Ban Qu

Final Report:
Heart Failure Prediction Analysis

Problem Statement

Cardiovascular diseases (CVDs) are the leading cause of death worldwide. To be specific, around 80 percent of deaths caused by CVDs are due to heart attacks and strokes, in which one-third of these deaths happen among people under 70 years old. Considering this, it is crucial to understand which factors contribute to heart failure for early detection.

The present study contains a dataset with 12 features that could be used to predict a potential heart failure. By using and analysing the data, various machine learning models are created to fit the current dataset and therefore to better predict heart failure for unseen patients.

K-Nearest Neighbor, Logistic Regression, Random Forest, Support Vector Machine, Naïve Bayes, and Gradient Boosting are selected in this study. The best performances were among Random Forest, KNN, and Gradient Boosting, and the latter two agree on the most predicting feature of heart failure as “ST_Slope” (the slope of the peak ST segment).

Data Wrangling

The raw dataset is derived from 5 datasets combined together that are related to heart. It is also the largest heart disease dataset available so far for research purposes. “HeartDisease” is the target variable. There is no missing data and duplications in this case.

Among the 11 predicting variables, five are converted from categorical to numeric forms: “Sex”, “ChestPainType”, “RestingRCG”, “ExerciseAngina”, “ST_Slope”. Since blood pressure cannot be 0, the related row is dropped. The new descriptions of data is as follows:

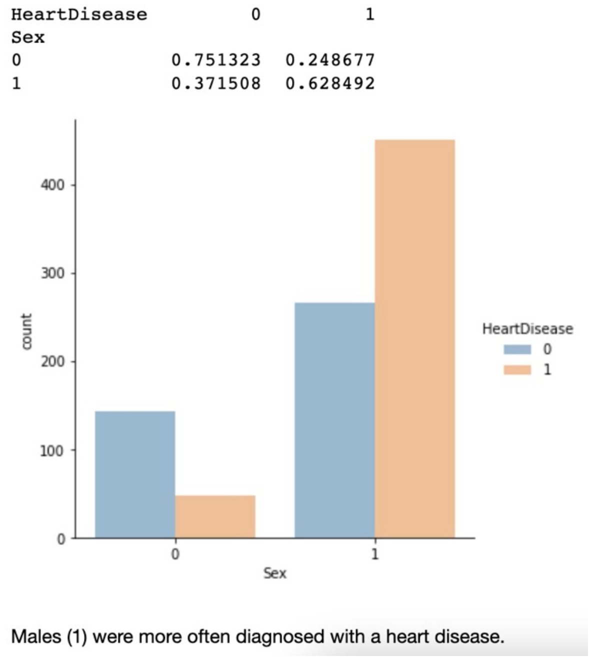
- **Age**: age of the patient [years]
- **Sex**: sex of the patient [0: Female, 1: Male]
- **ChestPainType**: chest pain type [0: Asymptomatic, 1: Atypical Angina, 2: Non-Anginal Pain, 3: Typical Angina]
- **RestingBP**: resting blood pressure [mm Hg]
- **Cholesterol**: serum cholesterol [mm/dl]
- **FastingBS**: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- **RestingECG**: resting electrocardiogram results [0: showing probable or definite left ventricular hypertrophy by Estes' criteria, 1: Normal, 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)]
- **MaxHR**: maximum heart rate achieved [Numeric value between 60 and 202]
- **ExerciseAngina**: exercise-induced angina [0: No, 1: Yes]
- **Oldpeak**: oldpeak = ST [Numeric value measured in depression]
- **ST_Slope**: the slope of the peak exercise ST segment [0: downsloping, 1: flat, 2: upsloping]
- **HeartDisease**: output class [1: heart disease, 0: Normal]

An outlier check with histogram indicates that “Oldpeak”, “Cholesterol”, “RestingBP”, and “MaxHR” require attention. All related outliers are dropped (i.e., out of 3 times of standard deviation).

The new data contains 12 columns and 905 observations. Under the target variable “HeartDisease”, 497 individuals (54.91%) have heart diseases whereas 408 (45.09%) does bot.

