COURSE TITLE: BIG DATA ANALYTICS TECHNIQUES

HEART DISEASE PREDICTION ANALYSIS AND VISUALIZATION

Team Members:

- Muhammad Saad
- Muhammad Hamza Ansari
- Saadullah Khan Warsi

Course Teachers:

- Sir Affan Alim
- Sir Arsalan

IBA CICT
Center for Information & Communication Technology
Institute of Business Administration

Prime Minister's Kamyab Jawan National Youth Development Programme

Motivation

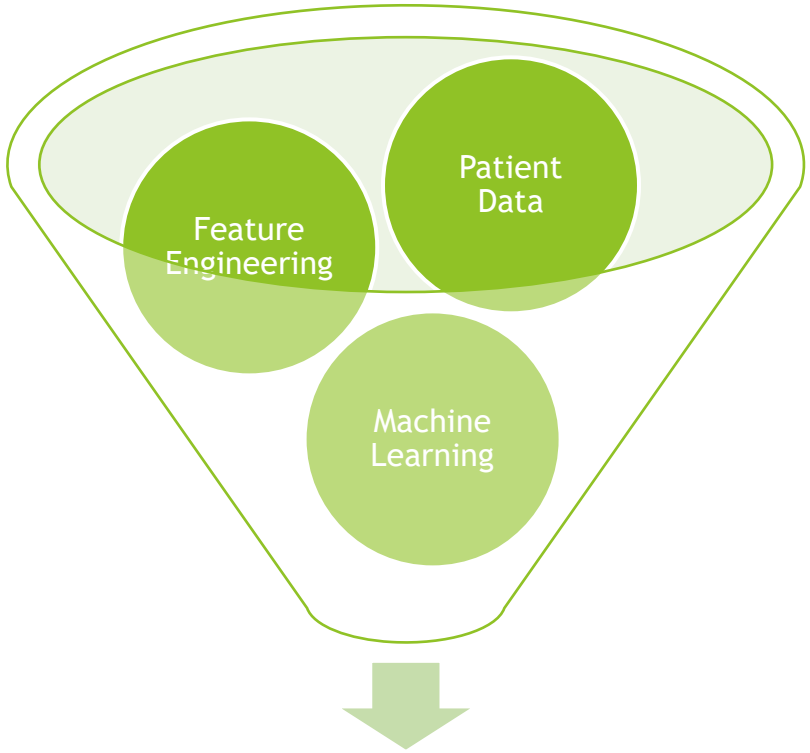
- Develop a Learning tool which can predict Heart Diseases based on a series of Data Sets accumulated by Doctors.
- The Learning tool can provide second thoughts or opinions to doctors about whether a heart disease is likely present or not.
- Same procedure can be applied to predict several other diseases based on their data sets.

Project Background

- Heart Failure is one of the leading medical issue in the current era. With the growing number of Heart Failure cases in the past, several data bases have been developed which contains important features (data) that leads to or may have cause Heart Failure.

