4/6/2022

Heart Attack or Myocardial Infarction Detection

CS 699 Data Mining

Osama Muhammad, Sravani Oruganti

Contents

1. **Statement** 4

2. **Description of dataset** 4

3. **Datamining tools** 9

   3.1 **Python** 9

   3.2 **R** 9

   3.3 **Excel** 9

4. **Classification Algorithms** 10

   4.1 **Gaussian Naïve Bayes** 10

   4.2 **Random Forest** 10

   4.3 **Logistic Regression** 10

   4.4 **Decision Tree** 10

   4.5 **Ada Boost** 11

5. **Attribute Selection Methods** 12

   5.1 **F – score** 12

   5.2 **Recursive Feature Elimination** 12

   5.3 **Select From Model** 12

   5.4 **Sequential Feature Selection** 12

   5.5 **Correlation based Feature Selection** 12

6. **Attributes selected by different methods** 13

7. **Data Mining Procedure** 14

   7.1 **Data Preparation** 14

   7.2 **Balancing the data set and Feature Selection** 17

   7.3 **Machine Learning** 17

8. **Data Mining Results & Evaluation** 18

   8.1 **Borderline SMOTE Balanced Data Set** 18

      8.1.1 **Ada Boost** 18

      8.1.2 **Decision Tree** 23

      8.1.3 **Logistic Regression** 28

      8.1.4 **Gaussian Naïve Bayesian** 33

      8.1.5 **Random Forest** 38

8.2 Best Model for Borderline SMOTE Balanced Data Set 43
8.3 SMOTE Balanced Data Set 44
8.3.1 Ada Boost 44
8.3.2 Decision Tree 49
8.3.3 Logistic Regression 54
8.3.4 Gaussian Naïve Bayesian 59
8.3.5 Random Forest 64
8.4 Best Model for SMOTE Balanced Data Set 69
8.5 Running the code 70
9. Conclusion 75
10. Contribution 75
11. References 76

1. Statement

This data contains U.S citizens risk behaviors and preventive health practices that could affect their health status. Our goal with this dataset is to analyze the data using multiple data mining tools with different algorithms and accurately predict the risk of Heart Attack. Our goal is to find an algorithm which can give us a good percentage of true positives and minimize the number of false negatives.

2. Description of dataset

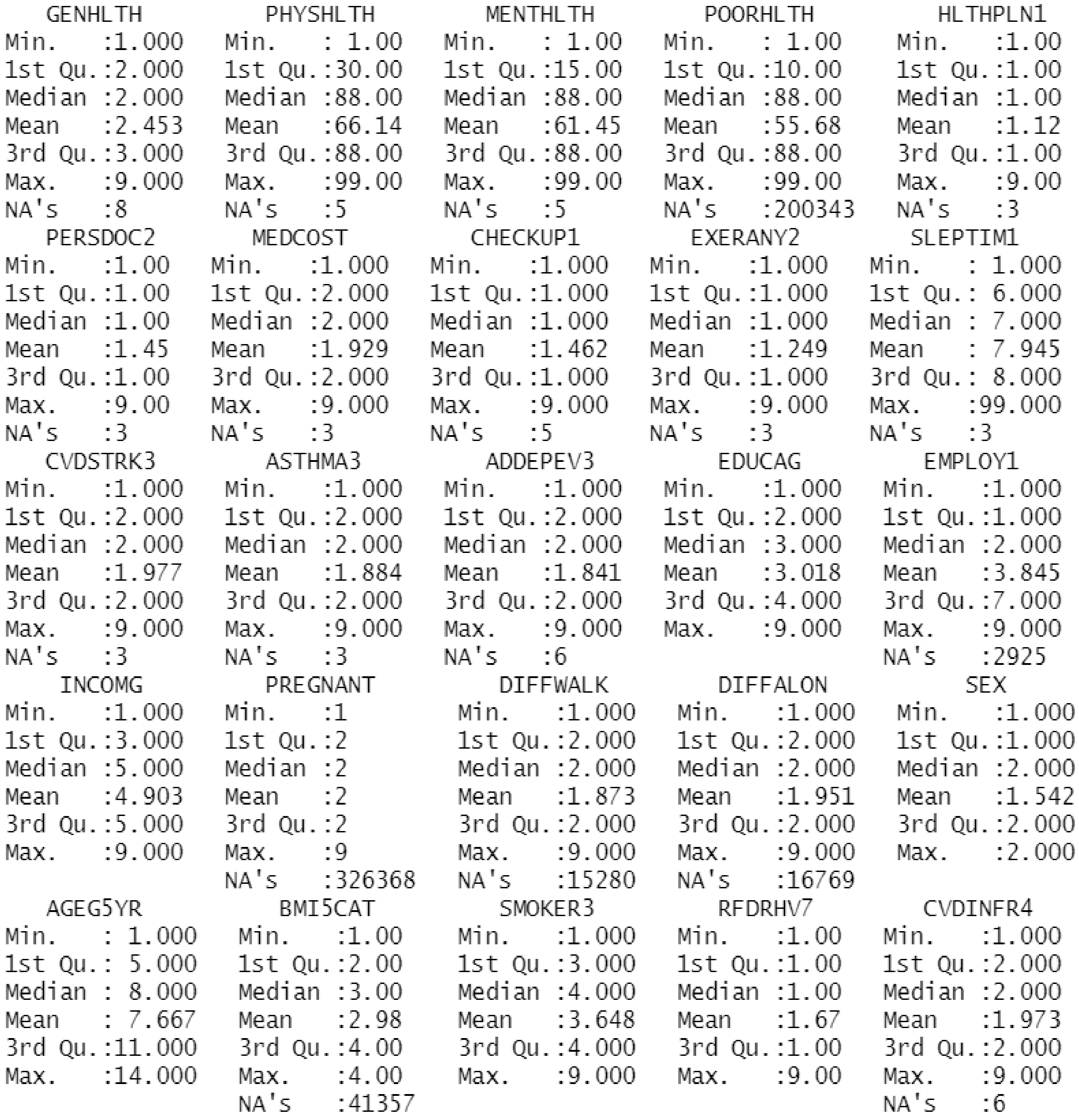
This dataset has been taken from CDC website https://www.cdc.gov/brfss/annual_data/annual_2020.html. The original dataset had many variables which were not related to heart attack detection, so we removed those variables. The selected dataset has 25 variables in total of which 1 is the class or target variable. The dataset has 401,950 rows of data. The following table contains the description of each of the variables (the variable highlighted in **yellow** is the class variable):

Attribute Name	Attribute Description
GENHLTH	Would you say that in general your health is: 1: Excellent, 2: Very Good, 3: Good, 4: Fair, 5: Poor, 7: Don't know / Not sure, 9: Refused, BLANK: Not asked or missing
PHYSHLTH	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? 1 – 30: Number of days, 88: None, 77: Don't know / Not sure, 99: Refused, BLANK: Not asked or missing
MENTHLTH	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? 1 – 30: Number of days, 88: None, 77: Don't know / Not sure, 99: Refused, BLANK: Not asked or missing
POORHLTH	During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities, such as self-care, work, or recreation? 1 – 30: Number of days, 88: None, 77: Don't know / Not sure, 99: Refused, BLANK: Not asked or missing
HLTHPLN1	Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service? 1: Yes, 2: No, 7: Don't know / Not sure, 9: Refused, BLANK: Not asked or missing
PERSDOC2	Do you have one person you think of as your personal doctor or health care provider? (If ´No´ ask ´Is there more than one or is there no person who you think of as your personal doctor or health care provider? ´.) 1: Yes, 2: No, 7: Don't know / Not sure, 9: Refused, BLANK: Not asked or missing

MEDCOST	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 1: Yes, 2: No, 7: Don’t know / Not sure, 9: Refused, BLANK: Not asked or missing
CHECKUP1	About how long has it been since you last visited a doctor for a routine checkup? [A routine checkup is a general physical exam, not an exam for a specific injury, illness, or condition.] 1: Within past year (anytime less than 12 months ago), 2: Within past 2 years (1 year but less than 2 years ago), 3: Within past 5 years (2 years but less than 5years ago), 4: 5 or more years ago, 7: Don’t know / Not sure, 8: Never, 9: Refused, BLANK: Not asked or missing
EXERANY2	During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise? 1: Yes, 2: No, 7: Don’t know / Not sure, 9: Refused, BLANK: Not asked or missing
SLEPTIM1	On average, how many hours of sleep do you get in a 24-hour period? 1 – 24: Number of hours [1-24], 77: Don’t know/Not Sure, 99: Refused, BLANK: Missing
CVDSTRK3	(Ever told) (you had) a stroke. 1: Yes, 2: No, 7: Don’t know / Not sure, 9: Refused, BLANK: Not asked or missing
ASTHMA3	(Ever told) (you had) asthma? 1: Yes, 2: No, 7: Don’t know / Not sure, 9: Refused, BLANK: Not asked or missing
ADDEPEV3	(Ever told) (you had) a depressive disorder (including depression, major depression, dysthymia, or minor depression)? 1: Yes, 2: No, 7: Don’t know / Not sure, 9: Refused, BLANK: Not asked or missing
EDUCAG	What is the highest grade or year of school you completed? 1: Did not graduate High School, 2: Graduated High School, 3: Attended College or Technical School, 4: Graduated from College or Technical School, 9: Don’t know/Not sure/Missing
EMPLOY1	Are you currently…? 1: Employed for wages, 2: Self-employed, 3: Out of work for 1 year or more, 4: Out of work for less than 1 year, 5: A homemaker, 6: A student, 7: Retired, 8: Unable to work, 9: Refused, BLANK: Not asked or missing
INCOMG	Income categories 1: Less than $15,000, 2: $15,000 to less than $25,000, 3: $25,000 to less than $35,000, 4: $35,000 to less than $50,000, 5: $50,000 or more, 9: Don’t know/Not sure/Missing
PREGNANT	To your knowledge, are you now pregnant? 1: Yes, 2: No, 7: Don’t know/Not Sure, 9: Refused, BLANK: Not asked or missing
DIFFWALK	Do you have serious difficulty walking or climbing stairs? 1: Yes, 2: No, 7: Don’t know/Not Sure, 9: Refused, BLANK: Not asked or missing
DIFFALON	Because of a physical, mental, or emotional condition, do you have difficulty doing errands alone such as visiting a doctor´s office or shopping? 1: Yes, 2: No, 7: Don’t know/Not Sure, 9: Refused, BLANK: Not asked or missing
SEX	Calculated sex variable, 1: Mal2, 2: Female

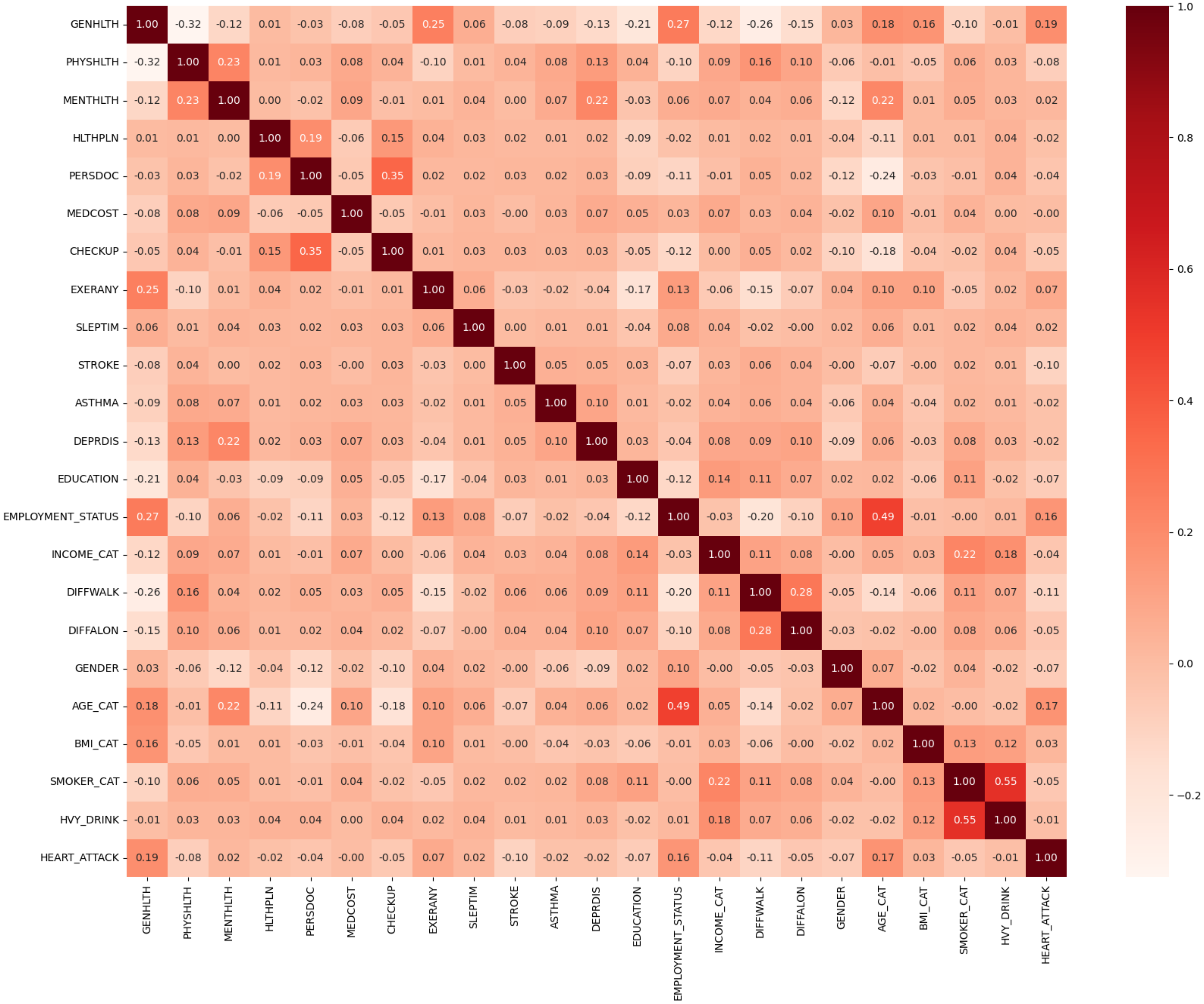
AGEG5YR	Fourteen-level age category 1: Age 18 to 24, 2: Age 25 to 29, 3: Age 30 to 34, 4: Age 35 to 39, 5: Age 40 to 44, 6: Age 45 to 49, 7: Age 50 to 54, 8: Age 55 to 59, 9: Age 60 to 64, 10: Age 65 to 69, 11: Age 70 to 74, 12: Age 75 to 79, 13: Age 80 or older, 14: Don’t know/Refused/Missing
BMI5CAT	Four-categories of Body Mass Index (BMI) 1: Underweight, 2: Normal Weight, 3: Overweight, 4: Obese, BLANK: Don’t know/Refused/Missing
SMOKER3	Four-level smoker status: Every day smoker, Someday smoker, Former smoker, Non-smoker 1: Current smoker -now smokes every day, 2: Current smoker - now smokes some days, 3: Former smoker, 4: Never smoked, 9: Don’t know/Refused/Missing
RFDRHV7	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 1: No, 2: Yes, 9: Don’t know/Refused/Missing
CVDINFR4	(Ever told) you had a heart attack, also called a myocardial infarction? 1: Yes, 2: No, 7: Don’t know / Not sure, 9: Refused, BLANK: Not asked or missing

Below screenshot shows summary of the dataset that is produced using R:



```
   GENHLTH          PHYSHLTH          MENTHLTH          POORHLTH             HLTHPLN1
 Min.   :1.000   Min.   : 1.00   Min.   : 1.00   Min.   : 1.00     Min.   :1.00
 1st Qu.:2.000   1st Qu.:30.00   1st Qu.:15.00   1st Qu.:10.00     1st Qu.:1.00
 Median :2.000   Median :88.00   Median :88.00   Median :88.00     Median :1.00
 Mean   :2.453   Mean   :66.14   Mean   :61.45   Mean   :55.68     Mean   :1.12
 3rd Qu.:3.000   3rd Qu.:88.00   3rd Qu.:88.00   3rd Qu.:88.00     3rd Qu.:1.00
 Max.   :9.000   Max.   :99.00   Max.   :99.00   Max.   :99.00     Max.   :9.00
 NA's   :8       NA's   :5       NA's   :5       NA's   :200343    NA's   :3
   PERSDOC2        MEDCOST          CHECKUP1         EXERANY2         SLEPTIM1
 Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 1.000
 1st Qu.:1.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 6.000
 Median :1.00   Median :2.000   Median :1.000   Median :1.000   Median : 7.000
 Mean   :1.45   Mean   :1.929   Mean   :1.462   Mean   :1.249   Mean   : 7.945
 3rd Qu.:1.00   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.: 8.000
 Max.   :9.00   Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :99.000
 NA's   :3      NA's   :3       NA's   :5       NA's   :3       NA's   :3
   CVDSTRK3         ASTHMA3          ADDEPEV3          EDUCAG           EMPLOY1
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000
 Median :2.000   Median :2.000   Median :2.000   Median :3.000   Median :2.000
 Mean   :1.977   Mean   :1.884   Mean   :1.841   Mean   :3.018   Mean   :3.845
 3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:4.000   3rd Qu.:7.000
 Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :9.000
 NA's   :3       NA's   :3       NA's   :6                       NA's   :2925
    INCOMG           PREGNANT          DIFFWALK          DIFFALON          SEX
 Min.   :1.000   Min.   :1          Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:3.000   1st Qu.:2          1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000
 Median :5.000   Median :2          Median :2.000   Median :2.000   Median :2.000
 Mean   :4.903   Mean   :2          Mean   :1.873   Mean   :1.951   Mean   :1.542
 3rd Qu.:5.000   3rd Qu.:2          3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.000
 Max.   :9.000   Max.   :9          Max.   :9.000   Max.   :9.000   Max.   :2.000
                 NA's   :326368     NA's   :15280   NA's   :16769
   AGEG5YR          BMI5CAT           SMOKER3          RFDRHV7         CVDINFR4
 Min.   : 1.000   Min.   :1.00     Min.   :1.000   Min.   :1.00    Min.   :1.000
 1st Qu.: 5.000   1st Qu.:2.00     1st Qu.:3.000   1st Qu.:1.00    1st Qu.:2.000
 Median : 8.000   Median :3.00     Median :4.000   Median :1.00    Median :2.000
 Mean   : 7.667   Mean   :2.98     Mean   :3.648   Mean   :1.67    Mean   :1.973
 3rd Qu.:11.000   3rd Qu.:4.00     3rd Qu.:4.000   3rd Qu.:1.00    3rd Qu.:2.000
 Max.   :14.000   Max.   :4.00     Max.   :9.000   Max.   :9.00    Max.   :9.000
                  NA's   :41357                                    NA's   :6
```

Below plot is a correlation heat map of the entire data set. From this we can see that the highest correlation among the attributes is 0.55. Also, the predictors do not have very high correlation with the target variable. The predictor variable having the highest correlation value with the target variable (HEART_ATTACK) is GENHLTH and the value is 0.19.



3. Datamining tools

3.1 Python

Python is an open-source language and easy to learn. It is a large library of packages that can help to create data models from scratch. Python can be used to customize a software to an organization’s specifications. Python offers libraries such as Data manipulation, Data visualization, Statistics, Machine learning, etc. Python has libraries like NumPy, Pandas, Matplotlib, etc. for data analysis. It is made for carrying out repetitive tasks and data manipulations. As data mining involves a lot of repetition tasks, it is important to use a tool which handles that.

3.2 R

R is an open-source programming language for statistical computing and graphics. R is one of the 5 languages with an Apache Spark API. R is mostly used among data miners and statisticians for data analysis. It has a large, coherent, integrated collection of intermediate tools for data analysis. R has many libraries that can be used to implement various statistical techniques, spatial and time series analysis, classification, clustering, etc. Another advantage of R is static graphics. It can generate publication-quality graphs with mathematical symbols. R has a group of packages called Tidyverse, which is popularly used for datamining tasks like data import, cleaning, transformation, and visualization. It also has packages for dynamic and interactive graphics.

3.3 Excel

Microsoft Excel is a spread sheet developed to store any kind of data in grid of cells arranged in numbered rows. Excel has many functions like arithmetic operations, display data using graphs, histograms, charts, etc. We can perform arithmetic operations like addition, subtraction, multiplication, division, etc. One can create their own formula to get desired output from a given data.

4. Classification Algorithms

4.1 Gaussian Naïve Bayes

Bayes theorem is used to calculate conditional probability and applied in machine learning for probability. Naïve Bayes classifier is based on bayes theorem. One assumption for naïve bayes classifier is strong independence among the features. This classifier is efficiently trained and need a training data set to estimate the parameters required for classification. It has a simple design and implementation and can be applied in many real-life situations. Gaussian Naïve Bayes is a type of Naïve Bayes algorithm that follows Gaussian normal distribution and supports continuous data. Naïve Bayes makes an assumption that all the attributes are independent of each other.

4.2 Random Forest

Random forest is a machine learning algorithm which combines the output of multiple decision trees to get a single output. It handles both classification and regression problems. It is easy to use and flexible. Random forest builds multiple decision trees using bagging method and merges them together to get single, accurate and stable prediction. Bagging method is a combination of learning models to increase the overall result. The goal of random forest is to reduce variance by averaging multiple decision trees, trained on different parts of the same training data set. Random forest gives output similar to K-fold cross validation.

4.3 Logistic Regression

Logistic Regression is a statistical model that uses a logistic function to model a binary dependent variable along with many more complex extensions. In logistic regression classification, the output can take only discrete values for a given set of inputs. Setting a threshold value is an important aspect of logistic regression and is dependent on the classification problem itself. The threshold value is mainly affected by the values of precision and recall. In applications where we want to reduce the number of false negatives without necessarily reducing the number of false positives, we choose a threshold value with low precision and high recall. In applications where we want to reduce the number of false positives without necessarily reducing number of false negatives, we choose threshold value with high precision and low recall. Logistic regression can be classified as binomial, multinomial and ordinal.

4.4 Decision Tree

Decision tree is like a flow chart in which internal node represents a test on an attribute. Each branch represents an outcome of a test, and each leaf node represents a class label. Paths from root to leaf represents classification rules. One of the important features of decision tree classifier is the capability of capturing descriptive decision-making knowledge from given data. The first node of decision tree is called a root node. We add a node to the tree every time when a question is asked. The result of questions splits the dataset based on the value of a feature and creates new nodes. If there are no further questions or we decide to stop the process after a split, the last nodes are created, and they are called leaf nodes. The goal of decision tree is to continue to split the feature space and apply rules until there are no more rule to apply or no data point left.

4.5 Ada Boost

Ada Boost, also known as adaptive boosting, is one of the boosting classifiers. It combines multiple classifiers to improve the performance of weak classifiers. It is the first boosting algorithm developed for binary classification. The concept of Ada boost is to set the weights of classifiers and training the data sample in each iteration in a way that it ensures the accurate predictions of unusual observations. Ada boost randomly selects the training dataset and iteratively trains the model by selecting the training set based on the accurate prediction of the last training. Observations that are wrongly classified are assigned higher weight so that these observations will get high probably for classification in the next iteration. In each iteration, the weights are assigned to trained classifier based on the accuracy. More accurate classifier gets high weight. This process continues until the complete training data fits without any error or until the maximum specified number of estimators are reached. To classify, perform a vote across all the learning algorithms that were built.

5. Attribute Selection Methods

5.1 F – score

F-score is a measure of model's accuracy on a dataset. It is used to evaluate binary classification systems. It combines precision and recall of a model, and it is defined as harmonic mean of the model's precision and recall. It is commonly used to evaluate information retrieval systems in machine learning models. It is a feature selection method in which it scores each of the features individually. F-score can be adjusted to give more importance to precision over recall or vice-versa.

5.2 Recursive Feature Elimination

Recursive Feature Elimination (RFE) is most popular and easy to use method because it is effective in selecting those features from a dataset that are more relevant in predicting the target variable. This method eliminates the weakest features until the specified number of features are reached. It ranks features by the model's "coef" or "feature importance" attributes and then eliminated a minimum number of features per loop thus removing dependencies and collinearities present in the model and increase the model efficiency.

5.3 Select From Model

Select From Model uses a classifier to calculate the feature importance or coef of the features. A threshold value is used to discard the features. The default threshold value is the mean of feature importance. A feature is kept if the feature importance value is greater than the threshold value. Else, it is discarded.

5.4 Sequential Feature Selection

Sequential Feature Selection adds and removes features from a dataset sequentially. The goal of this algorithm is to improve computational efficiency and reduce the generalization error by removing the irrelevant features or noise. It evaluates each feature separately and selects certain number of features based on individual scores from all the features. In Forward Sequential selection features are added to an empty set until the addition of extra features does not reduce the criterion. Sequentially Backward selection picks all the features from input data and combines them in a set and sequentially removes them from the set until removal of features increases criterion.

5.5 Correlation based Feature Selection

Correlation based feature selection evaluates feature subsets based on correlations. Correlation refers to how close two variables have a linear relationship with each other. When two features have high correlation, one feature among them can be dropped because it gives the same effect as the other.

