Predicting Heart Disease Using Supervised Learning

November 17, 2019

Road Map

- Motivation
- Results
- Method
    - Exploratory Data analysis
    - Data Preparation
    - Modelling, Classification and Evaluation
- Conclusion and next steps

Motivation

Heart disease is one of the major causes of death globally. Reliable and accurate detection of heart disease is one of the crucial steps in stopping preventable deaths caused by heart disease. With the advent of data science tools, there are many methods that can advance early detection and prognosis of cardiovascular diseases.

Advancements in heart disease detection relying on indicators that are statistically determined to play a significant role in heart disease related deaths could save lives.

1:4

Deaths are caused by heart disease annually

735,000

People have heart disease every year

Source: Center for Disease Control and Prevention

Data set

Data Title: Predicting Heart Disease

Data Source: Cleveland Heart Disease Database via the UCI Machine Learning repository

Shape: (180 x 15)

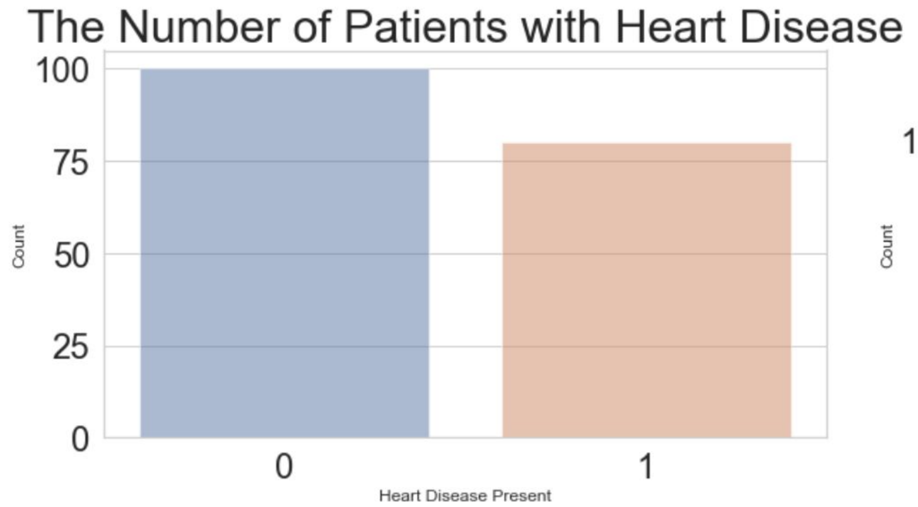
Outcome Variable : Heart Disease Present