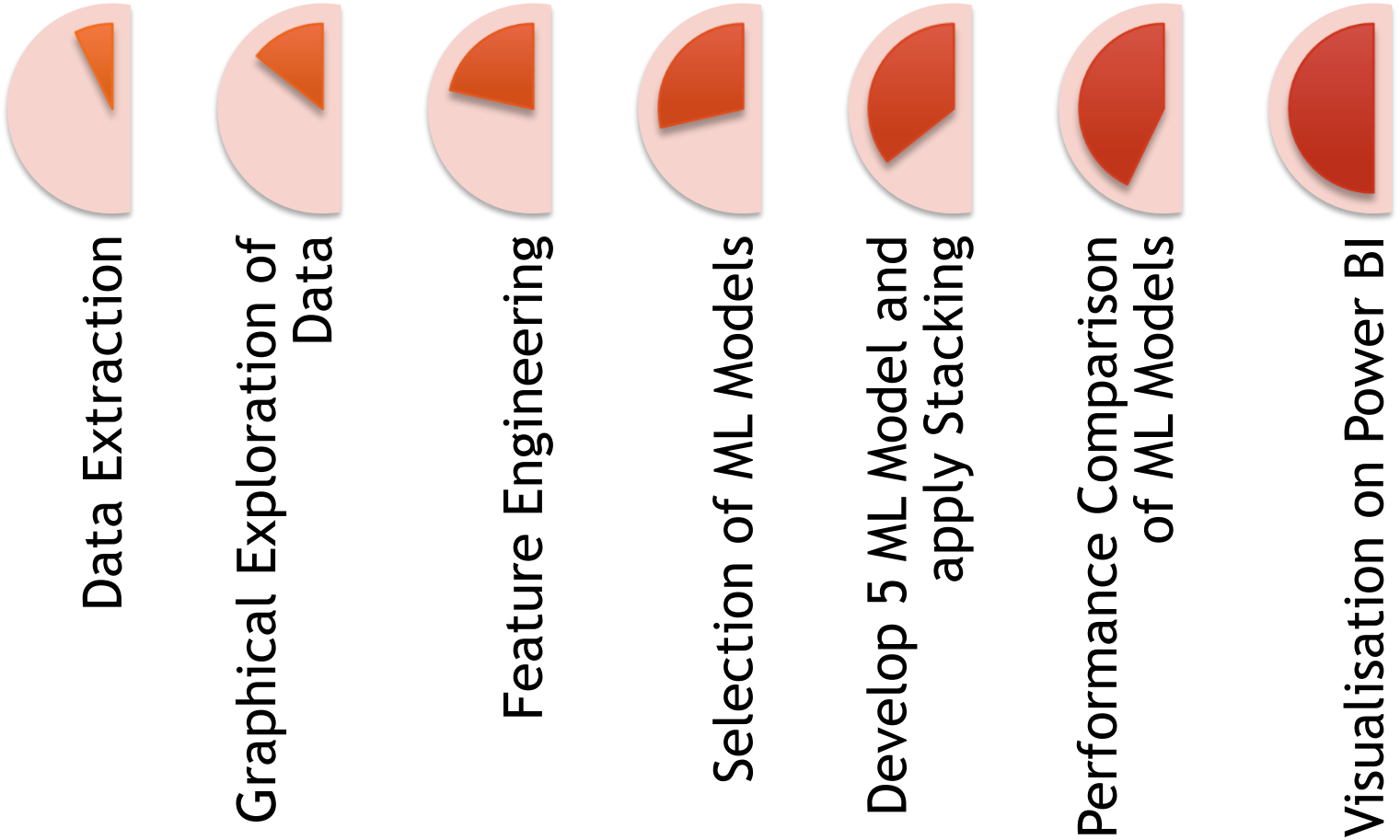
- With the data sets available on Kaggle/Other Web Platforms we can perform Machine Learning on the medical database and develop Models that can predict possible heart failure for a patient at a critical time.

Project Overview



Heart Disease (Yes or NO)