6. Attributes selected by different methods

We ran the models with 5 random splits into the train and test sets and each of the splits has resulted in different set of attributes. The attributes mentioned below are the ones that were selected when the splitting algorithm was run with random state 600. We used different random states during the 5 splits to make sure that the results were reproducible. The following are the selected attributes:

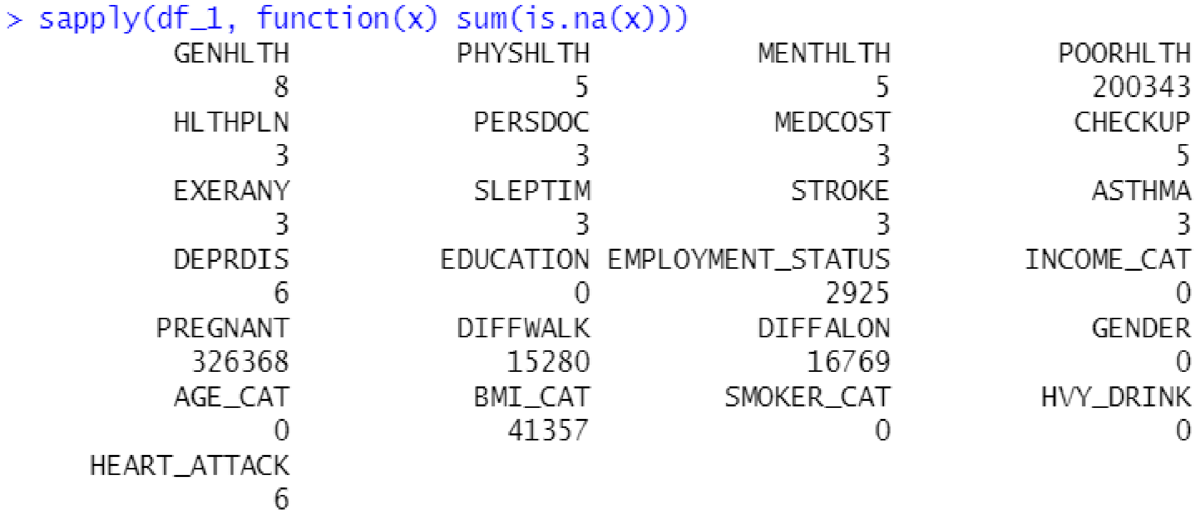
Balancing Method	**Classifier**	**Features Selected**
Borderline SMOTE	Correlation Based Feature Selection	STROKE', 'EXERANY', 'EDUCATION', 'CHECKUP', ‘ PHYSHLTH', 'GENDER', 'DIFFWALK', 'EMPLOYMENT_STATUS', 'GENHLTH', 'AGE_CAT'
	F Score Feature Selection	GENHLTH', 'PHYSHLTH', 'CHECKUP', 'EXERANY', 'STROKE', 'EDUCATION', 'EMPLOYMENT_STATUS', 'DIFFWALK', 'GENDER', 'AGE_CAT'
	Forward Sequential Feature Selection	GENHLTH', 'PERSDOC', 'CHECKUP', 'STROKE', 'EDUCATION', 'EMPLOYMENT_STATUS', 'GENDER', 'AGE_CAT', 'SMOKER_CAT', 'HVY_DRINK'
	Recursive Feature Elimination	GENHLTH', 'PERSDOC', 'CHECKUP', 'STROKE', 'ASTHMA', 'EDUCATION', 'DIFFWALK', 'GENDER', 'AGE_CAT', 'SMOKER_CAT'
	Select From Model Feature Selection	GENHLTH', 'PHYSHLTH', 'EXERANY', 'SLEPTIM', 'STROKE', 'EDUCATION', 'INCOME_CAT', 'GENDER', 'AGE_CAT', 'SMOKER_CAT'
SMOTE	Correlation Based Feature Selection	EXERANY', 'EDUCATION', 'GENDER', 'STROKE', 'CHECKUP', 'PHYSHLTH', 'DIFFWALK', 'EMPLOYMENT_STATUS', 'GENHLTH', 'AGE_CAT'
	F Score Feature Selection	GENHLTH', 'PHYSHLTH', 'CHECKUP', 'EXERANY', 'STROKE', 'EDUCATION', 'EMPLOYMENT_STATUS', 'DIFFWALK', 'GENDER', 'AGE_CAT'
	Forward Sequential Feature Selection	GENHLTH', 'HLTHPLN', 'PERSDOC', 'CHECKUP', 'STROKE', 'EDUCATION', 'GENDER', 'AGE_CAT', 'BMI_CAT', 'SMOKER_CAT'
	Recursive Feature Elimination	GENHLTH', 'PERSDOC', 'MEDCOST', 'CHECKUP', 'STROKE', 'ASTHMA', 'DEPRDIS', 'EDUCATION', 'DIFFWALK', 'AGE_CAT'
	Select From Model Feature Selection	GENHLTH', 'PHYSHLTH', 'EXERANY', 'SLEPTIM', 'STROKE', 'EDUCATION', 'INCOME_CAT', 'GENDER', 'AGE_CAT', 'SMOKER_CAT'

7. Data Mining Procedure

7.1 Data Preparation

The data preparation of the data set was done using R. We mainly used the R Tidyverse library for doing the data preparation. The following are the steps followed during data preparation:

1. We loaded the data set in R which is a CSV file named “Heart Attack Detection Dataset.csv”.
2. Then we renamed the columns in the data set so it would be easy to understand what each column represents.
3. Then we check the dimensions of the data set which were 401958 rows and 25 columns including the target variable.
4. Then we checked for null values in the entire dataset. The results are shown below:



```
> sapply(df_1, function(x) sum(is.na(x)))
          GENHLTH          PHYSHLTH          MENTHLTH          POORHLTH 
                8                 5                 5            200343 
          HLTHPLN           PERSDOC           MEDCOST           CHECKUP 
                3                 3                 3                 5 
          EXERANY           SLEPTIM            STROKE            ASTHMA 
                3                 3                 3                 3 
          DEPRDIS         EDUCATION EMPLOYMENT_STATUS        INCOME_CAT 
                6                 0              2925                 0 
         PREGNANT          DIFFWALK          DIFFALON            GENDER 
           326368             15280             16769                 0 
          AGE_CAT           BMI_CAT        SMOKER_CAT         HVY_DRINK 
                0             41357                 0                 0 
     HEART_ATTACK 
                6 
```

5. We filtered the data for rows with class attribute HEART_ATTACK = 1 & HEART_ATTACK = 2.
6. We can see from the above results that the columns POORHLTH & PREGNANT have a very large number of null values. Therefore, we just removed those columns from the dataset. For the rest of the columns with the null values, we imputed the median value of those columns in place of the null values with respect to the class of each null value.
7. We swapped the labels of class HEART_ATTACK = 1 & HEART_ATTACK = 2 so that HEART_ATTACK = 1 meant “no risk of a heart attack” and HEART_ATTACK = 2 meant “at risk of a heart attack”. This produced the full clean data set.

The cleaned data set has 377918 rows for class HEART_ATTACK = 1 and 21957 rows for class HEART_ATTACK = 2 and 23 rows including the target variable. The following table below shows the final set of attributes and their description:

Attribute Name	Attribute Description
GENHLTH	Would you say that in general your health is: 1: Excellent, 2: Very Good, 3: Good, 4: Fair, 5: Poor, 7: Don't know / Not sure, 9: Refused
PHYSHLTH	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? 1 – 30: Number of days, 88: None, 77: Don't know / Not sure, 99: Refused
MENTHLTH	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good? 1 – 30: Number of days, 88: None, 77: Don't know / Not sure, 99: Refused
HLTHPLN	Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service? 1: Yes, 2: No, 7: Don't know / Not sure, 9: Refused
PERSDOC	Do you have one person you think of as your personal doctor or health care provider? (If ´No´ ask ´Is there more than one or is there no person who you think of as your personal doctor or health care provider? ´.) 1: Yes, 2: No, 7: Don't know / Not sure, 9: Refused
MEDCOST	Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 1: Yes, 2: No, 7: Don't know / Not sure, 9: Refused
CHECKUP	About how long has it been since you last visited a doctor for a routine checkup? [A routine checkup is a general physical exam, not an exam for a specific injury, illness, or condition.] 1: Within past year (anytime less than 12 months ago), 2: Within past 2 years (1 year but less than 2 years ago), 3: Within past 5 years (2 years but less than 5years ago), 4: 5 or more years ago, 7: Don't know / Not sure, 8: Never, 9: Refused
EXERANY	During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise? 1: Yes, 2: No, 7: Don't know / Not sure, 9: Refused
SLEPTIM	On average, how many hours of sleep do you get in a 24-hour period? 1 – 24: Number of hours [1-24], 77: Don't know/Not Sure, 99: Refused
STROKE	(Ever told) (you had) a stroke. 1: Yes, 2: No, 7: Don't know / Not sure, 9: Refused
ASTHMA	(Ever told) (you had) asthma? 1: Yes, 2: No, 7: Don't know / Not sure, 9: Refused

DEPRDIS	(Ever told) (you had) a depressive disorder (including depression, major depression, dysthymia, or minor depression)? 1: Yes, 2: No, 7: Don’t know / Not sure, 9: Refused
EDUCATION	What is the highest grade or year of school you completed? 1: Did not graduate High School, 2: Graduated High School, 3: Attended College or Technical School, 4: Graduated from College or Technical School, 9: Don’t know/Not sure/Missing
EMPLOYMENT_STATUS	Are you currently…? 1: Employed for wages, 2: Self-employed, 3: Out of work for 1 year or more, 4: Out of work for less than 1 year, 5: A homemaker, 6: A student, 7: Retired, 8: Unable to work, 9: Refused
INCOME_CAT	Income categories 1: Less than $15,000, 2: $15,000 to less than $25,000, 3: $25,000 to less than $35,000, 4: $35,000 to less than $50,000, 5: $50,000 or more, 9: Don’t know/Not sure/Missing
DIFFWALK	Do you have serious difficulty walking or climbing stairs? 1: Yes, 2: No, 7: Don’t know/Not Sure, 9: Refused
DIFFALON	Because of a physical, mental, or emotional condition, do you have difficulty doing errands alone such as visiting a doctor´s office or shopping? 1: Yes, 2: No, 7: Don’t know/Not Sure, 9: Refused
GENDER	Sex variable, 1: Male, 2: Female
AGE_CAT	Fourteen-level age category 1: Age 18 to 24, 2: Age 25 to 29, 3: Age 30 to 34, 4: Age 35 to 39, 5: Age 40 to 44, 6: Age 45 to 49, 7: Age 50 to 54, 8: Age 55 to 59, 9: Age 60 to 64, 10: Age 65 to 69, 11: Age 70 to 74, 12: Age 75 to 79, 13: Age 80 or older, 14: Don’t know/Refused/Missing
BMI_CAT	Four-categories of Body Mass Index (BMI) 1: Underweight, 2: Normal Weight, 3: Overweight, 4: Obese
SMOKER_CAT	Four-level smoker status: Every day smoker, Someday smoker, Former smoker, Non-smoker 1: Current smoker -now smokes every day, 2: Current smoker -now smokes some days, 3: Former smoker, 4: Never smoked, 9: Don’t know/Refused/Missing
HVY_DRINK	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 1: No, 2: Yes, 9: Don’t know/Refused/Missing
HEART_ATTACK	(Ever told) you had a heart attack, also called a myocardial infarction? 1: Yes, 2: No

7.2 Balancing the data set and Feature Selection

Python was used for this part of the project. The cleaned dataset was loaded into python and then min max scaled between 1 and 2. Then the data was split into training and testing parts in the ratio of 66% to 34% respectively.

Then we performed balancing of the training dataset using the imbalanced learn library in python. The training data set was highly imbalanced as the original data set was also highly unbalanced. Therefore, we used two oversampling techniques (SMOTE & Borderline SMOTE) to balance the training dataset.

Once the training dataset was balanced, we performed 5 different feature selection techniques on the training dataset which are as follows:

1. Correlation based feature selection
2. F Score feature selection
3. Forward sequential feature selection
4. Recursive feature elimination
5. Select from model feature selection

These techniques produced 5 different feature selected data sets which are then used in machine learning.

7.3 Machine Learning

We used each of the 5 feature selected data sets with 5 different classifiers. The 5 classifiers that we used are:

1. Ada Boost with Logistic Regression as base estimator
2. Decision Tree Classifier
3. Logistic Regression
4. Gaussian Naïve Bayesian
5. Random Forest Classifier

These classifiers were fit on the balanced feature selected training datasets and then tested on the independent unbalanced testing data set. Then the metrics were calculated for each of these classifiers.

8. Data Mining Results & Evaluation

We ran each algorithm 5 times for each feature selection method. The columns random state shows the different values that were used to run each time. These values were set so that the results are reproducible. Then we averaged the results over the 5 runs of the algorithms. The confusion matrices were also averaged so the values have been rounded to the nearest whole number. As we used two balanced methods, we analyzed results from both the methods.

8.1 Borderline SMOTE Balanced Data Set

8.1.1 Ada Boost

This classifier was implemented in python with Logistic Regression as the base estimator. The default max number of base estimators, 50 were used to run this algorithm.

Feature Selection	Class	Accuracy	TPR	FPR	Precision	Recall	F1 Score	MCC	ROC
Correlation based feature selection	Heart Attack = 1	0.943644361	0.998132171	0.001867829	0.945299146	0.998132171	0.970997406	0.01382565	0.512735356
	Heart Attack = 2	0.943644361	0.004583292	0.995416708	0.125296482	0.004583292	0.008839774	0.01382565	0.512735356
	Weighted	0.943644361	0.943644361	0.056355639	0.900326299	0.943644361	0.918231565	0.01382565	0.512735356
F Score feature selection	Heart Attack = 1	0.941989438	0.996269284	0.003730716	0.945303092	0.996269284	0.970117092	0.010193589	0.57751485
	Heart Attack = 2	0.941989438	0.006521094	0.993478906	0.091946692	0.006521094	0.012177457	0.010193589	0.57751485
	Weighted	0.941989438	0.941989438	0.058010562	0.898502344	0.941989438	0.917581872	0.010193589	0.57751485
Forward Sequential Feature Selection	Heart Attack = 1	0.942549905	0.996795631	0.003204369	0.945390942	0.996795631	0.970411522	0.016775978	0.653372377
	Heart Attack = 2	0.942549905	0.007678544	0.992321456	0.129027063	0.007678544	0.014134905	0.016775978	0.653372377
	Weighted	0.942549905	0.942549905	0.057450095	0.900620089	0.942549905	0.917966984	0.016775978	0.653372377
Recursive Feature Elimination	Heart Attack = 1	0.941318643	0.995303511	0.004696489	0.945487607	0.995303511	0.969751133	0.022000973	0.639595686
	Heart Attack = 2	0.941318643	0.011036088	0.988963912	0.144768779	0.011036088	0.019508946	0.022000973	0.639595686
	Weighted	0.941318643	0.941318643	0.058681357	0.901598677	0.941318643	0.917631732	0.022000973	0.639595686
Select from Model	Heart Attack = 1	0.910297298	0.955370616	0.044629384	0.950100717	0.955370616	0.952437056	0.090141901	0.699211019
	Heart Attack = 2	0.910297298	0.133029812	0.866970188	0.151948227	0.133029812	0.126473479	0.090141901	0.699211019
	Weighted	0.910297298	0.910297298	0.089702702	0.906328444	0.910297298	0.907152289	0.090141901	0.699211019

Correlation Based Feature Selection

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	128262	240
HEART ATTACK = 2	7422	34

F Score Feature Selection

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	128022	479
HEART ATTACK = 2	7408	49

Forward Sequential Feature Selection

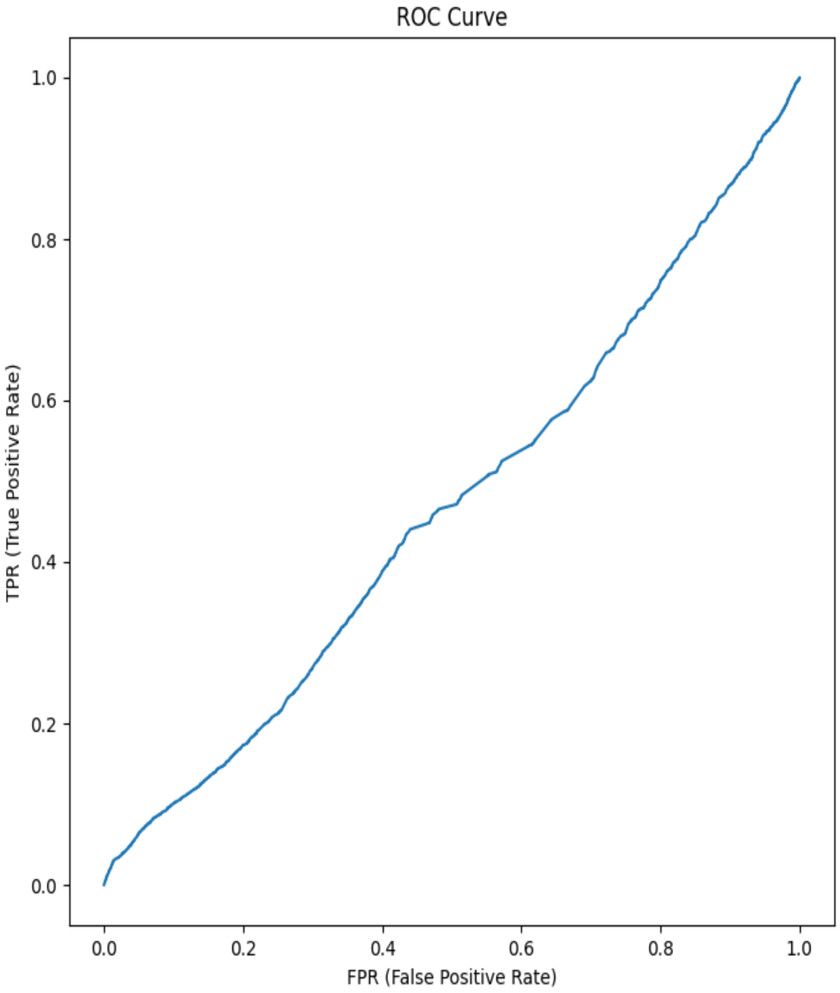
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	128090	412
HEART ATTACK = 2	7399	57

Recursive Feature Elimination

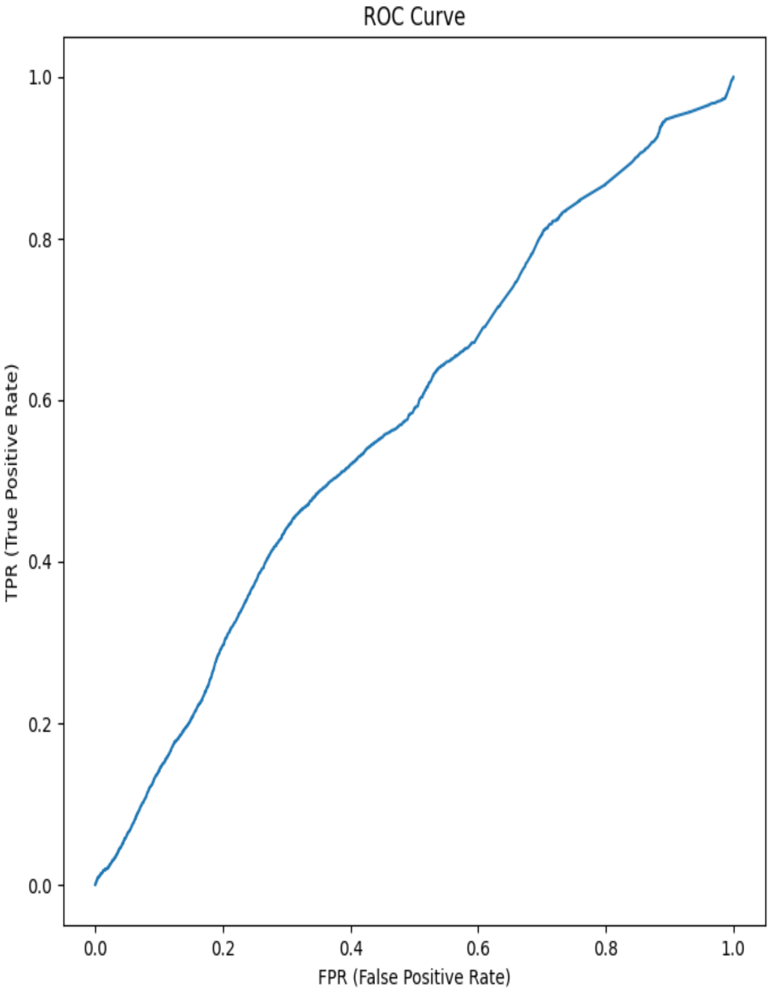
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	127898	604
HEART ATTACK = 2	7374	82

Select From Model Feature Selection

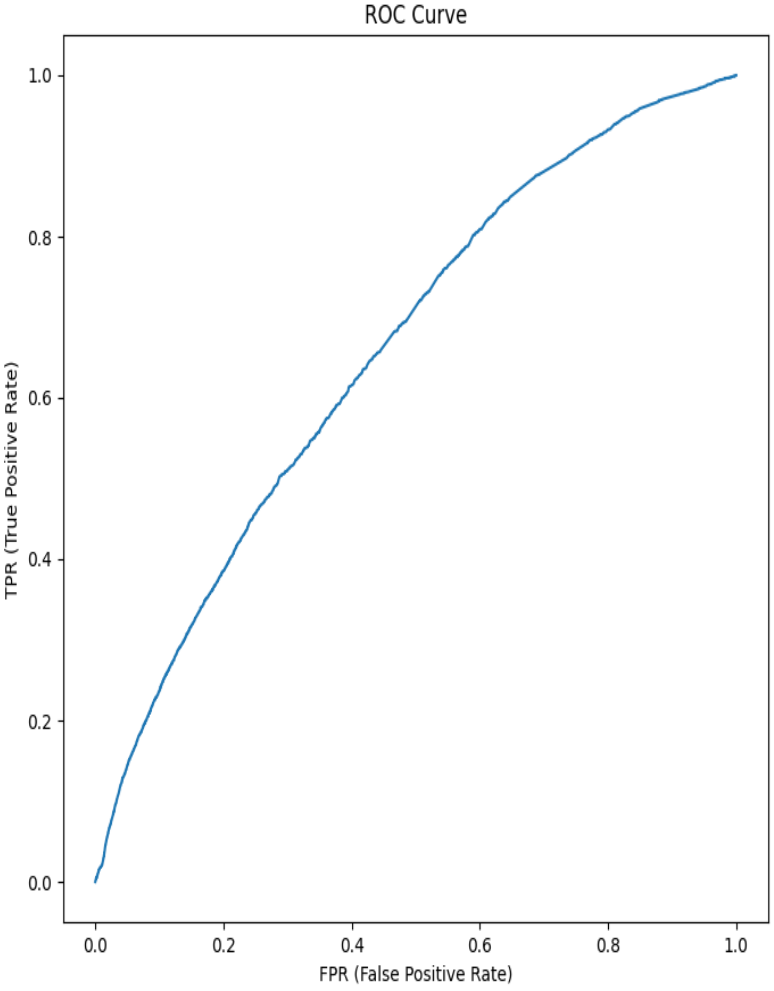
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	122768	5734
HEART ATTACK = 2	6462	994



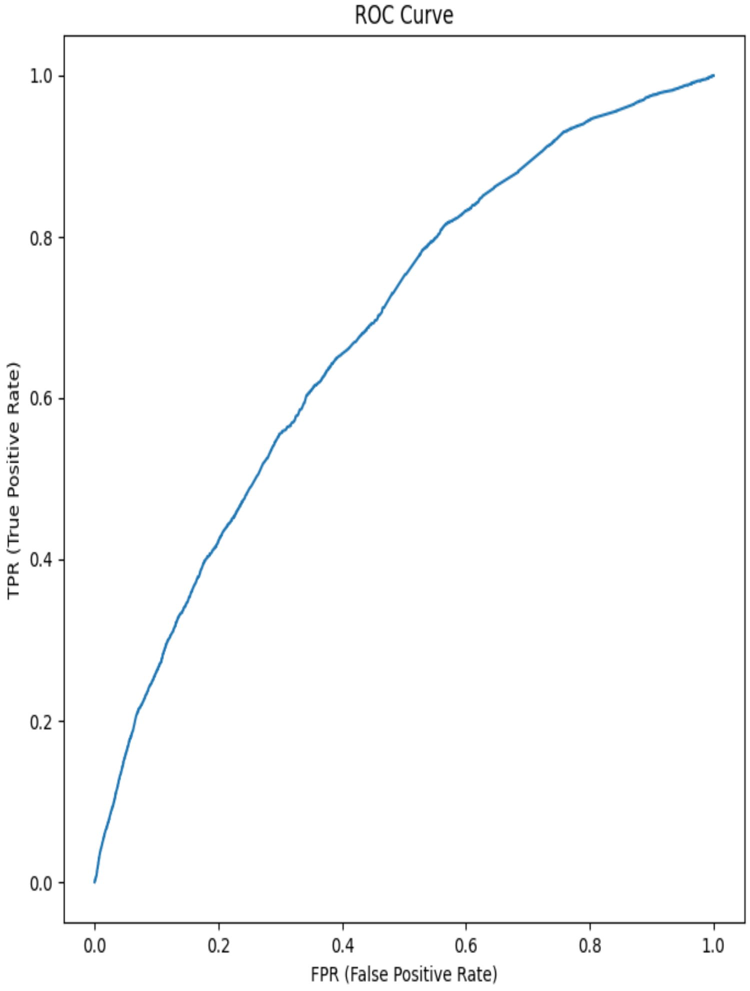
Correlation Based Feature Selection ROC Curve



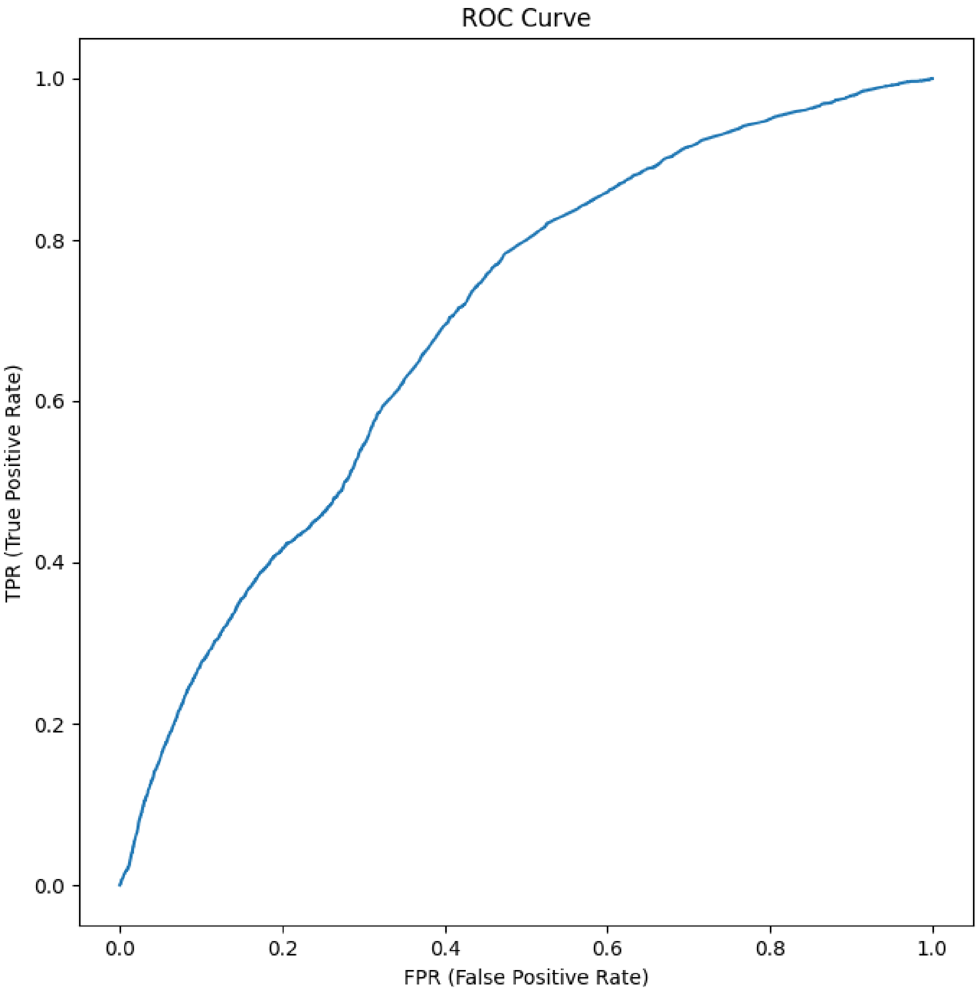
F Score Feature Selection ROC Curve



Forward Sequential Feature Selection ROC Curve



Recursive Feature Elimination



Select from Model Feature Selection

8.1.2 Decision Tree

This classifier was implemented in python using entropy as the measure of quality of the split

