Heart Disease

Muhammad Faraz Akram

Introduction

- Dataset

  Heart Diseases

- Data source

  CDC Official Website

- Reason

  To find which Age, Gender, Race is more at risk of getting Heart Disease

- Expectation

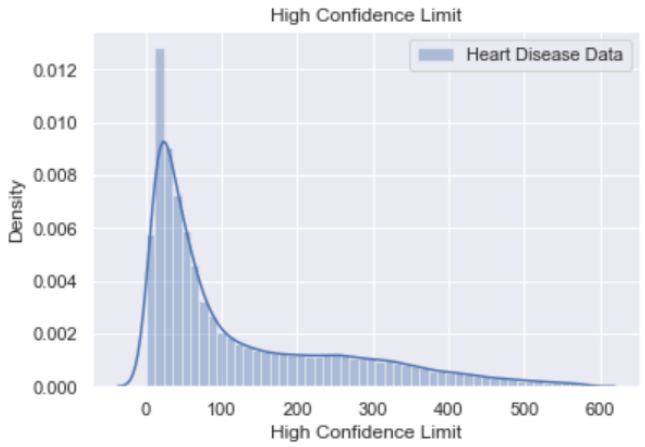
  To run classification models on the Dataset at hand

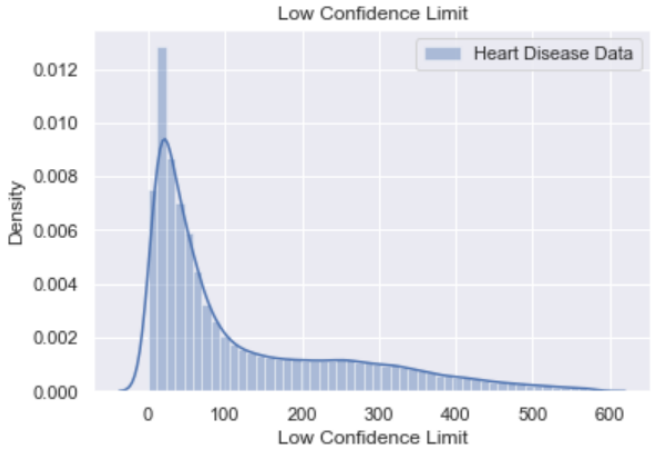
Cleaning of the Dataset

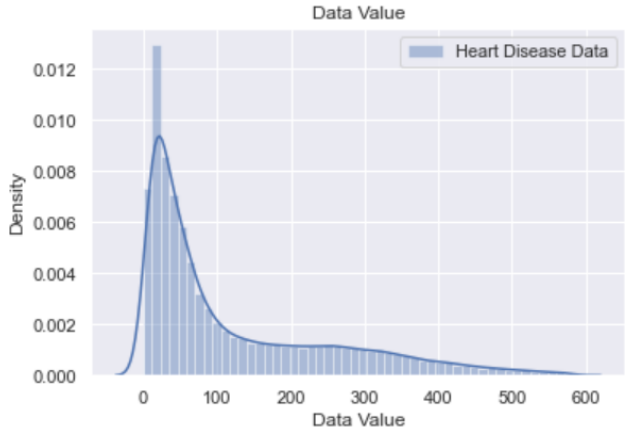
- Missing Values
- Target Values
- Unneeded Columns
- Plan of Action
- Unique Values
- Fixing Outliers