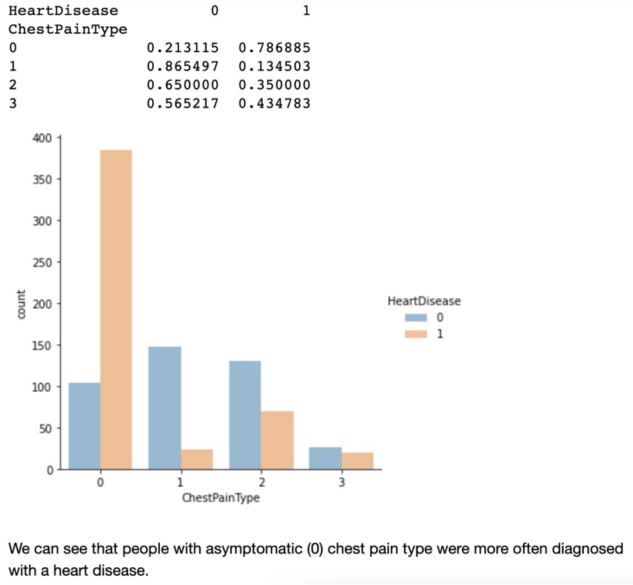
Exploratory Data Analysis

Using Searborn and Matplotlib, some findings are worth noticing.



```
HeartDisease         0         1
Sex
0             0.751323  0.248677
1             0.371508  0.628492
```

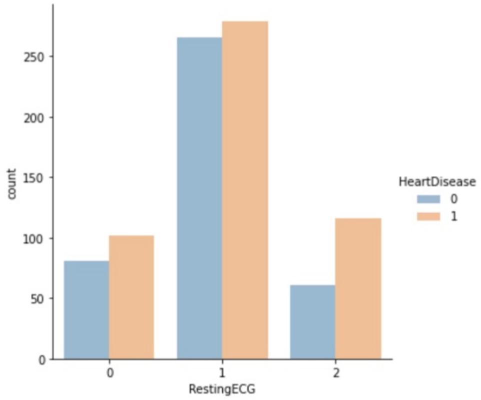
Males (1) were more often diagnosed with a heart disease.



```
HeartDisease         0         1
ChestPainType
0             0.213115  0.786885
1             0.865497  0.134503
2             0.650000  0.350000
3             0.565217  0.434783
```

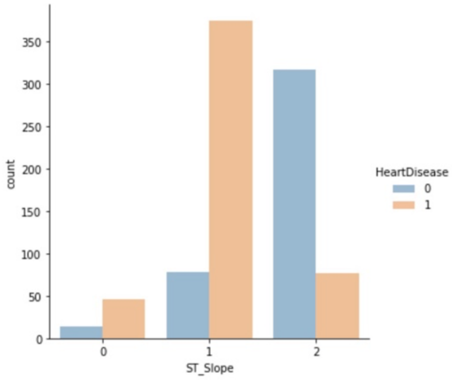
We can see that people with asymptomatic (0) chest pain type were more often diagnosed with a heart disease.

```
HeartDisease         0         1
RestingECG
0             0.442623  0.557377
1             0.488073  0.511927
2             0.344633  0.655367
```



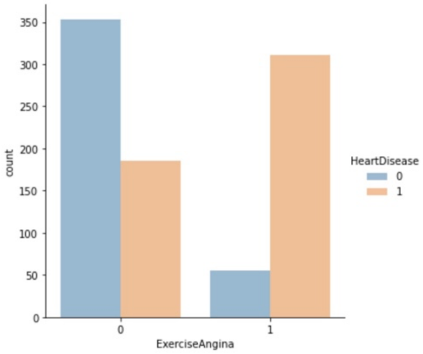
People having ST-T wave abnormality (2) were more often diagnosed with a heart disease.

```
HeartDisease         0         1
ST_Slope
0             0.233333  0.766667
1             0.172566  0.827434
2             0.804071  0.195929
```



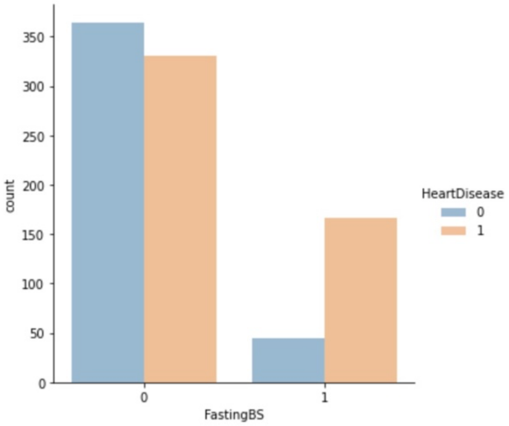
People with a flat (1) ST slope of peak exercise were more often diagnosed with a heart disease.

```
HeartDisease          0         1
ExerciseAngina
0              0.654917  0.345083
1              0.150273  0.849727
```

