Brain Stroke Prediction

Hardik Jatolia , Govind Mali ,Anshul Tomar

B21EE022, B21EE021, B21EE083

❖ Problem Statement :

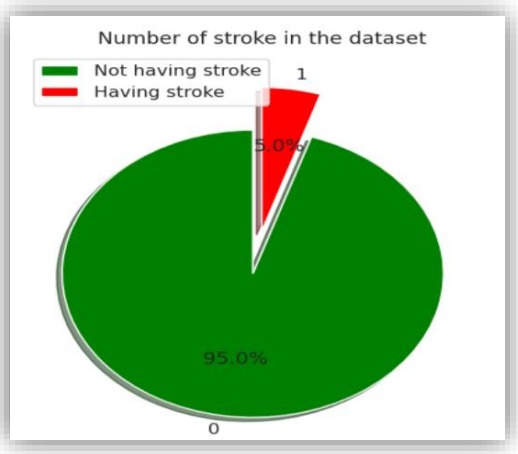
Stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. The given Dataset is used to predict whether a patient is likely to get a stroke based on the input parameters like gender, age, various diseases, and smoking status.

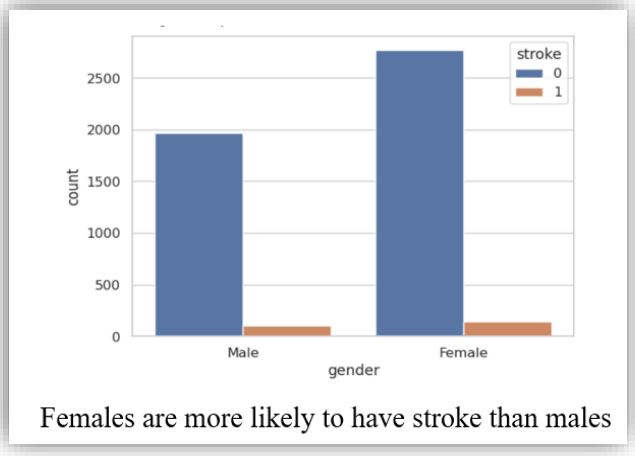
❖ Dataset Description :

The given dataset has 11 columns which includes gender, age, hypertension (1 if person has hypertension otherwise 0), heart disease (1 if person has otherwise 0), whether person is married or not, work type of person, residence type of person, person's average glucose level, person's BMI, smoking status of person and the label column.

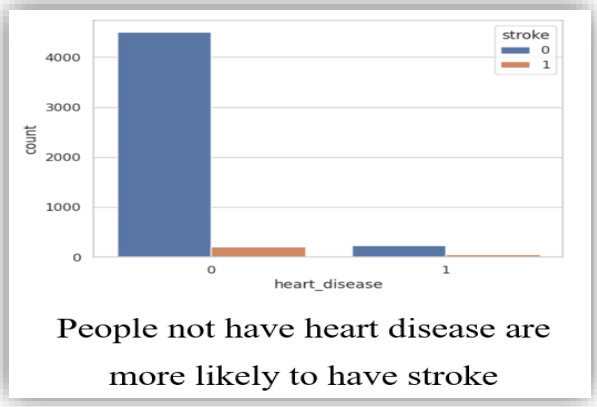
❖ Pre-processing and Exploratory Data Analysis :

- First we checked for null values and found none.
- Checked the datatype of each column and then stored categorical and numerical columns in two separate lists.
- We plotted a pie chart for the 'stroke' column and we found that only 5% of the people in the data have stroke. This shows that data is highly imbalanced.
- We also plotted graphs for categorical and numerical columns and analysed them thoroughly to see how many have strokes in each feature



