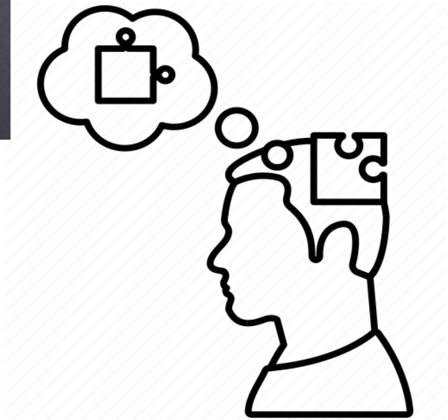
Capstone Project

Cardiovascular Risk Prediction

Mind Benders Team Members

Abdul Aziz
Pooja Yadav
G M Sravya Sree
Abdullah Bin Mohammed



- Introduction to the Subject of Cardiovascular Risk Prediction.
- Problem Statement.
- Data Overview.
- Exploratory Data Analysis.
- Machine Learning Models
- Conclusions.
- Future Work.
- Reference

About Cardiovascular Risk Prediction

- The Framingham Heart Study is a long-term, ongoing cardiovascular cohort study of residents of the city of Framingham, Massachusetts.
- The study began in 1948 with 5,209 adult subjects from Framingham, and is now on its third generation of participants.
- Prior to the study almost nothing was known about the epidemiology of hypertensive or arteriosclerotic cardiovascular disease.
- Much of the now-common knowledge concerning heart disease, such as the effects of diet, exercise, and common medications such as aspirin, is based on this longitudinal study.
- It is a project of the National Heart, Lung, and Blood Institute, in collaboration with (since 1971) Boston University. Various health professionals from the hospitals and universities of Greater Boston staff the project.

Problem Statement

- The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts.
- The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD).
- The dataset provides the patients' information. It includes over 3,000 records and 17 attributes.

