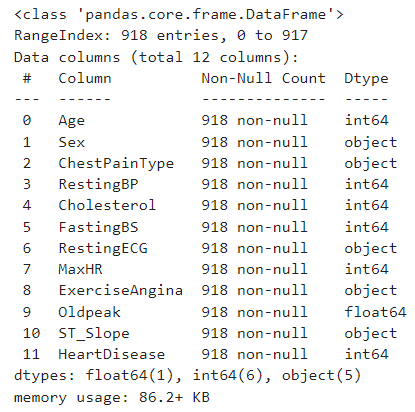
Summary of the Classification dataset   
– Heart Disease Prediction

**About the data:**

Link: [Heart Disease Prediction Dataset](https://www.kaggle.com/datasets/krishujeniya/heart-diseae)  
Data Info:  
 

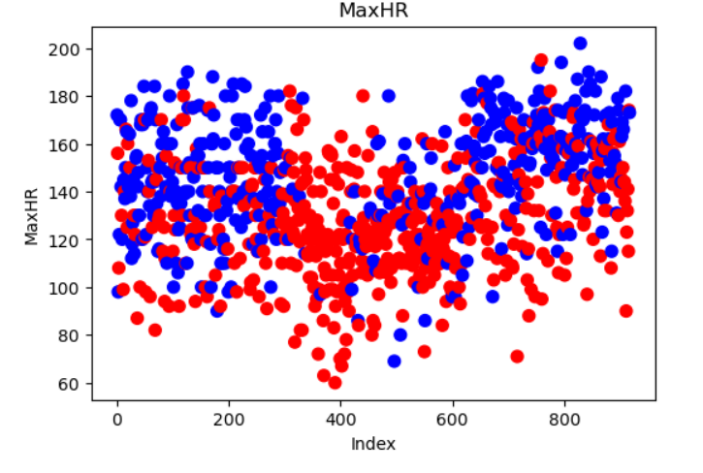
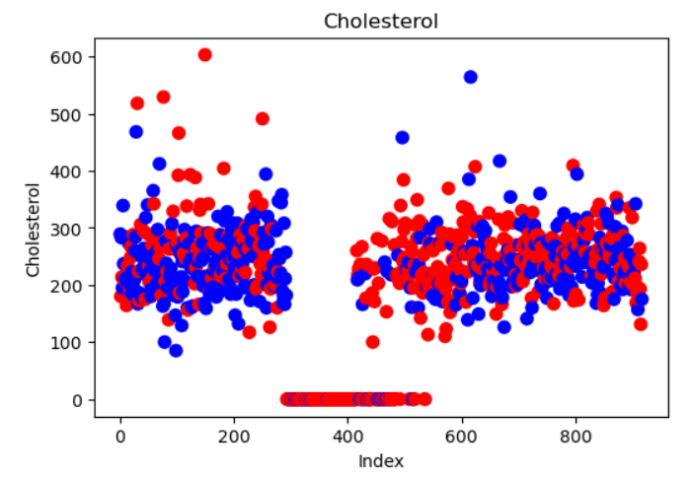
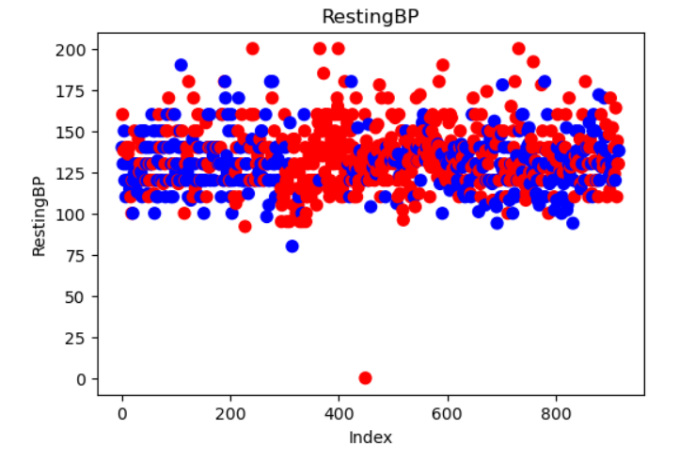
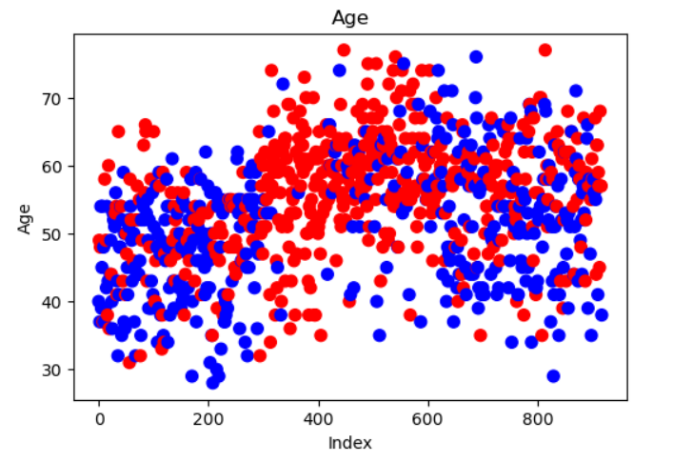
Columns:  
1) Age: The age of the individual in years.  
2) Sex: The gender of the individual (M->Male, F->Female).  
3) Chest Pain Type: Type of chest pain experienced (ASY->asymptomatic, TA->typical angina, ATA->atypical angina, NAP->non-anginal pain).  
4) Resting BP: Resting blood pressure in mm Hg at the time of admission.   
5) Cholesterol: Serum cholesterol level in mg/dL.  
6) Fasting BS: Fasting blood sugar level (>120 mg/dL = 1, otherwise = 0).  
7) Resting ECG: Results of the resting electrocardiogram (Normal->normal, ST->ST-T wave abnormality, LVH->probable/definite left ventricular hypertrophy).  
8) Max HR: Maximum heart rate achieved during exercise.  
9) Exercise Angina: Indication of exercise-induced angina (Y->Yes, N->No).  
10) Old Peak: ST depression induced by exercise relative to rest, indicating ischemia.  
11) ST Slope: The slope of the peak exercise ST segment (Down->down sloping, Flat->flat, Up->upsloping).  
12) Heart Disease: Target variable indicating the presence of heart disease (1 = Yes, 0 = No).