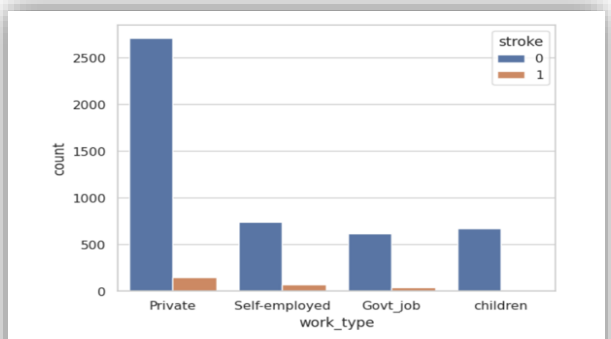


Females are more likely to have stroke than males

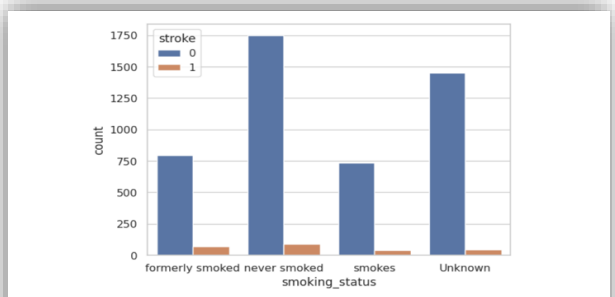


People not have heart disease are more likely to have stroke

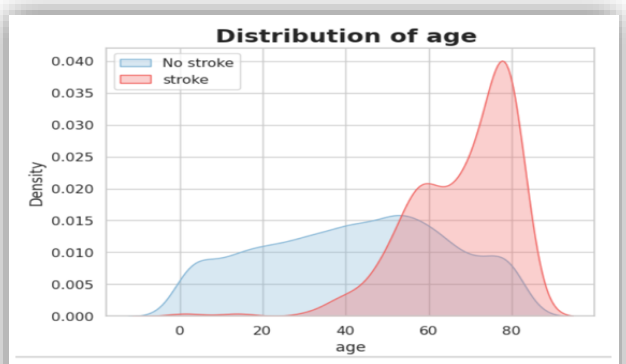
Chance of stroke is more if married than if not married



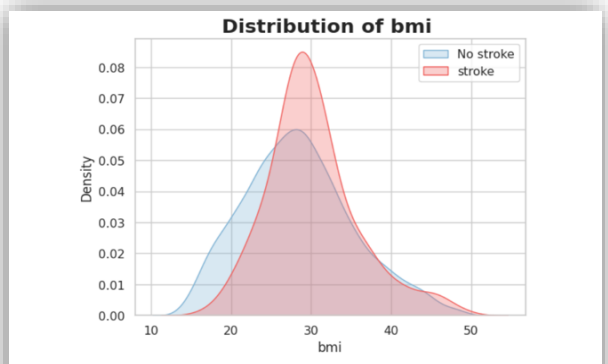
People who work are likely to get stroke



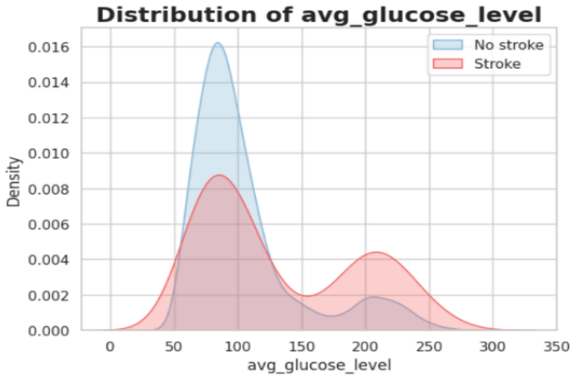
People who smoke are likely to get stroke (since formerly smoked + smokes > never smokes)



People with more age are likely to get stroke than people with less age



There's no significant relation b/w BMI and stroke. Similar BMI has stroke as well no stroke



There's very little relation b/w avg_glucose_level and stroke. Similar glucose level has stroke as well as no stroke. If a person has avg_glucose_level > 150, he/she has very likely to have stroke.

❖ Feature Engineering :

- From the list of categorial columns we performed label encoding using label encoder.
- We removed stroke , hypertension, heart_disease columns from numerical column as they are categorial having int values and then standardize the rest numerical using StandardScalar().
- Then we split the data in 80:20 using stratified split method of sklearn library.

❖ Metric Selection :

- Now, since the dataset is imbalanced and it is the binary classification problem, the best metric for evaluation would be Area under the ROC curve (AUC). We can also use precision and recall, but AUC combines both of these metrices.
- Also, for this specific problem statement Recall is more important metric for evaluation since we do not want any true case i.e. person having stroke to be predicted as false case i.e. person with no stroke.