Feature Selection	Class	Accuracy	TPR	FPR	Precision	Recall	F1 Score	MCC	ROC
Correlation based feature selection	Heart Attack = 1	0.368472617	0.348251467	0.651748533	0.942249441	0.348251467	0.46053387	0.021764693	0.530846926
	Heart Attack = 2	0.368472617	0.713680245	0.286319755	0.060409347	0.713680245	0.110564444	0.021764693	0.530846926
	Weighted	0.368472617	0.368472617	0.631527383	0.893896239	0.368472617	0.44145991	0.021764693	0.530846926
F Score feature selection	Heart Attack = 1	0.610020742	0.627077794	0.372922206	0.937958858	0.627077794	0.743691443	-0.027591147	0.470803203
	Heart Attack = 2	0.610020742	0.313790602	0.686209398	0.049192303	0.313790602	0.084235987	-0.027591147	0.470803203
	Weighted	0.610020742	0.610020742	0.389979258	0.889220582	0.610020742	0.707593075	-0.027591147	0.470803203
Forward Sequential Feature Selection	Heart Attack = 1	0.680352756	0.704337489	0.295662511	0.94164905	0.704337489	0.796269083	-0.007632153	0.486413521
	Heart Attack = 2	0.680352756	0.268458523	0.731541477	0.058852476	0.268458523	0.090271269	-0.007632153	0.486413521
	Weighted	0.680352756	0.680352756	0.319647244	0.89324416	0.680352756	0.757524885	-0.007632153	0.486413521
Recursive Feature Elimination	Heart Attack = 1	0.526452287	0.527989715	0.472010285	0.945780352	0.527989715	0.667219014	0.013359905	0.51312783
	Heart Attack = 2	0.526452287	0.498337831	0.501662169	0.060990579	0.498337831	0.107606474	0.013359905	0.51312783
	Weighted	0.526452287	0.526452287	0.473547713	0.897257369	0.526452287	0.63658044	0.013359905	0.51312783
Select from Model	Heart Attack = 1	0.640587534	0.652842971	0.347157029	0.951157354	0.652842971	0.771315026	0.041820438	0.541617839
	Heart Attack = 2	0.640587534	0.430152752	0.569847248	0.069925432	0.430152752	0.11964779	0.041820438	0.541617839
	Weighted	0.640587534	0.640587534	0.359412466	0.90282836	0.640587534	0.73556036	0.041820438	0.541617839

Correlation Based Feature Selection

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	44764	83738
HEART ATTACK = 2	2123	5333

F Score Feature Selection

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	80593	47909
HEART ATTACK = 2	5112	2344

Forward Sequential Feature Selection

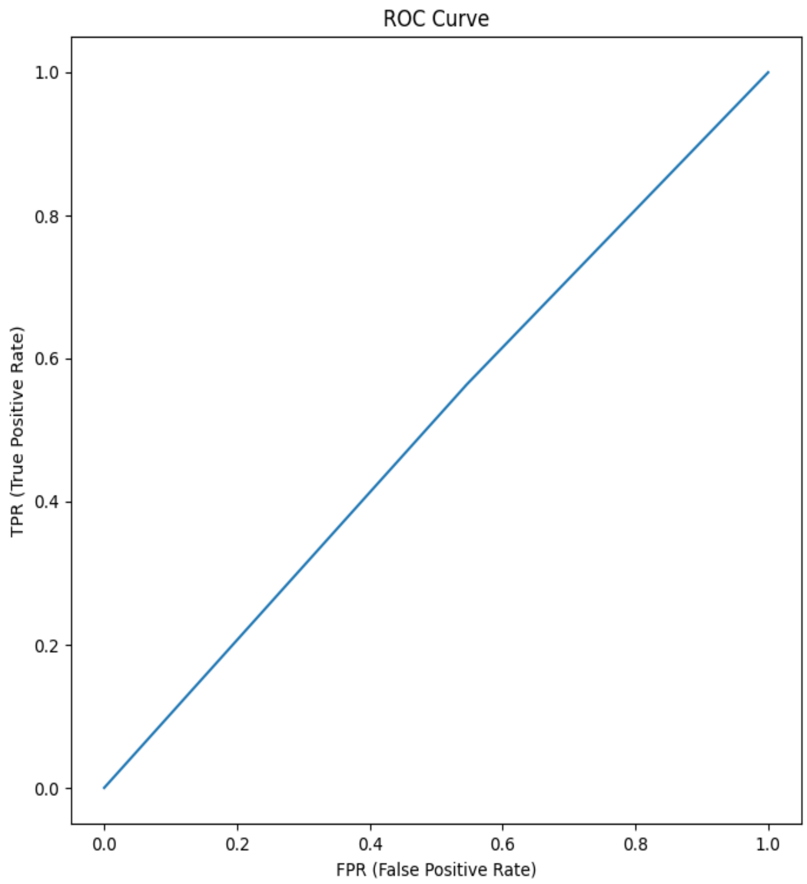
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	90502	38000
HEART ATTACK = 2	5458	1998

Recursive Feature Elimination

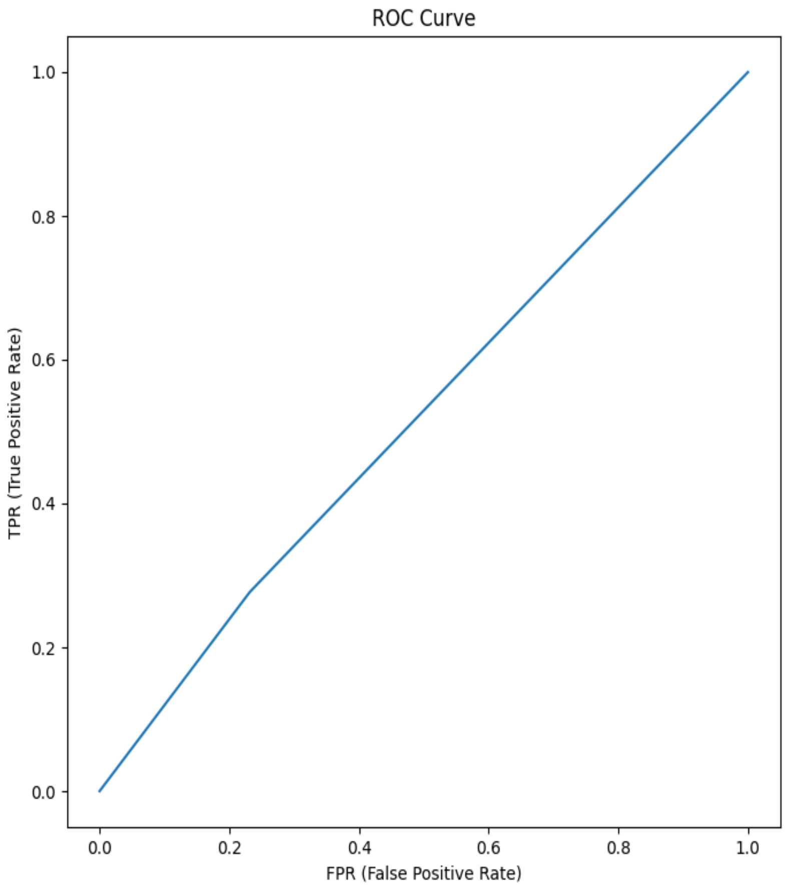
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	67857	60645
HEART ATTACK = 2	3738	3718

Select From Model Feature Selection

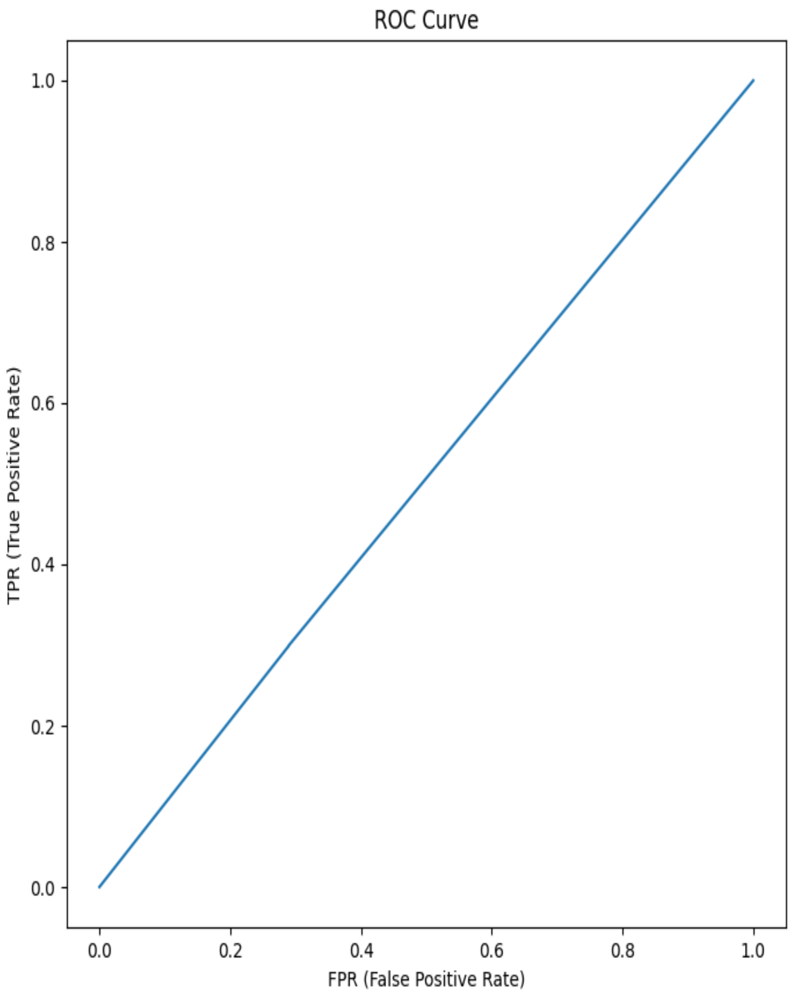
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	83889	44612
HEART ATTACK = 2	4253	3204



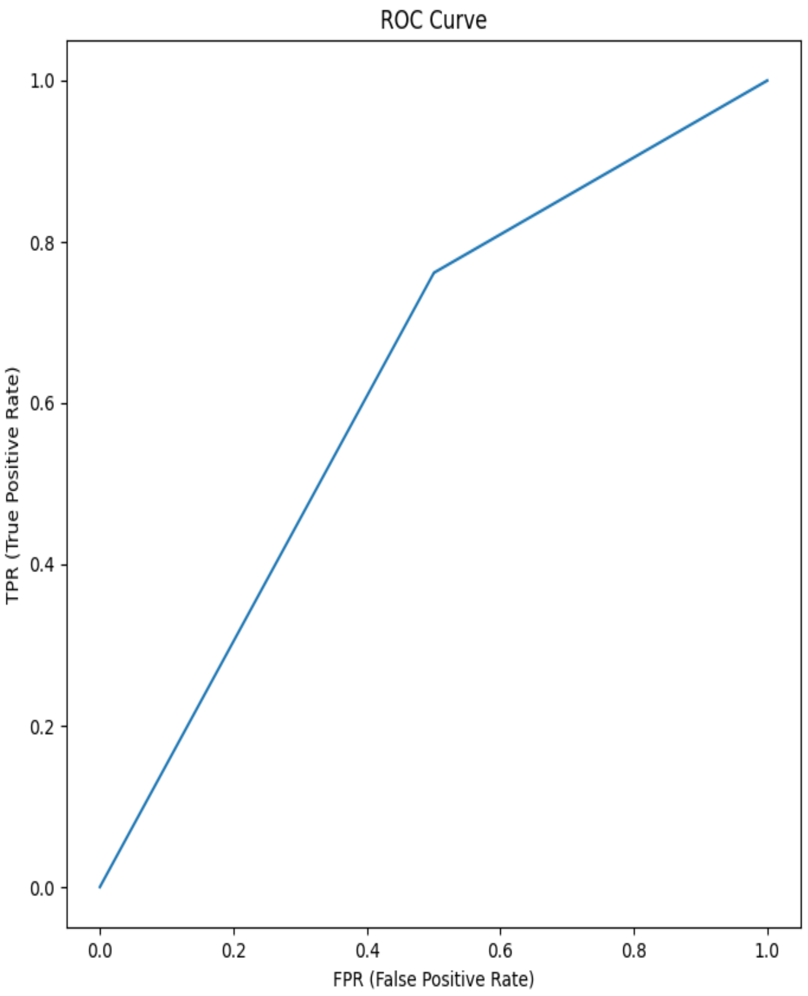
Correlation Based Feature Selection ROC Curve



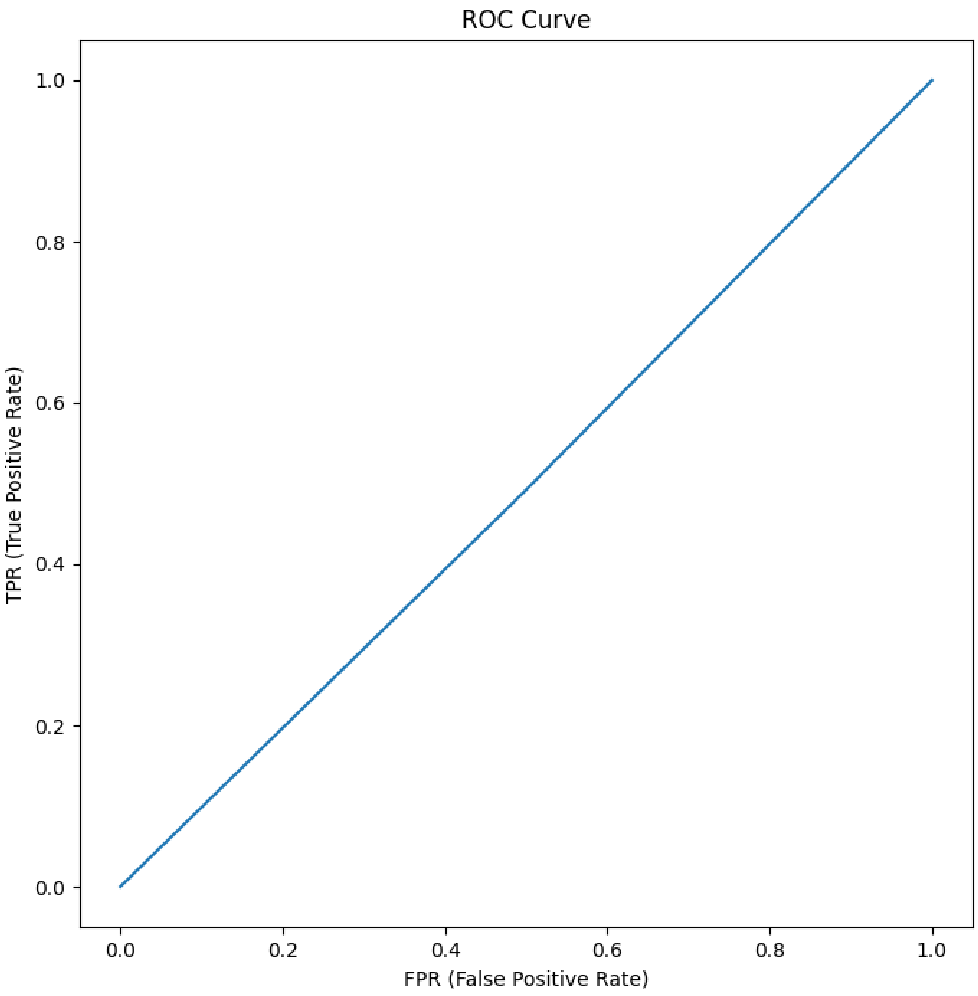
F Score Feature Selection ROC Curve



Forward Sequential Feature Selection ROC Curve



Recursive Feature Elimination



Select from Model Feature Selection

8.1.3 Logistic Regression

This classifier was implemented in python using the default parameters.

Feature Selection	Class	Accuracy	TPR	FPR	Precision	Recall	F1 Score	MCC	ROC
Correlation based feature selection	Heart Attack = 1	0.93962253	0.993566294	0.006433706	0.945339286	0.993566294	0.968852836	0.00945789	0.389521608
	Heart Attack = 2	0.93962253	0.009889616	0.990110384	0.080841391	0.009889616	0.017544069	0.00945789	0.389521608
	Weighted	0.93962253	0.93962253	0.06037747	0.897932099	0.93962253	0.916685141	0.00945789	0.389521608
F Score feature selection	Heart Attack = 1	0.940818488	0.994228711	0.005771289	0.945917731	0.994228711	0.969471512	0.040896709	0.619528511
	Heart Attack = 2	0.940818488	0.020331557	0.979668443	0.169298485	0.020331557	0.036262895	0.040896709	0.619528511
	Weighted	0.940818488	0.940818488	0.059181512	0.903325357	0.940818488	0.918292963	0.040896709	0.619528511
Forward Sequential Feature Selection	Heart Attack = 1	0.938725195	0.991176266	0.008823734	0.94651709	0.991176266	0.968330086	0.058188162	0.659068422
	Heart Attack = 2	0.938725195	0.034735477	0.965264523	0.186567093	0.034735477	0.057973982	0.058188162	0.659068422
	Weighted	0.938725195	0.938725195	0.061274805	0.904839791	0.938725195	0.918406879	0.058188162	0.659068422
Recursive Feature Elimination	Heart Attack = 1	0.938391268	0.990967618	0.009032382	0.946374602	0.990967618	0.968157631	0.052314496	0.65337832
	Heart Attack = 2	0.938391268	0.032245039	0.967754961	0.171954622	0.032245039	0.054204339	0.052314496	0.65337832
	Weighted	0.938391268	0.938391268	0.061608732	0.903903064	0.938391268	0.918036988	0.052314496	0.65337832
Select from Model	Heart Attack = 1	0.9199238	0.96464528	0.03535472	0.951317313	0.96464528	0.95792908	0.129149269	0.731736896
	Heart Attack = 2	0.9199238	0.149063378	0.850936622	0.195817836	0.149063378	0.168717102	0.129149269	0.731736896
	Weighted	0.9199238	0.9199238	0.0800762	0.909888	0.9199238	0.914652965	0.129149269	0.731736896

Correlation Based Feature Selection

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	127675	827
HEART ATTACK = 2	7382	74

F Score Feature Selection

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	127760	742
HEART ATTACK = 2	7304	152

Forward Sequential Feature Selection

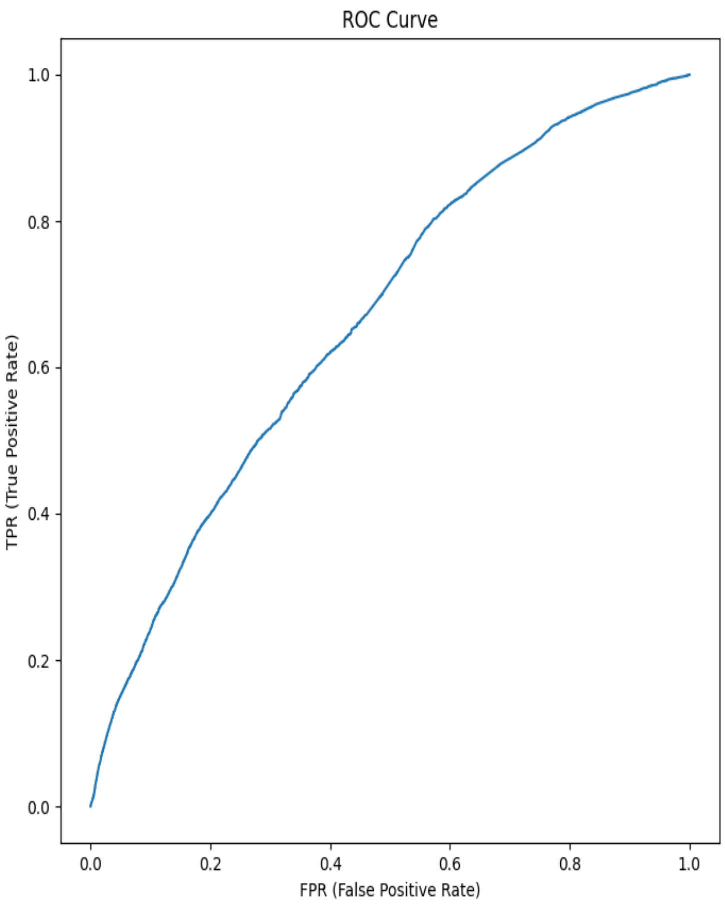
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	127368	1134
HEART ATTACK = 2	7197	259

Recursive Feature Elimination

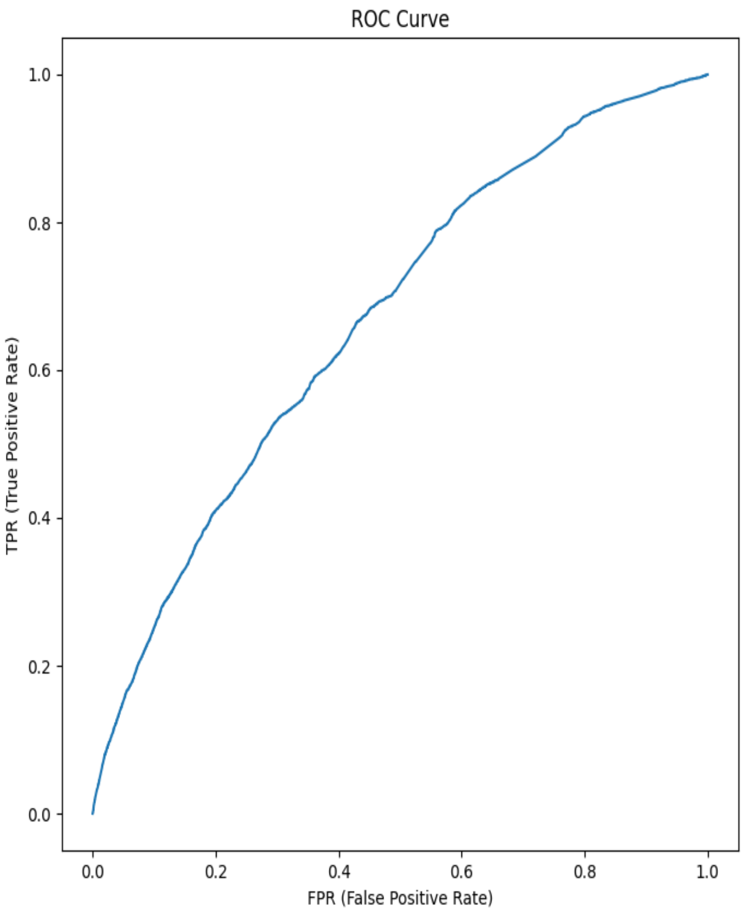
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	127341	1160
HEART ATTACK = 2	7216	241

Select From Model Feature Selection

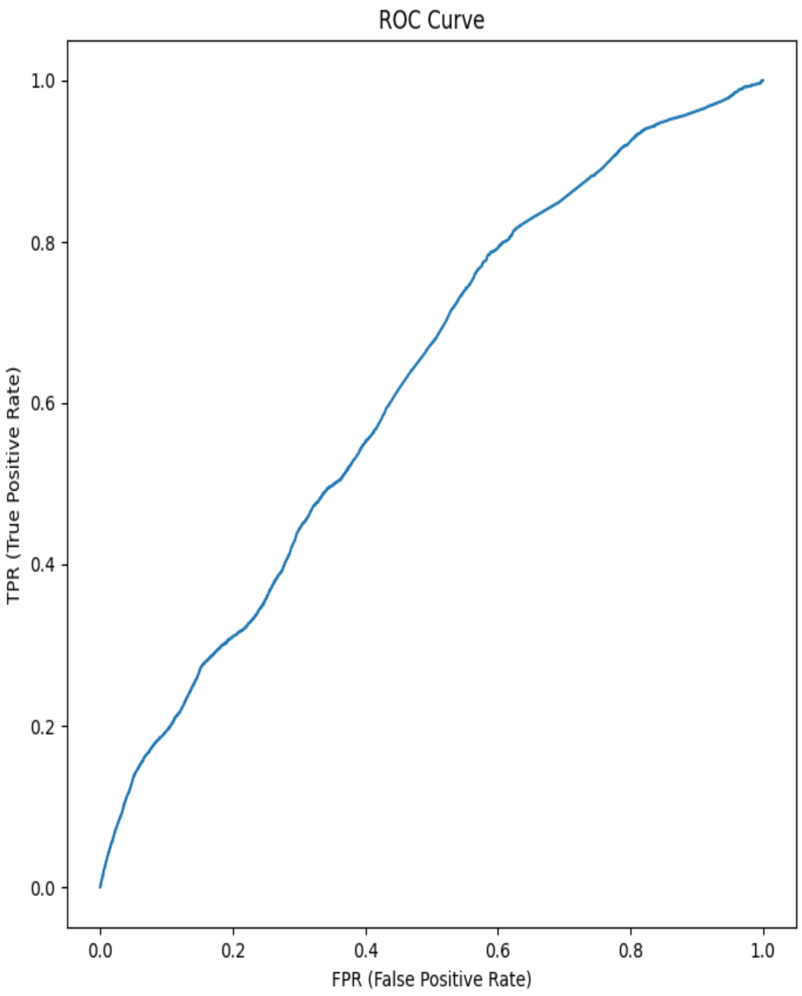
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	123959	4543
HEART ATTACK = 2	6344	1112



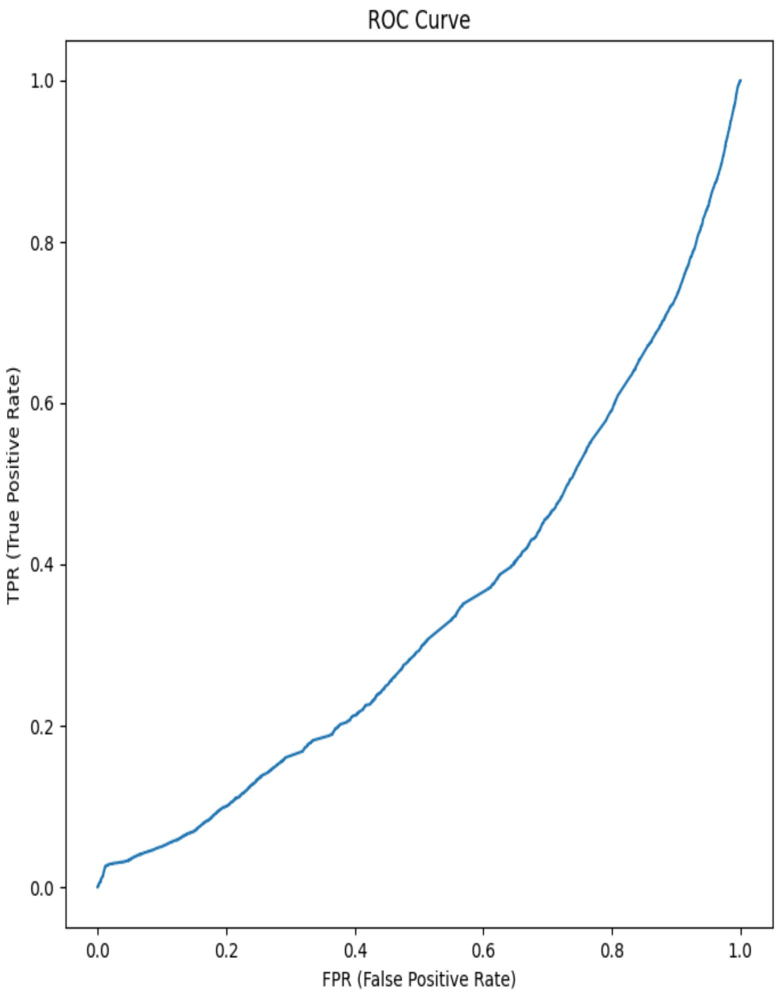
Correlation Based Feature Selection ROC Curve