**Pre-Processing:**

Encoding the categorical fields into integer data:

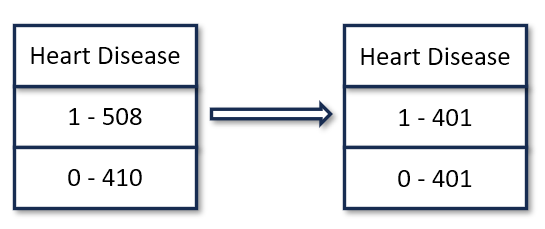
1. Sex:
   1. M – 1 b. F – 0
2. Chest Pain Type:
   1. ASY – 0 b. ATA – 1 c. NAP – 2 d. TA – 3
3. Resting ECG:
   1. LVH – 0 b. Normal – 1 c. ST – 2
4. Exercise Angina:
   1. N – 0 b. Y – 1
5. ST Slope:
   1. Down – 0 b. Flat – 1 c. Up – 2

**Plotting the data:**

1. Plotting all the fields against the index, using blue to represent data points without heart disease and red for those with heart disease:   
   

**Each field shows a consistent pattern or symmetry with respect to heart disease, indicating that all fields contribute valuable information. Therefore, no field can be removed from the dataset.**

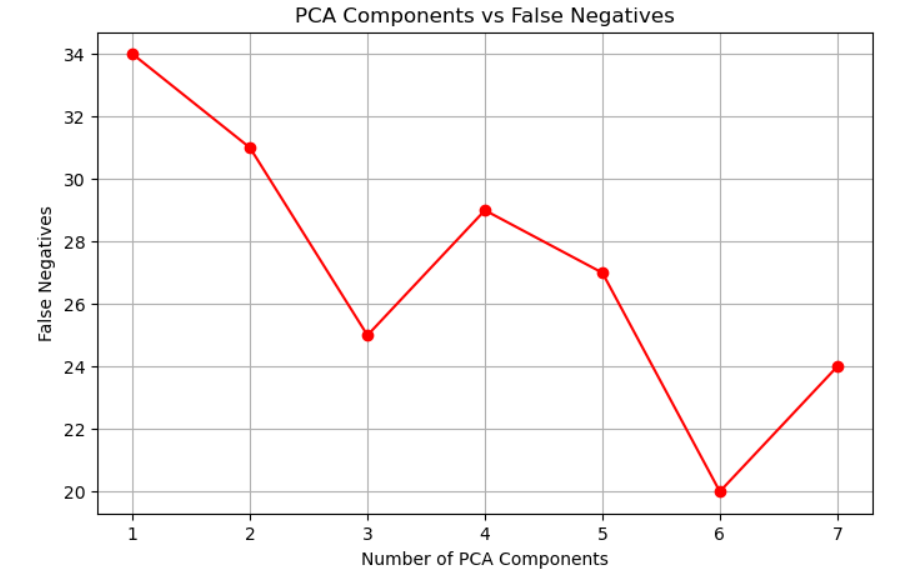
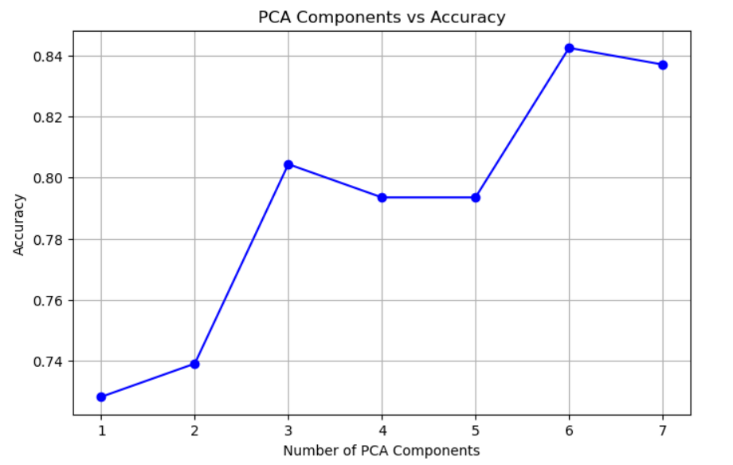
**Applying SMOTE** (Synthetic Minority Oversampling Technique)   
(to balance the dataset, ensuring an equal number of data points for cases with and without heart disease)



**Trying PCA and BERUTA for further Pre-processing:**

1. Selected Features after BERUTA:

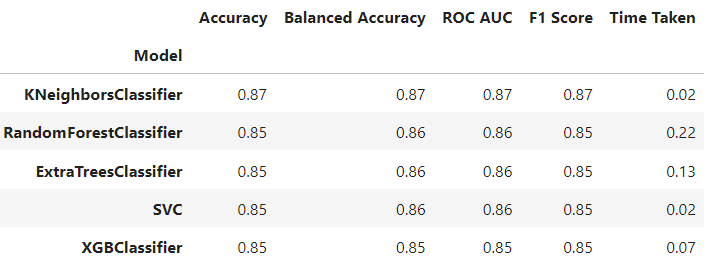
|  |  |  |
| --- | --- | --- |
| Age | Chest Pain Type | Cholesterol |
| Max HR | Exercise Angina | Old Peak |
| ST Slope |  |  |
|  |  |  |

1. Plot of Accuracy and False negatives Vs n-components of PCA**:  
   **

**Selected data after preprocessing:** Balanced data after applying PCA (n\_components=5)

**Model Selection:**

1. Used lazy predict to get some of the best possible models that can be used



1. Hyperparameter tuned the top models and got the best model as:
   1. Random Forest Classifier
   2. Parameters:
      1. max\_depth: None
      2. max\_features: 'sqrt'
      3. min\_samples\_leaf: 1
      4. min\_samples\_split: 10
      5. n\_estimators: 200
   3. Threshold of 0.32
   4. Accuracy: 0.87
   5. Number of False Negatives 5

Overall Result:  
Got by Random Forest Classifier to get maximum accuracy of 0.87 and minimum false negatives as 5.