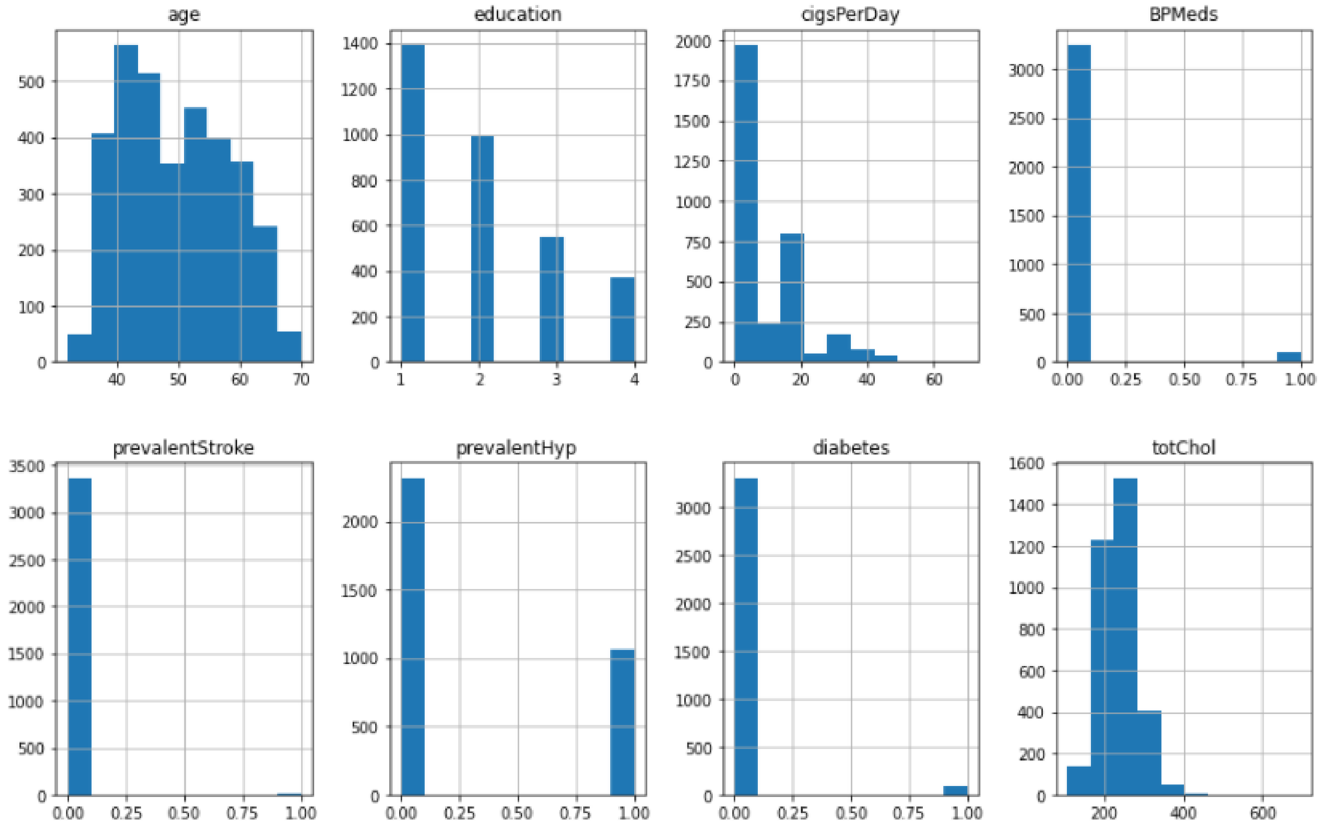
Data Overview

Observations:3390
Features:17

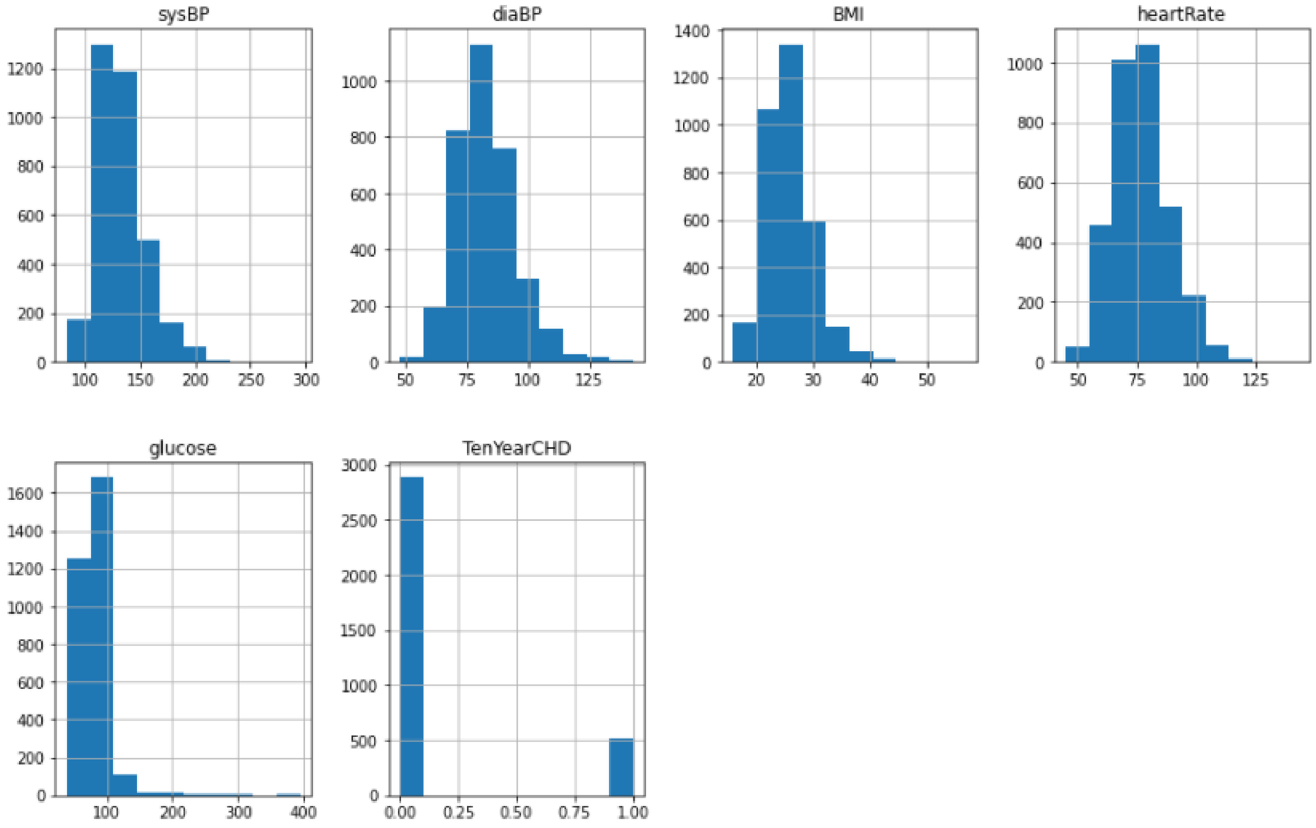
Numerical	Categorical
ID	SEX
AGE	IS_SMOKING
EDUCATION	
CIGS_ PER_DAY	
BP_MEDS	
PREVALENT STROKE	
PREVALENT_HYP	
DIABETES	
TOT_CHOL	
SYS_BP	
DIA_BP	
BMI	
HEART_RATE	
GLUCOSE	

Exploratory Data Analysis

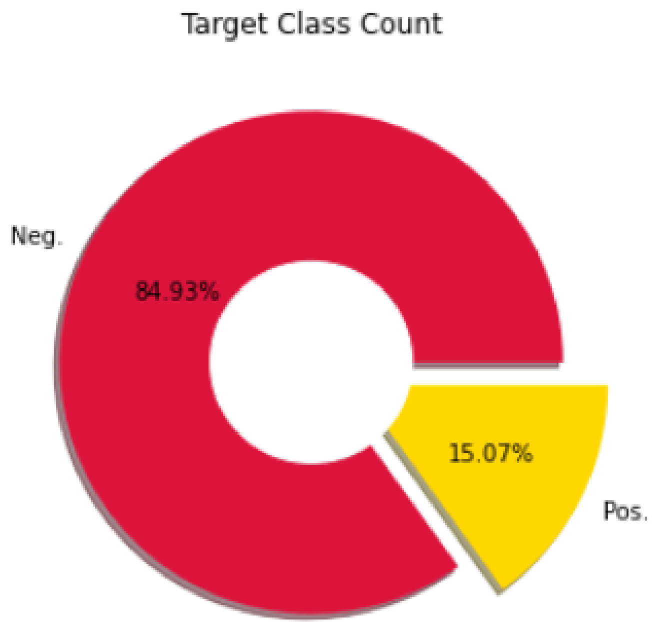
Understanding distribution of data before imputation



EDA (Continued)