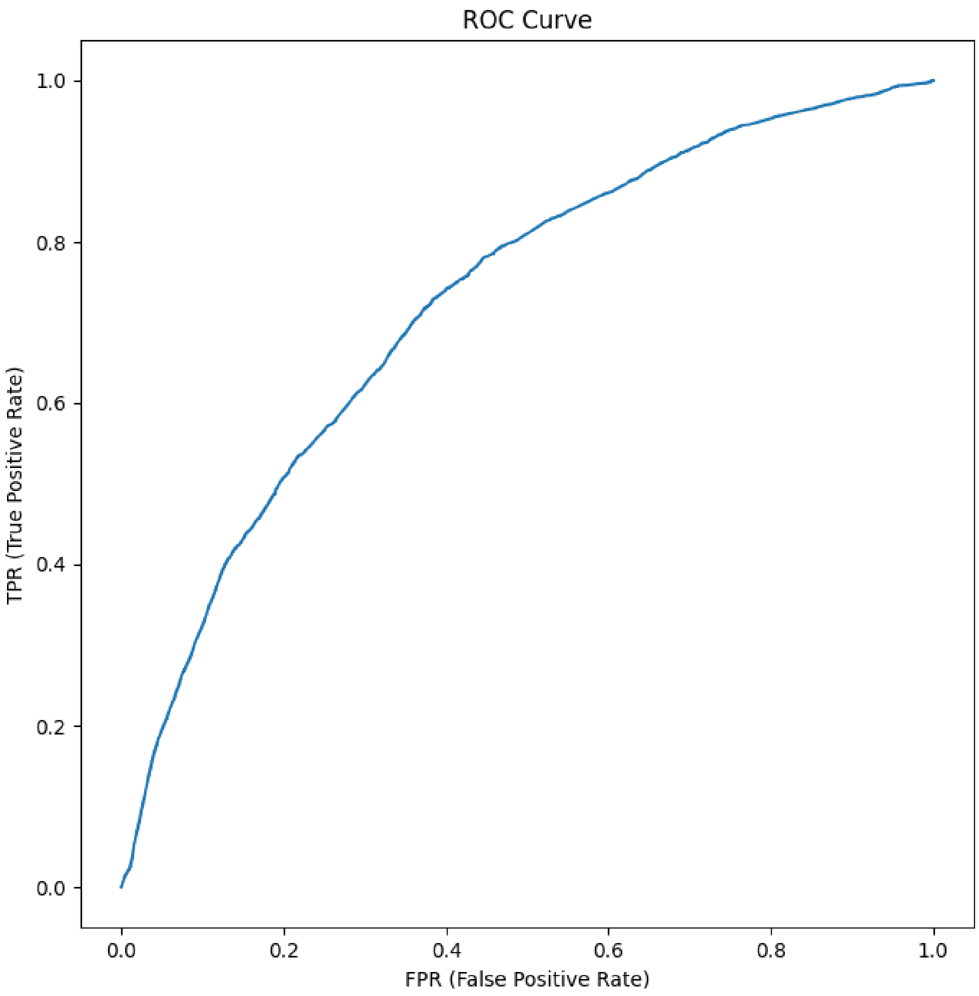
F Score Feature Selection ROC Curve



Forward Sequential Feature Selection ROC Curve



Recursive Feature Elimination



Select from Model Feature Selection

8.1.4 Gaussian Naïve Bayesian

This classifier was implemented in python using the default parameters.

Feature Selection	Class	Accuracy	TPR	FPR	Precision	Recall	F1 Score	MCC	ROC
Correlation based feature selection	Heart Attack = 1	0.51676253	0.509793769	0.490206231	0.963158641	0.509793769	0.642735997	0.069862471	0.632093558
	Heart Attack = 2	0.51676253	0.631766939	0.368233061	0.071393146	0.631766939	0.126100009	0.069862471	0.632093558
	Weighted	0.51676253	0.51676253	0.48323747	0.914243024	0.51676253	0.614523307	0.069862471	0.632093558
F Score feature selection	Heart Attack = 1	0.935856662	0.989450341	0.010549659	0.945244413	0.989450341	0.966842238	0.00364669	0.490675735
	Heart Attack = 2	0.935856662	0.012194868	0.987805132	0.062843518	0.012194868	0.020416392	0.00364669	0.490675735
	Weighted	0.935856662	0.935856662	0.064143338	0.896853737	0.935856662	0.914939725	0.00364669	0.490675735
Forward Sequential Feature Selection	Heart Attack = 1	0.937830801	0.99161649	0.00838351	0.945287658	0.99161649	0.967892173	0.005213157	0.597317896
	Heart Attack = 2	0.937830801	0.010835585	0.989164415	0.067050468	0.010835585	0.017908071	0.005213157	0.597317896
	Weighted	0.937830801	0.937830801	0.062169199	0.897127403	0.937830801	0.915796238	0.005213157	0.597317896
Recursive Feature Elimination	Heart Attack = 1	0.924314862	0.97614705	0.02385295	0.945528203	0.97614705	0.960575299	0.011313978	0.596688478
	Heart Attack = 2	0.924314862	0.030832348	0.969167652	0.073616947	0.030832348	0.042016894	0.011313978	0.596688478
	Weighted	0.924314862	0.924314862	0.075685138	0.897706531	0.924314862	0.910207336	0.011313978	0.596688478
Select from Model	Heart Attack = 1	0.924314862	0.97614705	0.02385295	0.945528203	0.97614705	0.960575299	0.011313978	0.596688478
	Heart Attack = 2	0.924314862	0.030832348	0.969167652	0.073616947	0.030832348	0.042016894	0.011313978	0.596688478
	Weighted	0.924314862	0.924314862	0.075685138	0.897706531	0.924314862	0.910207336	0.011313978	0.596688478

Correlation Based Feature Selection

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	65529	62973
HEART ATTACK = 2	2727	4729

F Score Feature Selection

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	127146	1356
HEART ATTACK = 2	7365	91

Forward Sequential Feature Selection

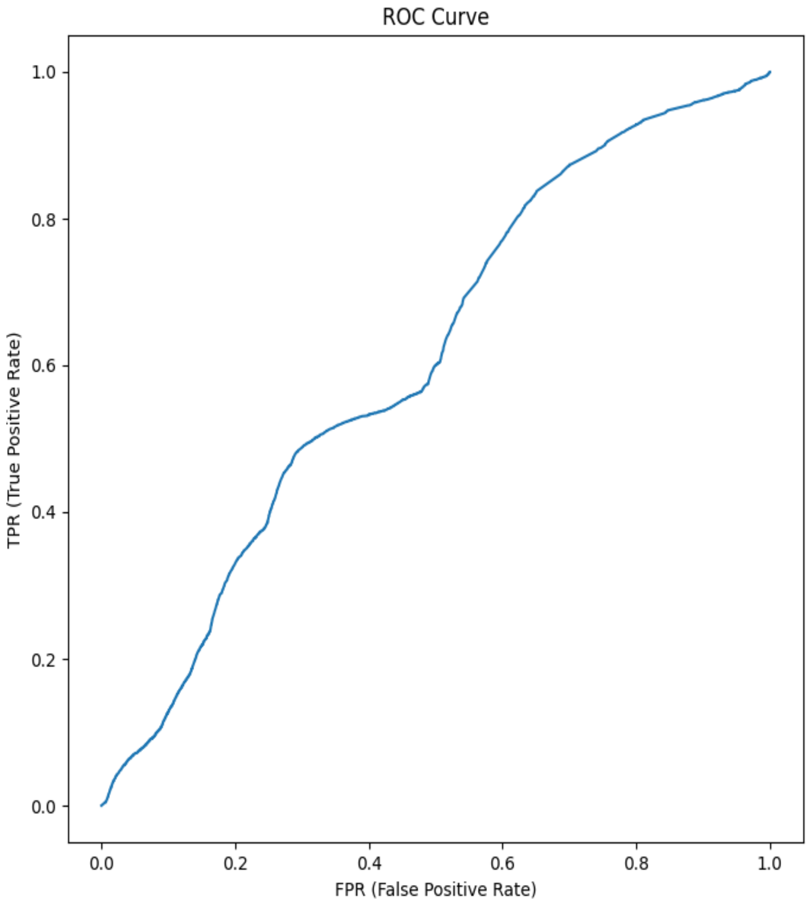
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	127425	1077
HEART ATTACK = 2	7375	81

Recursive Feature Elimination

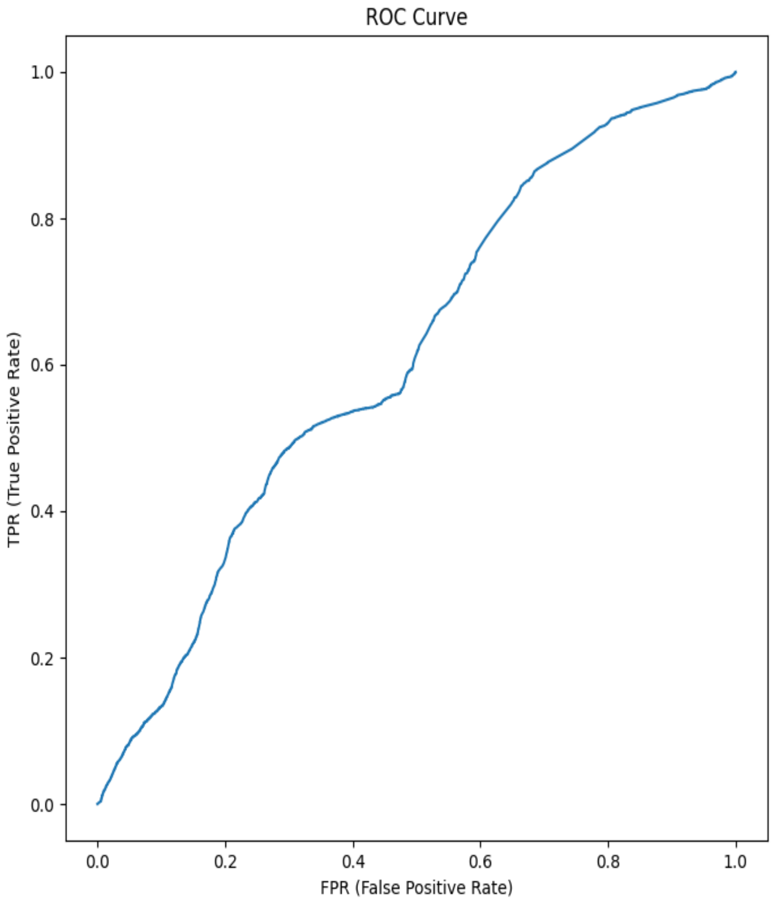
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	125437	3064
HEART ATTACK = 2	7226	231

Select From Model Feature Selection

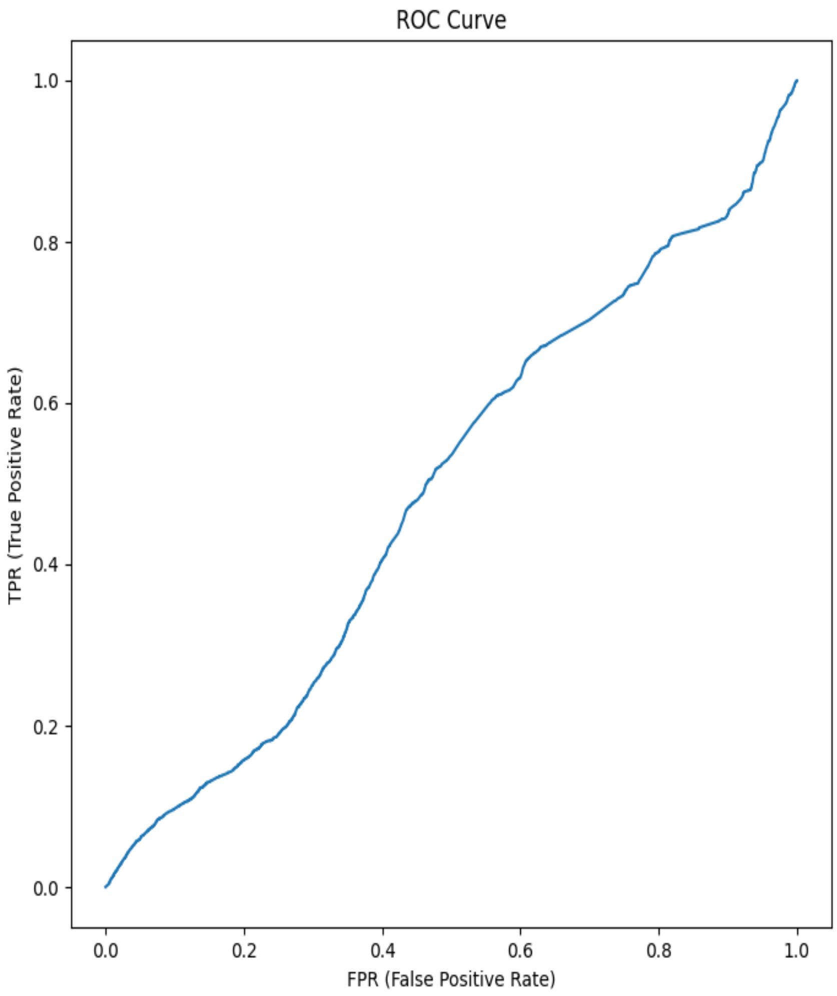
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	125437	3064
HEART ATTACK = 2	7226	231



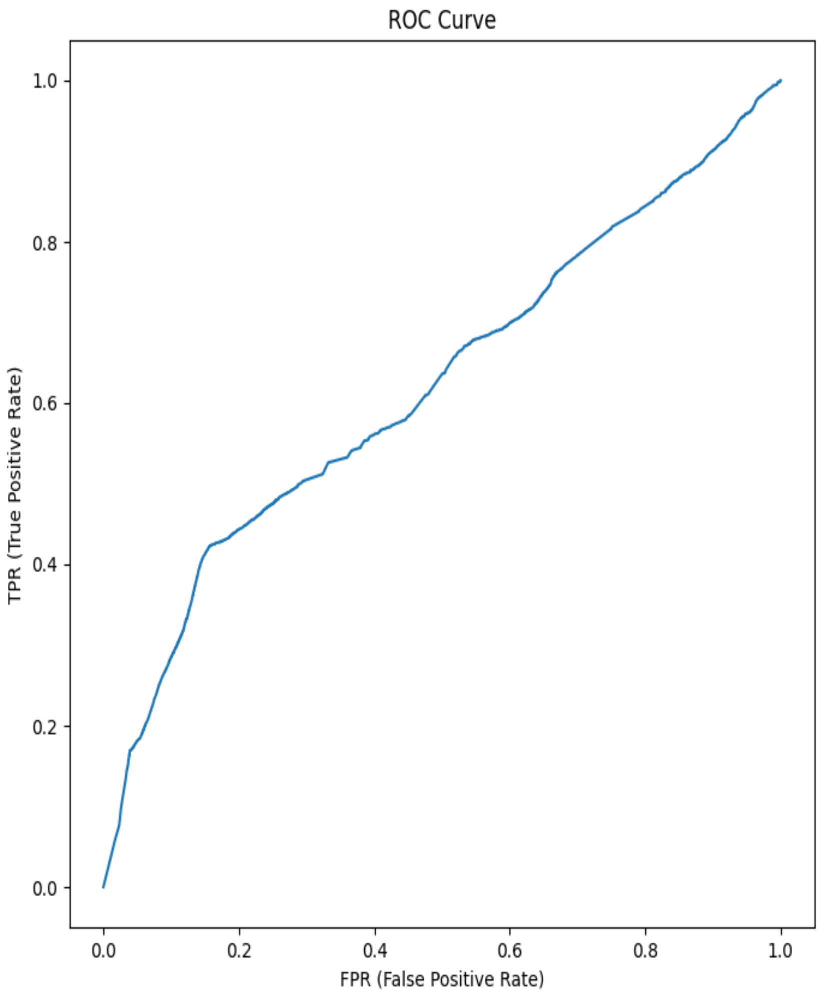
Correlation Based Feature Selection ROC Curve



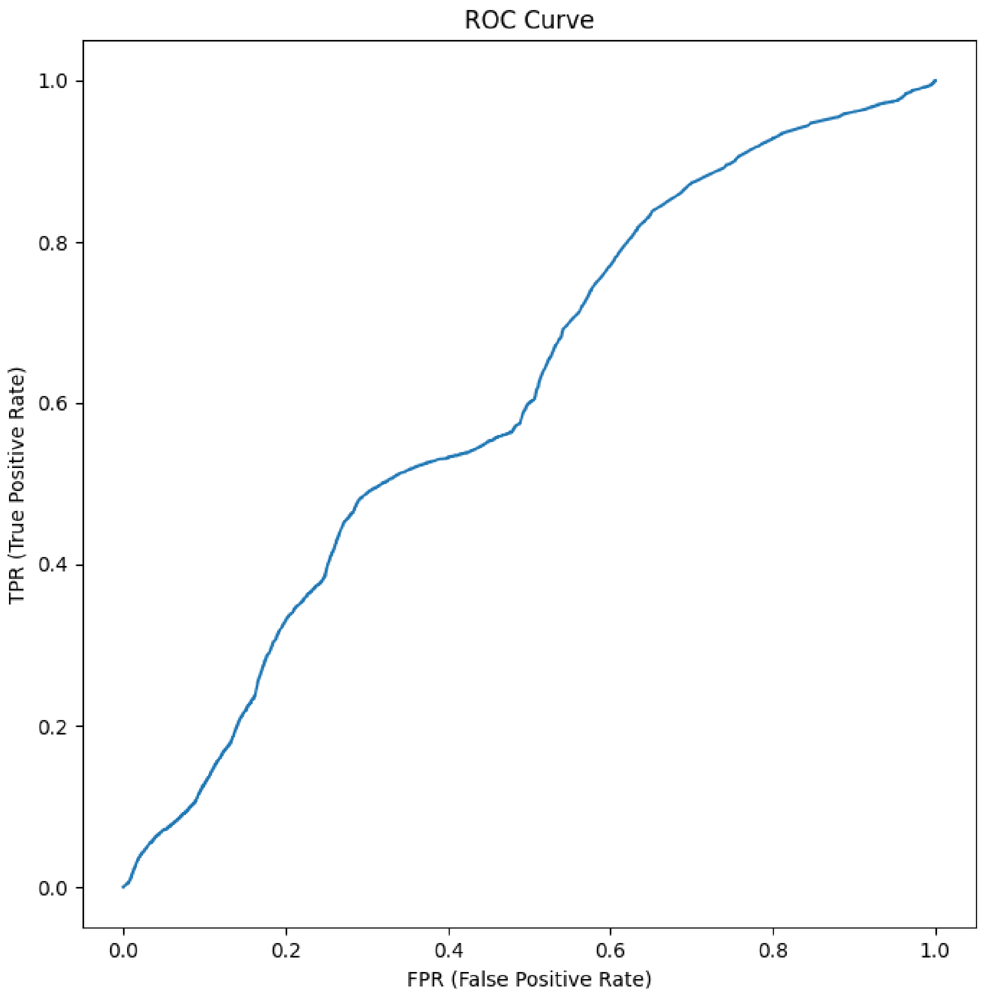
F Score Feature Selection ROC Curve



Forward Sequential Feature Selection ROC Curve



Recursive Feature Elimination



Select from Model Feature Selection

8.1.5 Random Forest

This classifier was implemented in python using entropy as the measure of quality of the split. The default number of trees, 100 were used to build this model.

Feature Selection	Class	Accuracy	TPR	FPR	Precision	Recall	F1 Score	MCC	ROC
Correlation based feature selection	Heart Attack = 1	0.797455096	0.837469496	0.162530504	0.94167275	0.837469496	0.88590295	-0.034908929	0.432081655
	Heart Attack = 2	0.797455096	0.106761219	0.893238781	0.036199842	0.106761219	0.053260208	-0.034908929	0.432081655
	Weighted	0.797455096	0.797455096	0.202544904	0.892018454	0.797455096	0.840263844	-0.034908929	0.432081655
F Score feature selection	Heart Attack = 1	0.845005075	0.887896961	0.112103039	0.94471569	0.887896961	0.912795619	-0.004409428	0.485727929
	Heart Attack = 2	0.845005075	0.105202035	0.894797965	0.052386576	0.105202035	0.061438169	-0.004409428	0.485727929
	Weighted	0.845005075	0.845005075	0.154994925	0.895782493	0.845005075	0.866118095	-0.004409428	0.485727929
Forward Sequential Feature Selection	Heart Attack = 1	0.836158225	0.87502868	0.12497132	0.94769769	0.87502868	0.908232402	0.029514606	0.563582495
	Heart Attack = 2	0.836158225	0.166104822	0.833895178	0.074772597	0.166104822	0.096205604	0.029514606	0.563582495
	Weighted	0.836158225	0.836158225	0.163841775	0.899831739	0.836158225	0.863705443	0.029514606	0.563582495
Recursive Feature Elimination	Heart Attack = 1	0.828903044	0.867023518	0.132976482	0.947415603	0.867023518	0.904710287	0.028490227	0.563454276
	Heart Attack = 2	0.828903044	0.171358739	0.828641261	0.07418109	0.171358739	0.100373258	0.028490227	0.563454276
	Weighted	0.828903044	0.828903044	0.171096956	0.899519985	0.828903044	0.860606181	0.028490227	0.563454276
Select from Model	Heart Attack = 1	0.777178246	0.803290358	0.196709642	0.953902596	0.803290358	0.871608275	0.071393258	0.600906931
	Heart Attack = 2	0.777178246	0.326892863	0.673107137	0.085376017	0.326892863	0.134642127	0.071393258	0.600906931
	Weighted	0.777178246	0.777178246	0.222821754	0.906272322	0.777178246	0.831197088	0.071393258	0.600906931

Correlation Based Feature Selection

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	107621	20881
HEART ATTACK = 2	6656	800

F Score Feature Selection

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	114098	14403
HEART ATTACK = 2	6670	787

Forward Sequential Feature Selection

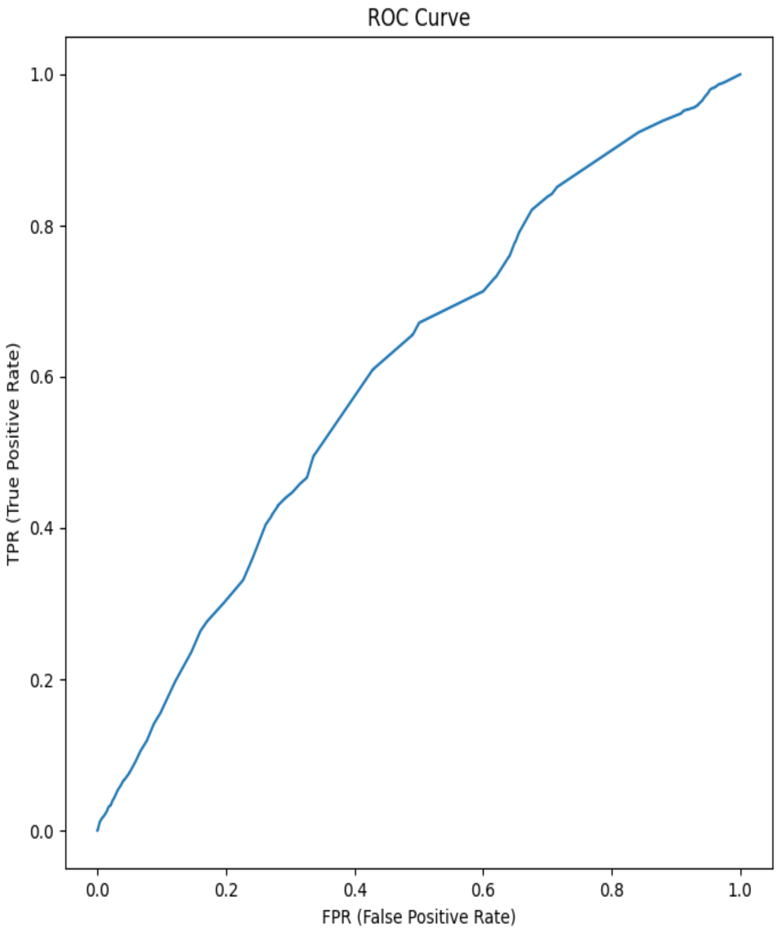
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	112443	16059
HEART ATTACK = 2	6217	1239

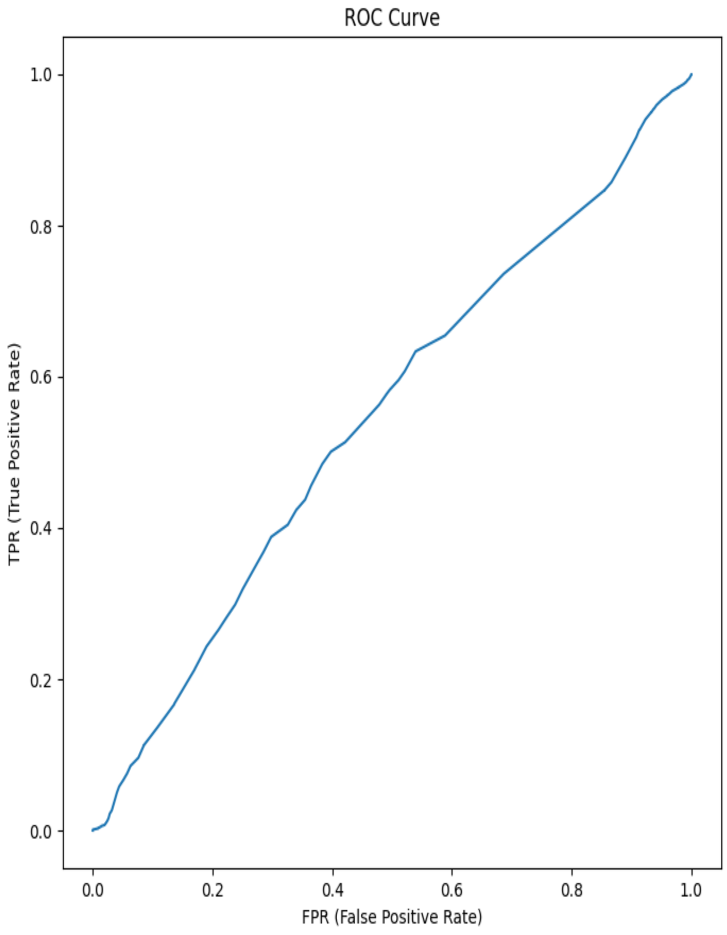
Recursive Feature Elimination

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	111417	17085
HEART ATTACK = 2	6177	1279