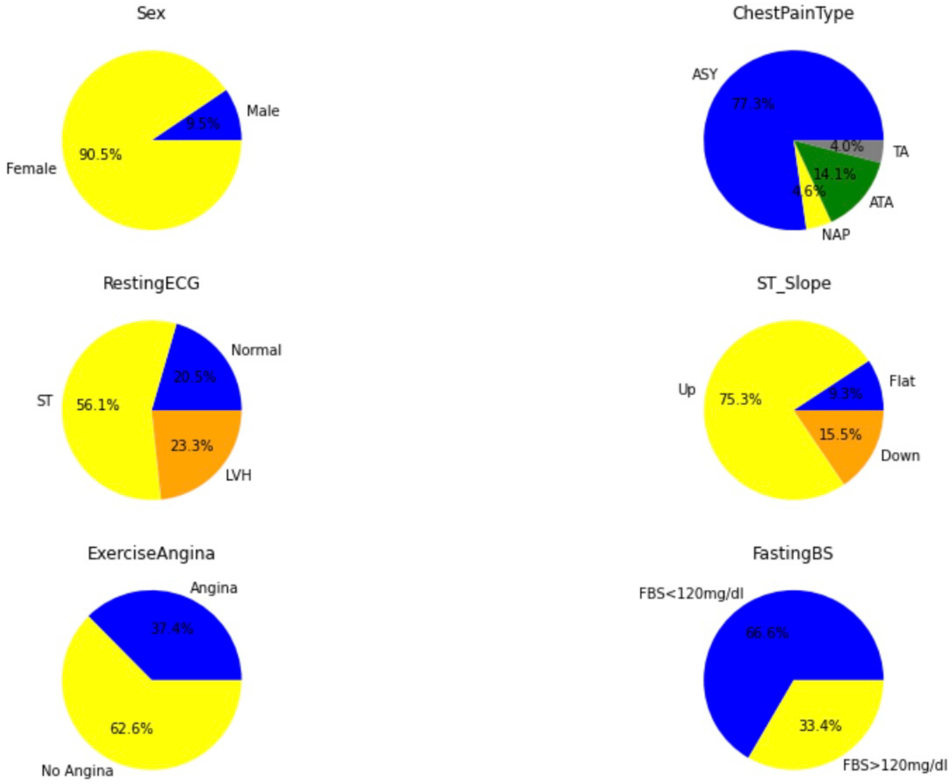


People with exercise-induced angina (1) were more often diagnosed with a heart disease.

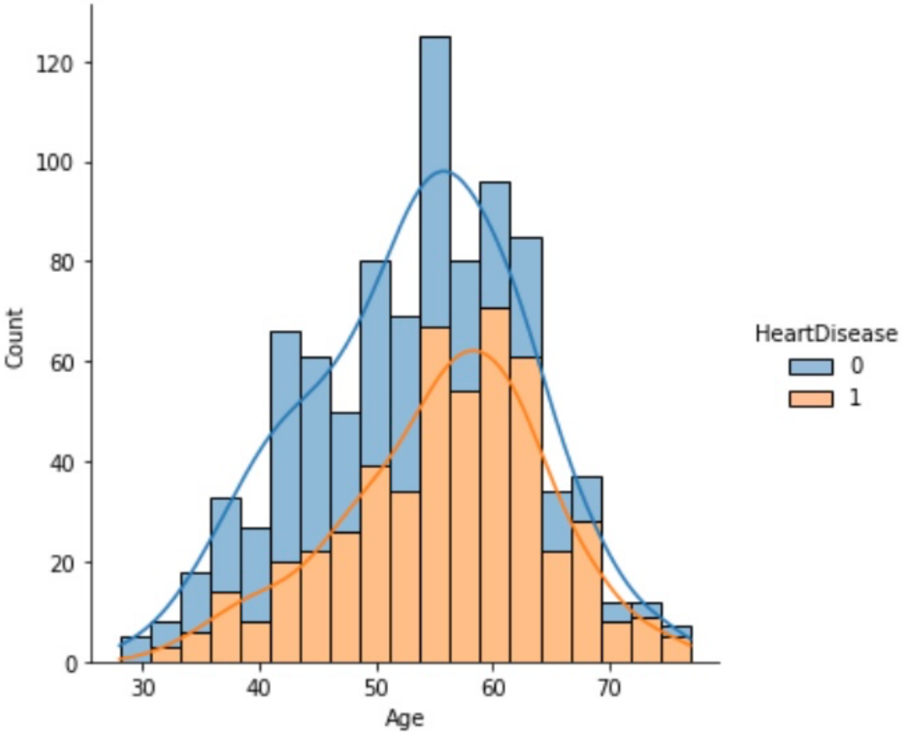
```
HeartDisease         0         1
FastingBS
0             0.523741  0.476259
1             0.209524  0.790476
```