Feature Engineering

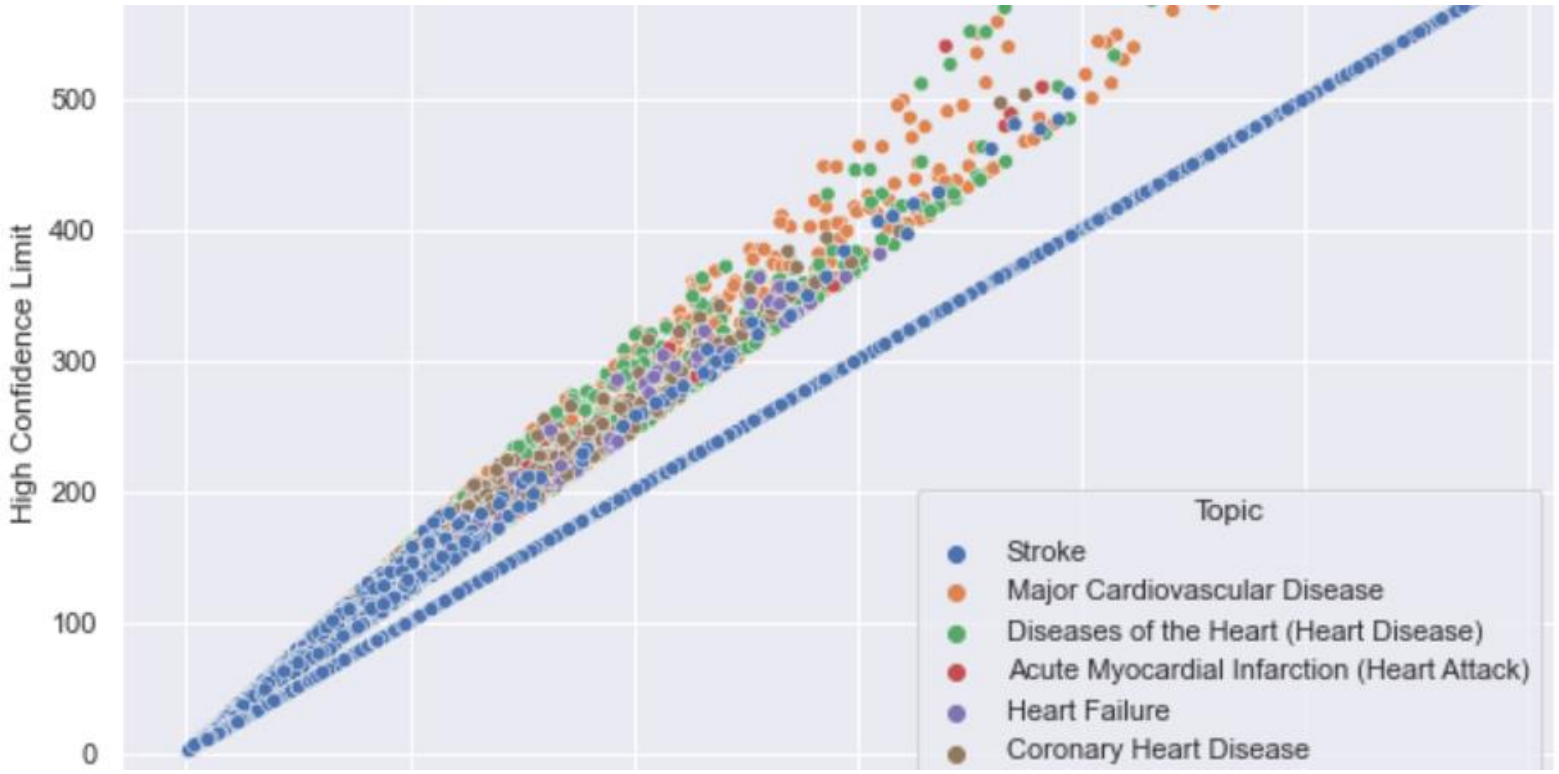
Diseases	Dummying up Target Variable (6 Types of Heart Conditions)
Categories	Dummying up Categories (Age, Race, Gender,..)
Binary Data	Converting Columns into Binary Data