Select From Model Feature Selection

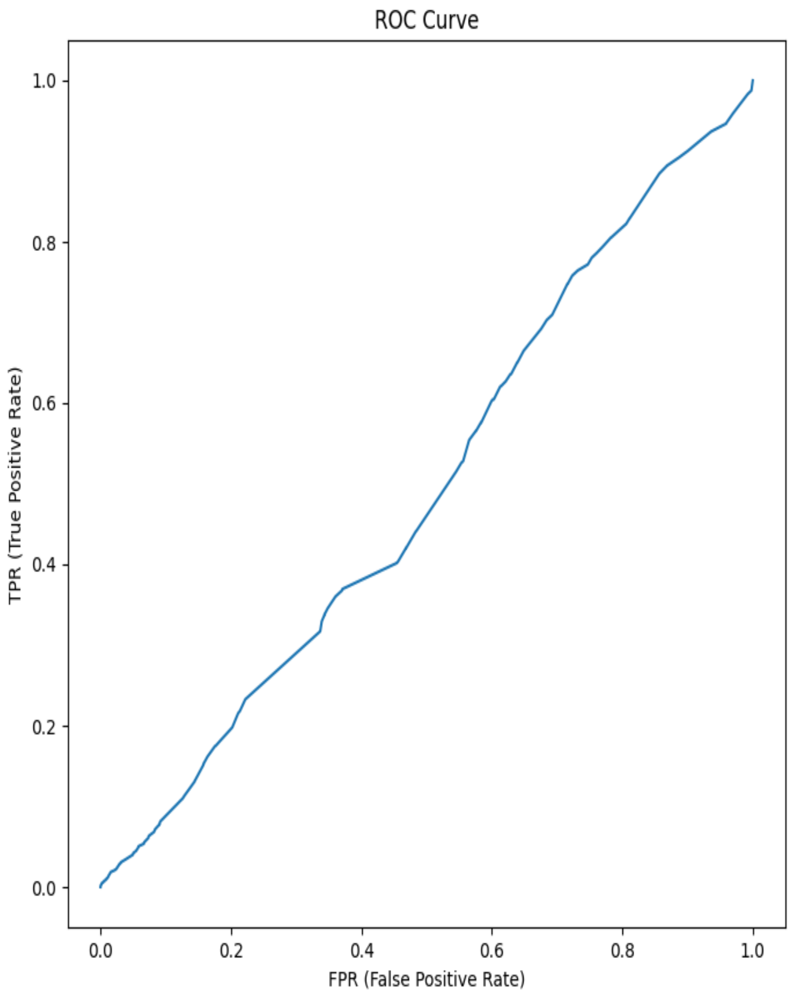
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	103225	25277
HEART ATTACK = 2	5017	2439

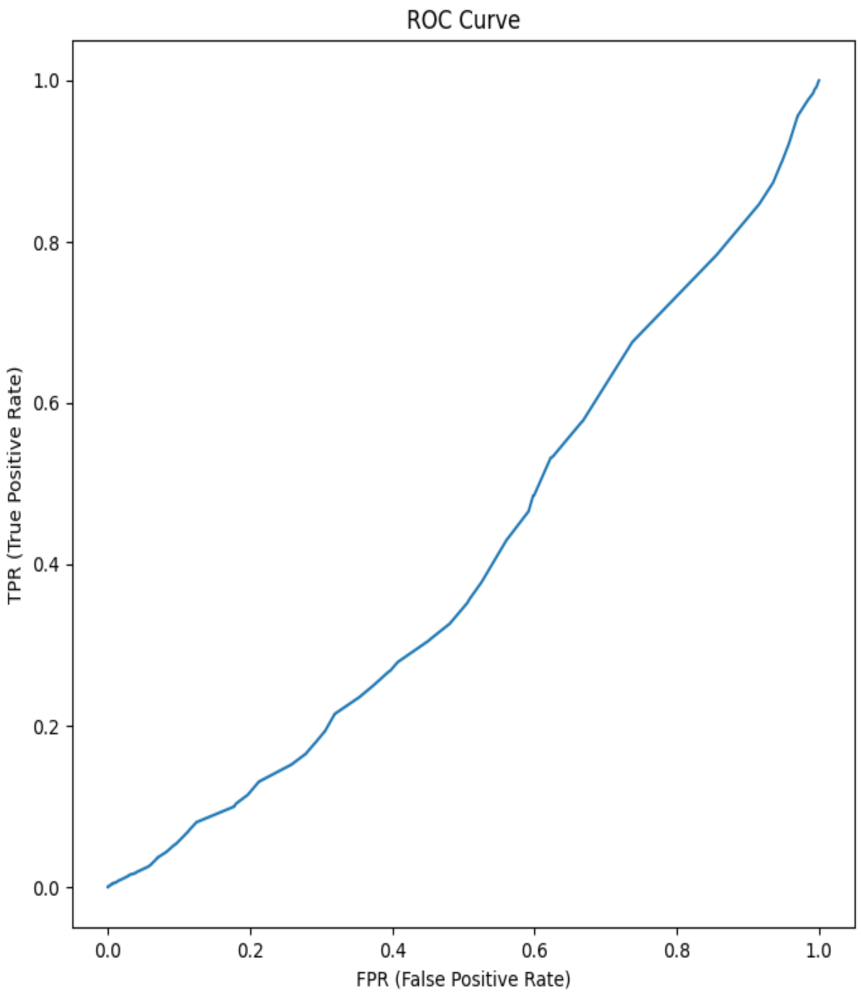




Correlation Based Feature Selection ROC Curve

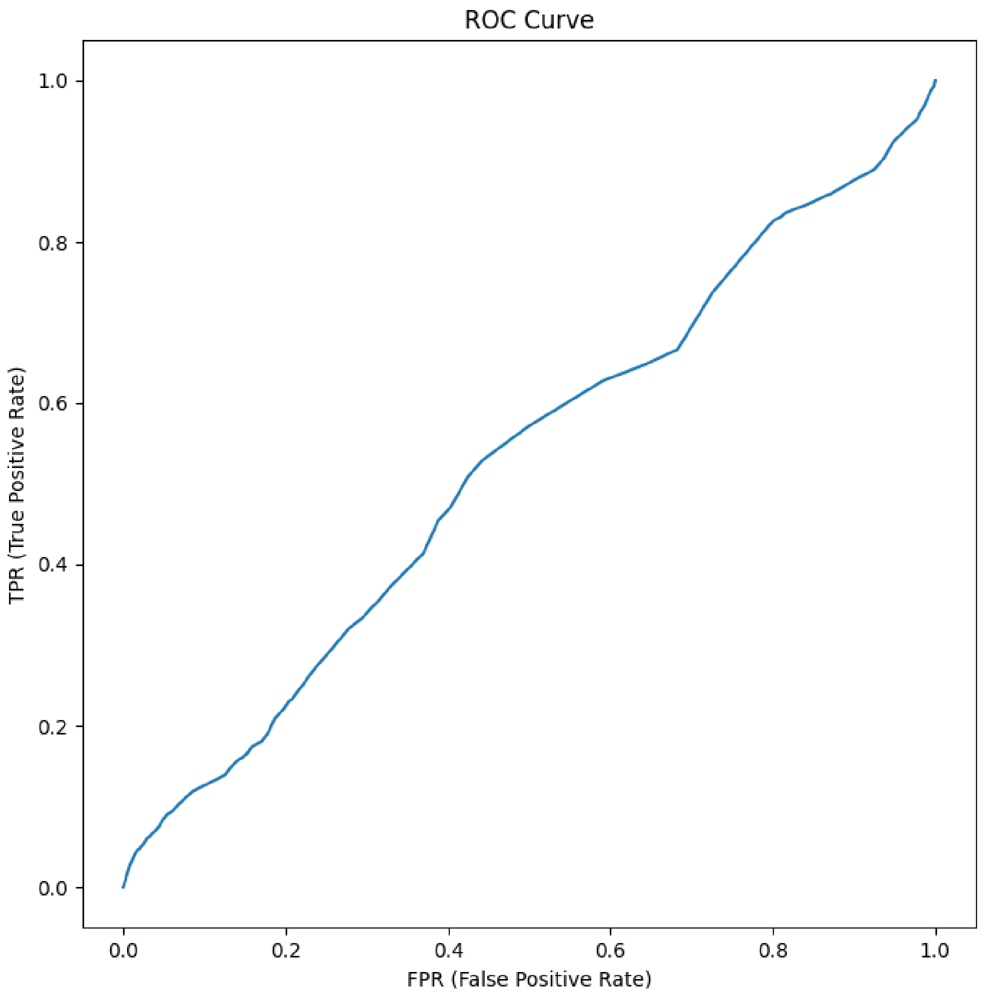
F Score Feature Selection ROC Curve





Forward Sequential Feature Selection ROC Curve

Recursive Feature Elimination



Select from Model Feature Selection

8.2 Best Model for Borderline SMOTE Balanced Data Set

We first look at the best model for each machine learning classifier from the feature selected data sets. The following models were chosen:

The above models were chosen because they have a positive MCC score and the ROC area of each one of these models is greater than 0.5. Having a MCC score of greater than 0 means that the predictions are better than random. These models also have a better TPR for Heart Attack = 2 class which is the class of concern to us. We want to be able to be better at predicting a patient with heart attack chances but at the same time we do not want to misclassify too many patients with no risk of heart attack as those having a risk of a heart attack. These models also had much better accuracy than others.

Classifier	Feature Selection	Class	Accuracy	TPR	FPR	Precision	Recall	F1 Score	MCC	ROC
Ada Boost	Recursive Feature Elimination	Heart Attack = 1	0.941318643	0.995303511	0.004696489	0.945487607	0.995303511	0.969751133	0.022000973	0.639595686
		Heart Attack = 2	0.941318643	0.011036088	0.988963912	0.144768779	0.011036088	0.019508946	0.022000973	0.639595686
		Weighted	0.941318643	0.941318643	0.058681357	0.901598677	0.941318643	0.917631732	0.022000973	0.639595686
Decision Tree	Select from Model	Heart Attack = 1	0.640587534	0.652842971	0.347157029	0.951157354	0.652842971	0.771315026	0.041820438	0.541617839
		Heart Attack = 2	0.640587534	0.430152752	0.569847248	0.069925432	0.430152752	0.11964779	0.041820438	0.541617839
		Weighted	0.640587534	0.640587534	0.359412466	0.90282836	0.640587534	0.73556036	0.041820438	0.541617839
Logistic Regression	Select from Model	Heart Attack = 1	0.9199238	0.96464528	0.03535472	0.951317313	0.96464528	0.95792908	0.129149269	0.731736896
		Heart Attack = 2	0.9199238	0.149063378	0.850936622	0.195817836	0.149063378	0.168717102	0.129149269	0.731736896
		Weighted	0.9199238	0.9199238	0.0800762	0.909888	0.9199238	0.914652965	0.129149269	0.731736896
Gaussian Naïve Bayesian	Recursive Feature Elimination	Heart Attack = 1	0.924314862	0.97614705	0.02385295	0.945528203	0.97614705	0.960575299	0.011313978	0.596688478
		Heart Attack = 2	0.924314862	0.030832348	0.969167652	0.073616947	0.030832348	0.042016894	0.011313978	0.596688478
		Weighted	0.924314862	0.924314862	0.075685138	0.897706531	0.924314862	0.910207336	0.011313978	0.596688478
Random Forest	Select from Model	Heart Attack = 1	0.777178246	0.803290358	0.196709642	0.953902596	0.803290358	0.871608275	0.071393258	0.600906931
		Heart Attack = 2	0.777178246	0.326892863	0.673107137	0.085376017	0.326892863	0.134642127	0.071393258	0.600906931
		Weighted	0.777178246	0.777178246	0.222821754	0.906272322	0.777178246	0.831197088	0.071393258	0.600906931

Now, we will select the best model from these 5 models. **We have selected the Logistic Regression Model with Select from Model feature selection as the best model.** The reason for this being that this model has the best ROC area under curve and the best MCC. Although, it has slightly lower accuracy than other models, but it is also better than the other models at predicting the Heart Attack = 2 (TPR = 0.149063378) class which is of concern to us as we want to be able to identify patient with risk of a heart attack. At the same time, it does not misclassify a lot of the Heart Attack = 1 class (TPR = 0.96464528).

8.3 SMOTE Balanced Data Set

8.3.1 Ada Boost

This classifier was implemented in python with Logistic Regression as the base estimator. The default max number of base estimators, 50 were used to run this algorithm.

Feature Selection	Class	Accuracy	TPR	FPR	Precision	Recall	F1 Score	MCC	ROC
Correlation based feature selection	Heart Attack = 1	0.943895909	0.998404437	0.001595563	0.94530761	0.998404437	0.971130645	0.015895248	0.583381099
	Heart Attack = 2	0.943895909	0.004478198	0.995521802	0.144906683	0.004478198	0.008656026	0.015895248	0.583381099
	Weighted	0.943895909	0.943895909	0.056104091	0.901396561	0.943895909	0.918347218	0.015895248	0.583381099
F Score feature selection	Heart Attack = 1	0.941849689	0.996071584	0.003928416	0.945337477	0.996071584	0.970041476	0.012266354	0.579878388
	Heart Attack = 2	0.941849689	0.007379195	0.992620805	0.098273264	0.007379195	0.013726906	0.012266354	0.579878388
	Weighted	0.941849689	0.941849689	0.058150311	0.898881302	0.941849689	0.917595413	0.012266354	0.579878388
Forward Sequential Feature Selection	Heart Attack = 1	0.942333662	0.996614509	0.003385491	0.945340977	0.996614509	0.970296033	0.01313136	0.700430025
	Heart Attack = 2	0.942333662	0.006907136	0.993092864	0.116275926	0.006907136	0.012183363	0.01313136	0.700430025
	Weighted	0.942333662	0.942333662	0.057666338	0.899878271	0.942333662	0.917747393	0.01313136	0.700430025
Recursive Feature Elimination	Heart Attack = 1	0.943954751	0.998365781	0.001634219	0.945396378	0.998365781	0.97115917	0.024017229	0.631823228
	Heart Attack = 2	0.943954751	0.006225691	0.993774309	0.18033645	0.006225691	0.012034592	0.024017229	0.631823228
	Weighted	0.943954751	0.943954751	0.056045249	0.903438401	0.943954751	0.918558979	0.024017229	0.631823228
Select from Model	Heart Attack = 1	0.915259124	0.961845377	0.038154623	0.949270616	0.961845377	0.955267138	0.076989577	0.702672513
	Heart Attack = 2	0.915259124	0.112024146	0.887975854	0.141530531	0.112024146	0.109501674	0.076989577	0.702672513
	Weighted	0.915259124	0.915259124	0.084740876	0.904972904	0.915259124	0.908893848	0.076989577	0.702672513

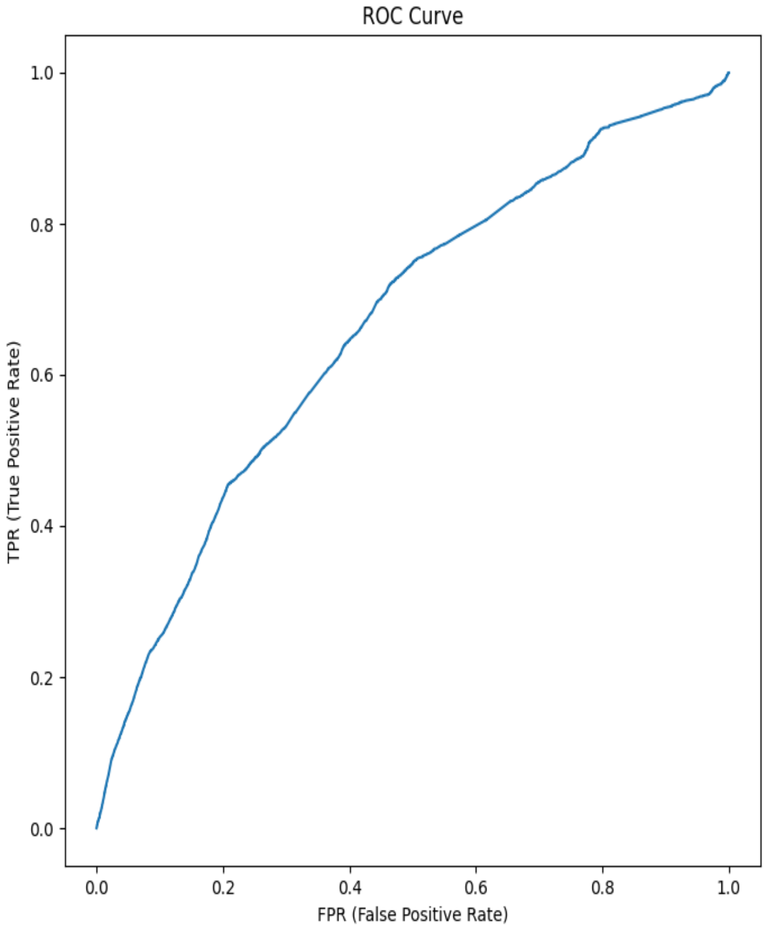
Correlation Based Feature Selection		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	128297	205
HEART ATTACK = 2	7423	33

F Score Feature Selection		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	127997	505
HEART ATTACK = 2	7401	55

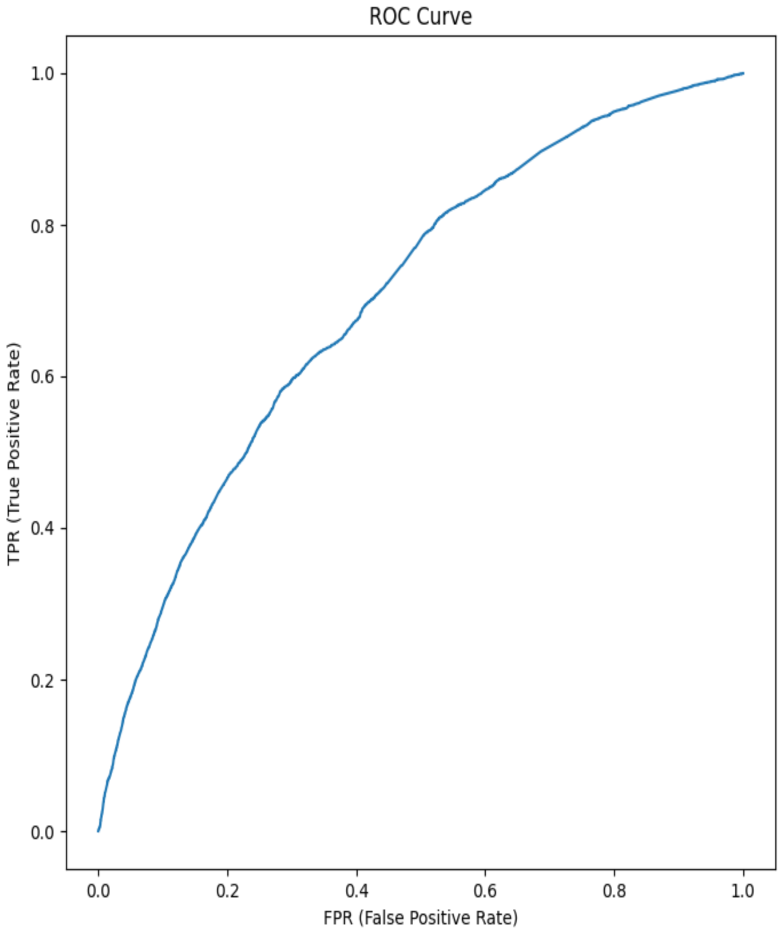
Forward Sequential Feature Selection		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	128067	435
HEART ATTACK = 2	7405	51

Recursive Feature Elimination		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	128292	210
HEART ATTACK = 2	7410	46

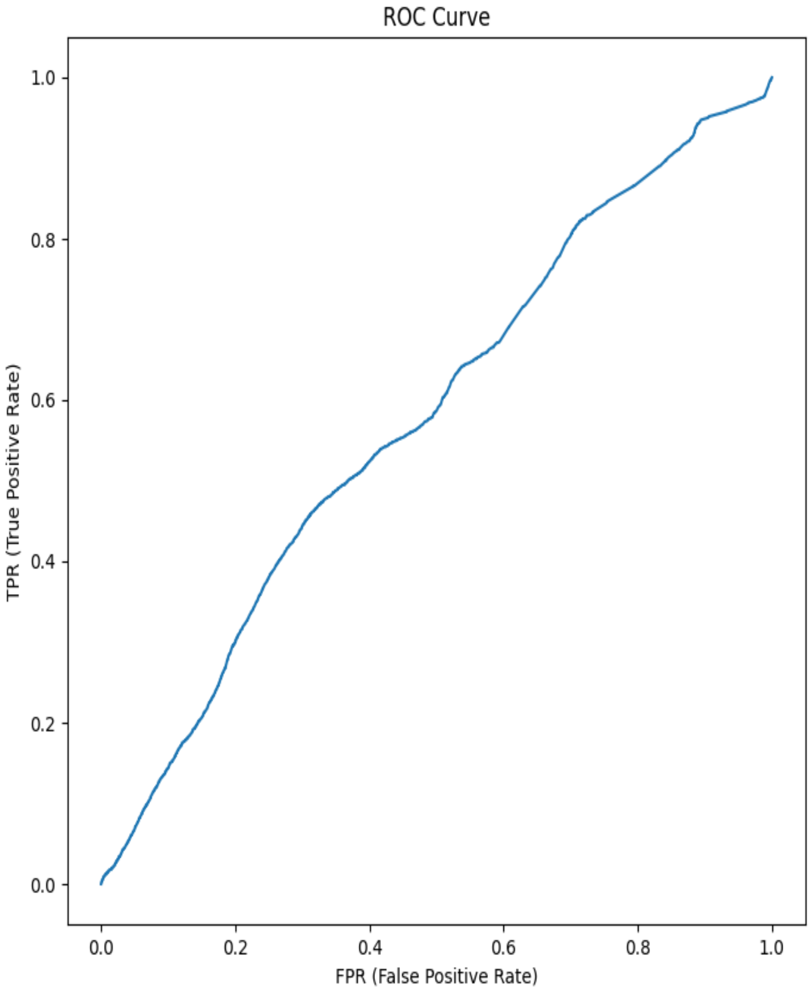
Select From Model Feature Selection		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	123600	4902
HEART ATTACK = 2	6619	837



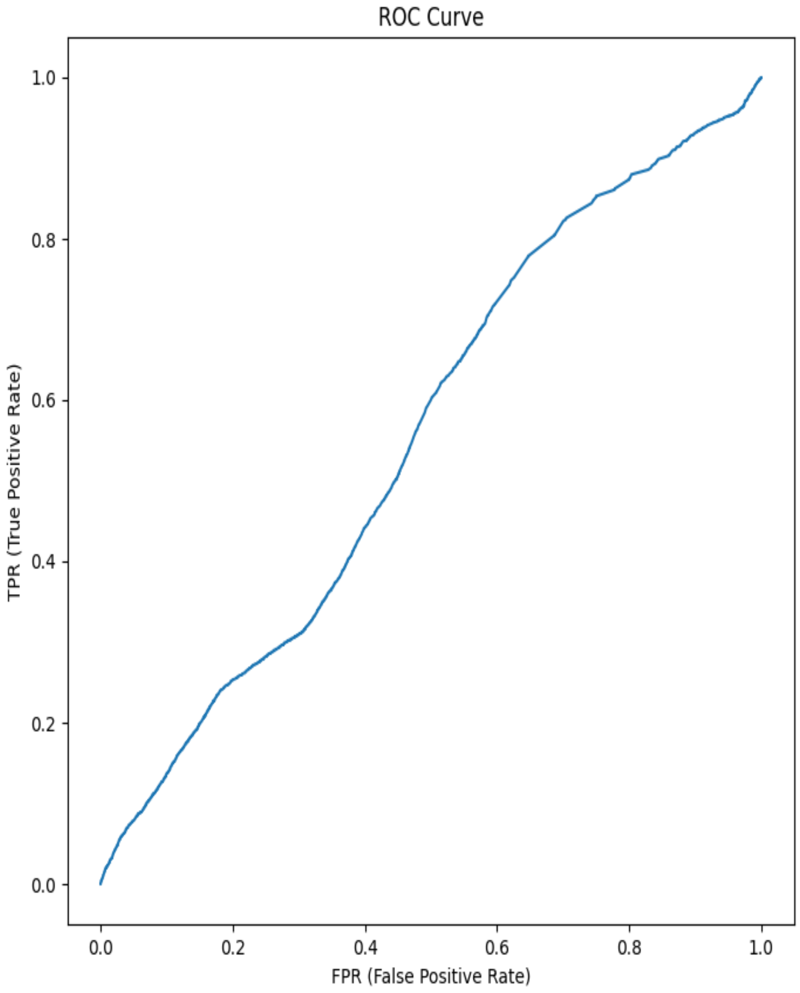
Correlation Based Feature Selection ROC Curve



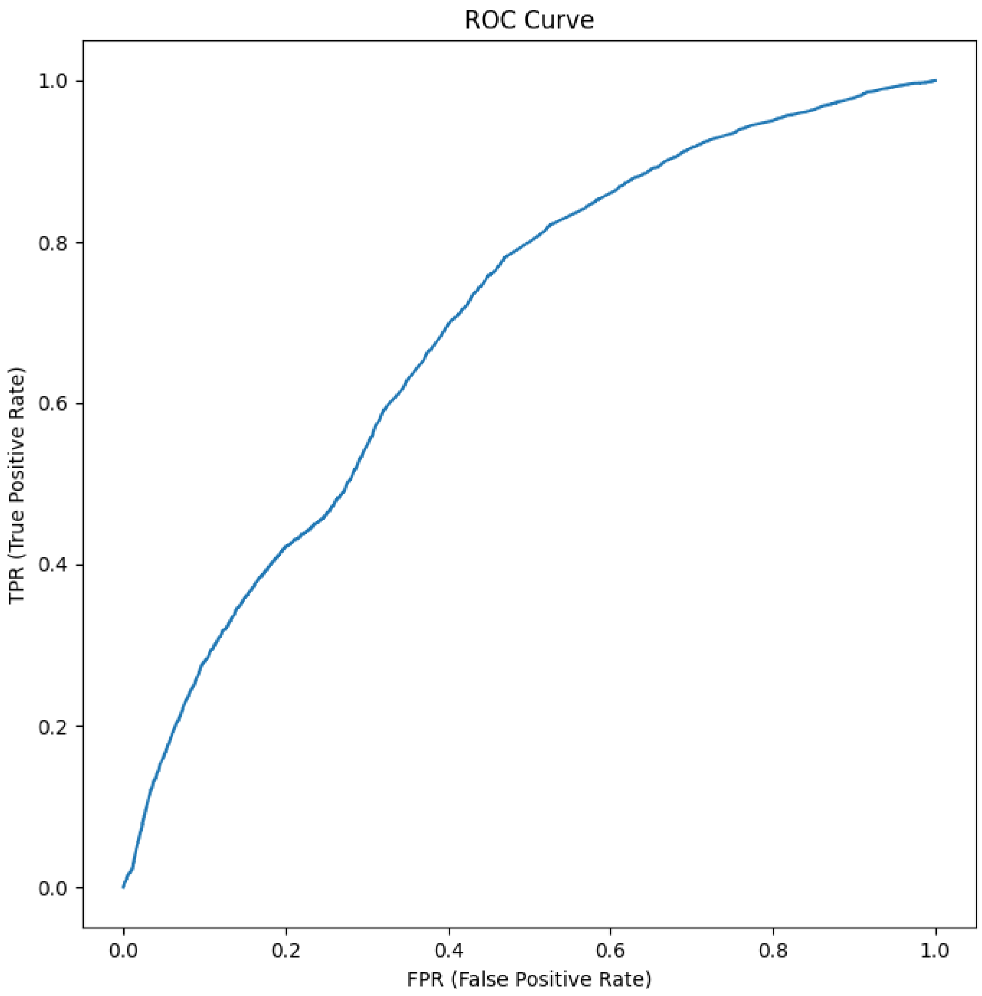
F Score Feature Selection ROC Curve



Forward Sequential Feature Selection ROC Curve



Recursive Feature Elimination



Select from Model Feature Selection

8.3.2 Decision Tree

This classifier was implemented in python using entropy as the measure of quality of the split.