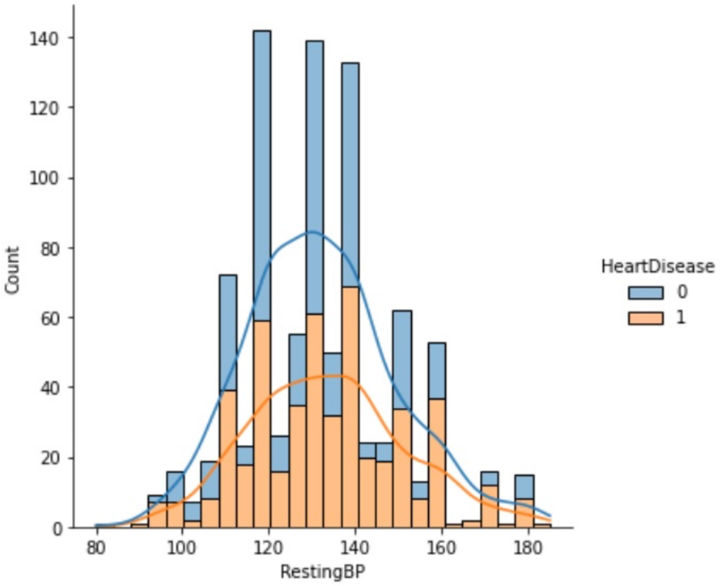
People with fasting blood sugar > 120 mg/dl (1) were more often diagnosed with a heart disease.



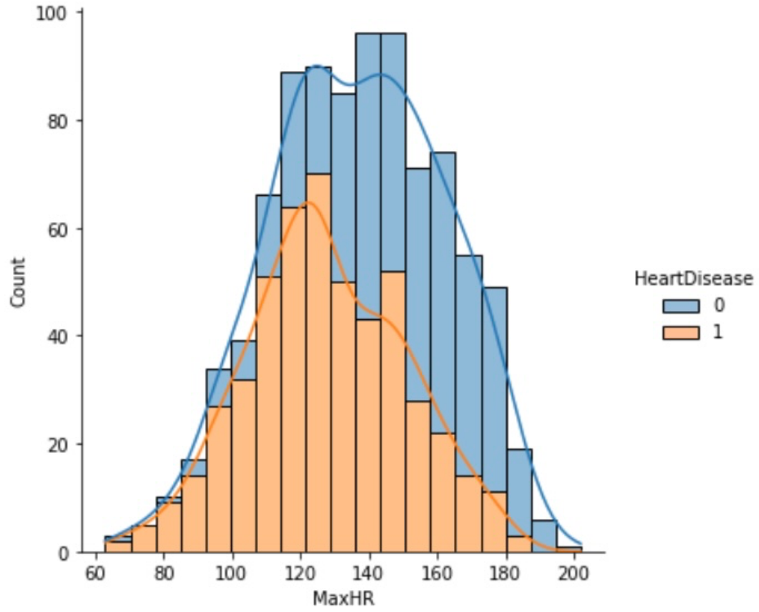
- The majority of people diagnosed with heart diseases is Male
- The majority of people diagnosed with heart diseases has chest pain type as 'Asymptomatic'
- The majority of people diagnosed with heart diseases has resting ECG as 'having ST-T wave abnormality'
- The majority of people diagnosed with heart diseases has slope of the peak exercise ST segmentas as 'Up'
- The majority of people diagnosed with heart diseases does not have exercise-induced Angima
- The majority of people diagnosed with heart diseases has fasting blood sugar less than 120mg/dl