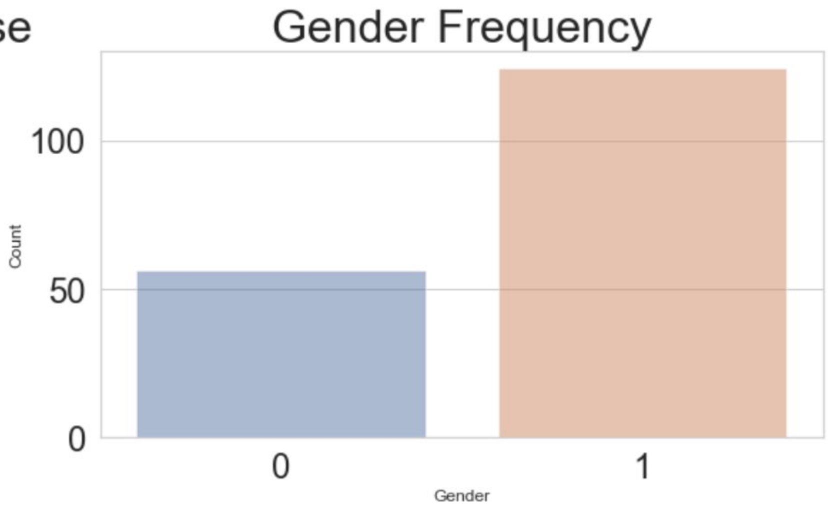
Features to be dropped : patient_id

Features

	Attribute	Description	Type
0	Age	Age in years	Continuous
1	Sex	Female (0), Male (1)	Categorical
2	Chest Pain Type	(1): typical angina,(2): atypical angina, (3): non-anginal pain, (4): asymptomatic	Categorical
3	Resting Blood Pressure	Measured in mm Hg, upon admission to hospital. Typically above 80 mm Hg	Continuous
4	Serum Cholesterol	Measured in mg/dl. It is the amount of cholesterol particles in the blood	Continuous
5	Fasting Blood Pressure	Measured in mg/dl. It indicates how well the body is managing blood sugar (>120).	Continuous
6	Resting electrocardiographic results	(0) normal,(1) having ST-T wave abnormality, (2) showing probable or definite left ventricular hypertrophy by Estes criteria	Categorical
7	Maximum heart rate achieved	Measured in beats per minute	Continuous
8	Exercise-induced angina	(0) not present, (1) present	Categorical
9	Oldpeak = ST depression induced by exercise relative to rest	On an ECG Plot: ST Segment abnormality indicates heart disease.	Continuous
10	Slope of peak exercise st segment	(1): upsloping,(2): flat,(3): downsloping	Categorical
11	Number of major vessels colored by flourosopy	0-3	Categorical
12	Thal	Refers to a blood disorder called thalassemia ((3) normal;(6) = fixed defect;(7) = reversable defect)	Categorical