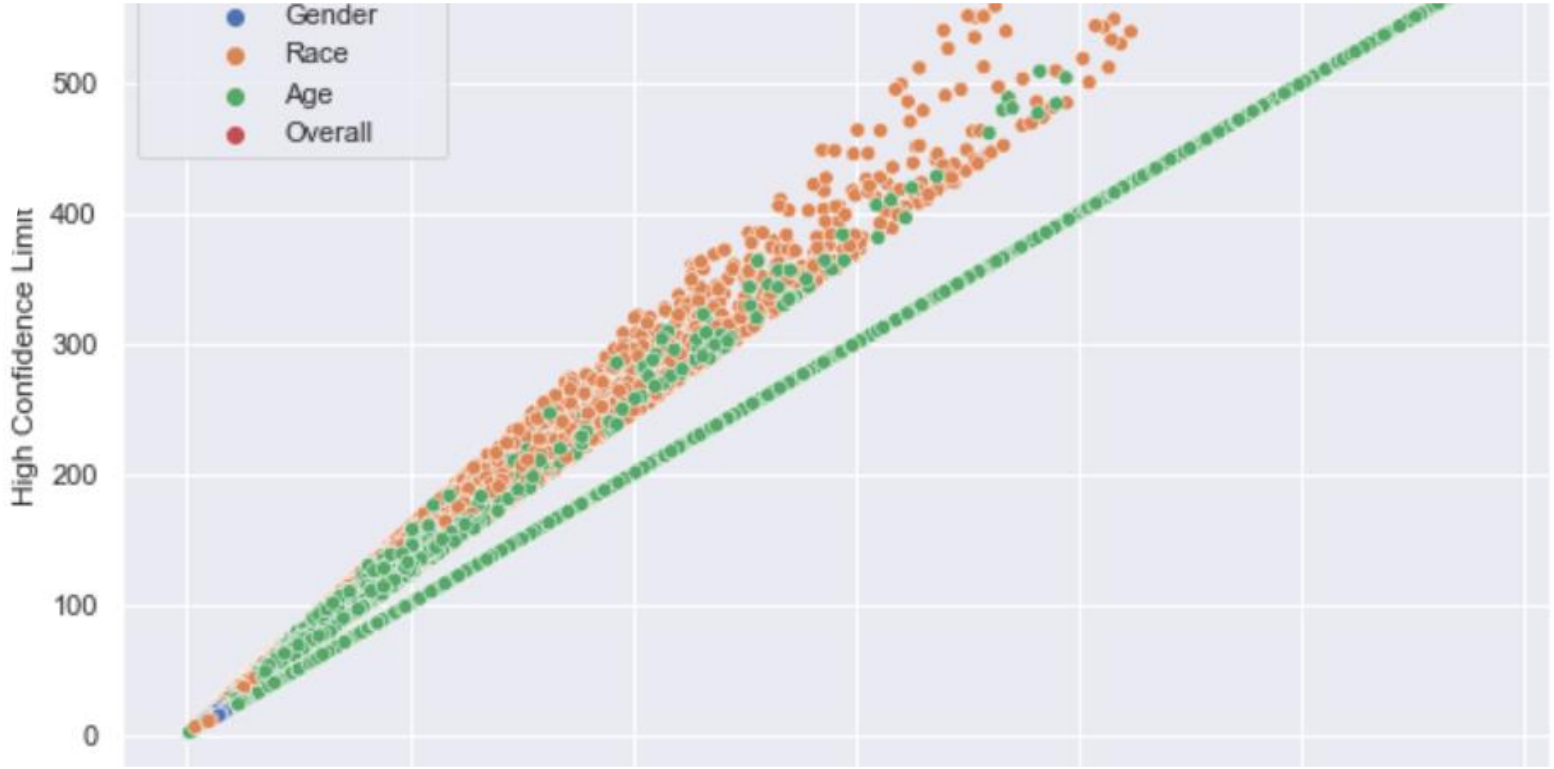




Distribution



Scatterplot for Confidence Limits



Scatterplot for Confidence Limits

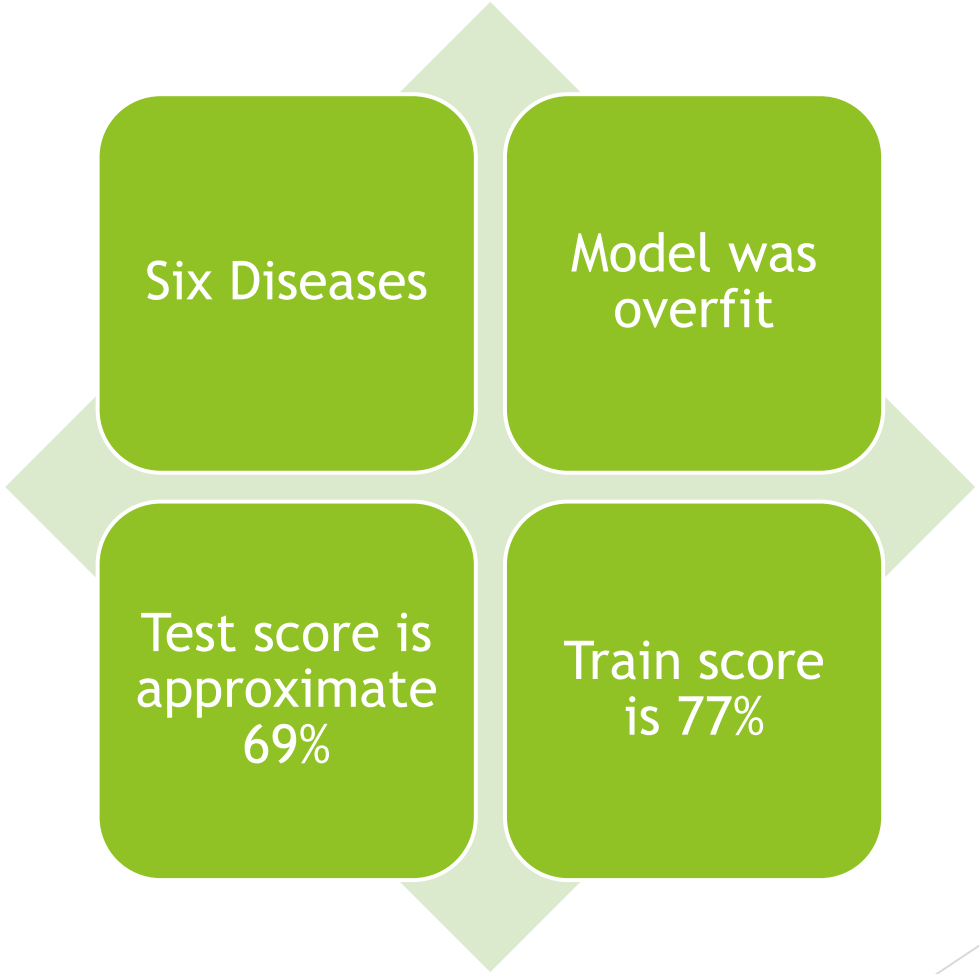
Types of Heart Condition

- Stroke
- Cardiovascular Disease
- Heart Attack
- Heart Failure
- Coronary Heart Disease
- Other Heart Diseases

Modeling

One Vs Rest Classification	Multiclassification Model used for more than two classes. (6 diseases)
Random Forest Classification	Using it to predict One Disease at a time (Binary Data)
Extra Tree Classification	Using it to predict One Disease at a time (Binary Data)