Feature Selection	Class	Accuracy	TPR	FPR	Precision	Recall	F1 Score	MCC	ROC
Correlation based feature selection	Heart Attack = 1	0.327321673	0.305250414	0.694749586	0.948860002	0.305250414	0.400174037	0.004793428	0.504230474
	Heart Attack = 2	0.327321673	0.703218762	0.296781238	0.054155784	0.703218762	0.099439713	0.004793428	0.504230474
	Weighted	0.327321673	0.327321673	0.672678327	0.899791389	0.327321673	0.383813658	0.004793428	0.504230474
F Score feature selection	Heart Attack = 1	0.429919534	0.427920142	0.572079858	0.930220056	0.427920142	0.57640923	-0.050797608	0.446463024
	Heart Attack = 2	0.429919534	0.465307847	0.534692153	0.045590604	0.465307847	0.082672478	-0.050797608	0.446463024
	Weighted	0.429919534	0.429919534	0.570080466	0.881704639	0.429919534	0.549306752	-0.050797608	0.446463024
Forward Sequential Feature Selection	Heart Attack = 1	0.528092499	0.533295031	0.466704969	0.937497783	0.533295031	0.658013549	-0.015043831	0.486272887
	Heart Attack = 2	0.528092499	0.439120875	0.560879125	0.054244103	0.439120875	0.09506401	-0.015043831	0.486272887
	Weighted	0.528092499	0.528092499	0.471907501	0.889053806	0.528092499	0.627100404	-0.015043831	0.486272887
Recursive Feature Elimination	Heart Attack = 1	0.451274658	0.443705899	0.556294101	0.933524525	0.443705899	0.561061427	0.006325978	0.514042861
	Heart Attack = 2	0.451274658	0.583836127	0.416163873	0.059204522	0.583836127	0.106316954	0.006325978	0.514042861
	Weighted	0.451274658	0.451274658	0.548725342	0.885563456	0.451274658	0.53603736	0.006325978	0.514042861
Select from Model	Heart Attack = 1	0.463094485	0.45480204	0.54519796	0.951618537	0.45480204	0.593006675	0.030341864	0.53129727
	Heart Attack = 2	0.463094485	0.607772787	0.392227213	0.063128287	0.607772787	0.113216569	0.030341864	0.53129727
	Weighted	0.463094485	0.463094485	0.536905515	0.902892084	0.463094485	0.56665121	0.030341864	0.53129727

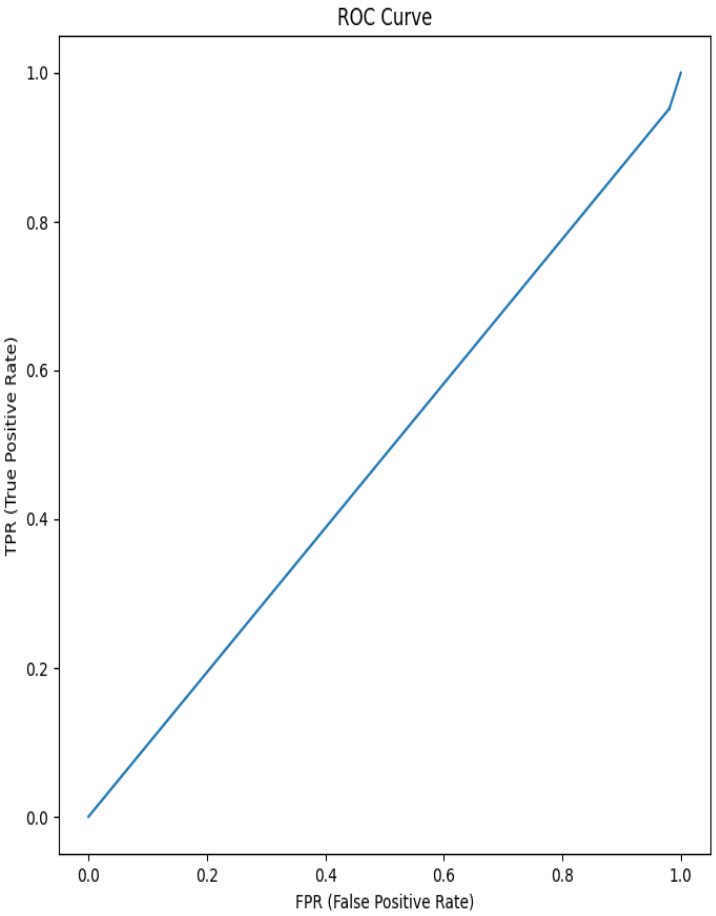
Correlation Based Feature Selection		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	39240	89262
HEART ATTACK = 2	2194	5262

F Score Feature Selection		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	54985	73517
HEART ATTACK = 2	3990	3466

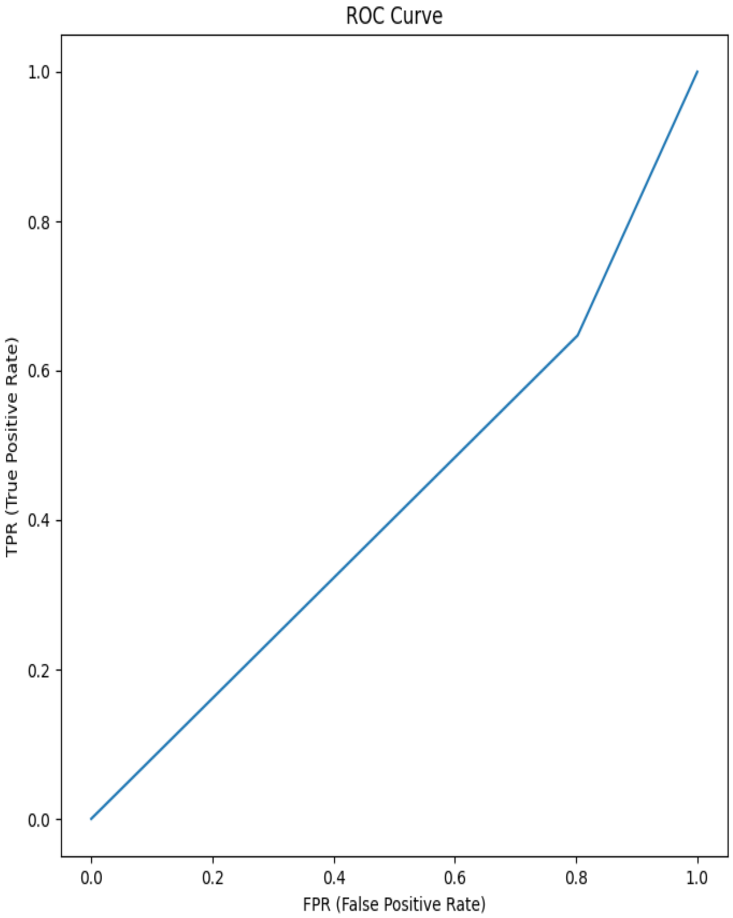
Forward Sequential Feature Selection		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	68525	59977
HEART ATTACK = 2	4183	3273

Recursive Feature Elimination		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	57010	71492
HEART ATTACK = 2	3112	4344

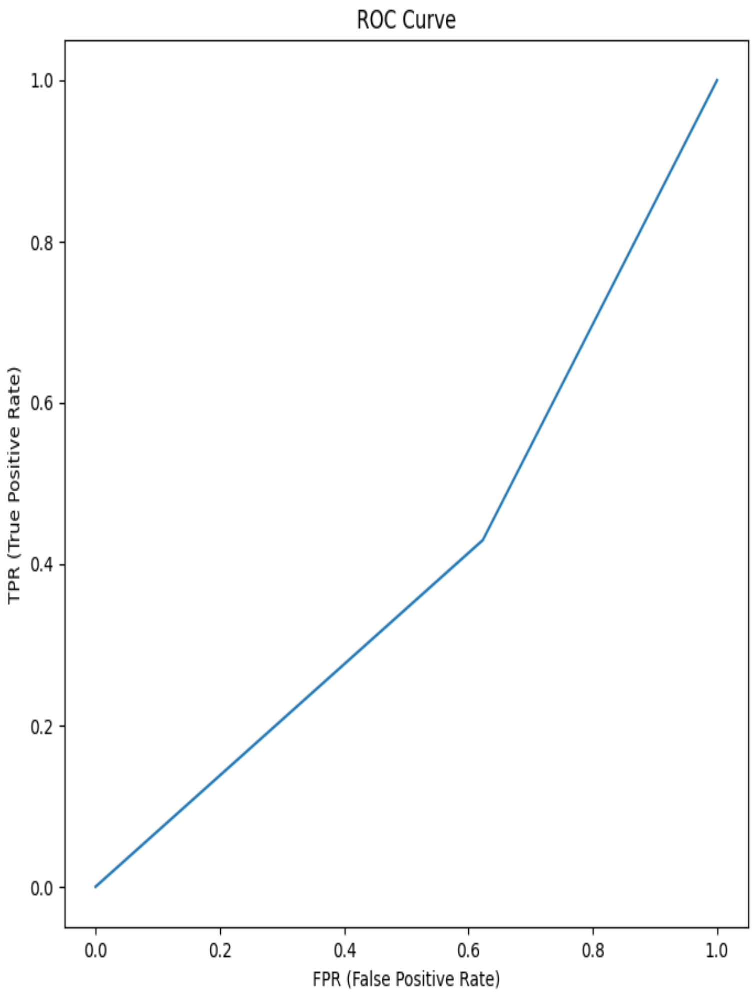
Select From Model Feature Selection		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	58439	70063
HEART ATTACK = 2	2934	4522



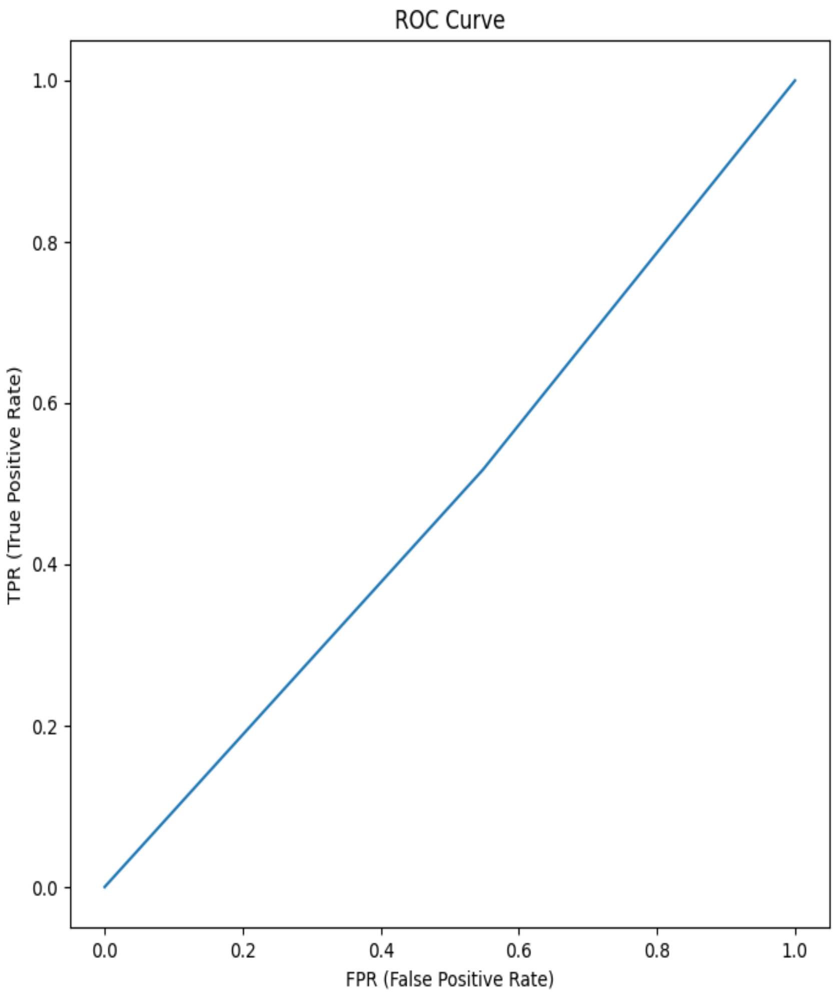
Correlation Based Feature Selection ROC Curve



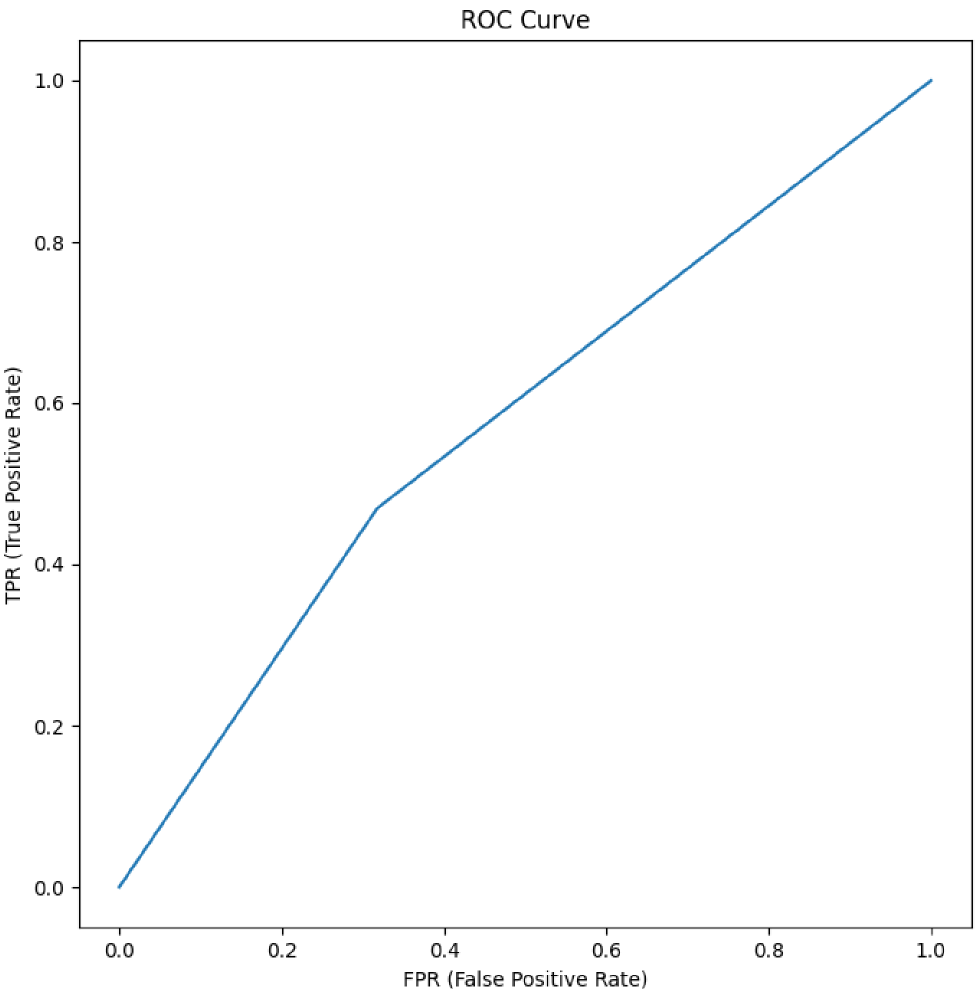
F Score Feature Selection ROC Curve



Forward Sequential Feature Selection ROC Curve



Recursive Feature Elimination



Select from Model Feature Selection

8.3.3 Logistic Regression

This classifier was implemented in python using the default parameters.

Feature Selection	Class	Accuracy	TPR	FPR	Precision	Recall	F1 Score	MCC	ROC
Correlation based feature selection	Heart Attack = 1	0.939925565	0.993660139	0.006339861	0.945552382	0.993660139	0.969006476	0.020211067	0.555098065
	Heart Attack = 2	0.939925565	0.013902739	0.986097261	0.112105045	0.013902739	0.024199227	0.020211067	0.555098065
	Weighted	0.939925565	0.939925565	0.060074435	0.899846398	0.939925565	0.917188268	0.020211067	0.555098065
F Score feature selection	Heart Attack = 1	0.92568146	0.975048673	0.024951327	0.947818235	0.975048673	0.961240309	0.069293863	0.625568559
	Heart Attack = 2	0.92568146	0.074835996	0.925164004	0.148487005	0.074835996	0.099468677	0.069293863	0.625568559
	Weighted	0.92568146	0.92568146	0.07431854	0.903982558	0.92568146	0.913981198	0.069293863	0.625568559
Forward Sequential Feature Selection	Heart Attack = 1	0.930821283	0.980358199	0.019641801	0.948203107	0.980358199	0.964010279	0.087814234	0.689557136
	Heart Attack = 2	0.930821283	0.07699684	0.92300316	0.186903503	0.07699684	0.108602659	0.087814234	0.689557136
	Weighted	0.930821283	0.930821283	0.069178717	0.906449559	0.930821283	0.917102204	0.087814234	0.689557136
Recursive Feature Elimination	Heart Attack = 1	0.924169229	0.973034647	0.026965353	0.94809654	0.973034647	0.960392896	0.07304206	0.655215776
	Heart Attack = 2	0.924169229	0.081769912	0.918230088	0.150837778	0.081769912	0.104565068	0.07304206	0.655215776
	Weighted	0.924169229	0.924169229	0.075830771	0.904372082	0.924169229	0.913468204	0.07304206	0.655215776
Select from Model	Heart Attack = 1	0.891469424	0.928317529	0.071682471	0.95563816	0.928317529	0.941629602	0.153130997	0.72649508
	Heart Attack = 2	0.891469424	0.255907016	0.744092984	0.17385881	0.255907016	0.202943629	0.153130997	0.72649508
	Weighted	0.891469424	0.891469424	0.108530576	0.91276306	0.891469424	0.901130137	0.153130997	0.72649508

Correlation Based Feature Selection

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	127687	815
HEART ATTACK = 2	7353	103

F Score Feature Selection

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	125296	3206
HEART ATTACK = 2	6898	558

Forward Sequential Feature Selection

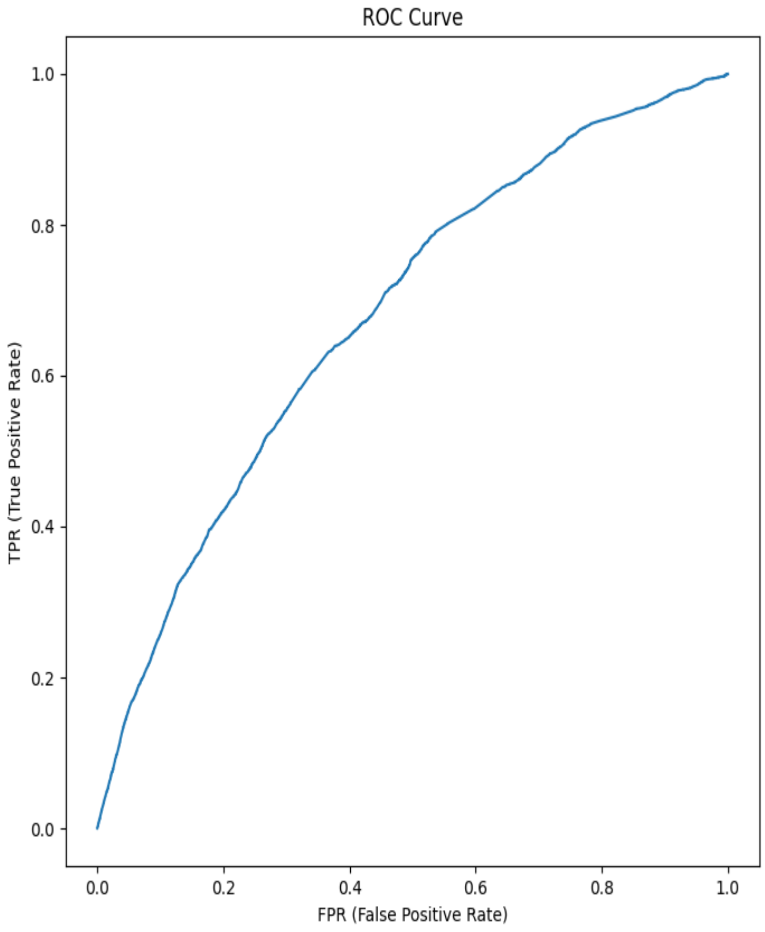
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	125978	2524
HEART ATTACK = 2	6881	575

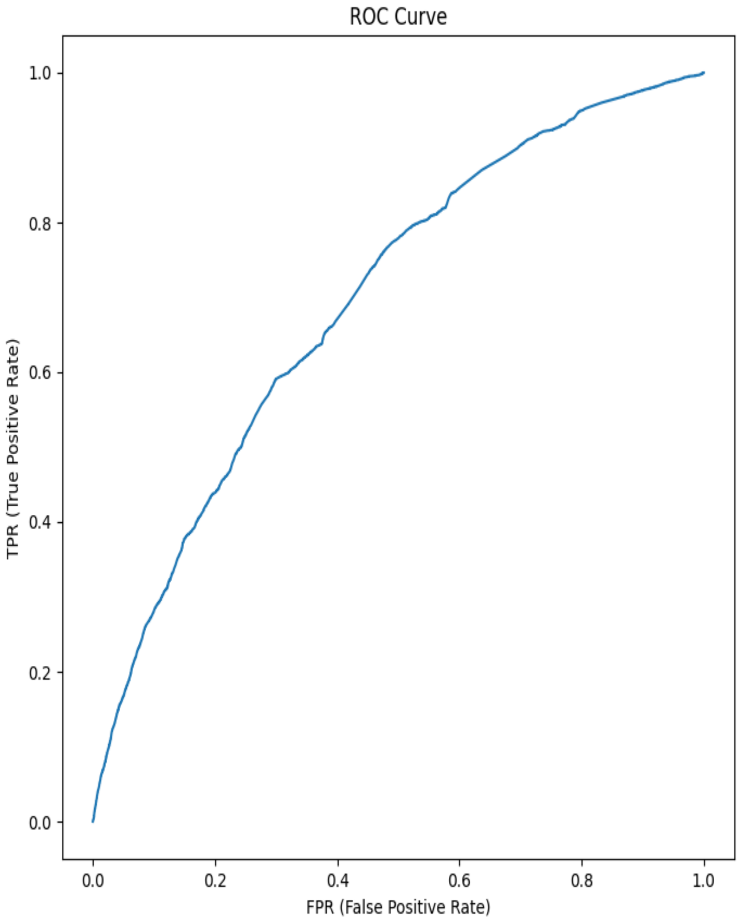
Recursive Feature Elimination

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	125037	3465
HEART ATTACK = 2	6845	611

Select From Model Feature Selection

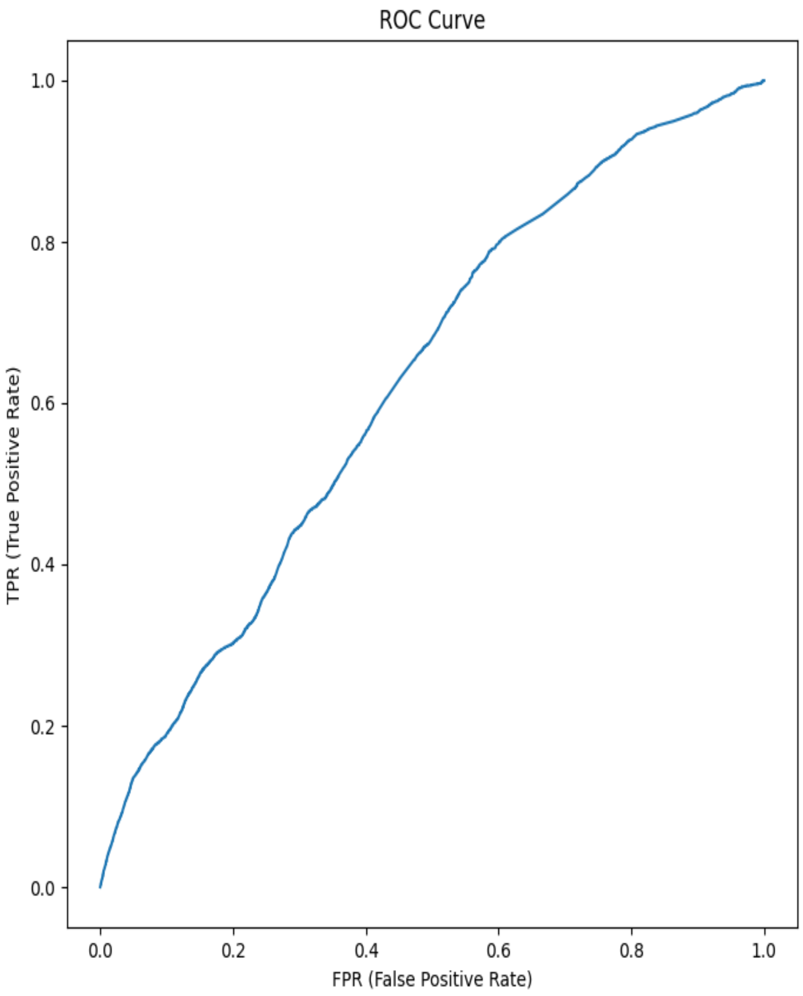
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	119291	9211
HEART ATTACK = 2	5545	1911



