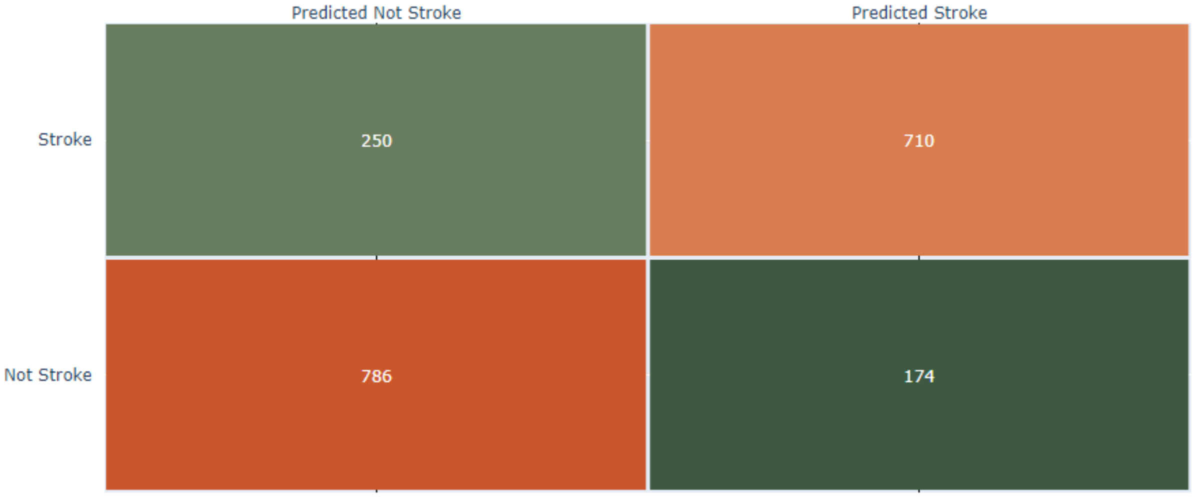
	Predicted Not Stroke	Predicted Stroke
Stroke	493	467
Not Stroke	855	105



Prediction On Upscaled Data With Decision Forest Model Confusion Matrix

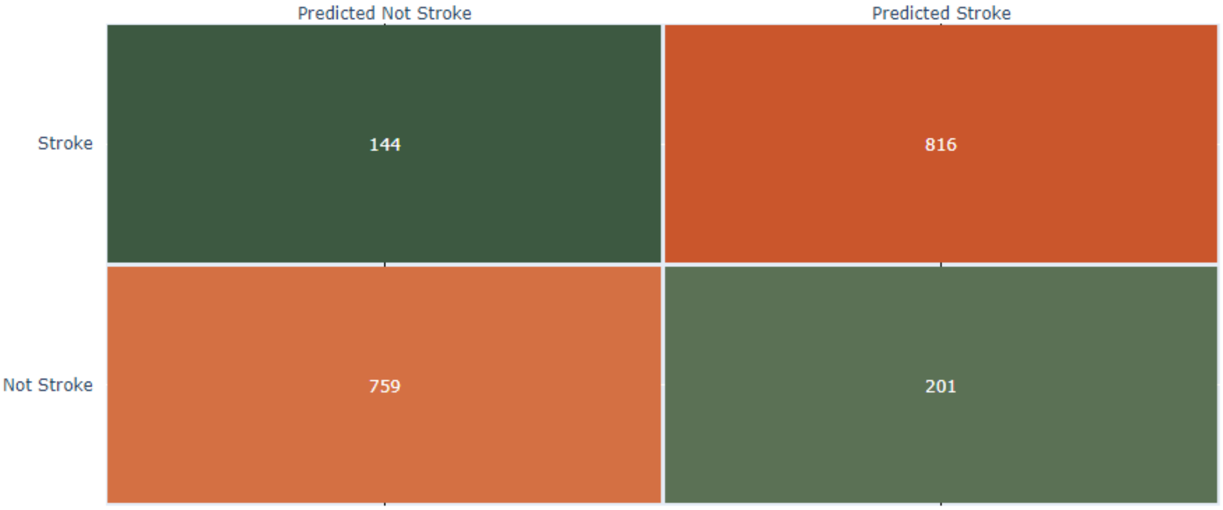
	Predicted Not Stroke	Predicted Stroke
Stroke	493	467
Not Stroke	855	105