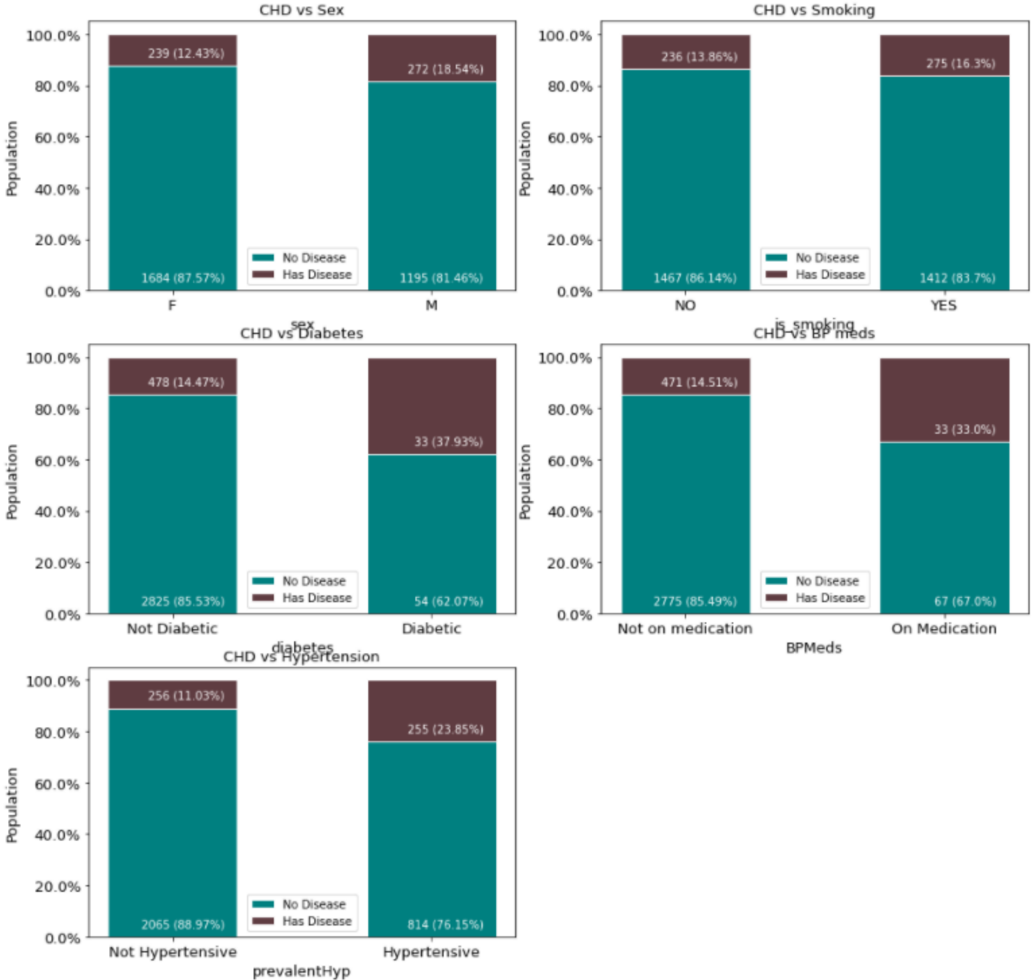


Target Class count, to check balance in data collection



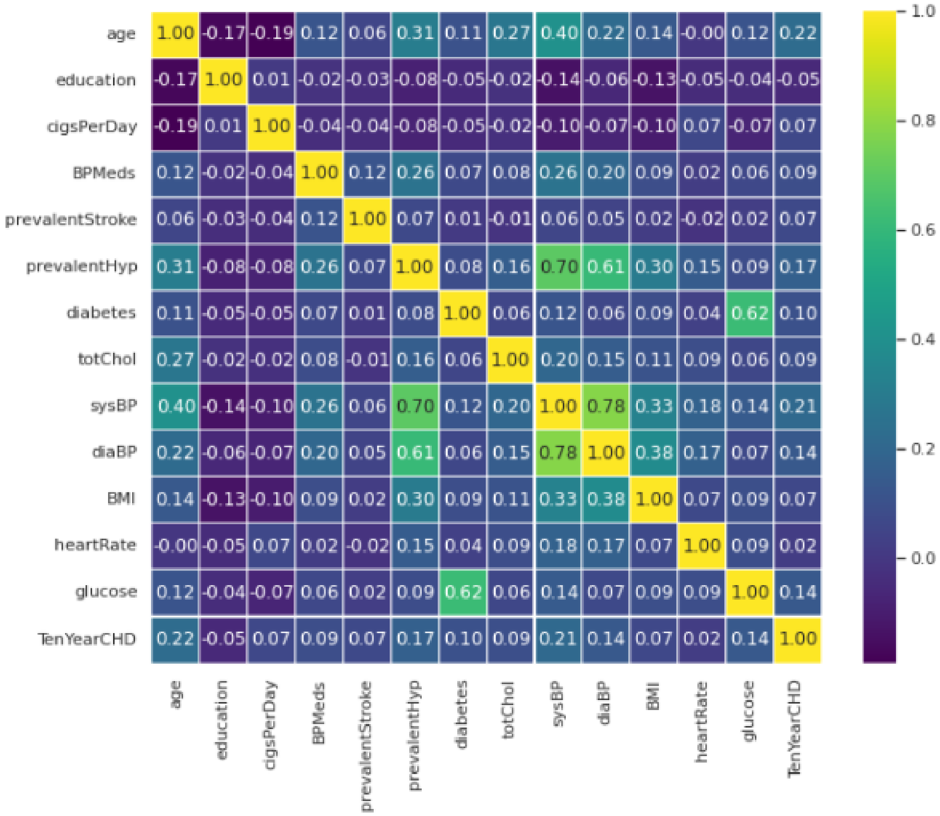
Data is highly imbalanced

stacked bar charts



- Slightly more males are suffering from CHD than females.
- The percentage of people who have CHD is almost equal between smokers and non smokers.
- The percentage of people who have CHD is higher among the diabetic, and those with prevalent hypertension as compared to those who don’t have similar morbidities.
- A larger percentage of the people who have CHD are on blood pressure medication.