Project Overview

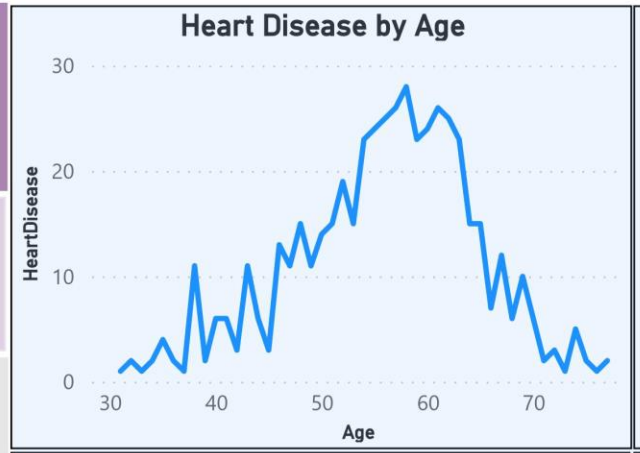


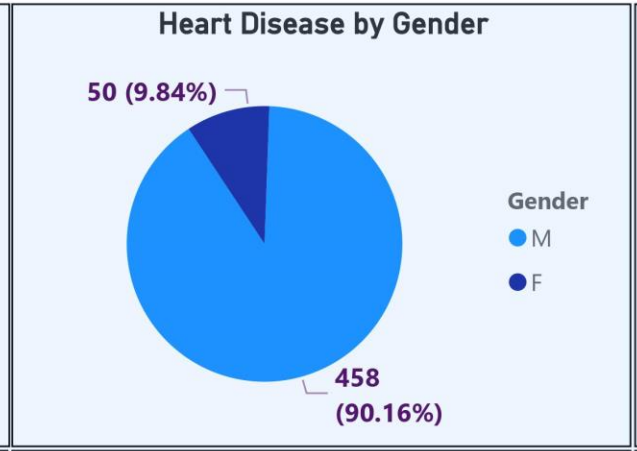
· HEART DISEASE PREDICTION MODELING

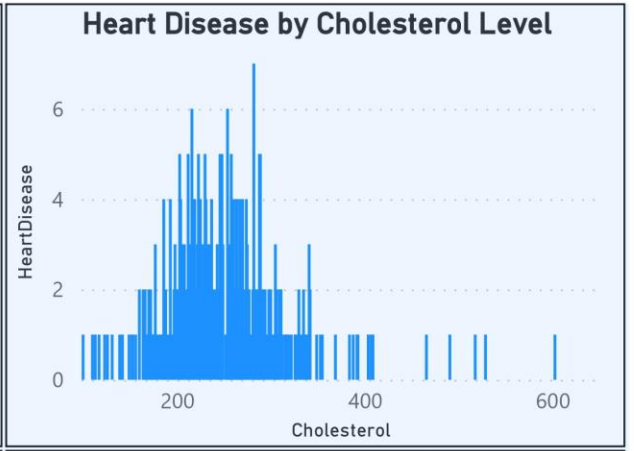
Critical Features

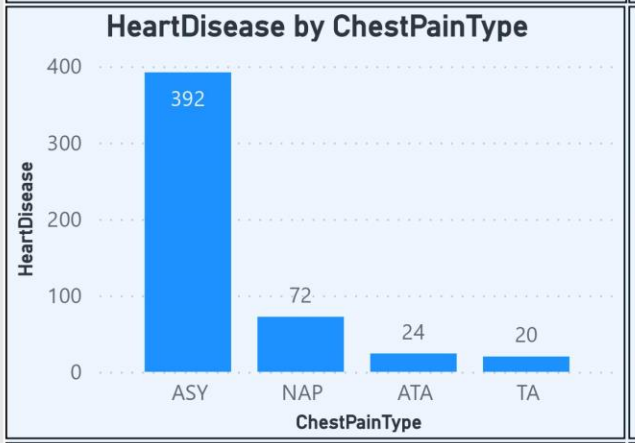
11

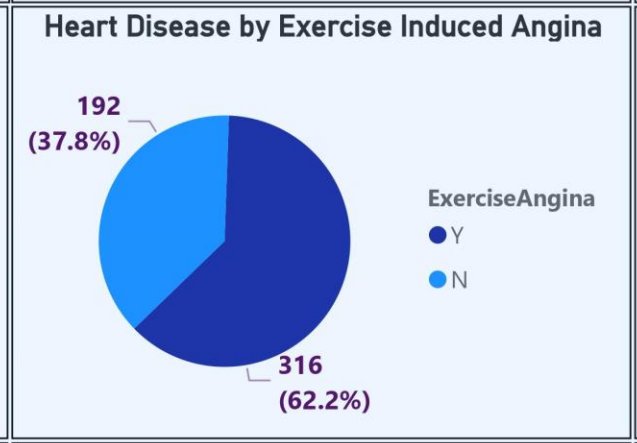
- Age
- Gender
- Chest Pain Type
- Cholesterol
- Exercise Engina
- Resting BP
- Resting ECG
- Max Heart Rate
- Fasting Blood Sugar
- Old Peak
- ST_Slope