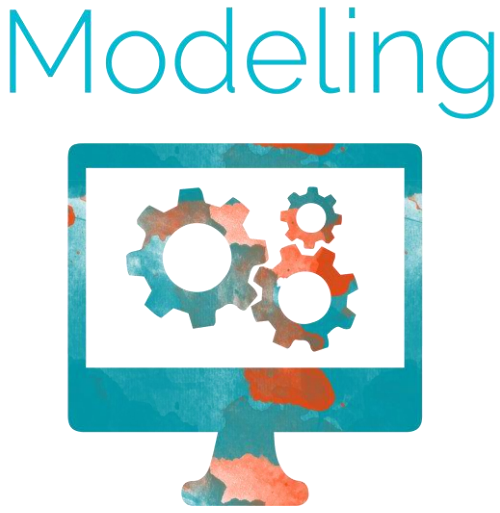
One Vs Rest Classification



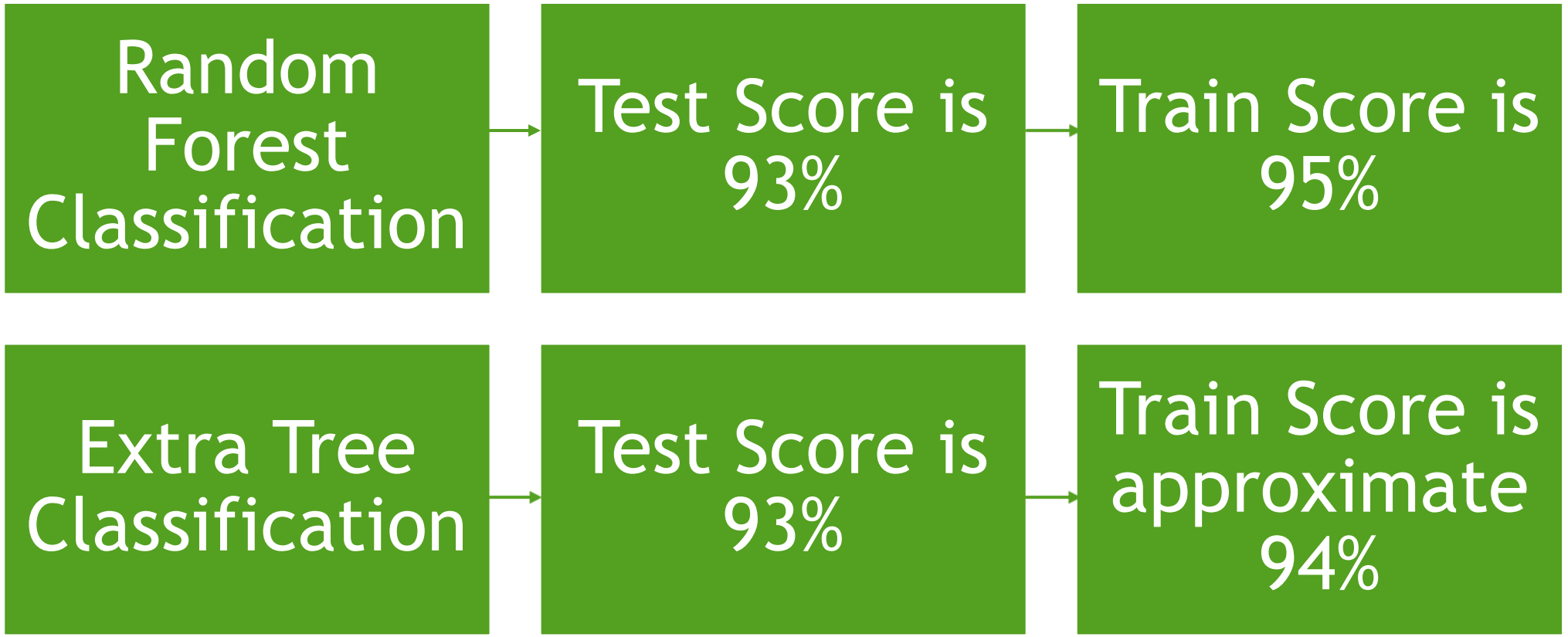
Modeling for Stroke Disease

- Random Forest Classification
- Test Score is approximate 81%
- Train Score is 86%
- Overfit

- Extra Tree Classification
- Test Score is 80%
- Train Score is approximate 84%
- Overfit



Modeling for Cardiovascular Disease



Modeling Coronary Heart Disease

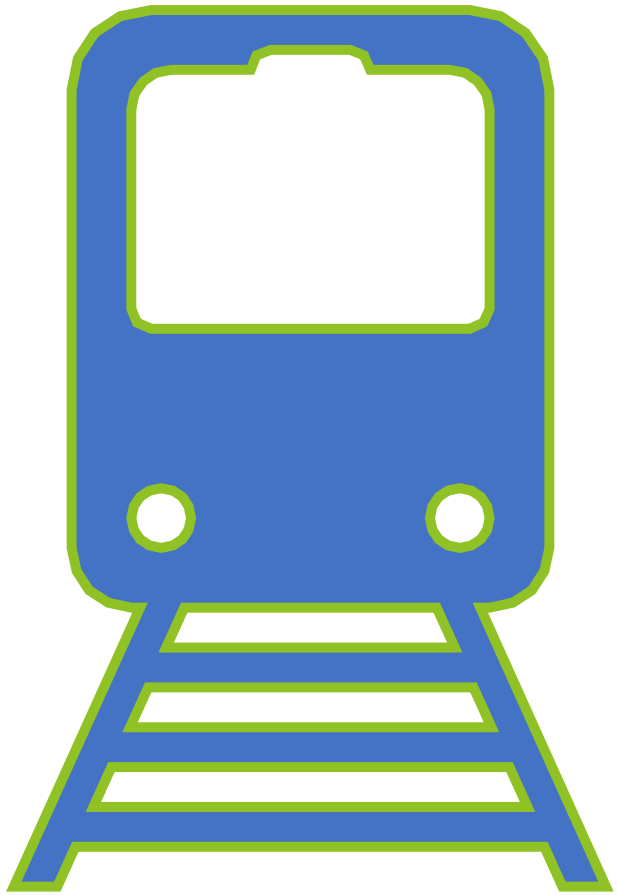
- Random Forest Classification
- Test Score is 94%
- Train Score is 95%