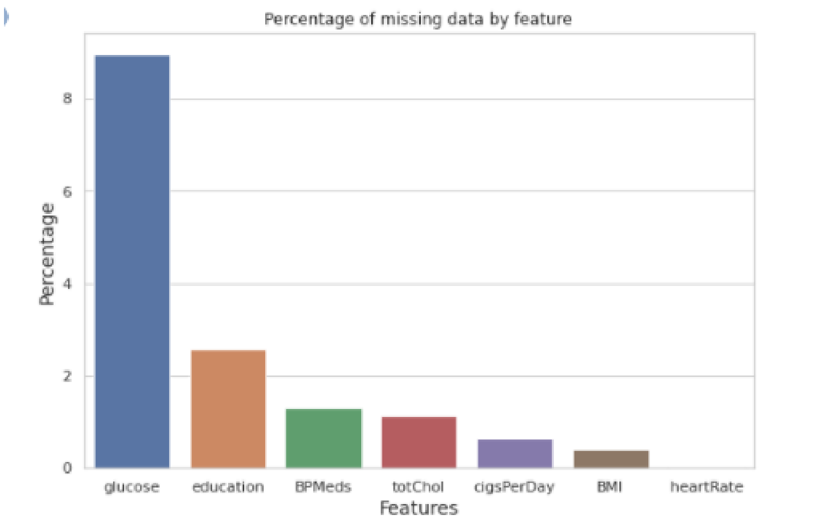
Correlation Matrix



Features that are highly correlated:

- Systolic and diastolic blood pressures
- Cigarette smoking and the number of cigarettes smoked per day

Data Cleaning

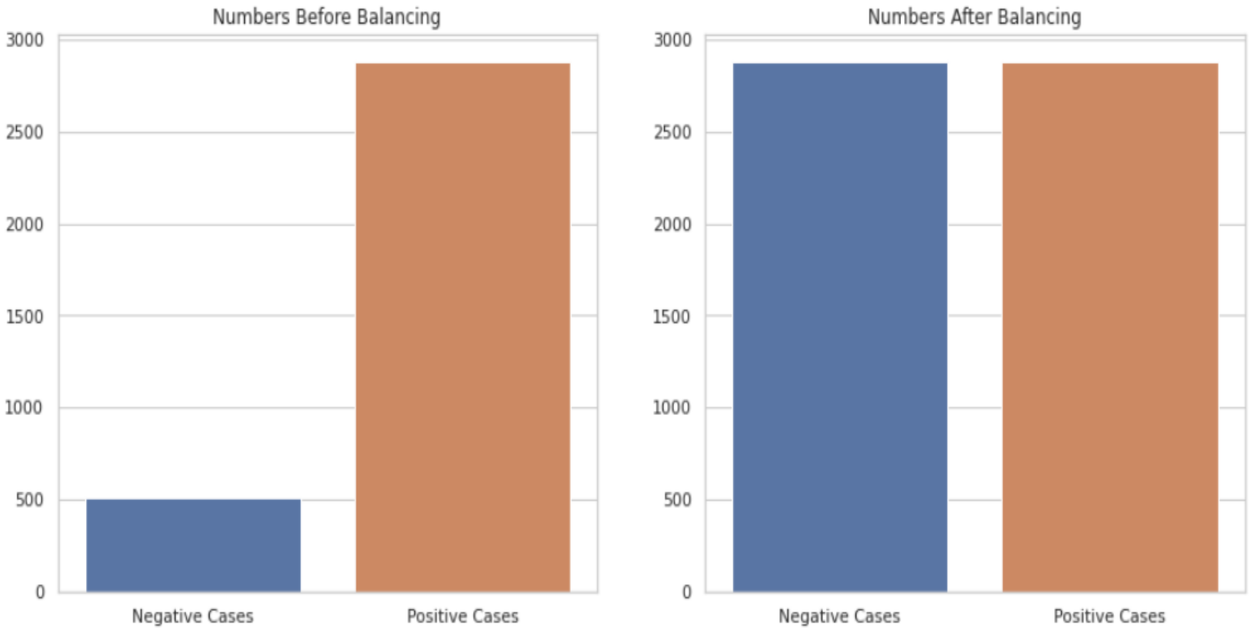


- We used mean, median and KNN imputer to fill the null values.

Cleaned Dataset

	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	64.0	2.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	221.0	148.0	85.0	25.38	90.0	80.0	1.0
1	36.0	4.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	212.0	168.0	98.0	29.77	72.0	75.0	0.0
2	46.0	1.0	0.0	1.0	10.0	0.0	0.0	0.0	0.0	250.0	116.0	71.0	20.35	88.0	94.0	0.0
3	50.0	1.0	1.0	1.0	20.0	0.0	0.0	1.0	0.0	233.0	158.0	88.0	28.26	68.0	94.0	1.0
4	64.0	1.0	0.0	1.0	30.0	0.0	0.0	0.0	0.0	241.0	136.5	85.0	26.42	70.0	77.0	0.0

Data Balancing using SMOTE model



- ❑ Shape:

Original dataset shape : 3390
Resampled dataset shape : 5758

- ❑ Number of values for class 1&0:

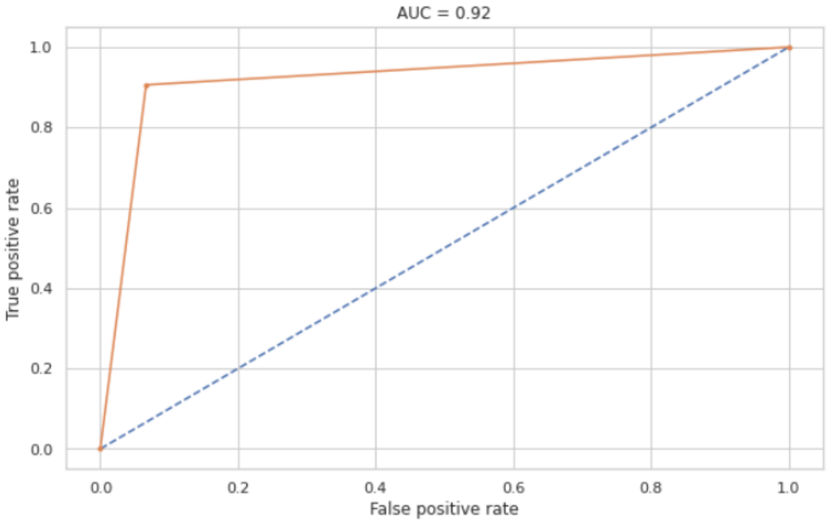
Before : {1.0: 511, 0.0: 2879}
After : {1.0: 2879, 0.0: 2879}