Prediction On Upscaled Data With NN Model Confusion Matrix



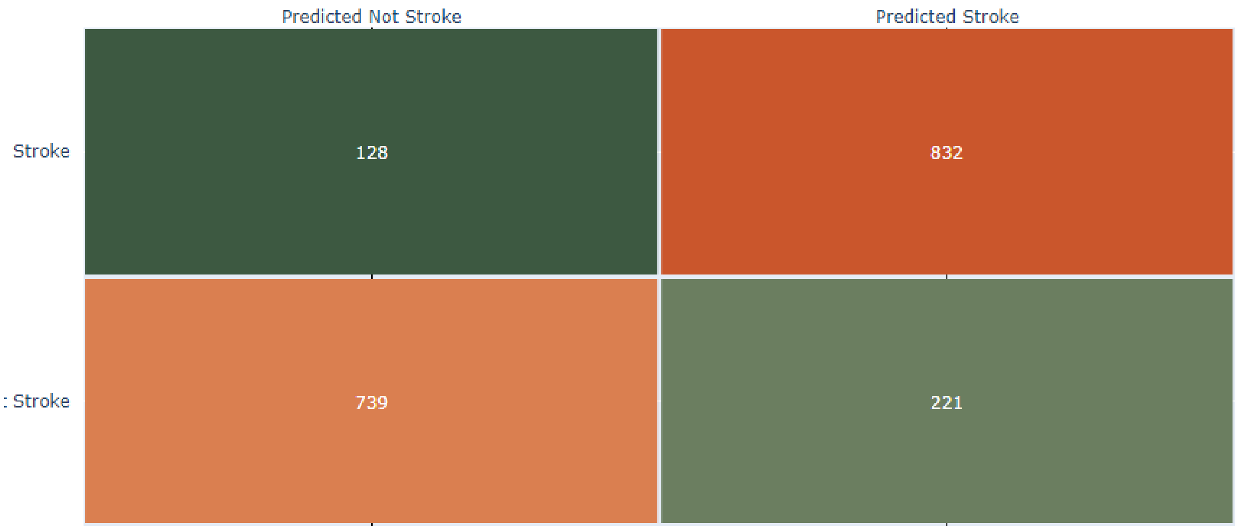
	Predicted Not Stroke	Predicted Stroke
Stroke	250	710
Not Stroke	786	174

Prediction On Upscaled Data With SVM Model Confusion Matrix



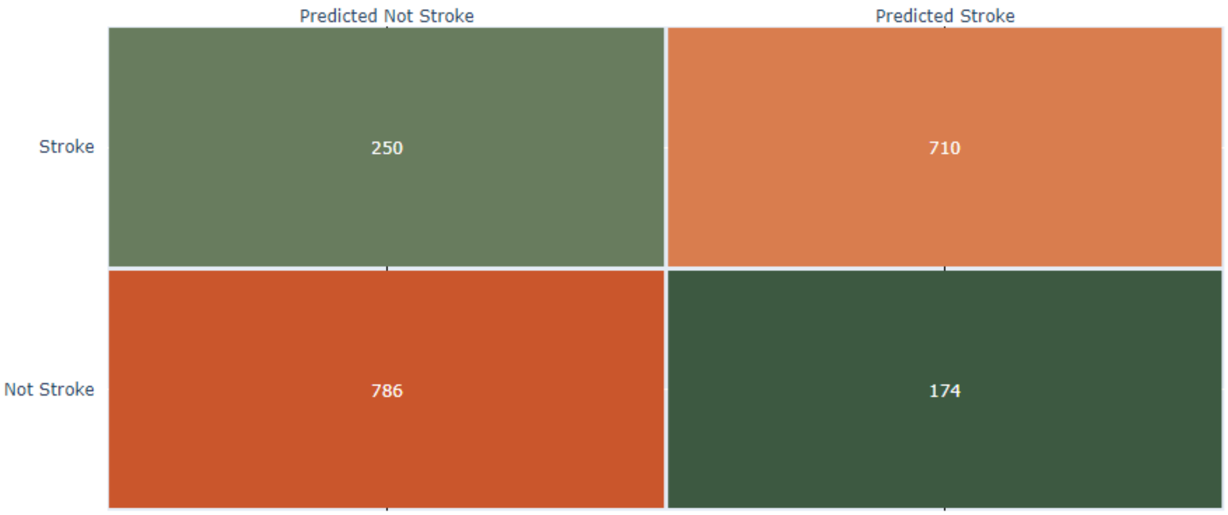
	Predicted Not Stroke	Predicted Stroke
Stroke	144	816
Not Stroke	759	201