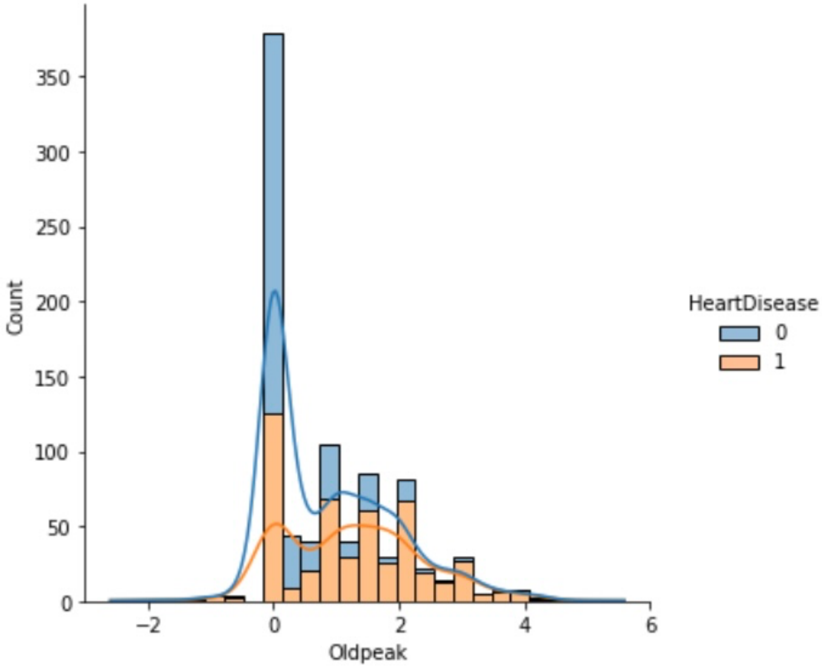
Older people were more often diagnosed with a heart disease.



People with higher Resting Blood Pressure were more often diagnosed with a heart disease.



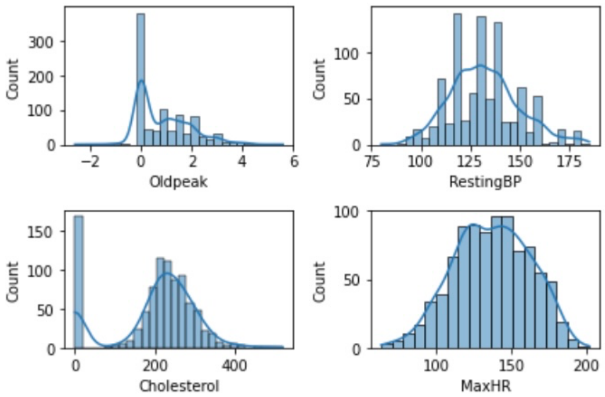
People with lower Maximum Heart Rate were more often diagnosed with a heart disease.

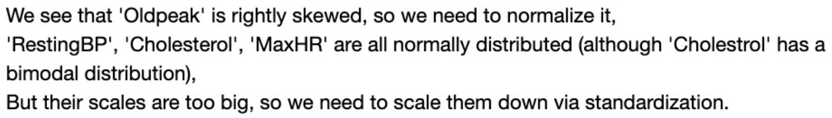


People with higher depression values were more often diagnosed with a heart disease.

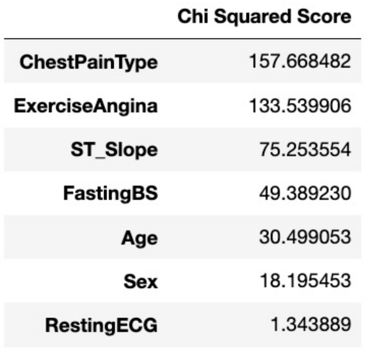
Pre-Processing

The variable “Age’ is further modified into groups (i.e., 5 groups with a group range of 10 years). Some columns have much higher ranges than others, and therefore the data is scaled.



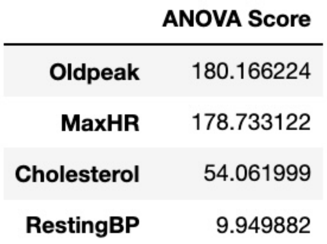

We see that 'Oldpeak' is rightly skewed, so we need to normalize it,
'RestingBP', 'Cholesterol', 'MaxHR' are all normally distributed (although 'Cholestrol' has a bimodal distribution),
But their scales are too big, so we need to scale them down via standardization.

With the scaled data, feature selection is performed before modeling. All features are divided into categorical and numerical. For categorical features, a Chi Squared Test is performed which suggests that all categorical features are important except “RestingECG”.