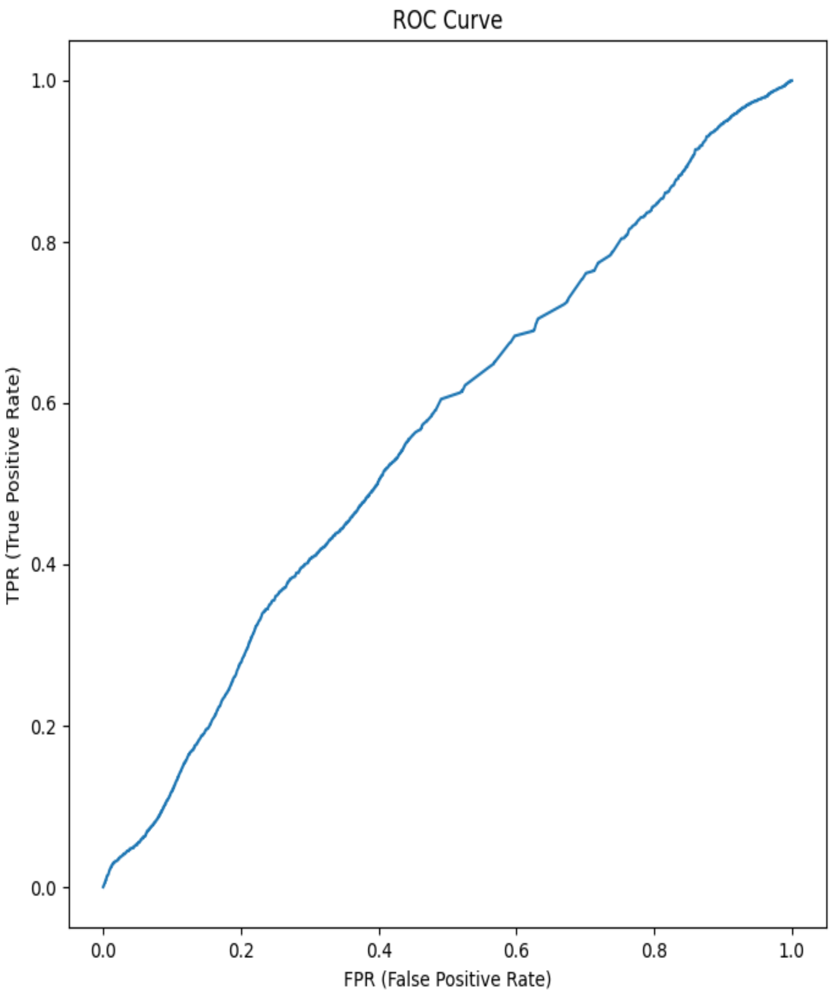


Correlation Based Feature Selection ROC Curve

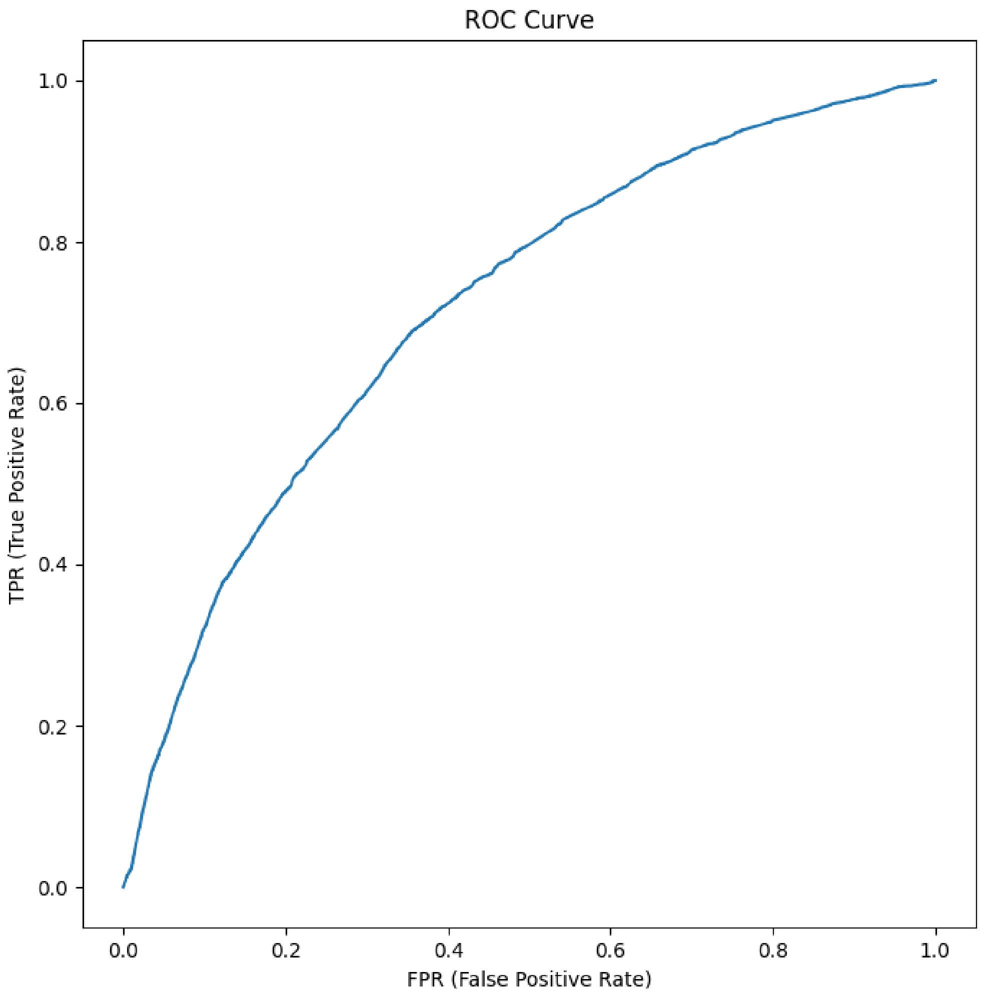
F Score Feature Selection ROC Curve



Forward Sequential Feature Selection ROC Curve



Recursive Feature Elimination



Select from Model Feature Selection

8.3.4 Gaussian Naïve Bayesian

This classifier was implemented in python using the default parameters.

Feature Selection	Class	Accuracy	TPR	FPR	Precision	Recall	F1 Score	MCC	ROC
Correlation based feature selection	Heart Attack = 1	0.593221436	0.597961847	0.402038153	0.955939566	0.597961847	0.698236374	0.059092199	0.588311161
	Heart Attack = 2	0.593221436	0.508529906	0.491470094	0.076929565	0.508529906	0.127456164	0.059092199	0.588311161
	Weighted	0.593221436	0.593221436	0.406778564	0.907725877	0.593221436	0.667002642	0.059092199	0.588311161
F Score feature selection	Heart Attack = 1	0.850155195	0.89339052	0.10660948	0.945064536	0.89339052	0.918501163	-0.001179156	0.501235911
	Heart Attack = 2	0.850155195	0.105010211	0.894989789	0.05406605	0.105010211	0.071379539	-0.001179156	0.501235911
	Weighted	0.850155195	0.850155195	0.149844805	0.896201432	0.850155195	0.872044426	-0.001179156	0.501235911
Forward Sequential Feature Selection	Heart Attack = 1	0.934903426	0.988040617	0.011959383	0.945534841	0.988040617	0.966315006	0.014756836	0.589052123
	Heart Attack = 2	0.934903426	0.019153443	0.980846557	0.086149322	0.019153443	0.030615629	0.014756836	0.589052123
	Weighted	0.934903426	0.934903426	0.065096574	0.898399865	0.934903426	0.914997838	0.014756836	0.589052123
Recursive Feature Elimination	Heart Attack = 1	0.87457303	0.920822836	0.079177164	0.945052032	0.920822836	0.932763005	-0.001346887	0.584556073
	Heart Attack = 2	0.87457303	0.077359496	0.922640504	0.053979697	0.077359496	0.063410847	-0.001346887	0.584556073
	Weighted	0.87457303	0.87457303	0.12542697	0.896183985	0.87457303	0.885089343	-0.001346887	0.584556073
Select from Model	Heart Attack = 1	0.87457303	0.920822836	0.079177164	0.945052032	0.920822836	0.932763005	-0.001346887	0.584556073
	Heart Attack = 2	0.87457303	0.077359496	0.922640504	0.053979697	0.077359496	0.063410847	-0.001346887	0.584556073
	Weighted	0.87457303	0.87457303	0.12542697	0.896183985	0.87457303	0.885089343	-0.001346887	0.584556073

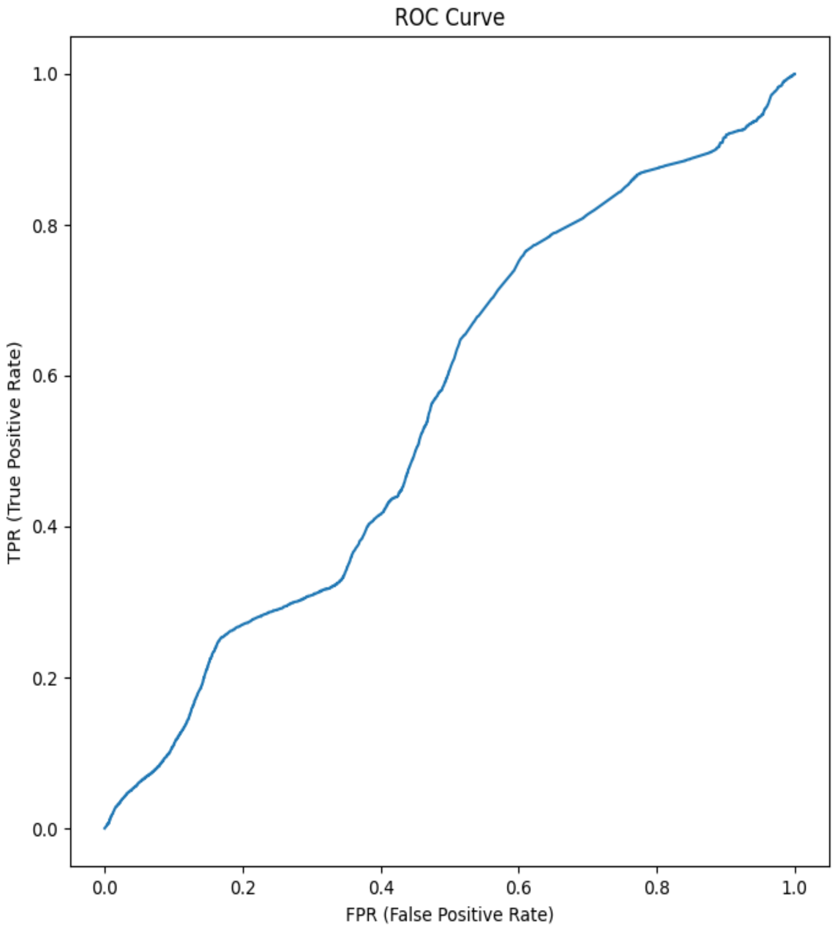
Correlation Based Feature Selection		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	76851	51651
HEART ATTACK = 2	3654	3802

F Score Feature Selection		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	114803	13699
HEART ATTACK = 2	6673	783

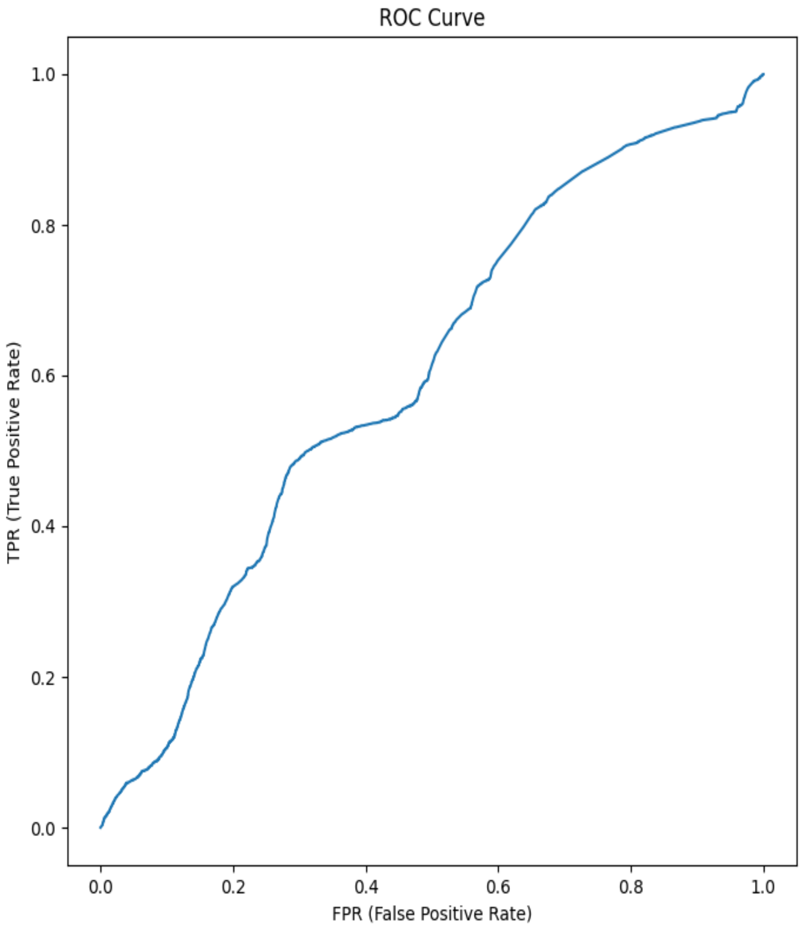
Forward Sequential Feature Selection		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	126965	1537
HEART ATTACK = 2	7314	142

Recursive Feature Elimination		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	118328	10174
HEART ATTACK = 2	6879	577

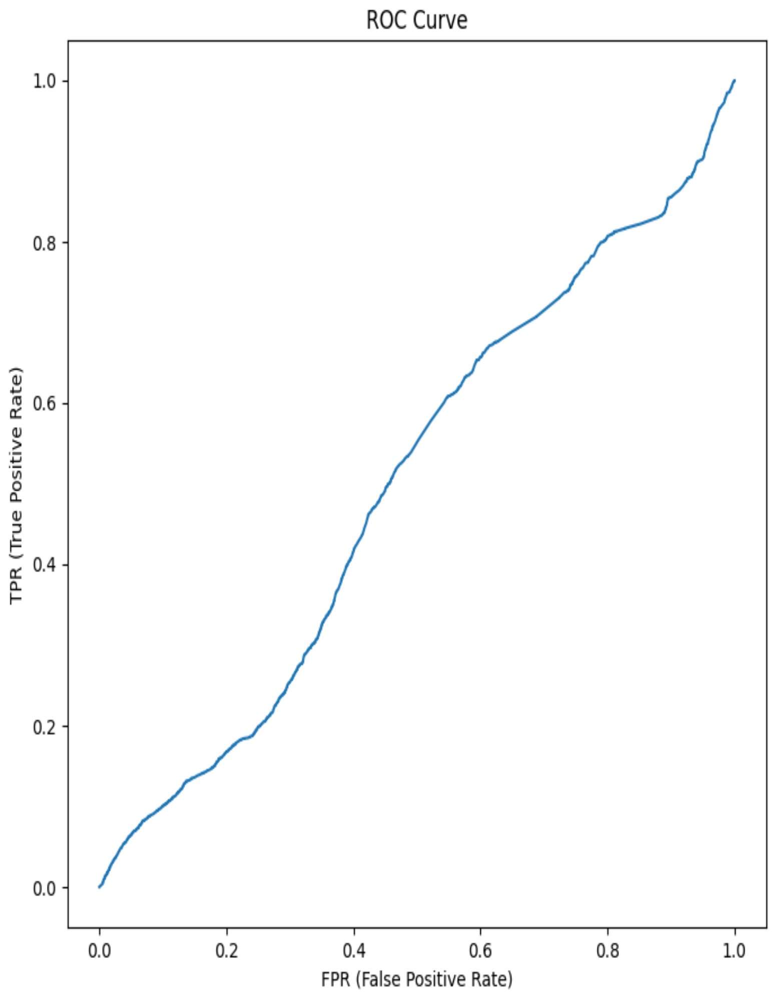
Select From Model Feature Selection		
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	118328	10174
HEART ATTACK = 2	6879	577



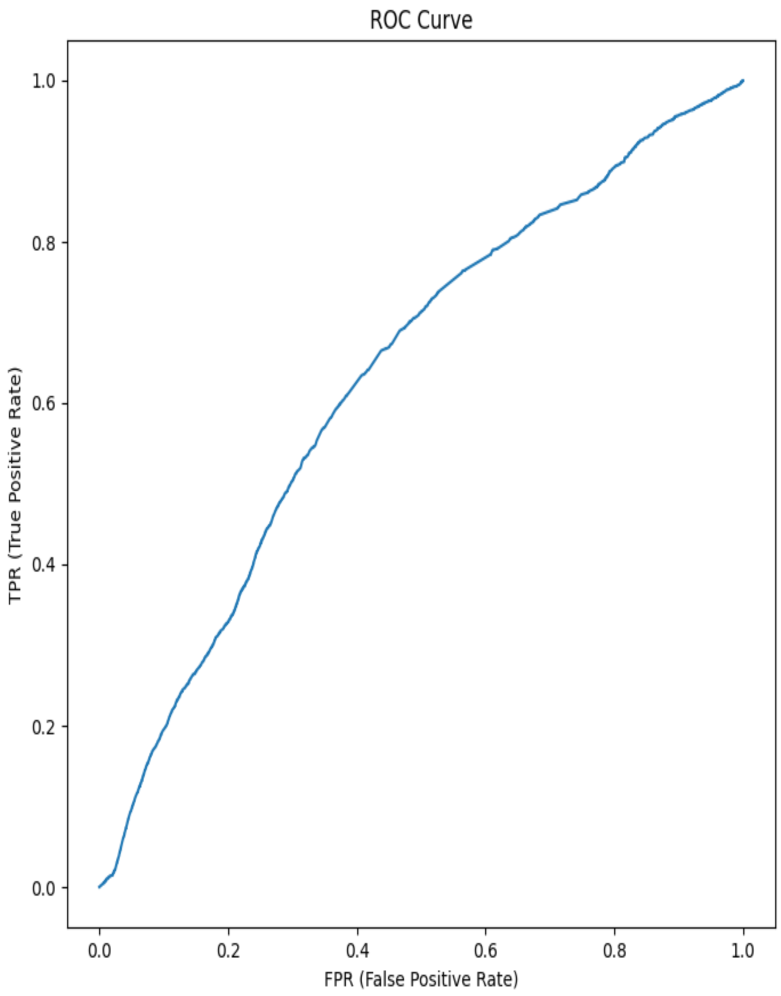
Correlation Based Feature Selection ROC Curve



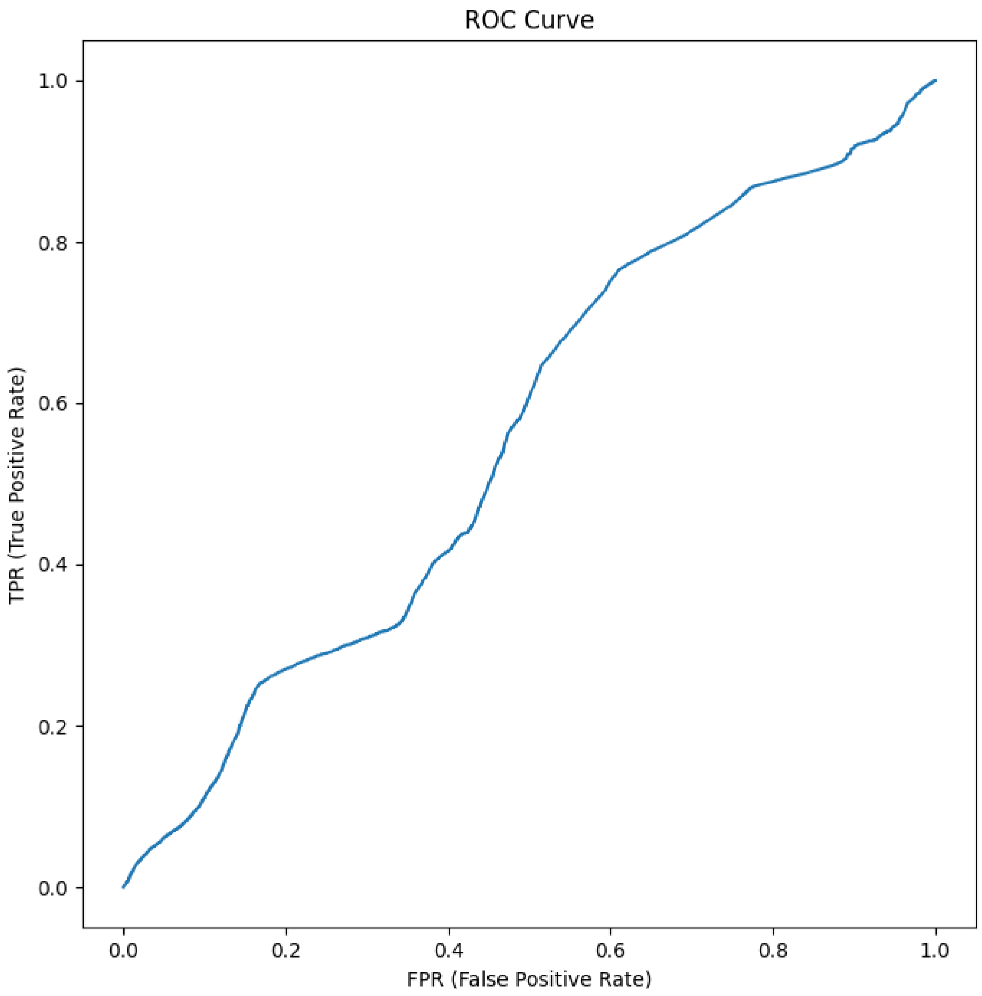
F Score Feature Selection ROC Curve



Forward Sequential Feature Selection ROC Curve



Recursive Feature Elimination



Select from Model Feature Selection

8.3.5 Random Forest

This classifier was implemented in python using entropy as the measure of quality of the split. The default number of trees, 100 were used to build this model.

Feature Selection	Class	Accuracy	TPR	FPR	Precision	Recall	F1 Score	MCC	ROC
Correlation based feature selection	Heart Attack = 1	0.49936451	0.498902735	0.501097265	0.944821975	0.498902735	0.643547883	0.004891152	0.502739416
	Heart Attack = 2	0.49936451	0.507213641	0.492786359	0.058497443	0.507213641	0.103590408	0.004891152	0.502739416
	Weighted	0.49936451	0.49936451	0.50063549	0.89621531	0.49936451	0.613947283	0.004891152	0.502739416
F Score feature selection	Heart Attack = 1	0.727112785	0.757211373	0.242788627	0.942079646	0.757211373	0.830697443	-0.016554567	0.478281542
	Heart Attack = 2	0.727112785	0.20786423	0.79213577	0.050379653	0.20786423	0.075889555	-0.016554567	0.478281542
	Weighted	0.727112785	0.727112785	0.272887215	0.893183917	0.727112785	0.78931956	-0.016554567	0.478281542
Forward Sequential Feature Selection	Heart Attack = 1	0.7371806	0.76359033	0.23640967	0.94923014	0.76359033	0.835047959	0.025591626	0.543785335
	Heart Attack = 2	0.7371806	0.280669931	0.719330069	0.066402759	0.280669931	0.10044698	0.025591626	0.543785335
	Weighted	0.7371806	0.7371806	0.2628194	0.90082409	0.7371806	0.794793721	0.025591626	0.543785335
Recursive Feature Elimination	Heart Attack = 1	0.559458068	0.565114091	0.434885909	0.937900204	0.565114091	0.647639433	0.012350493	0.541127136
	Heart Attack = 2	0.559458068	0.467077242	0.532922758	0.060961525	0.467077242	0.103535007	0.012350493	0.541127136
	Weighted	0.559458068	0.559458068	0.440541932	0.889799699	0.559458068	0.61764464	0.012350493	0.541127136
Select from Model	Heart Attack = 1	0.740769944	0.763996558	0.236003442	0.952375335	0.763996558	0.847600244	0.054819755	0.586494884
	Heart Attack = 2	0.740769944	0.340011244	0.659988756	0.076554944	0.340011244	0.124782534	0.054819755	0.586494884
	Weighted	0.740769944	0.740769944	0.259230056	0.904343254	0.740769944	0.807967736	0.054819755	0.586494884

Correlation Based Feature Selection

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	64113	64389
HEART ATTACK = 2	3677	3779

F Score Feature Selection

	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	97304	31197
HEART ATTACK = 2	5904	1553

Forward Sequential Feature Selection

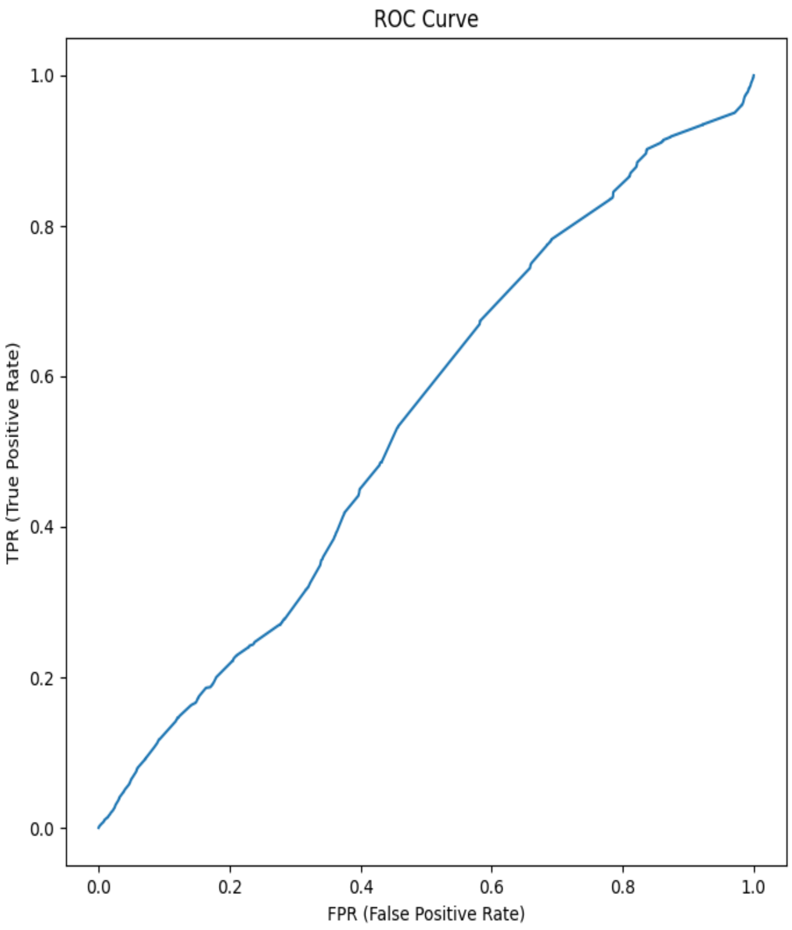
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	98126	30376
HEART ATTACK = 2	5356	2100