Prediction On UPscaled Data With Linear Regression Model Confusion Matrix



	Predicted Not Stroke	Predicted Stroke
Stroke	128	832
: Stroke	739	221

Prediction On Upscaled Data With KNN Confusion Matrix



	Predicted Not Stroke	Predicted Stroke
Stroke	250	710
Not Stroke	786	174

Section 4.2.2 Classification Performance Conclusion

1. In 10 fold cross validation testing we found that Random Forest was the best of all using f1 score s metric
2. This is seen when misclassification of stroke patients in original data was less for RF closely followed by DT similar to cross validation score
3. There are very high FP rates for all other models Mostly due to overfitting and underfitting of model
4. RF and DF performed best in terms of FN rates which is important in medical grounds

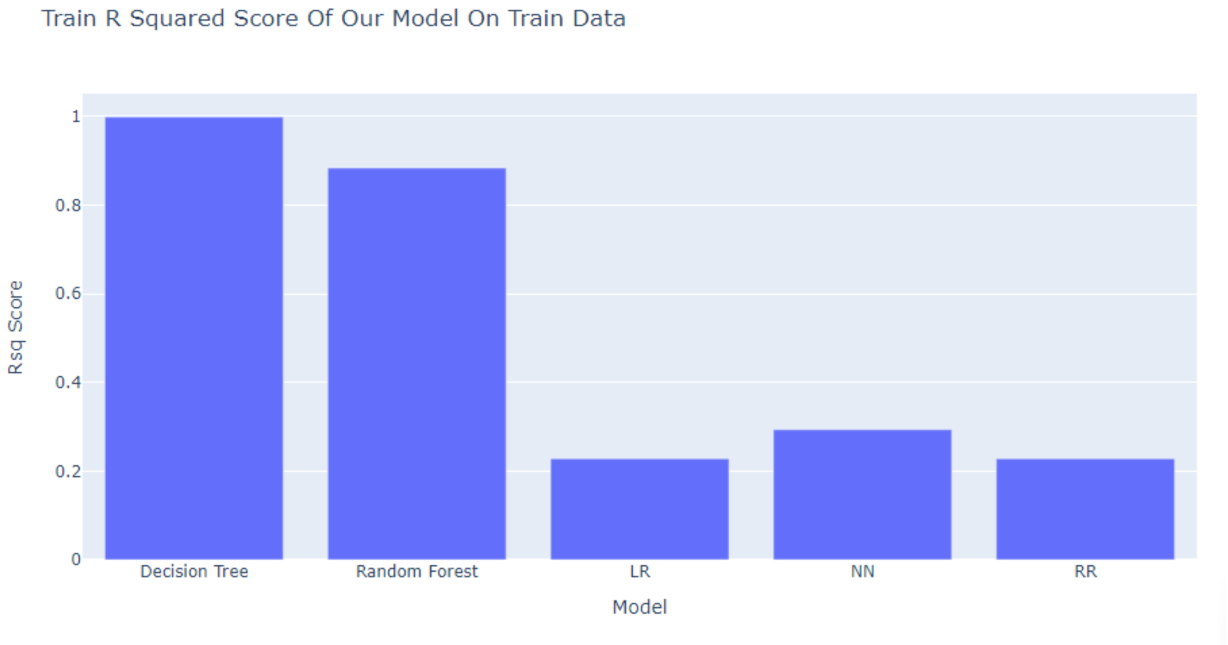
Section 4.3.1 Regression

Original data is used for predicting BMI of the patient using regression
Methods Used