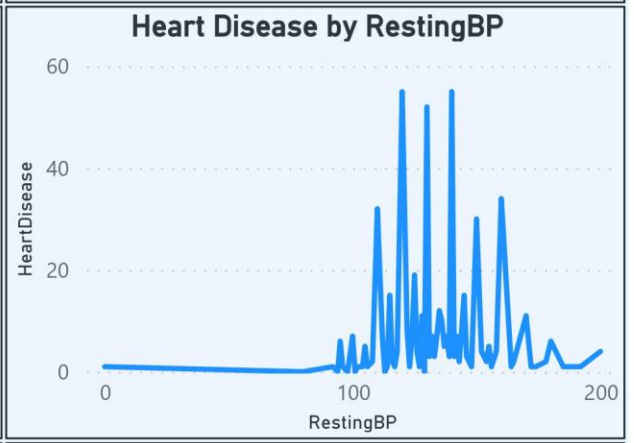


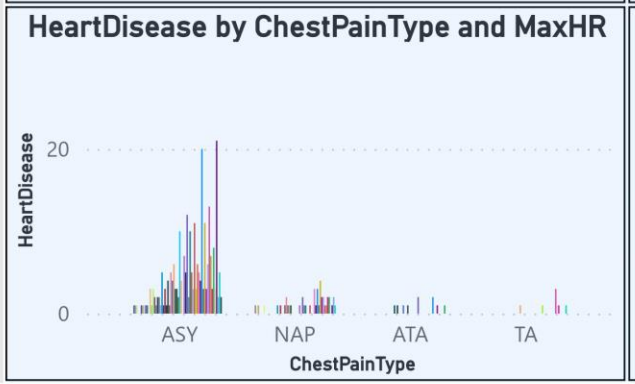


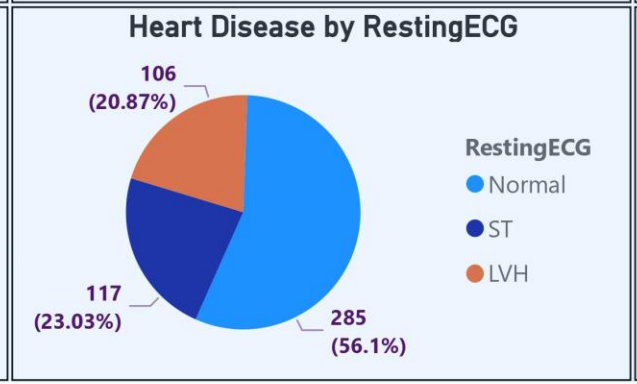


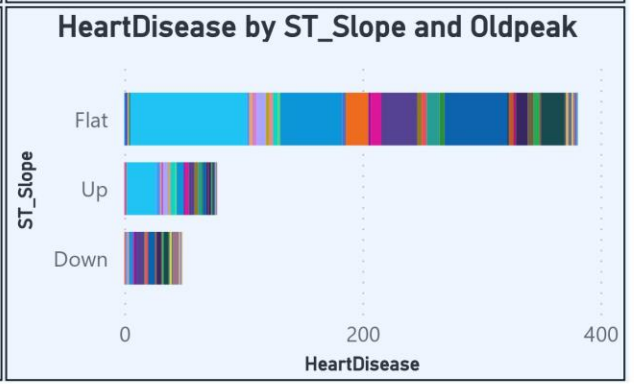












Data Information

```
In [34]: heart.info()
```

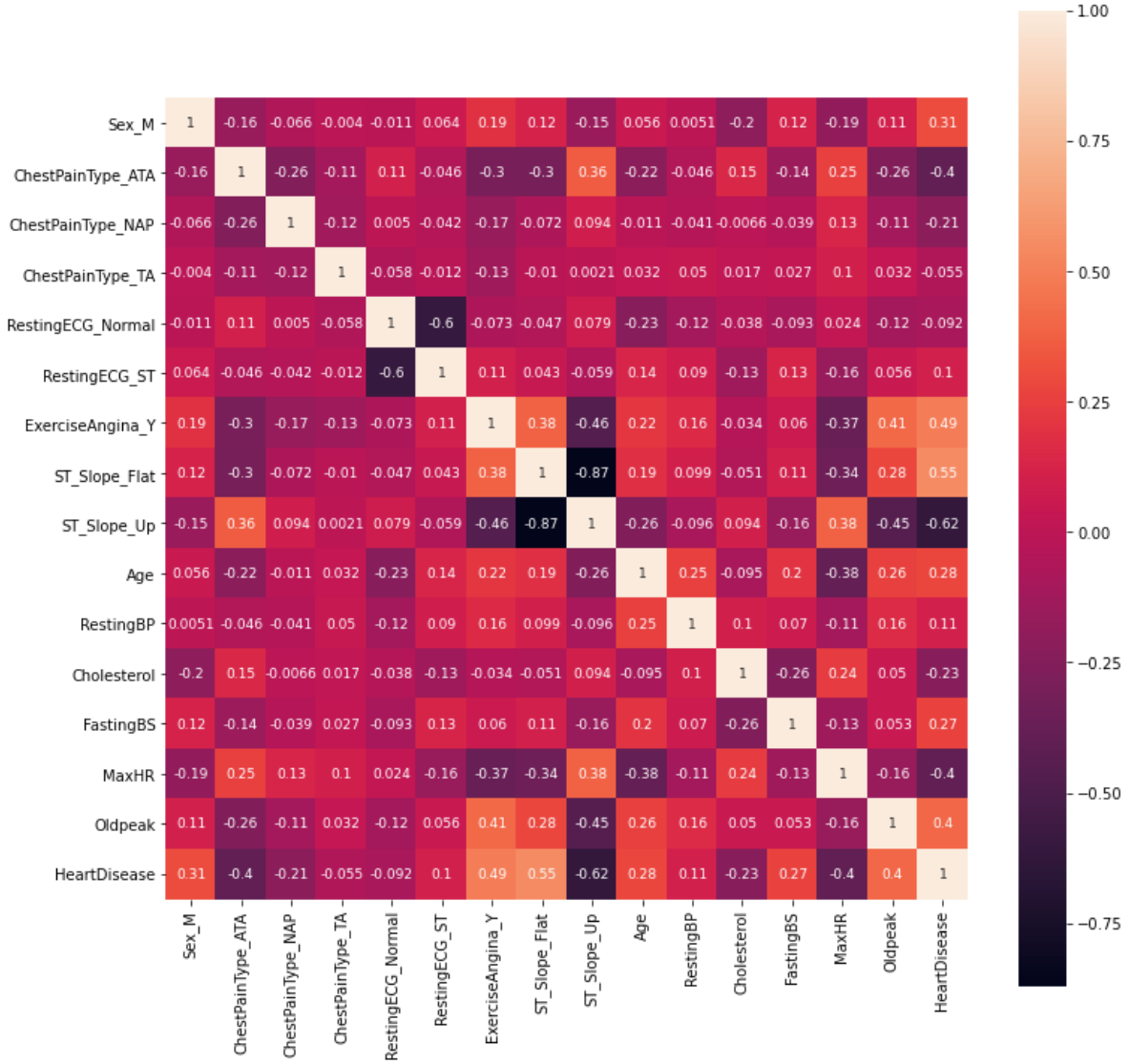
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Age             918 non-null    int64  
 1   Sex             918 non-null    object 
 2   ChestPainType   918 non-null    object 
 3   RestingBP       918 non-null    int64  