Recursive Feature Elimination

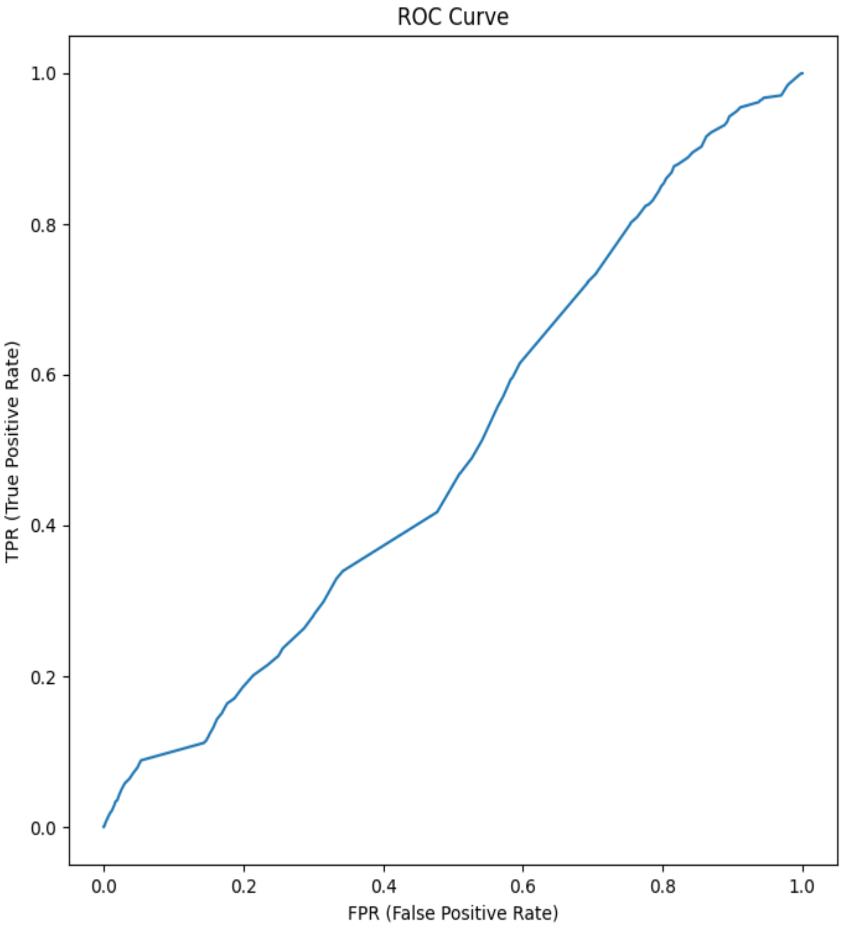
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	72599	55903
HEART ATTACK = 2	3992	3464

Select From Model Feature Selection

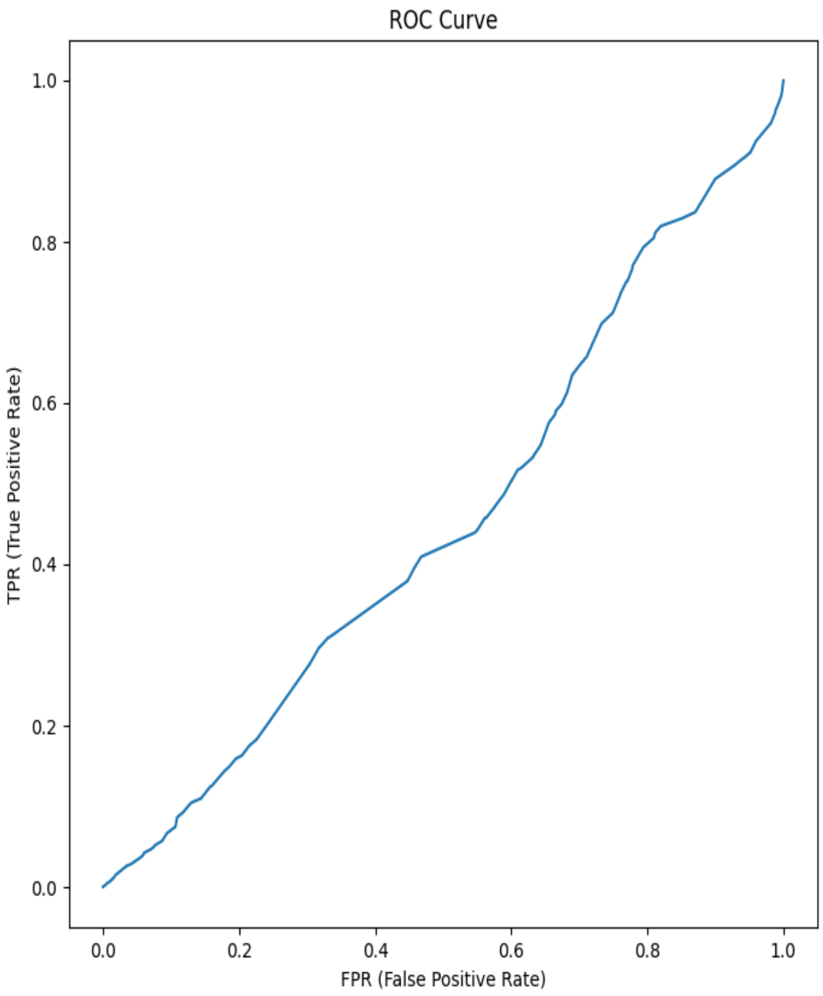
	HEART ATTACK = 1	HEART ATTACK = 2
HEART ATTACK = 1	98177	30325
HEART ATTACK = 2	4919	2537



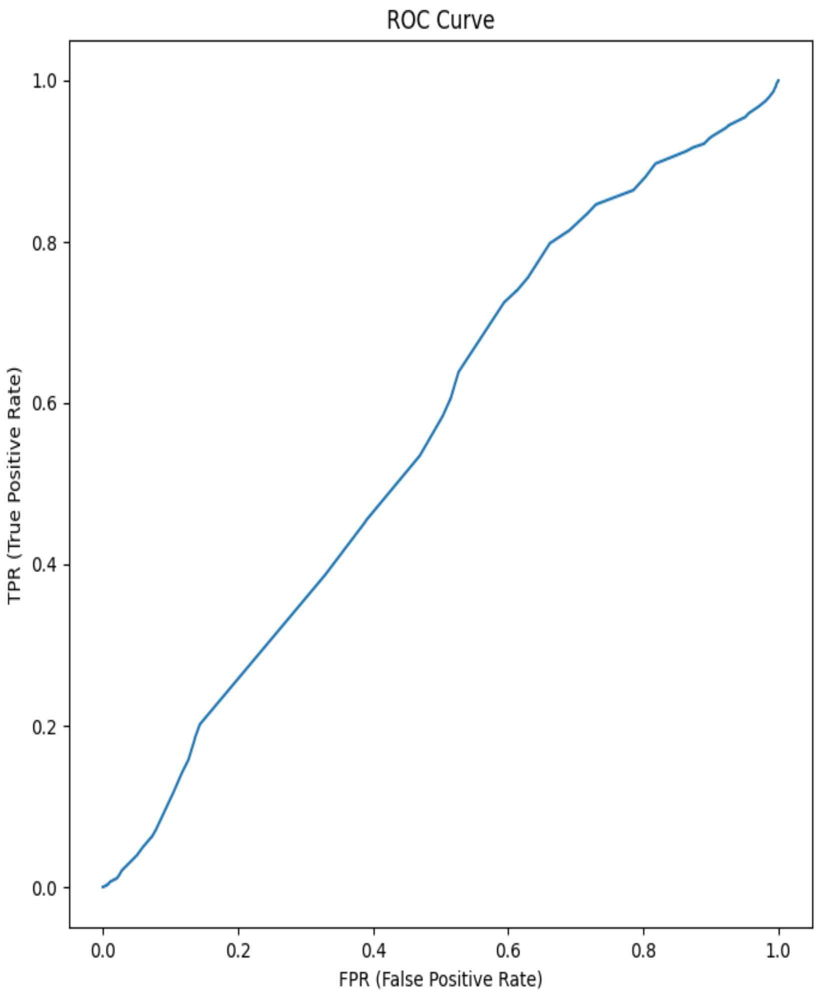
Correlation Based Feature Selection ROC Curve



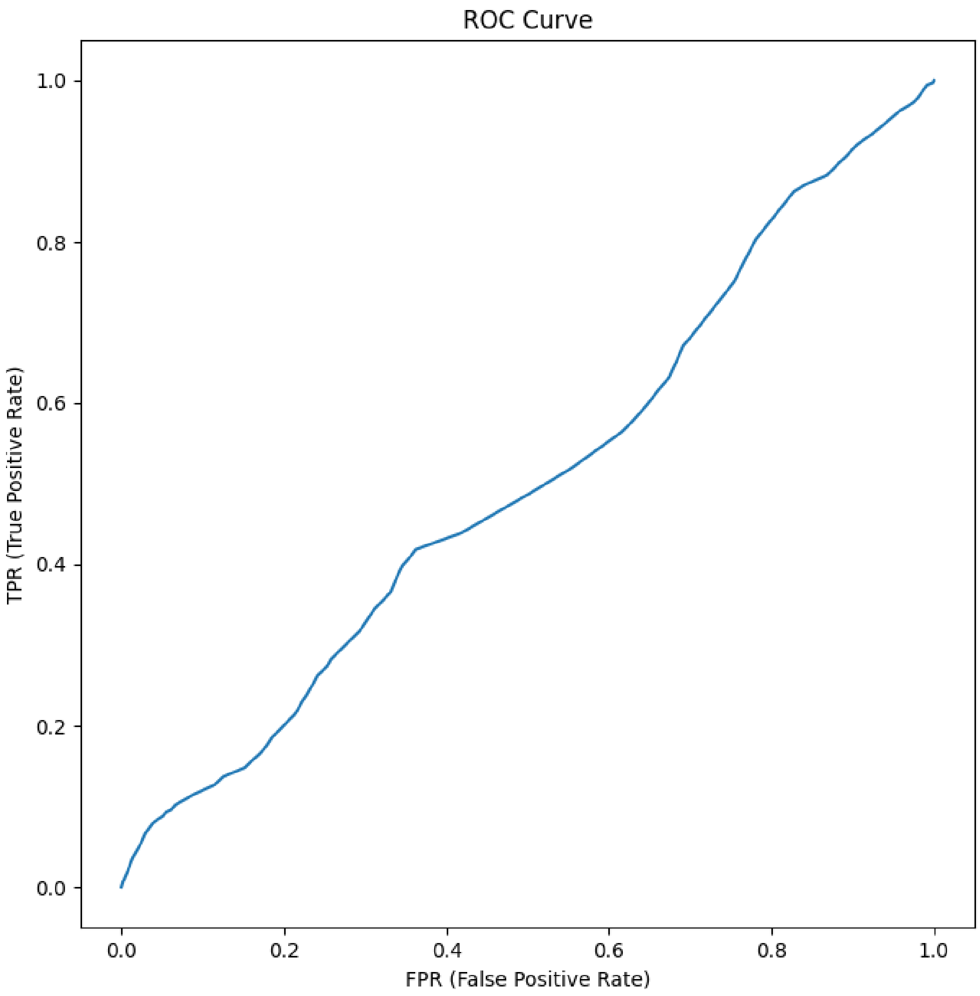
F Score Feature Selection ROC Curve



Forward Sequential Feature Selection ROC Curve



Recursive Feature Elimination



Select from Model Feature Selection

8.4 Best Model for SMOTE Balanced Data Set

We first look at the best model for each machine learning classifier from the feature selected data sets. The following models were chosen:

Classifier	Feature Selection	Class	Accuracy	TPR	FPR	Precision	Recall	F1 Score	MCC	ROC
Ada Boost	Select from Model	Heart Attack = 1	0.915259124	0.961845377	0.038154623	0.949270616	0.961845377	0.955267138	0.076989577	0.702672513
		Heart Attack = 2	0.915259124	0.112024146	0.887975854	0.141530531	0.112024146	0.109501674	0.076989577	0.702672513
		Weighted	0.915259124	0.915259124	0.084740876	0.904972904	0.915259124	0.908893848	0.076989577	0.702672513
Decision Tree	Select from Model	Heart Attack = 1	0.463094485	0.45480204	0.54519796	0.951618537	0.45480204	0.593006675	0.030341864	0.53129727
		Heart Attack = 2	0.463094485	0.607772787	0.392227213	0.063128287	0.607772787	0.113216569	0.030341864	0.53129727
		Weighted	0.463094485	0.463094485	0.536905515	0.902892084	0.463094485	0.56665121	0.030341864	0.53129727
Logistic Regression	Select from Model	Heart Attack = 1	0.891469424	0.928317529	0.071682471	0.95563816	0.928317529	0.941629602	0.153130997	0.72649508
		Heart Attack = 2	0.891469424	0.255907016	0.744092984	0.17385881	0.255907016	0.202943629	0.153130997	0.72649508
		Weighted	0.891469424	0.891469424	0.108530576	0.91276306	0.891469424	0.901130137	0.153130997	0.72649508
Gaussian Naïve Bayesian	Forward Sequential Feature Selection	Heart Attack = 1	0.934903426	0.988040617	0.011959383	0.945534841	0.988040617	0.966315006	0.014756836	0.589052123
		Heart Attack = 2	0.934903426	0.019153443	0.980846557	0.086149322	0.019153443	0.030615629	0.014756836	0.589052123
		Weighted	0.934903426	0.934903426	0.065096574	0.898399865	0.934903426	0.914997838	0.014756836	0.589052123
Random Forest	Select from Model	Heart Attack = 1	0.740769944	0.763996558	0.236003442	0.952375335	0.763996558	0.847600244	0.054819755	0.586494884
		Heart Attack = 2	0.740769944	0.340011244	0.659988756	0.076554944	0.340011244	0.124782534	0.054819755	0.586494884
		Weighted	0.740769944	0.740769944	0.259230056	0.904343254	0.740769944	0.807967736	0.054819755	0.586494884

The above models were chosen because they have a positive MCC score and the ROC area of each one of these models is greater than 0.5. Having a MCC score of greater than 0 means that the predictions are better than random. These models also have a better TPR for Heart Attack = 2 class which is the class of concern to us. We want to be able to be better at predicting a patient with heart attack chances but at the same time we do not want to misclassify too many patients with no risk of heart attack as those having a risk of a heart attack. These models also had much better accuracy than others.

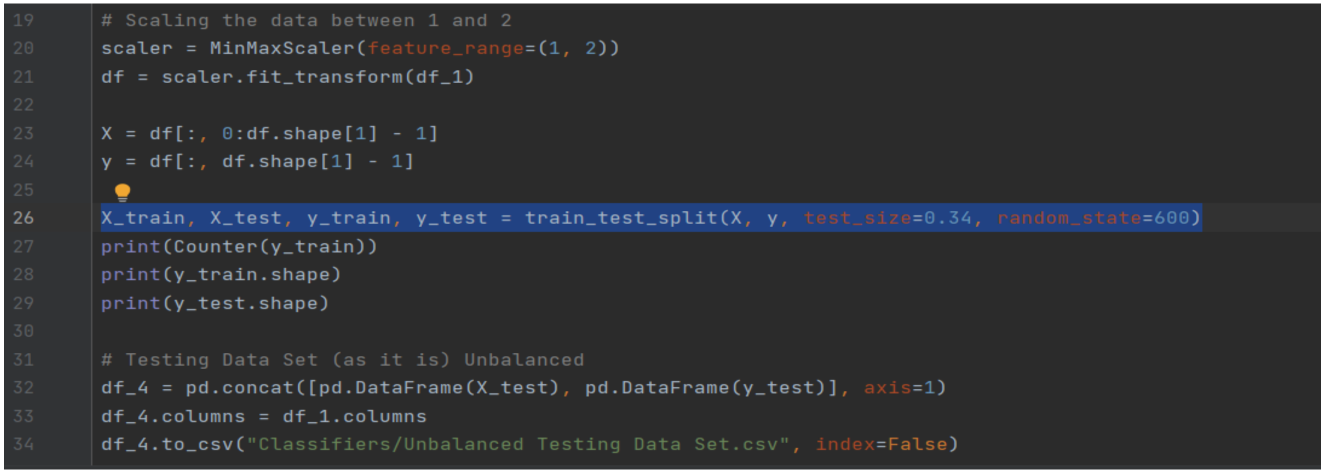
Now, we will select the best model from these 5 models. **We have selected the Logistic Regression Model with Select from Model feature selection as the best model.** The reason for this being that this model has the best ROC area under curve and the best MCC. Although, it has slightly lower accuracy than other models, but it is also better than the other models at predicting the Heart Attack = 2 (TPR = 0.255907016) class which is of concern to us as we want to be able to identify patient with risk of a heart attack. At the same time, it does not misclassify a lot of the Heart Attack = 1 class (TPR = 0.928317529).

8.5 Running the code

First, the data cleaning part was done using R. So, we first run the R Script named **“Data Preparation.R”**. Before running this, make sure to go to the **Session** tab in R, then **Set Working Directory** and select **To Source File Location**. Running this code will produce the cleaned data in a csv file named **“Full Clean Data.csv**”.

Then, we created a python file named **“Run Balancing & Feature Selection.py”**. This will run all the python files which contain the code to balance the data set and then produce the features selected data set. A CSV file named **“Features Selected.csv”** is also output which contains the list of features selected by each of the feature selection methods. The python files run in this code are:

1. Balancing the Dataset.py – This script contains the code to min max scale, split into train test sets and balance the training set. This script is used to generate the data sets for multiple runs. The screenshot below shows the random state values in the **train_test_split** function that can be changed to 200, 300, 400, 500 & 600 to reproduce the same results. We just need to change the random state values and then run the main scripts as mentioned in the **NOTE** section below.



```
# Scaling the data between 1 and 2
scaler = MinMaxScaler(feature_range=(1, 2))
df = scaler.fit_transform(df_1)

X = df[:, 0:df.shape[1] - 1]
y = df[:, df.shape[1] - 1]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.34, random_state=600)
print(Counter(y_train))
print(y_train.shape)
print(y_test.shape)

# Testing Data Set (as it is) Unbalanced
df_4 = pd.concat([pd.DataFrame(X_test), pd.DataFrame(y_test)], axis=1)
df_4.columns = df_1.columns
df_4.to_csv("Classifiers/Unbalanced Testing Data Set.csv", index=False)
```

2. Correlation Based FS Data Set.py – This script produces the correlation feature selected data set in **“Classifiers”** folder named **“Borderline SMOTE Correlation Selected Data Set.csv”** & **“SMOTE Correlation Selected Data Set.csv”**.
3. F Score Attribute Selection Data Set.py – This script produces the f score feature selected data set in **“Classifiers”** folder named **“Borderline SMOTE F Score Selected Data Set.csv”** & **“SMOTE F Score Selected Data Set.csv”**.
4. Forward SFS Data Set.py – This script produces the forward sequential feature selected data set in **“Classifiers”** folder named **“Borderline SMOTE Forward SFS Data Set.csv”** & **“SMOTE Forward SFS Data Set.csv”**.

5. RFE FS Data Set.py – This script produces the recursive feature elimination feature selected data set in **“Classifiers”** folder named **“Borderline SMOTE RFE Data Set.csv”** & **“SMOTE RFE Data Set.csv”**.
6. Select From Model Data Set.py – This script produces the select from model feature selected data set in **“Classifiers”** folder named **“Borderline SMOTE Select from Model Data Set.csv”** & **“SMOTE Select from Model Data Set.csv”**.

Now, to perform machine learning on the dataset, go inside the “**Classifiers”** folder and run the python script named **“Run Classification.py”**. This script will run all the python scripts for the classification models. The output of this script are the calculated metrics saved in the **“Borderline SMOTE Metrics.csv”** & **“SMOTE Metrics.csv”** files. ROC curves are also plotted, displayed, and saved. The python files run in this code are:

1. Ada Boost Correlation Feature selection.py – This script will run the Ada Boost classifier on the Correlation Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
2. Ada Boost F Score Feature Selection.py – This script will run the Ada Boost classifier on the F Score Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
3. Ada Boost Forward SFS.py – This script will run the Ada Boost classifier on the Forward Sequential Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
4. Ada Boost RFE Feature Selection.py – This script will run the Ada Boost classifier on the Recursive Feature Elimination Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
5. Ada Boost Select from Model Feature Selection.py – This script will run the Ada Boost classifier on the Select from Model Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
6. Decision Tree Correlation Feature selection.py – This script will run the Decision Tree classifier on the Correlation Feature Selected data set which was balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.

7. Decision Tree F Score Feature Selection.py – This script will run the Decision Tree classifier on the F Score Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
8. Decision Tree Forward SFS.py – This script will run the Decision Tree classifier on the Forward Sequential Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
9. Decision Tree RFE Feature Selection.py – This script will run the Decision Tree classifier on the Recursive Feature Elimination Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
10. Decision Tree Select from Model Feature Selection.py – This script will run the Decision Tree classifier on the Select from Model Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
11. Logistic Regression Correlation Feature selection.py – This script will run the Logistic Regression classifier on the Correlation Feature Selected data set which was balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
12. Logistic Regression F Score Feature Selection.py – This script will run the Logistic Regression classifier on the F Score Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
13. Logistic Regression Forward SFS.py – This script will run the Logistic Regression classifier on the Forward Sequential Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
14. Logistic Regression RFE Feature Selection.py – This script will run the Logistic Regression classifier on the Recursive Feature Elimination Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
15. Logistic Regression Select from Model Feature Selection.py – This script will run the Logistic Regression classifier on the Select from Model Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.

16. Naïve Bayes Correlation Feature selection.py – This script will run the Gaussian Naïve Bayesian classifier on the Correlation Feature Selected data set which was balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
17. Naïve Bayes F Score Feature Selection.py – This script will run the Gaussian Naïve Bayesian classifier on the F Score Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
18. Naïve Bayes Forward SFS.py – This script will run the Gaussian Naïve Bayesian classifier on the Forward Sequential Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
19. Naïve Bayes RFE Feature Selection.py – This script will run the Gaussian Naïve Bayesian classifier on the Recursive Feature Elimination Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
20. Naïve Bayes Select from Model Feature Selection.py – This script will run the Gaussian Naïve Bayesian classifier on the Select from Model Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
21. Random Forest Correlation Feature selection.py – This script will run the Random Forest classifier on the Correlation Feature Selected data set which was balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
22. Random Forest F Score Feature Selection.py – This script will run the Random Forest classifier on the F Score Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
23. Random Forest Forward SFS.py – This script will run the Random Forest classifier on the Forward Sequential Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.
24. Random Forest RFE Feature Selection.py – This script will run the Random Forest classifier on the Recursive Feature Elimination Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.

25. Random Forest Select from Model Feature Selection.py – This script will run the Random Forest classifier on the Select from Model Feature Selected data sets which were balanced by both SMOTE & Borderline SMOTE techniques. This will also display the metrics and save the ROC Curve.

To perform 10-fold cross validation, we need to run the script named **"Run 10 Cross Validation.py"**. The output of this script are the calculated metrics saved in a CSV file named **"10 CV Metrics.csv"** and the ROC curves which are plotted, displayed, and saved. This script will then run the following python files to perform the 10-fold cross validation:

1. Ada Boost 10 CV.py – This script runs 10-fold cross validation using Ada Boost classifier and then calculates the metrics and plots and save the ROC curves.
2. Decision Tree 10 CV.py – This script runs 10-fold cross validation using Decision Tree classifier and then calculates the metrics and plots and save the ROC curves.
3. Logistic Regression 10 CV.py – This script runs 10-fold cross validation using Logistic Regression classifier and then calculates the metrics and plots and save the ROC curves.
4. Naïve Bayes 10 CV.py – This script runs 10-fold cross validation using Naïve Bayes classifier and then calculates the metrics and plots and save the ROC curves.
5. Random Forest 10 CV.py – This script runs 10-fold cross validation using Random Forest classifier and then calculates the metrics and plots and save the ROC curves.

NOTE: Please run the scripts from the main scripts in the following order:

1. **Data Preparation.R**
2. **Run Balancing & Feature Selection.py**
3. **Run Classification.py**
4. **Run 10 Cross Validation.py**

We calculated the average over the runs in excel and displayed the averaged results in this report. The excel files in which the averages were performed are also included. These scripts will take quite some time to run as the data set is very large and the techniques are computation extensive.

9. Conclusion

We used two techniques to balance the data set, and both yielded the same results that Logistic Regression was the best classifier for this data set. We saw that different data set balancing techniques can give different results and can impact the metrics as well. Feature selection yielded different results for the two techniques used to balance the data set.

In this project, we followed the overall process of data mining from data preparation to machine learning. We learnt how to deal with unbalanced data sets by using different techniques. We also learnt about different feature selection methods and different classification algorithms. We saw that there is no one algorithm which always works the best. We need to experiment with different algorithms and then calculate the metrics to decide which one is the best for the data set. In this project, we only used 5 algorithms to test on the data set. But, to select the best algorithm, we need to experiment with more classifiers and other data set balancing techniques to get an evaluation of the best model to implement for the data set.

10. Contribution

Osama, Muhammad – Data preparation & Visualization, Correlation based feature selection, F Score feature selection, implementation of Ada Boost, Decision Tree & Logistic Regression Classifier. Explained these in the project report as well.
Oruganti, Sravani – Data preparation & Visualization, forward sequential feature selection, recursive feature elimination, select from model feature selection, implementation of Gaussian Naïve Bayesian & Random Forest Classifier. Explained these in the project report as well.

11. References

https://builtin.com/data-science/random-forest-algorithm

https://builtin.com/data-science/random-forest-algorithm

https://en.wikipedia.org/wiki/Random_forest

https://www.geeksforgeeks.org/understanding-logistic-regression/

https://web.stanford.edu/~jurafsky/slp3/5.pdf

https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575

https://www.sciencedirect.com/topics/computer-science/decision-tree-classifier

https://www.datacamp.com/community/tutorials/adaboost-classifier-python

https://medium.datadriveninvestor.com/data-science-adaboost-classifier-4879a45c4300

https://www.researchgate.net/publication/323119678_AdaBoost_classifier_an_overview

https://deepai.org/machine-learning-glossary-and-terms/f-score#:~:text=The%20F-score%2C%20also%20called%20the%20F1-score%2C%20is%20a,harmonic%20mean%20of%20the%20model%E2%80%99s%20precision%20and%20recall.

https://stats.stackexchange.com/questions/20341/the-disadvantage-of-using-f-score-in-feature-selection

https://www.simplilearn.com/recursive-feature-elimination-article

http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/

https://analyticsindiamag.com/a-complete-guide-to-sequential-feature-selection/#:~:text=%20A%20Complete%20Guide%20to%20Sequential%20Feature%20Selection,wrapper...%204%20Results%20of%20SFS.%20%20More%20