Method: Data Exploration



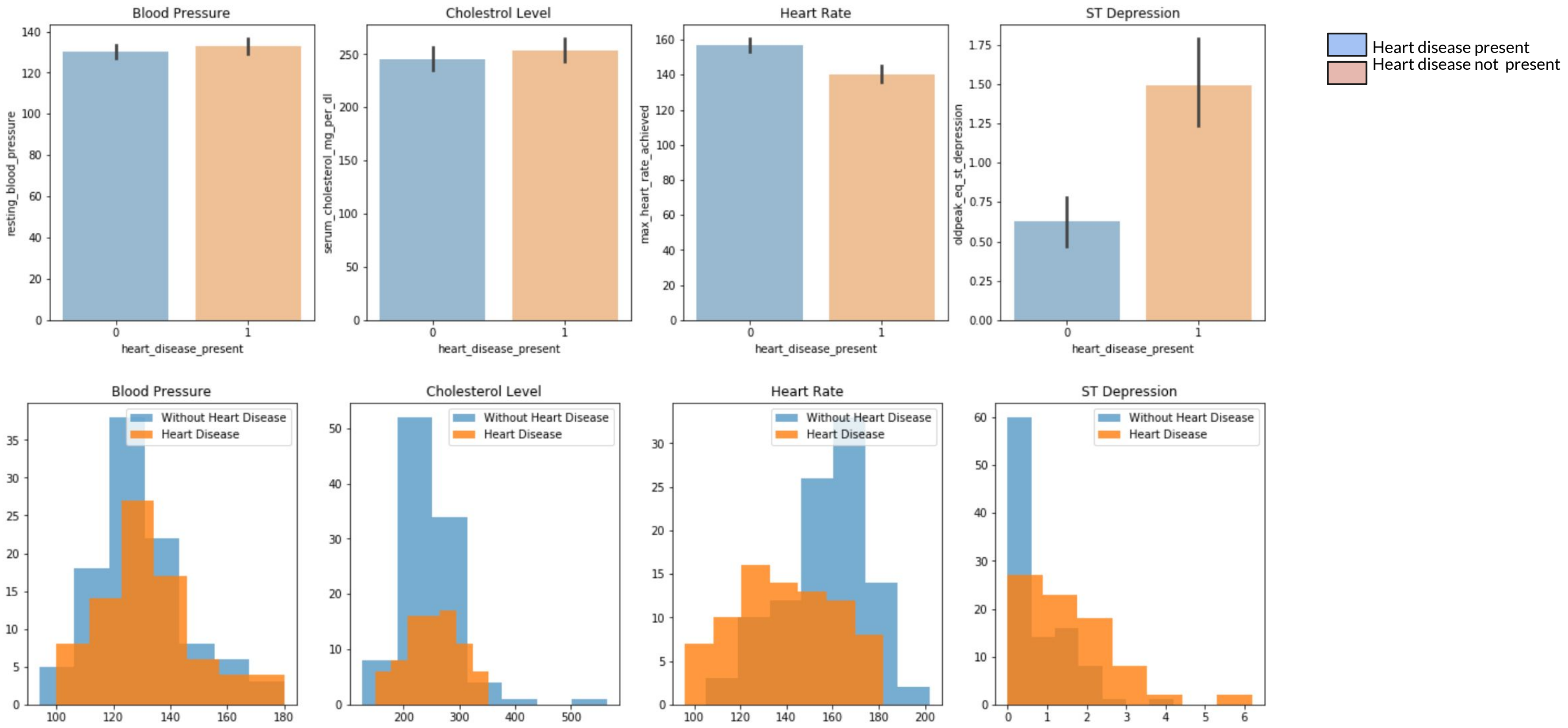


Heart disease present
Heart disease not present

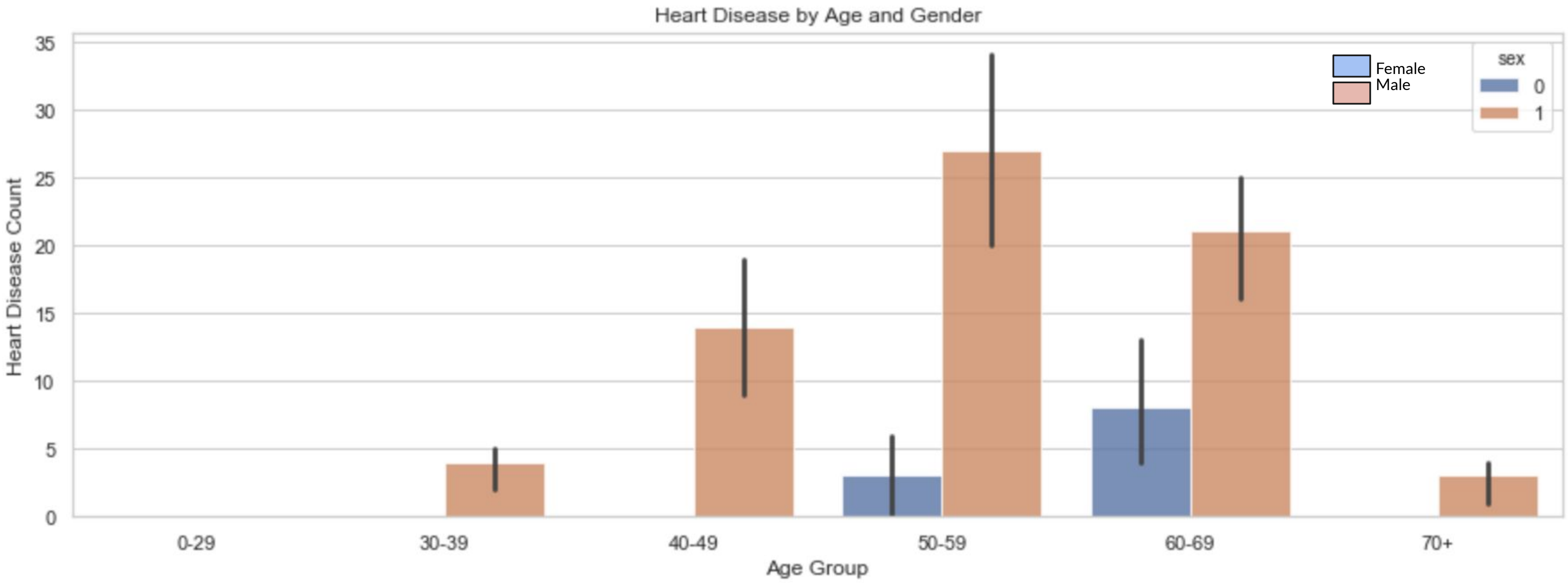
Female
Male

```
Heart Disease Value Counts:
 0    100
1     80
Name: heart_disease_present, dtype: int64
Gender Value Counts:
 1    124
0     56
Name: sex, dtype: int64
```