 4   Cholesterol     918 non-null    int64  
 5   FastingBS       918 non-null    int64  
 6   RestingECG      918 non-null    object 
 7   MaxHR           918 non-null    int64  
 8   ExerciseAngina  918 non-null    object 
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object 
 11  HeartDisease    918 non-null    int64  
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
```

Data Exploration: Checking Null Values

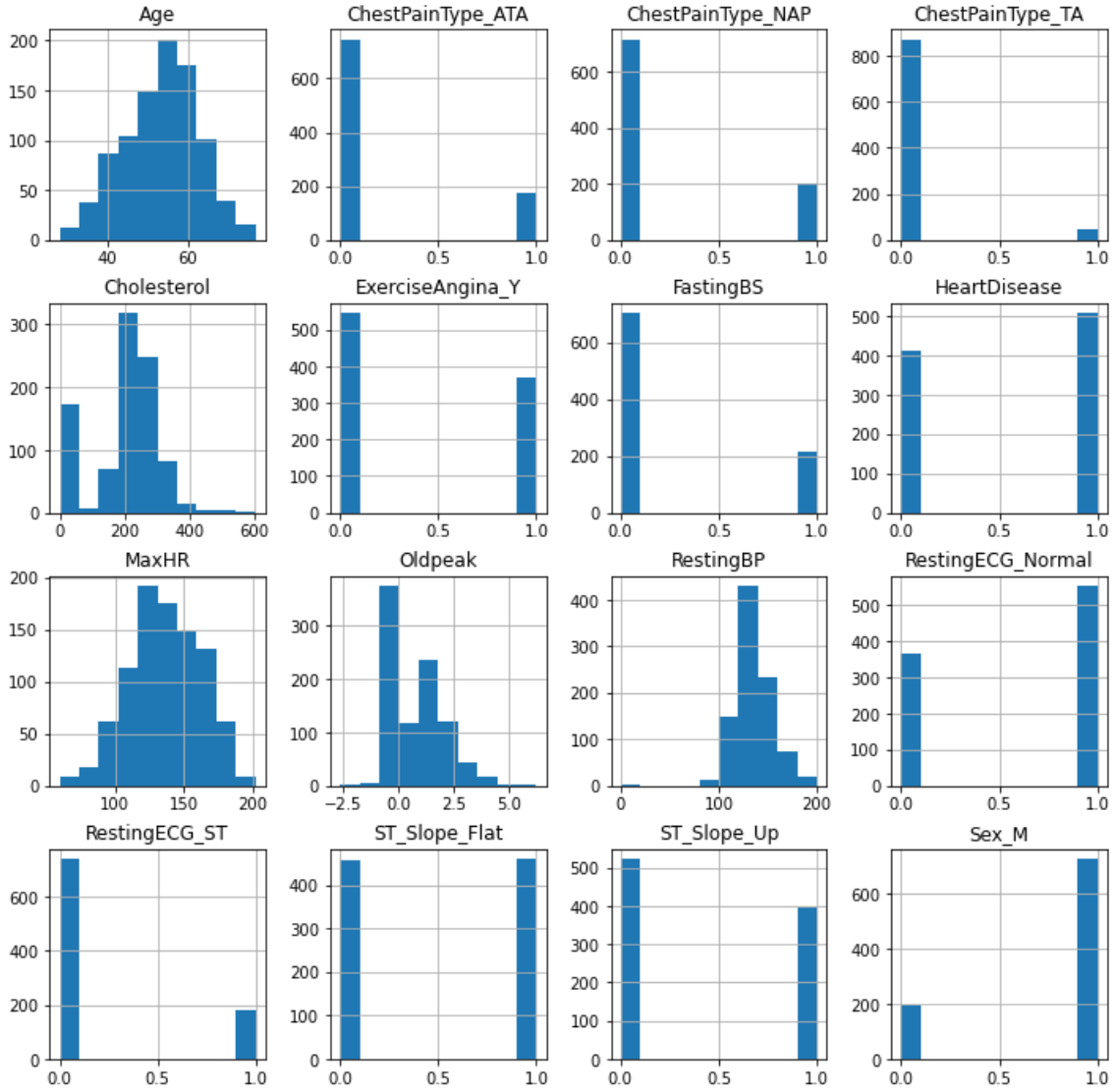
In [36]:
```
# Checking for Missing of Null Values in Data Set