	Chi Squared Score
ChestPainType	157.668482
ExerciseAngina	133.539906
ST_Slope	75.253554
FastingBS	49.389230
Age	30.499053
Sex	18.195453
RestingECG	1.343889

For numerical features, ANOVA test is performed and it suggests that all features are important except “RestingBP”.



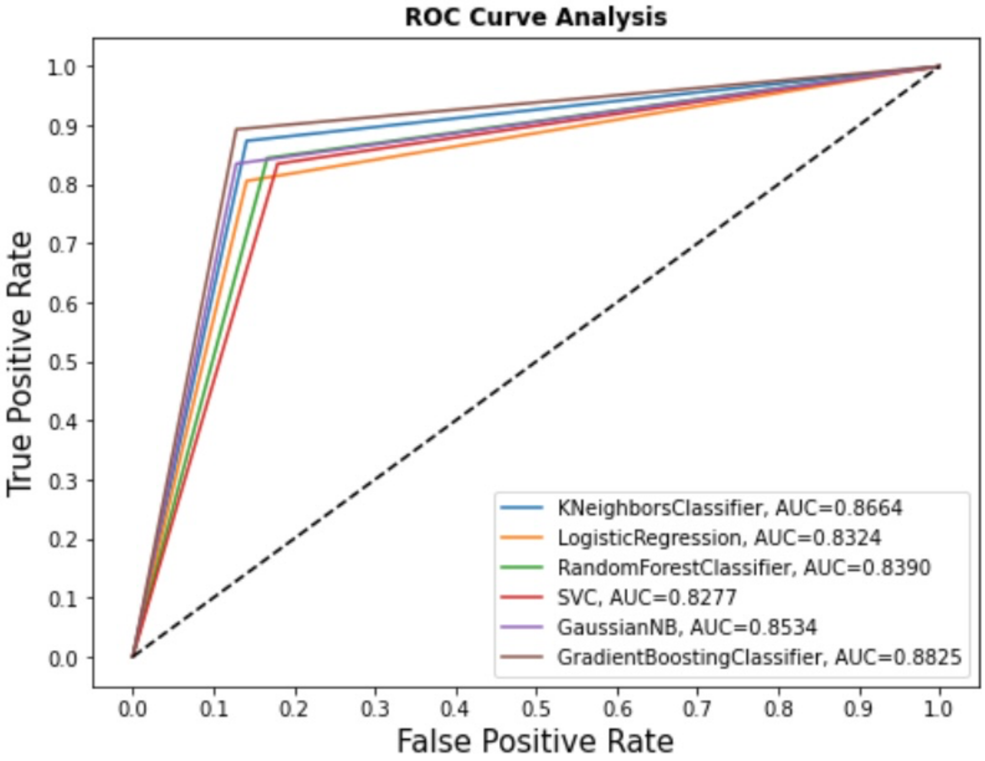
	ANOVA Score
Oldpeak	180.166224
MaxHR	178.733122
Cholesterol	54.061999
RestingBP	9.949882

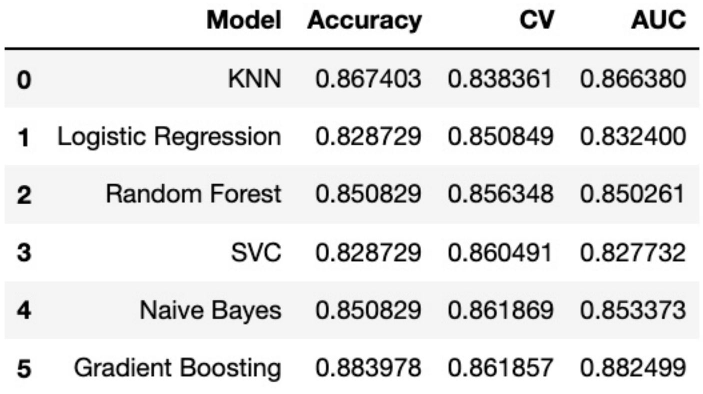
Therefore, the final features for modeling are: “ChestPainType”, “ExerciseAngina”, “ST_Slope”, “FastingBS”, “Age”, “Sex”, “Oldpeak”, “MaxHR”, and “Cholesterol”.

Finally, training data and testing data are split by 80% and 20%. X_train: (724, 9); X_test: (181, 9); y_train: (724,); y_test: (181,).