❖ Modelling the formed dataset without taking care of Imbalance :

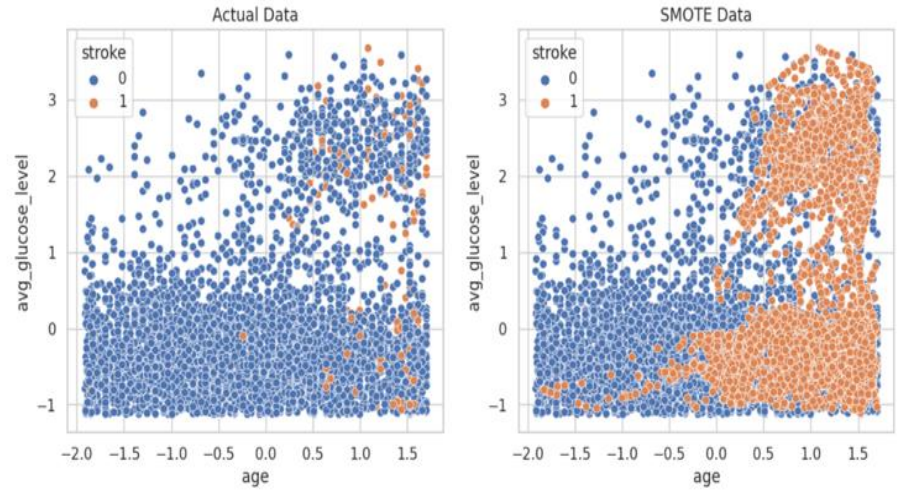
- Various classification models are fitted with the above dataset. These models include Logistic Regression, Random Forest Classifier, XGB Classifier, LGBM Classifier, Bernoulli NB, GaussianNB ,Gradient Boosting Classifier, Support Vector Machine, Decision Tree, AdaBoost, KNeighbours Classifier and the above metric is used to check which model works how. The result obtained is :

	Model	Precision	Recall	ROC_AUC_Score	F1_score
0	LogisticRegression	0.000000	0.000000	50.000000	0.000000
1	DecisionTreeClassifier	13.043478	12.000000	53.888068	12.500000
2	RandomForestClassifier	0.000000	0.000000	49.947202	0.000000
3	BernoulliNB	11.111111	4.000000	51.155227	5.882353
4	GaussianNB	13.274336	30.000000	59.825766	18.404908
5	Support Vector Machine	0.000000	0.000000	50.000000	0.000000
6	K-Nearest Neighbors	6.818182	6.000000	50.835269	6.382979
7	GradientBoostingClassifier	0.000000	0.000000	49.788807	0.000000
8	AdaBoost	0.000000	0.000000	49.841605	0.000000
9	XGBClassifier	0.000000	0.000000	49.419219	0.000000
10	LightGBM	0.000000	0.000000	49.683210	0.000000

- We can see that our models perform very poorly since the ROC_AUC_Score and the Recall is very low.
- This is due to data imbalance.
- So we have to overcome this problem. For that we can use oversampling techniques to handle this

❖ Making Dataset balanced with the help of Synthetic Minority Oversampling Technique (SMOTE) :

Smote Technique :- Oversample the examples in the minority class. This can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model. This can balance the class distribution but does not provide any additional information to the model.