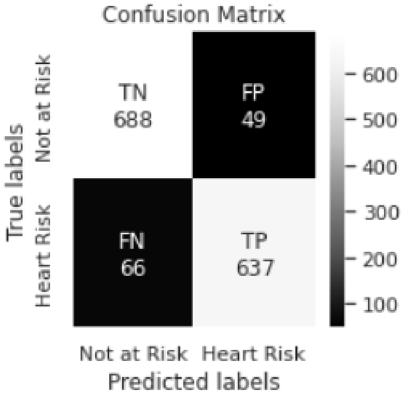
Machine Learning Models

Classification Models

1. SVC
2. K-NN
3. Decision Tree
4. Random Forest
5. Bagging Classifier
6. AdaBoost Classifier
7. XGB Classifier
8. AdaBoost Classifier with SVC
9. CatBoost Classifier
10. Stacking Classifier

SVM





- Kernel chosen for Support vector classifier is radial basis function(rbf).

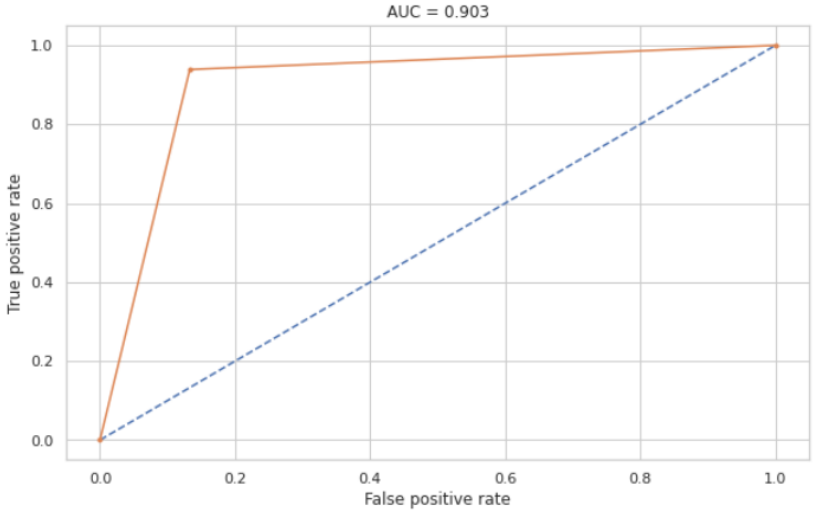
- Grid search cv to improve the performance by tuning the hyper parameters like C and gamma.

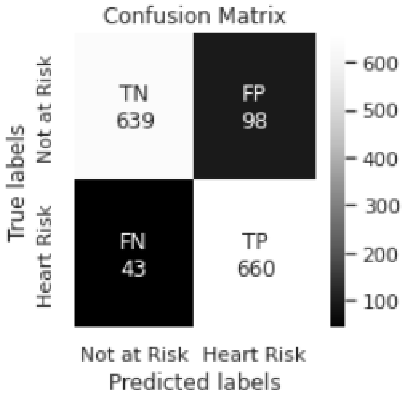
- ROC curve for the data is shown. A good value for area under curve is observed.

- Result:

ROC-AUC: 0.920
Precision: 0.929
Recall: 0.906
F1-Score 0.917
Accuracy 0.920
Cohen's Kappa Score 0.840

KNN

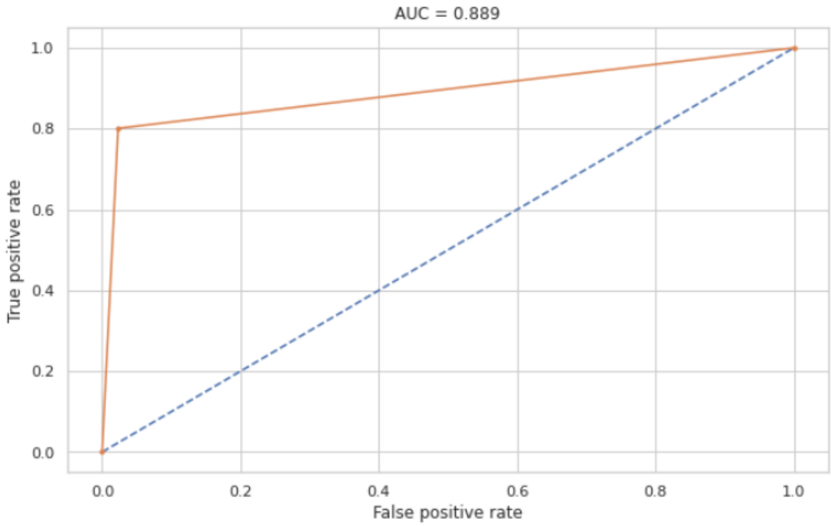


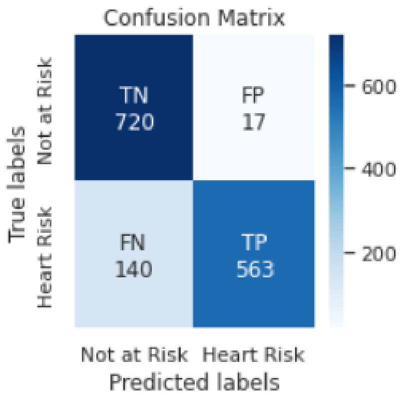


- Implemented Grid search with five fold cv to improve performance.
- Tuned hyper parameters n_neighbors, weights and metric
- ROC curve shows the area under curve is 0.903
- Result:

ROC-AUC: 0.903
Precision: 0.871
Recall: 0.939
F1-Score 0.903
Accuracy 0.902
Cohen's Kappa Score 0.840

DECISION TREE

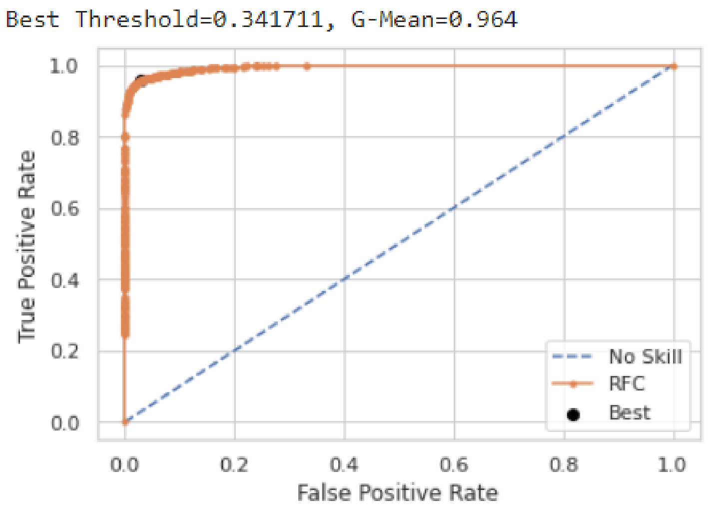


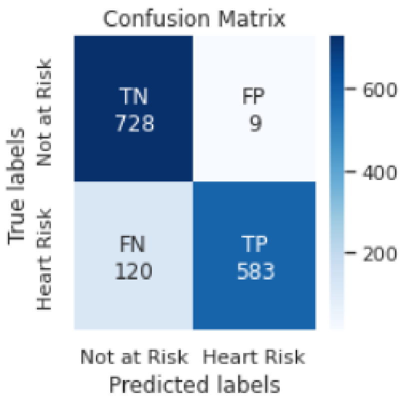


- Grid search four fold cv done to improve the performance
- Tuned the hyper parameters max_depth, min_samples_leaf and criterion.
- A ROC curve is plotted with the test data set predictions.
- Result:

ROC-AUC: 0.889
Precision: 0.971
Recall: 0.801
F1-Score 0.878
Accuracy 0.891
Cohen's Kappa Score 0.781

RANDOM FOREST

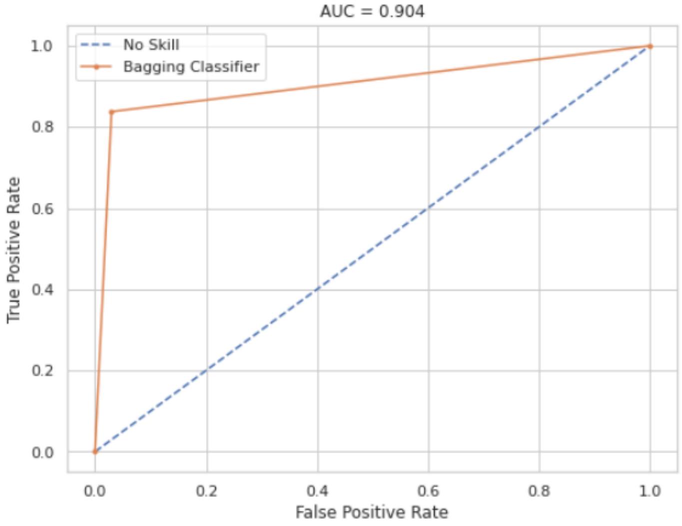


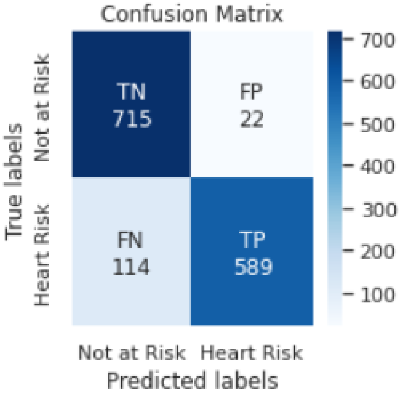


- Fitted a Random forest classifier that use ensembles of decision trees.
- Grid search with five cv done on hyper parameters n_estimators, max_depth, min_samples_leaf, min_samples_leaf and criterion.
- The optimal threshold for differentiating the class is identified with the help of the roc_curve.
- Result:

ROC-AUC: 0.891
Precision: 0.904
Recall: 0.871
F1-Score 0.887
Accuracy 0.892
Cohen's Kappa Score 0.783