heart.isnull().sum()
```

Out[36]:
```
Age               0
Sex               0
ChestPainType     0
RestingBP         0
Cholesterol       0
FastingBS         0
RestingECG        0
MaxHR             0
ExerciseAngina    0
Oldpeak           0
ST_Slope          0
HeartDisease      0
dtype: int64
```

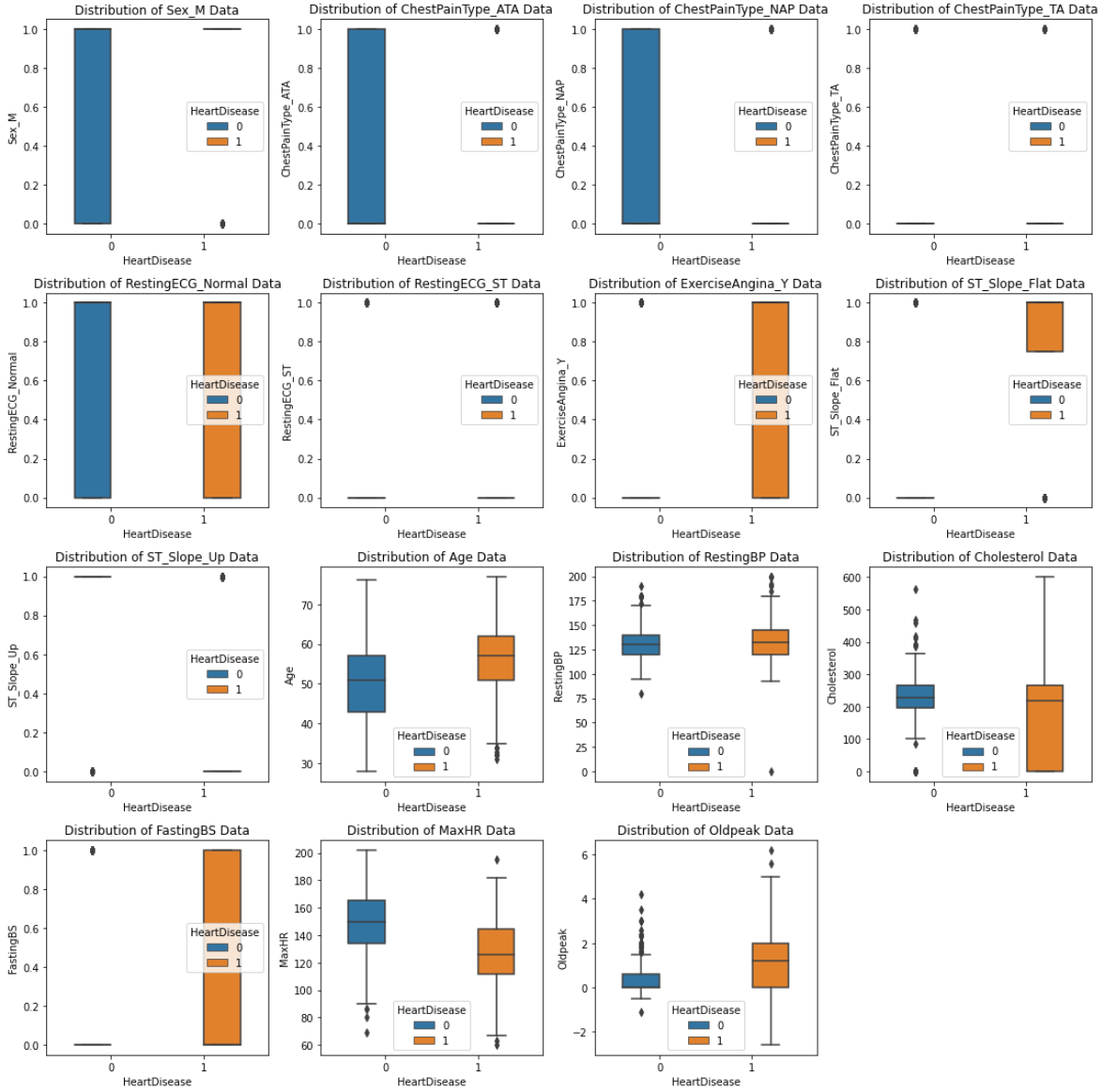
Data Exploration: Correlation



Data Exploration: Data Distribution



Data Exploration: Outliers



Machine Learning: Lazy Predict Method

Selection of Models

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
NearestCentroid	0.93	0.94	0.94	0.93	0.02
BernoulliNB	0.92	0.92	0.92	0.92	0.02
NuSVC	0.92	0.92	0.92	0.92	0.02
ExtraTreesClassifier	0.91	0.91	0.91	0.91	0.11
KNeighborsClassifier	0.91	0.91	0.91	0.91	0.01
RidgeClassifierCV	0.90	0.90	0.90	0.90	0.02
RidgeClassifier	0.90	0.90	0.90	0.90	0.02
LinearDiscriminantAnalysis	0.90	0.90	0.90	0.90	0.01
LinearSVC	0.90	0.90	0.90	0.90	0.04
LogisticRegression	0.90	0.90	0.90	0.90	0.01
CalibratedClassifierCV	0.90	0.90	0.90	0.90	0.11
SVC	0.90	0.90	0.90	0.90	0.02
RandomForestClassifier	0.90	0.90	0.90	0.90	0.14
SGDClassifier	0.89	0.89	0.89	0.89	0.01
LGBMClassifier	0.88	0.88	0.88	0.88	0.13
QuadraticDiscriminantAnalysis	0.88	0.88	0.88	0.88	0.01
XGBClassifier	0.88	0.88	0.88	0.88	0.06
PassiveAggressiveClassifier	0.88	0.87	0.87	0.88	0.02
GaussianNB	0.87	0.87	0.87	0.87	0.02
AdaBoostClassifier	0.87	0.87	0.87	0.87	0.09
BaggingClassifier	0.85	0.85	0.85	0.85	0.02
DecisionTreeClassifier	0.85	0.85	0.85	0.85	0.01

Machine Learning: NUSVC Model