Method: Data Exploration

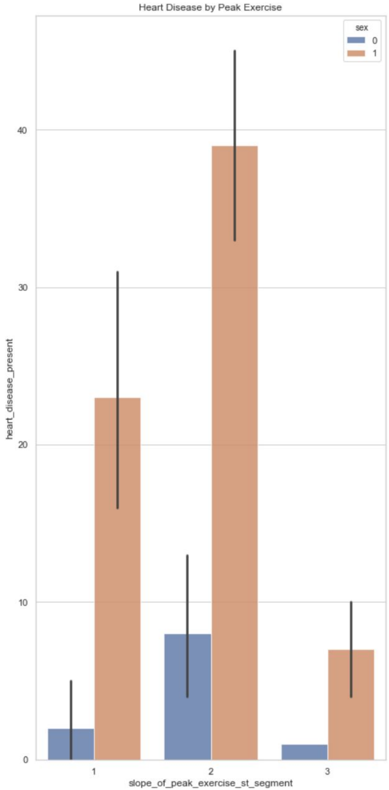


Method: Data Exploration

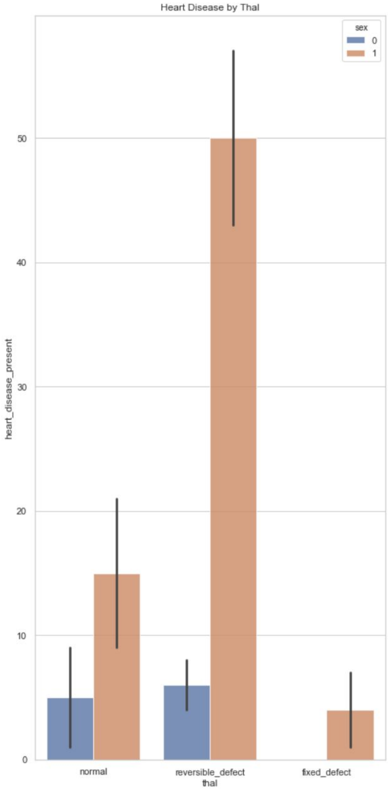


Method: Data Exploration

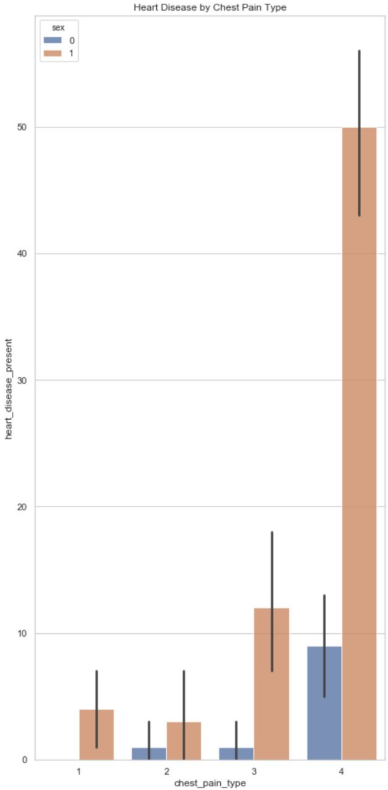
Heart Disease by
Peak Exercise ST Segment