BAGGING CLASSIFIER

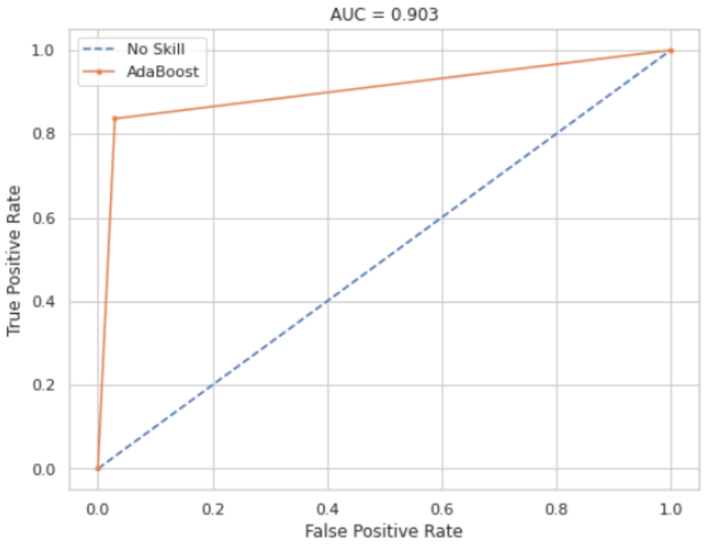


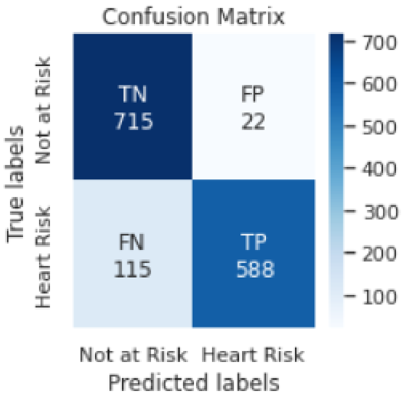


- Another ensemble method i.e Bagging Classifier with the default base classifier of Decision tree.
- Grid search using Repeated stratified five fold cross validation and scoring criterion as roc_auc to find the optimal number of estimators.
- Result:

ROC-AUC: 0.891
Precision: 0.904
Recall: 0.871
F1-Score 0.887
Accuracy 0.892
Cohen's Kappa Score 0.783

ADABOOST CLASSIFIER

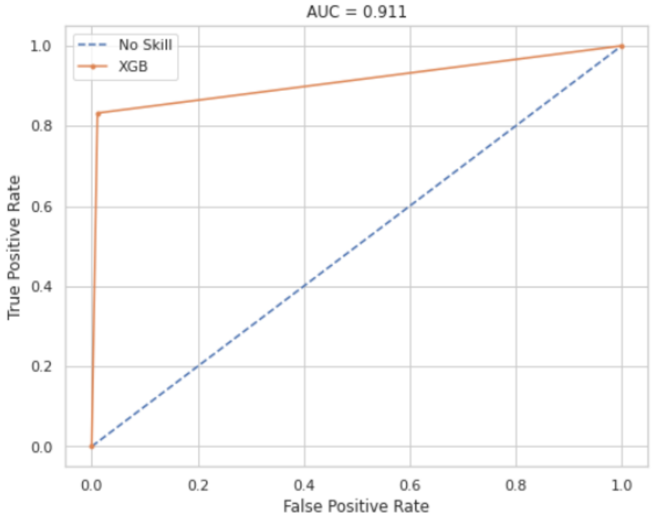


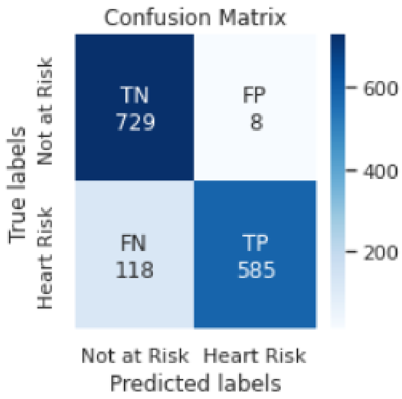


- The default number of estimators i.e 50 gives the optimal score.
- The plot shows the ROC curve for AdaBoost classifier. We get a good value for AUC around 0.9
- Result:

  ROC-AUC: 0.903
  Precision: 0.964
  Recall: 0.836
  F1-Score 0.896
  Accuracy 0.905
  Cohen's Kappa Score 0.809

XGBOOST

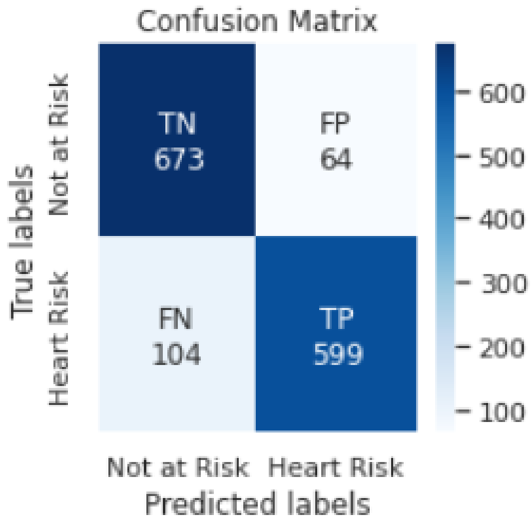




- Tuned hyper parameters using grid search three fold cv
- N_estimators, learning_rate, booster, gamma, alpha, reg_alpha and base score were tuned.
- The plot shows the ROC curve for XGB classifier, a good value for AUC around 0.9
- Result:

ROC-AUC: 0.911
Precision: 0.987
Recall: 0.832
F1-Score 0.903
Accuracy 0.912
Cohen's Kappa Score 0.824

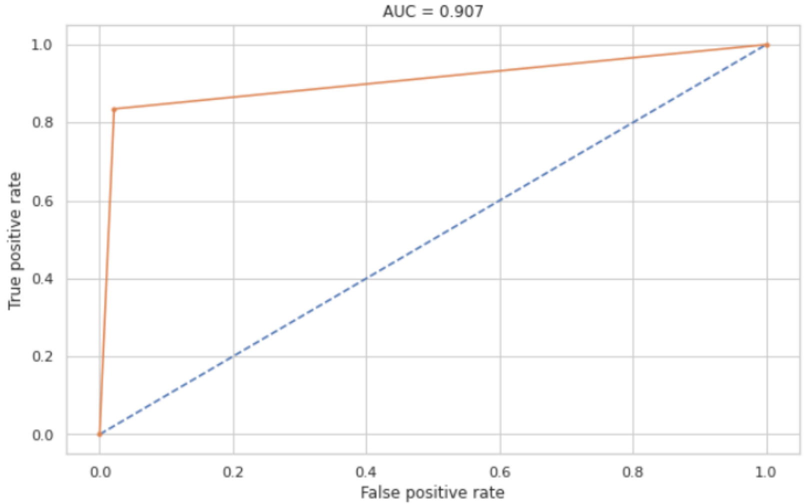
ADABOOST WITH SVC

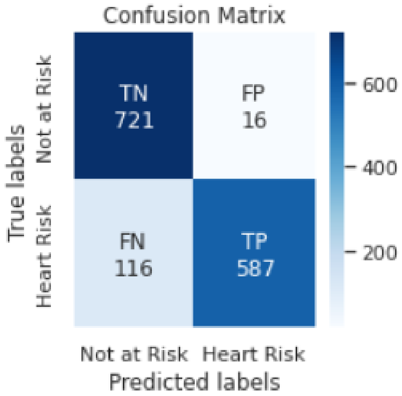


- Adaboost with base estimator Support Vector classifier
- Kernel used is Radial basis function kernel and a learning rate of 0.1 and number of estimators as 20.
- Result:

  ROC-AUC: 0.833
  Precision: 0.903
  Recall: 0.852
  F1-Score 0.877
  Accuracy 0.883
  Cohen's Kappa Score 0.766

CATBOOST Classifier

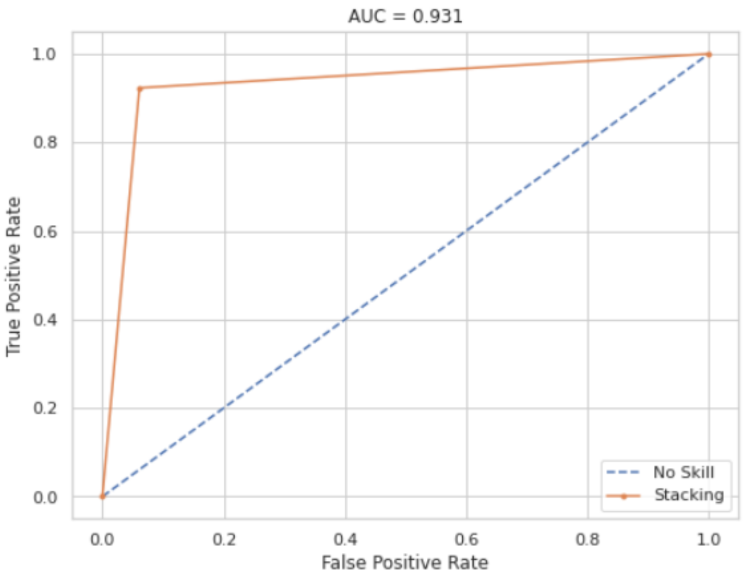


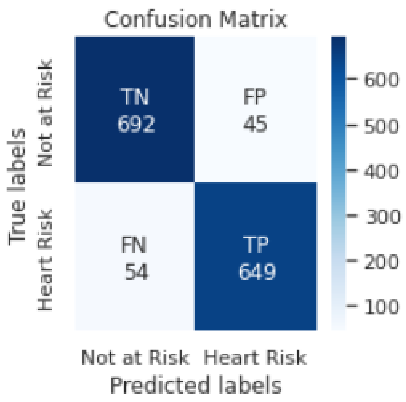


- CatBoost classifier used to fit the train data and predict the test data
- Mean cross validation accuracy is 90%.
- With this model as well we get a good Area under curve value.
- Result:

ROC-AUC: 0.907
Precision: 0.973
Recall: 0.835
F1-Score 0.899
Accuracy 0.908
Cohen's Kappa Score 0.816

STACKING





- Base models are KNN, Decision Tree and SVC
- Meta learner is Logistic Regression.
- Logistic Regression model is trained with input as predictions from base models.
- It has the best Cohen's Kappa score , an indication of the effectiveness of the model against varied datasets.
- Significant improvement in Recall score compared to others.
- Result:

  ROC-AUC: 0.931
  Precision: 0.935
  Recall: 0.923
  F1-Score 0.929
  Accuracy 0.931
  Cohen's Kappa Score 0.862

Conclusion:

	SVC	KNN	DecTree	RF	Bagging	AdaBoost	XGB	AdB_SVC	CatBoost	Stacking
ROC-AUC	0.919815	0.902931	0.888893	0.891180	0.903994	0.903282	0.910647	0.882612	0.906642	0.931064
Precision	0.928571	0.870712	0.970690	0.903988	0.963993	0.963934	0.986509	0.903469	0.973466	0.935159
Recall	0.906117	0.938834	0.800853	0.870555	0.837838	0.836415	0.832148	0.852063	0.834993	0.923186
F1-Score	0.917207	0.903491	0.877631	0.900386	0.896499	0.895659	0.902778	0.877013	0.898928	0.929134
Accuracy	0.920139	0.902083	0.890972	0.891667	0.905556	0.904861	0.912500	0.883333	0.908333	0.931250
Cohens Kappa Score	0.840099	0.804410	0.780939	0.820078	0.810434	0.809033	0.824268	0.766230	0.815961	0.862383

- After training and testing the data on multiple models ranging from simple to ensembled approaches, we have the result for different scores like ROC-AUC, Precision, Recall, F1, Accuracy and Cohen's Kappa score at a single place. On comparing the Cohen's kappa score we can observe that Stacking has the best value followed by Support Vector Classifier. Also if we go back to our problem statement, our goal is to predict the risk of heart disease. In this type of problem our priority should be to reduce the number of False Negatives or find maximum Recall score. If we misclassify someone as having no risk to heart disease, it can be highly detrimental, it can lead to loss of life. Stacking gives us an excellent Recall and at the same time doesn't compromise on Precision. If we require a model with more strict Recall values we can opt for KNN.

Scope of Improvement :

- For future work we need to look for more optimal methods to handle the data imbalance. We can also go for hyper parameter tuning with more range of values and more number of parameters. More sophisticated approach like Neural nets can be used. With stacking we can use more optimal version of the base models and we can check with more combinations of base models.

References

- Kaggle competition
- Analytics vidhya

Thank You!!