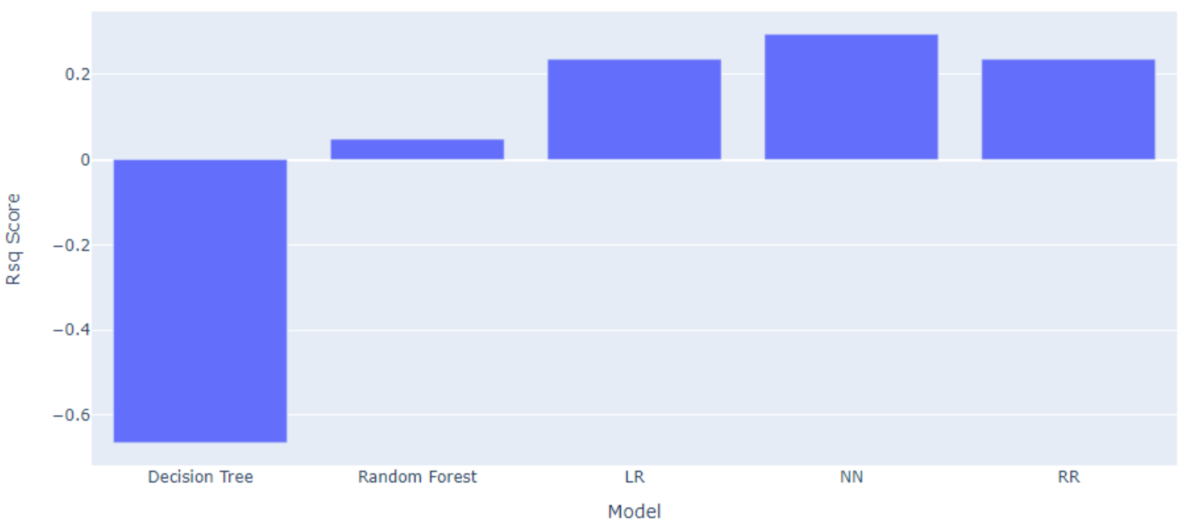
1. MLP Regressor: 100 nodes in one hidden layer 500 iteration
2. Random Forest Regressor
3. Decision Tree Regressor
4. Ridge Regression auto solver lambda=1
5. Linear Regressor

Input: 'age','hypertension','work_type','avg_glucose_level','stroke'