- Extra Tree Classification
- Test Score is 93%
- Train Score is 94%

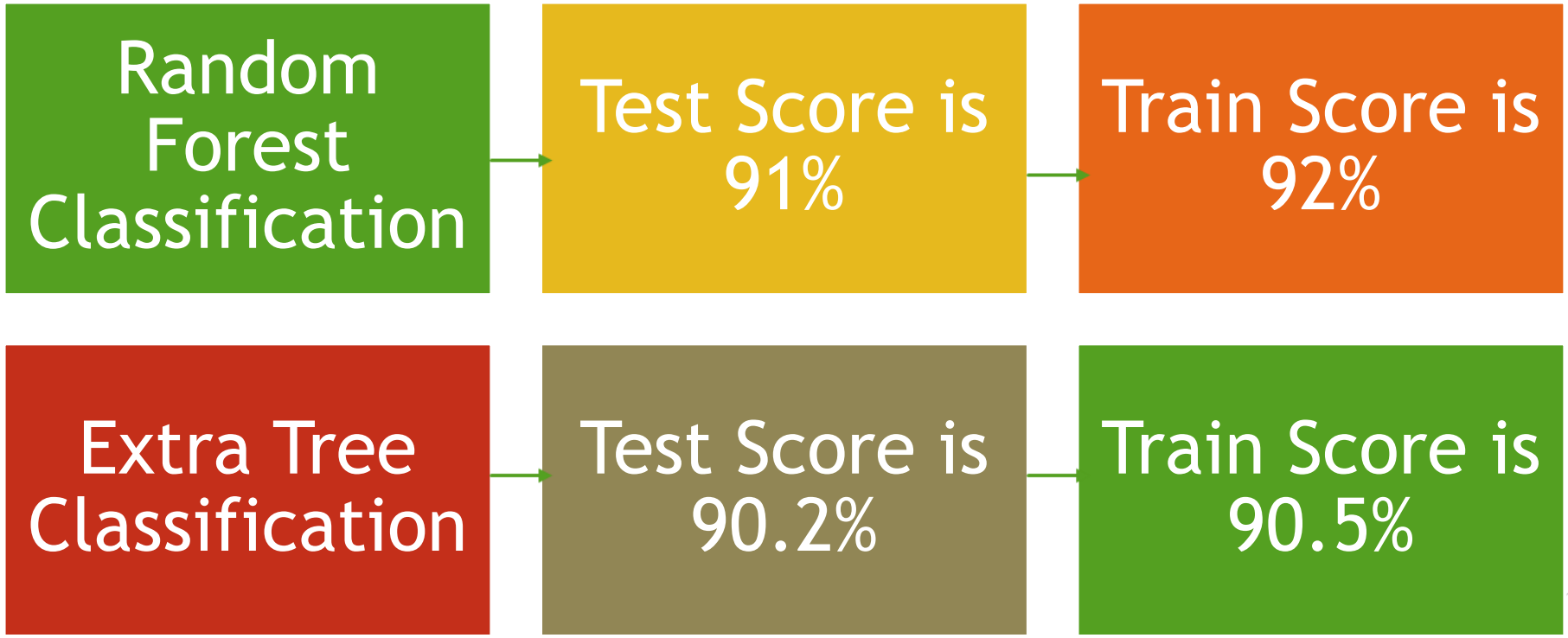
Modeling for Heart Attack

- Random Forest Classification
- Test Score is approximate 88%
- Train Score is 89%

- Extra Tree Classification
- Test Score is approximate 88%
- Train Score is approximate 89%