Heart Disease by
Thal



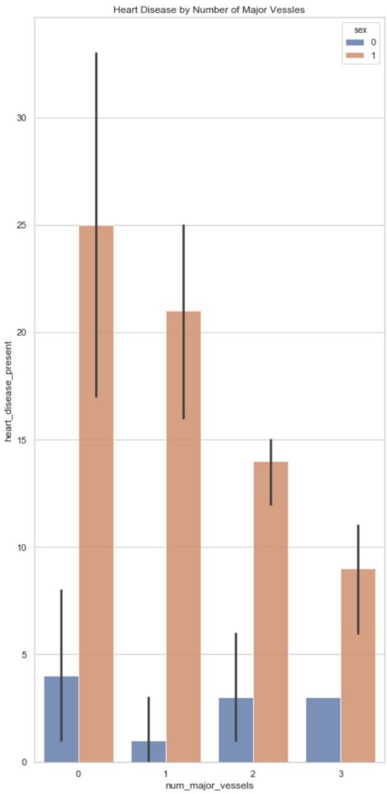
Heart Disease by
Chest Pain Type



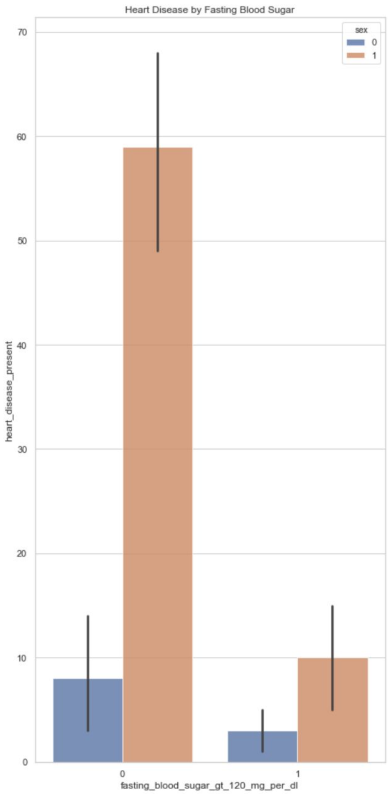
Heart disease present
Heart disease not present

Method: Exploring the data

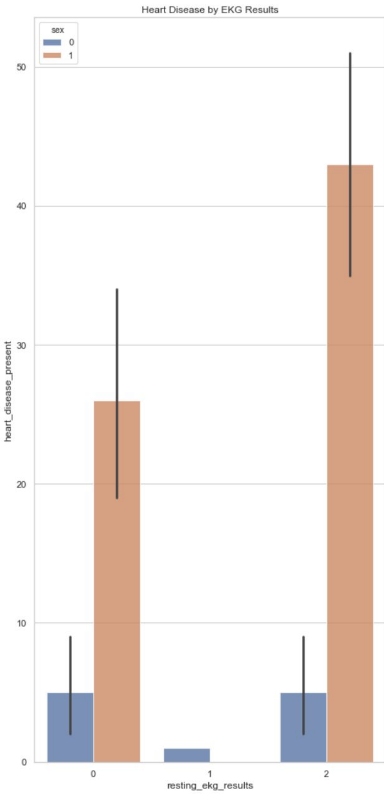
Heart Disease by
Number of Major Vessels