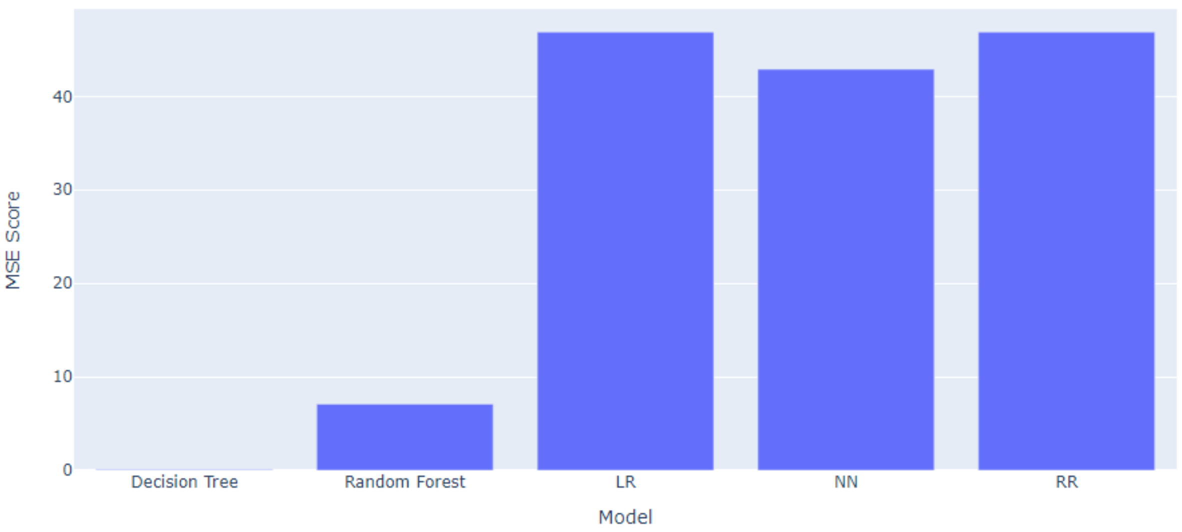
Section 4.3.2 Performance R Squared

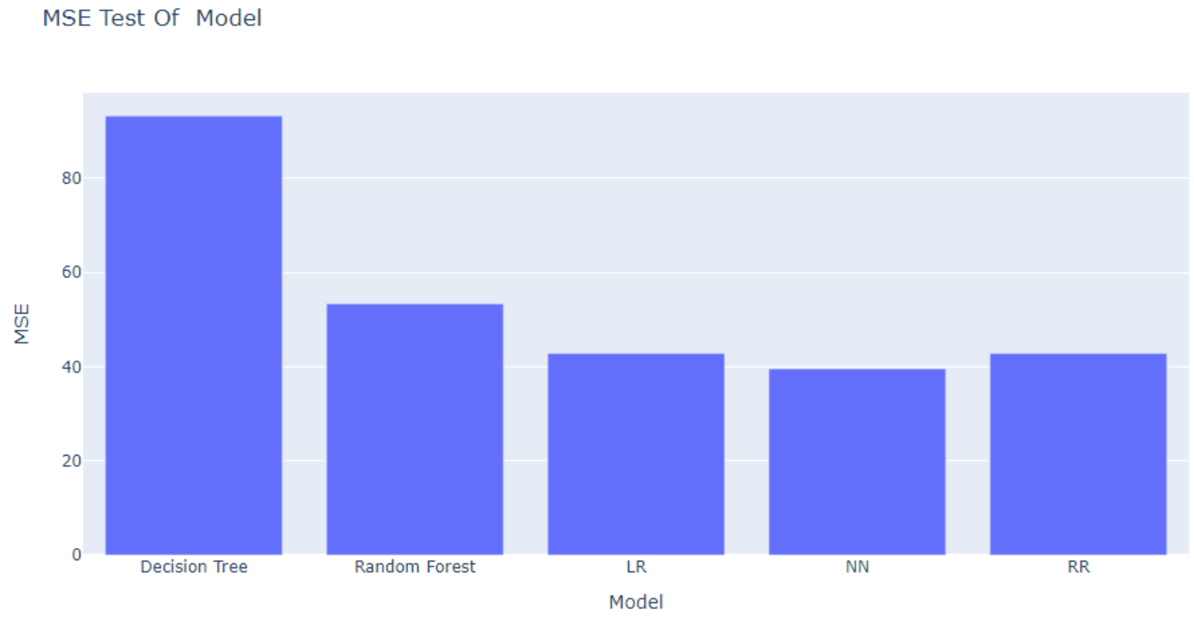


R Squared Score Of Our Model On Test Data



MSE TRAIN Of Our Model





Section 4.4 Inference of Regression

1. Regression has very low accuracy with Decision Model Overfitting data becoming worst model
2. From R squared comparatively good model is MLP regressor similarly visible in MSE test

Section 5 Comparison with classical/ linear methods and sequential approaches

Linear and Logistic Regression also gave worse results due to decision boundary being nonlinear
MLP and Tree method gave good results visible from the results

Section 6 Convince the manager or client of the impact of your solution on the workflow or finances

From the data and real life it is visile that people who are obese , old or having diabetes are prone to have stroke . Nowadays wearables have sensors for glucose levels. Combined with this the classifier model is highly useful for risk client as an early warning system

https://www.kaggle.com/fedesoriano/stroke-prediction-dataset