```
Model Performance NuSVC
Accuracy Train: 0.8514986376021798
MCC Train: 0.6994572721327352
F1 Score Train : 0.8706998813760378

Accuracy Test: 0.875
MCC Test: 0.746633218777735
F1 Score Test : 0.8899521531100477
```

	precision	recall	f1-score	support
0	0.83	0.88	0.86	77
1	0.91	0.87	0.89	107
accuracy			0.88	184
macro avg	0.87	0.88	0.87	184
weighted avg	0.88	0.88	0.88	184

Machine Learning: GRADIENT BOOSTING CLASSIFIER

```
Model Performance GRADIENT BOOSTING
Accuracy Train: 0.9373297002724795
MCC Train: 0.873132298894877
F1 Score Train : 0.9436274509803921

Accuracy Test: 0.9565217391304348
MCC Test: 0.9120038259206121
F1 Score Test : 0.9607843137254902
```

	precision	recall	f1-score	support
0	0.95	0.95	0.95	82
1	0.96	0.96	0.96	102
accuracy			0.96	184
macro avg	0.96	0.96	0.96	184
weighted avg	0.96	0.96	0.96	184

Machine Learning: XGBOOST

```
Model Performance XGBOOST
Accuracy Train: 0.9877384196185286
MCC Train: 0.9752077070247509
F1 Score Train : 0.9889025893958077

Accuracy Test: 0.9891304347826086
MCC Test: 0.978289060789934
F1 Score Test : 0.99009900990099
```

	precision	recall	f1-score	support
0	1.00	0.98	0.99	84
1	0.98	1.00	0.99	100
accuracy			0.99	184
macro avg	0.99	0.99	0.99	184
weighted avg	0.99	0.99	0.99	184

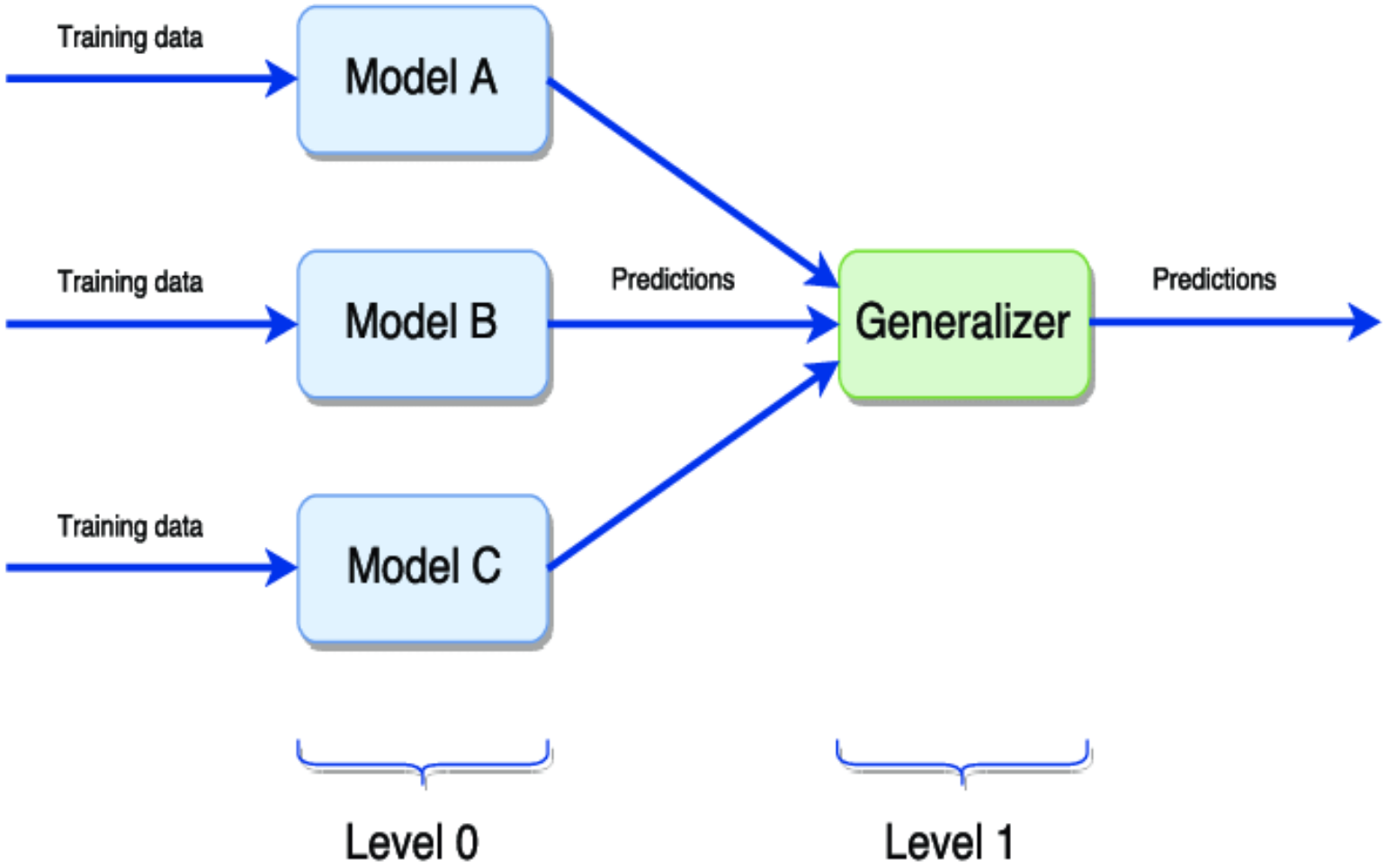
Machine Learning: EXTRA TRESS CLASSIFIER