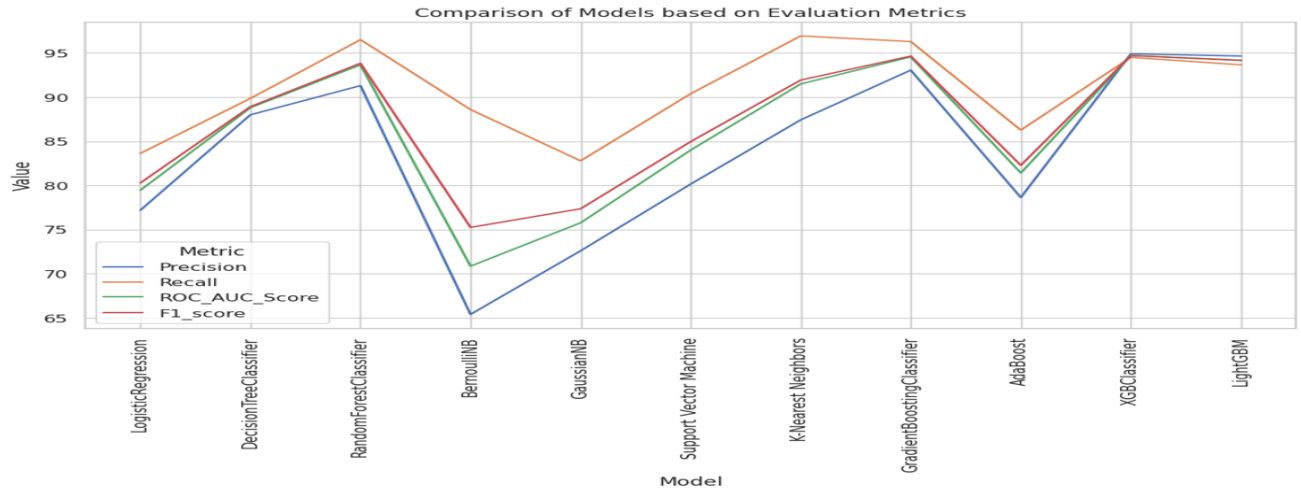
- Dataset is oversampled using SMOTE Technique and the final dataset look like the right plot of the dataset.
- Left image is the actual dataset.

❖ Modelling the dataset after SMOTE using the same models :

	Model	Precision	Recall	ROC_AUC_Score	F1_score
0	LogisticRegression	77.192982	83.632524	79.461457	80.283832
1	DecisionTreeClassifier	88.004137	89.862724	88.806758	88.923720
2	RandomForestClassifier	91.308691	96.515312	93.664203	93.839836
3	BernoulliNB	65.393609	88.595565	70.855333	75.246637
4	GaussianNB	72.592593	82.787751	75.765576	77.355698
5	Support Vector Machine	80.149813	90.390707	84.002112	84.962779
6	K-Nearest Neighbors	87.428571	96.937698	91.499472	91.937907
7	GradientBoostingClassifier	93.061224	96.304118	94.561774	94.654904
8	AdaBoost	78.633301	86.272439	81.414995	82.275932
9	XGBClassifier	94.909862	94.508976	94.720169	94.708995
10	LightGBM	94.663821	93.664203	94.192186	94.161359

- We can now clearly see that after applying SMOTE our models has performed much better than before.
- Now we have to choose the best models out of them

- For that we can compare the metrics of all the models and choose accordingly

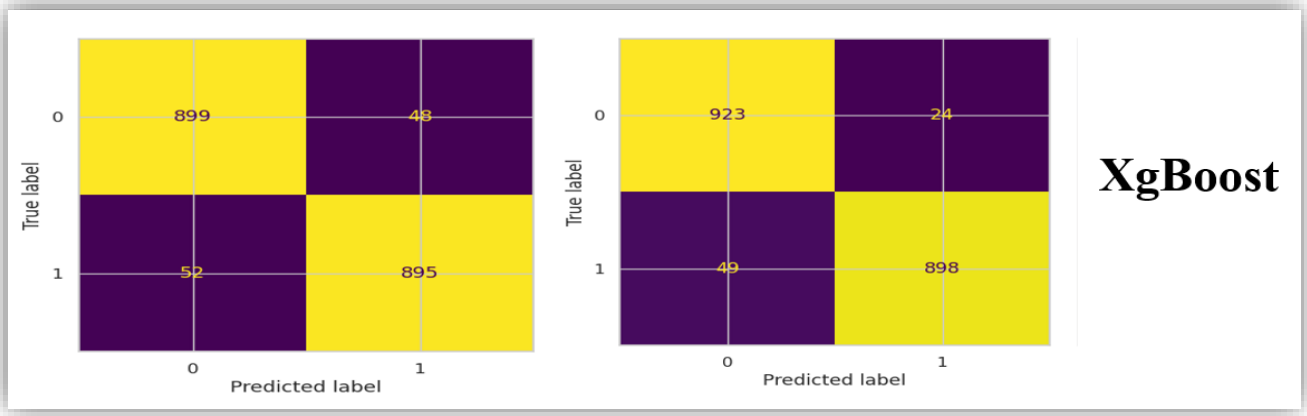


- Selecting models having high Recall Score and having high ROC_AUC score. Hence, RFC, KNN, GBC, XGB, LGBM are selected. So, these models are now tuned by the hypermeters which is followed by 10-fold cross validation.