Modeling for Heart Failure



Modeling for Other Heart Diseases

- Random Forest Classification
- Test Score is 90%
- Train Score is 93%

- Extra Tree Classification
- Test Score is 89%
- Train Score is 91%

Confusion Matrix

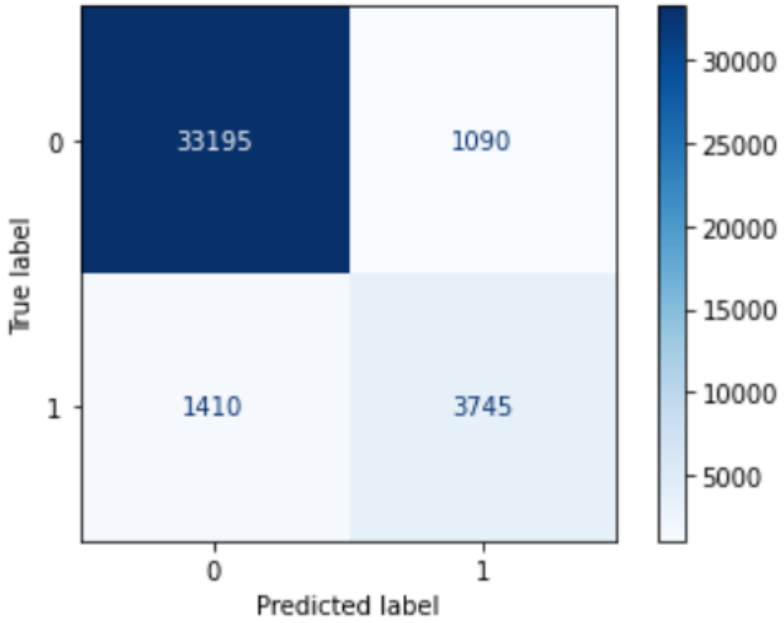
In this Confusion Matrix, we can better explain the outcome

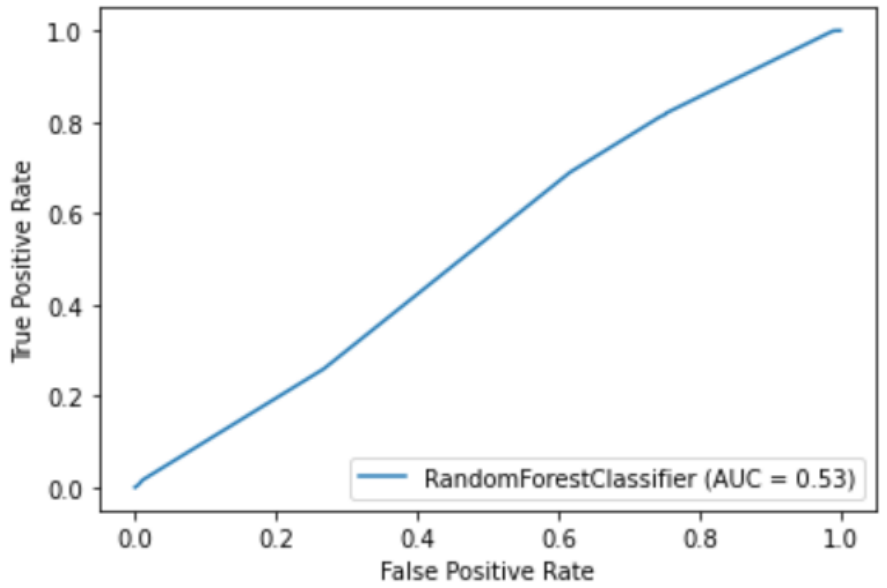
True negatives means that we correctly predicted the people who don't have Heart Disease

True positives means that we correctly predicted the people who have Heart Disease

False positives means that we incorrectly predicted the people who have Heart Disease(who don't have Heart Disease)

False negatives means that we incorrectly predicted the people who don't have Heart Disease(who have Heart Disease)





Confusion Matrix for Cardiovascular Disease

Conclusion

Values related to Stoke 62%

One Vs Rest Classification was Overfit

Models worked well

Success in True Negatives

Thank You for Listening