```
Model Performance EXTRA TRESS CLASSIFIER
Accuracy Train: 1.0
MCC Train: 1.0
F1 Score Train : 1.0

Accuracy Test: 0.8804347826086957
MCC Test: 0.7579384823238913
F1 Score Test : 0.8952380952380952

              precision    recall  f1-score   support

           0       0.83      0.89      0.86        76
           1       0.92      0.87      0.90       108

    accuracy                           0.88       184
   macro avg       0.88      0.88      0.88       184
weighted avg       0.88      0.88      0.88       184
```

Machine Learning: RANDOMFORESTCLASSIFIER

```
Model Performance RANDOM FORREST
Accuracy Train: 1.0
MCC Train: 1.0
F1 Score Train : 1.0

Accuracy Test: 0.8695652173913043
MCC Test: 0.7360114777618364
F1 Score Test : 0.8823529411764706

              precision    recall  f1-score   support

           0       0.85      0.85      0.85        82
           1       0.88      0.88      0.88       102

    accuracy                           0.87       184
   macro avg       0.87      0.87      0.87       184
weighted avg       0.87      0.87      0.87       184
```

Machine Learning: STACKING



Machine Learning: STACKING

```
Model Performance STACKING MODEL
Accuracy Train: 0.9959128065395095
MCC Train: 0.9917589340613957
MCC Train fold: 0.943706492883453
F1 Score Train : 0.996319018404908

Accuracy Test: 0.8913043478260869
MCC Test: 0.7800095648015304
MCC Test fold: 1.0
F1 Score Test : 0.9019607843137255
```

	precision	recall	f1-score	support
0	0.88	0.88	0.88	82
1	0.90	0.90	0.90	102
accuracy			0.89	184
macro avg	0.89	0.89	0.89	184
weighted avg	0.89	0.89	0.89	184

PREDICTION PERFORMANCE RESULTS COMPARISON OF MACHINE LEARNING MODELS FOR HEART DISEASE DETECTION

NU-SVC Model Results	Gradient Boosting Model Results	XG Boosting Model Results	ExtraTree Classifier Model	Random Forrest Model Results	Stacked ModelResults
F1 Score Test 0.89	F1 Score Test 0.96	F1 Score Test 0.99	F1 Score Train 1.00	F1 Score Test 0.88	F1 Score Test 0.90
F1 Score Train 0.87	F1 Score Train 0.94	F1 Score Train 0.99	Train Accuracy 1.00	F1 Score Train 1.00	F1 Score Train 0.99
Precision 0.87	Precision 0.96	Precision 0.99	F1 Score Test 0.89	Precision 0.87	Precision 0.89
Recall 0.87	Recall 0.96	Recall 0.99	Precision 0.88	Recall 0.87	Recall 0.89
Test Accuracy 0.88	Test Accuracy 0.95	Test Accuracy 0.99	Recall 0.88	Test Accuracy 0.87	Test Accuracy 0.89
Train Accuracy 0.85	Train Accuracy 0.94	Train Accuracy 0.99	Test Accuracy 0.88	Train Accuracy 1.00	Train Accuracy 0.99