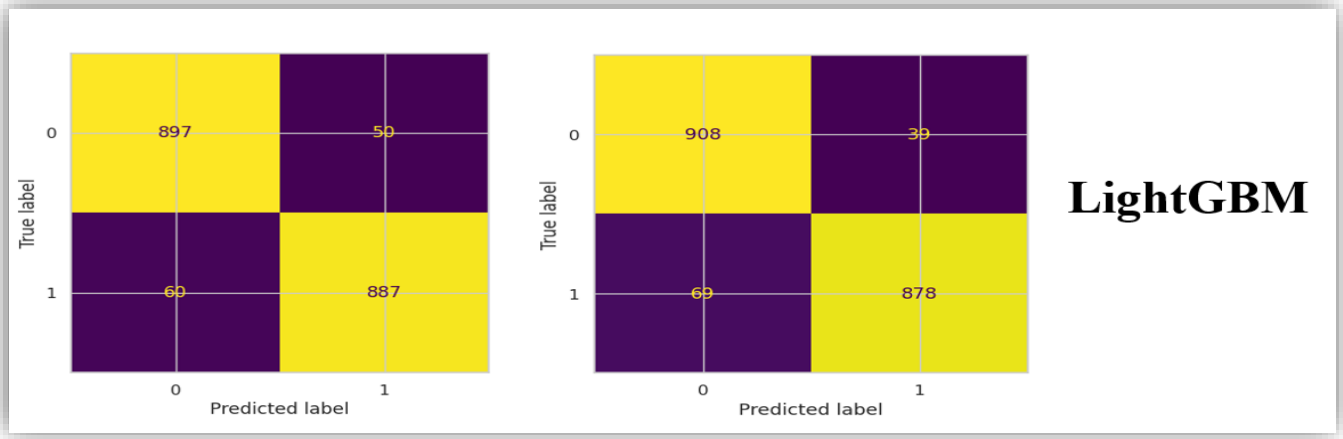
	Model	Cross Validation AUC ROC Score
0	RandomForestClassifier	0.984903
1	KNeighborsClassifier	0.917384
2	GradientBoostingClassifier	0.987715
3	XGBClassifier	0.948786
4	LightGBM	0.949842

- 10 – Fold Cross Validation Result:
- From the table on the left side that KNN do not work very well,
- Hence, we would take the rest 4 models for the final classification as the R0C_AUC score received is good to take.

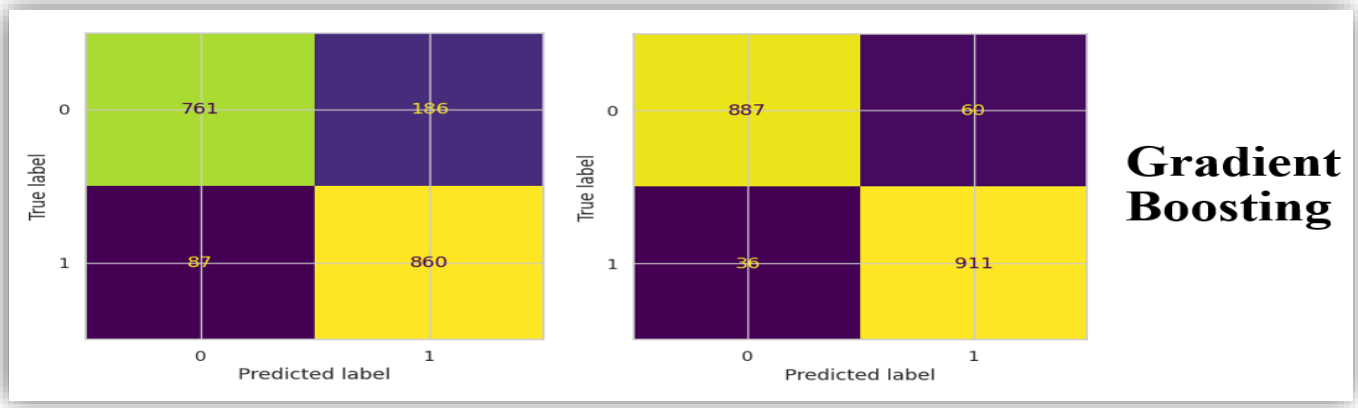
❖ **Comparison of Confusion Matrix of Top 4 Models before (Left side) and after (Right side) Hyperparameter Tuning:**