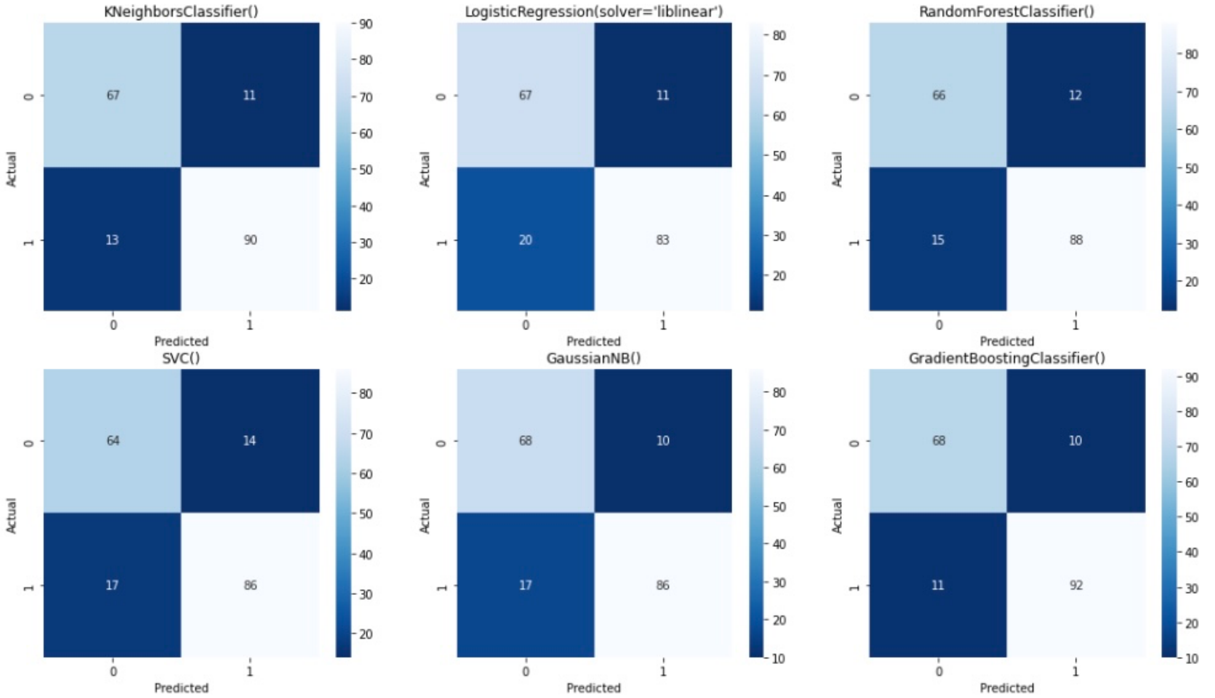
Modeling

There are 9 predictive features and 1 target feature (heart diseases: 0 and 1). Since this is a classification problem, the following machine learning models are selected: KNN, Logistic Regression, Random Forest, Support Vector Machine, Naïve Bayes, and Gradient Boosting. At first, a pipeline is built with default parameters for all models. The evaluation metrics used in this study are Accuracy Score, CV Score, ROC-AUC Score for model performance.





	Model	Accuracy	CV	AUC
0	KNN	0.867403	0.838361	0.866380
1	Logistic Regression	0.828729	0.850849	0.832400
2	Random Forest	0.850829	0.856348	0.850261
3	SVC	0.828729	0.860491	0.827732
4	Naive Bayes	0.850829	0.861869	0.853373
5	Gradient Boosting	0.883978	0.861857	0.882499



From the above evaluations, we can notice that the Gradient Boosting has the highest accuracy score and AUC score, followed by KNN. Logistic Regression and SVC seem to be not doing well. On the other hand, Random Forest has the highest Cross validation score, followed by Gradient Boosting. But remember we only train each model with its default parameters; model performance might be different if applying different parameters. So, the next step is hyperparameter tuning.

To optimize model performance, three models are selected for hyperparameter tuning: Random Forest, Gradient Boosting, and KNN.

	Accuracy Score	Cross Validation Score	AUC Score
Random Forest	0.8619	0.9235	0.9349
Gradient Boosting	0.8564	0.5	0.9349
KNN	0.8785	0.9229	0.9349

From the results after tuning, the highest accuracy score is KNN (0.8785) and the highest cv score is Random Forest (0.9235).

Random Forest and Gradient Boosting both agree on the most predicting feature: 'ST_Slope'.