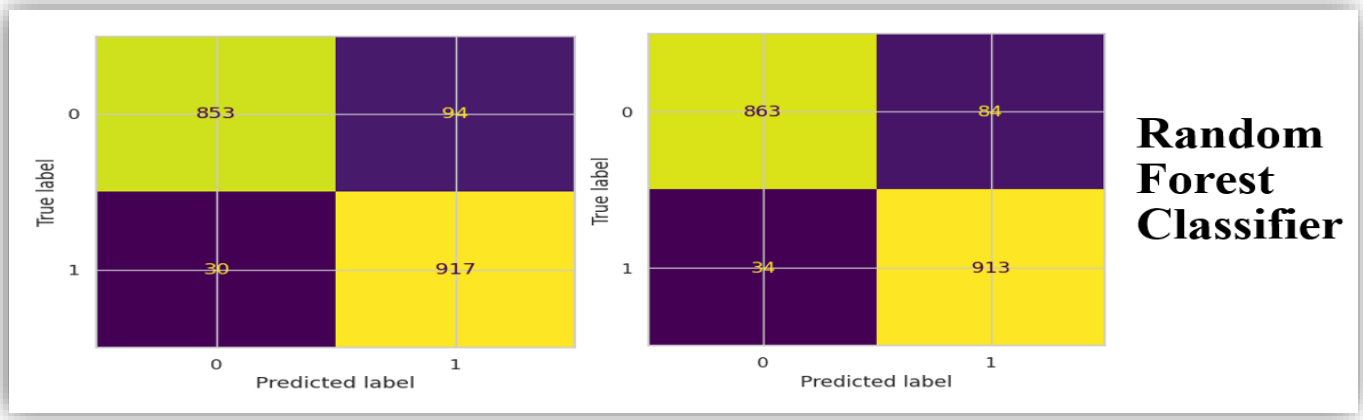
XgBoost



LightGBM



Gradient Boosting



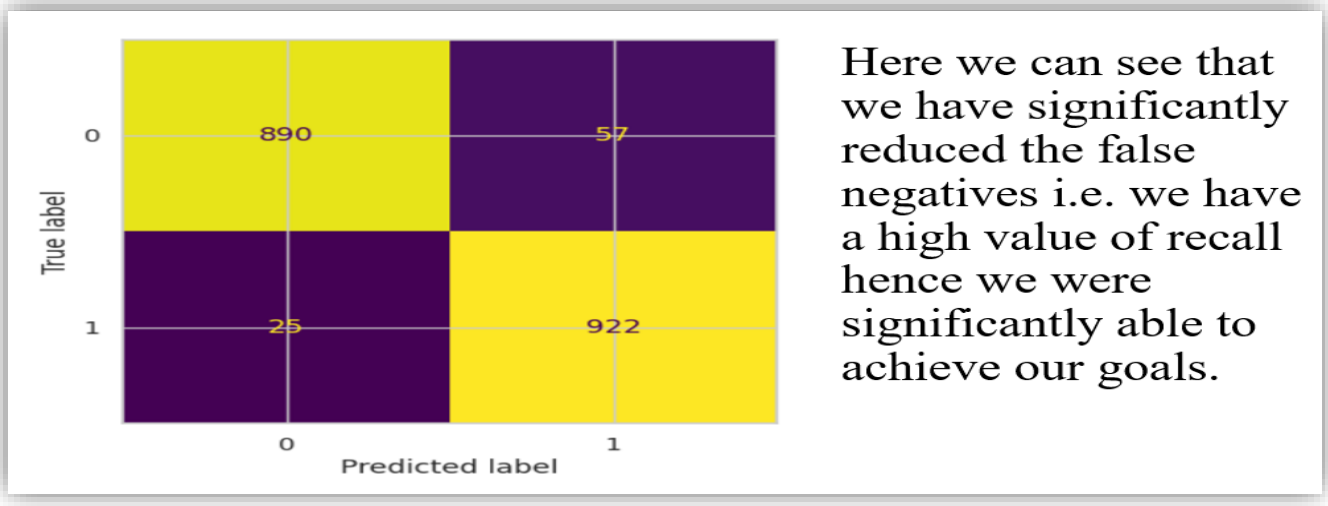
Random Forest Classifier

❖ Justification why these classifiers yield good results :

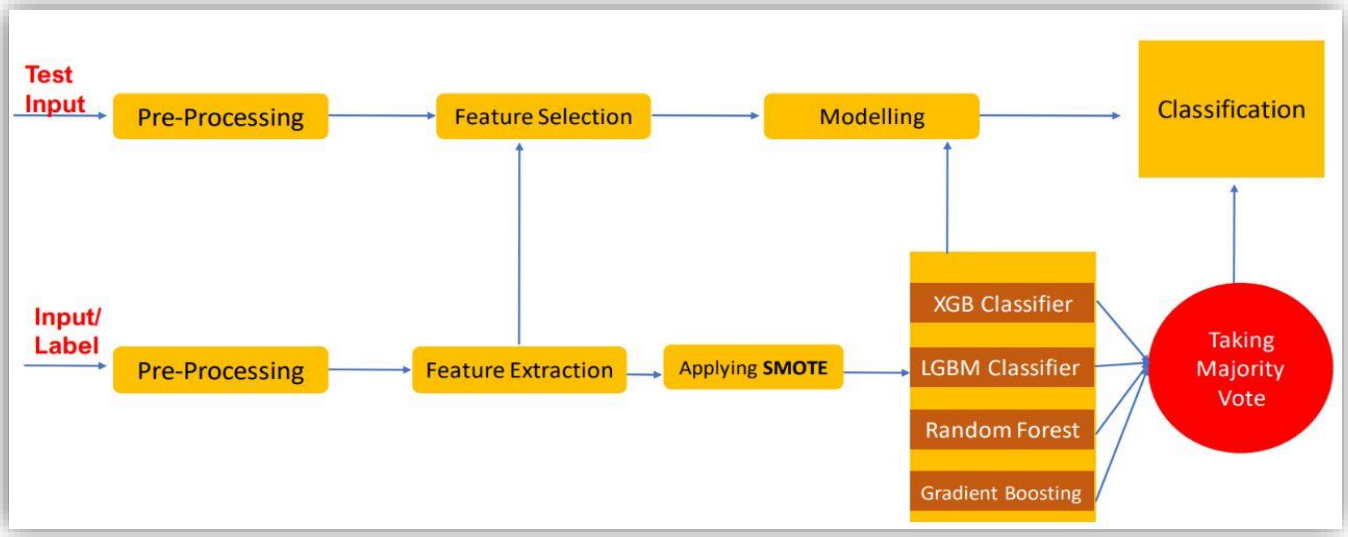
- So, all the models selected uses ensemble learning technique where RFC uses bagging to give accuracy. In RFC, bagging is just like training a bunch of individual decision trees in a parallel way. Other is boosting technique which the other 3 classifiers uses and it's about training a bunch of individual models in sequential manner by learning from the mistakes done by the previous models.

❖ Majority Voting :

- Final Prediction is received by taking the majority vote of the above chosen classifiers.
- Final Confusion Matrix we got using Majority voting is :



❖ End -To – End Pipeline for the given Dataset :



The final ROC_AUC score for the test data received is 95.67053%.

❖ Deployed the model and created the Website : Github Link

We have deployed the above model on Github and created a website of the same where user can input the parameters asked example gender, age etc. Based on the input parameters received, model predicts the output of getting stroke and not getting stroke and show the result on the website. We used pickle library to export our models and in backend we used these models to predict on the basis of the inputs that we received from the frontend.

❖ Contribution of members :

- Hardik Jatolia (B21EE022) : Pre- processing, exploratory data analysis, pipeline diagram, report, Website.
- Govind Mali (B21EE021): Feature Engineering, modelling with SMOTE, Metric evaluation, report, Website
- Anshul Tomar (B21EE083): Modelling without SMOTE, Hyper Tuned parameters, Confusion Matrix Comparision, n-fold cross validation, report