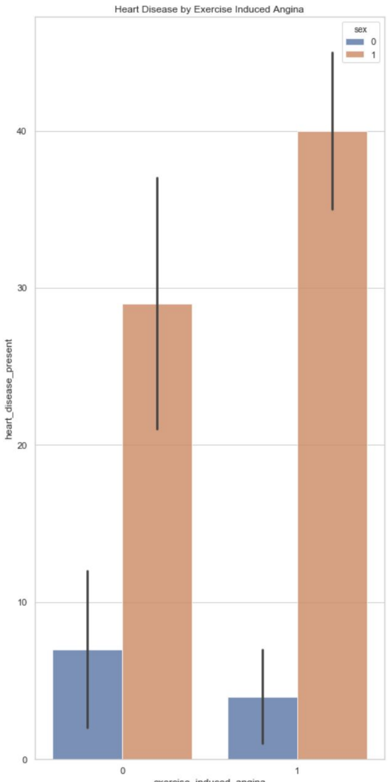
Heart Disease by
Fasting Blood Sugar



Heart Disease by
EKG Results

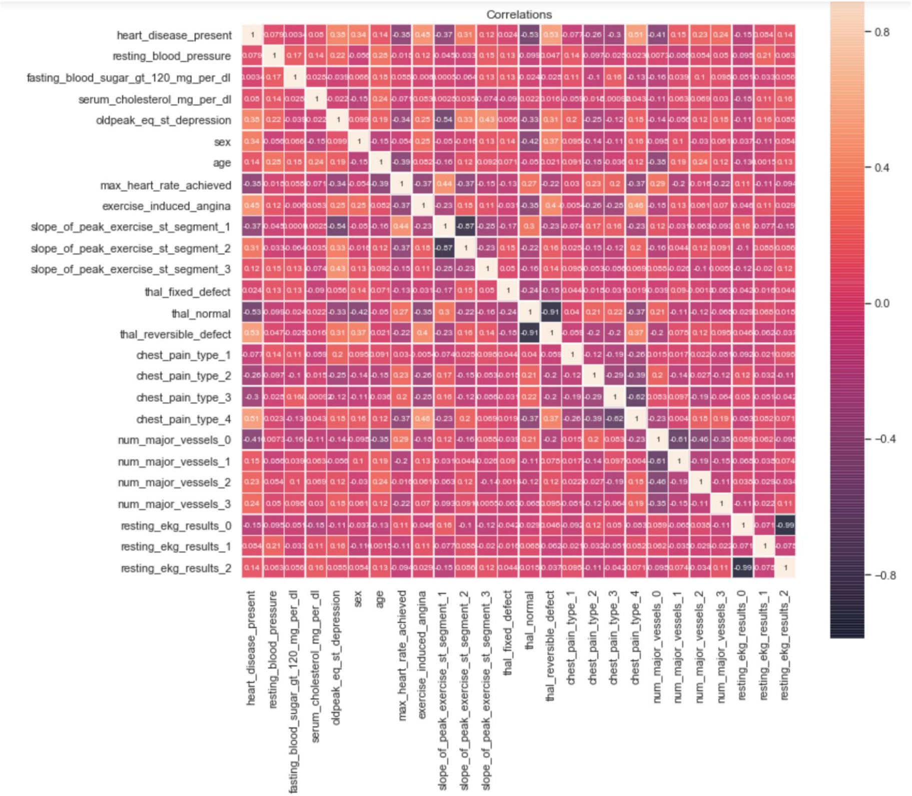


Heart Disease by
Exercise Induced Angina



Heart disease present
Heart disease not present

Method: Exploring the data Correlations



Method: Data Preparation

The following categorical features were binarized using get_dummies:

- Slope of peak exercise
- Thal
- Chest pain type
- Number of major vessels
- Resting EKG results

Method: Modelling & Evaluation

Outcome Variable: Categorical	Data Splitting	Applying Classification Models	Model Tuning	Evaluation
Presence of Heart disease is indicated by binary (0,1)	Data sets are split into Train and Test Sets using train_test_split from sklearn.model_selectio n.	In order to predict heart disease, classification methods will be used (Naive Bayes, Logistic Regression, KNN Classifier, Decision Tree, Random Forest, Bagging , Support Vector Models, Gradient Boosting Model, Ada Boosting and Stacking.	Each model requires different type of tuning.	To evaluate the performance of the different classifiers accuracy scores are used (metrics.accuracy_scor e), as well as classification reports (classification_report).

Method: Naive Bayes

- Data Set Features: mixed data types (categorical and continuous). The following process is applied:
  a. Data is divided according to data type.
  b. For categorical data, the Bernoulli model is used (BernoulliNB).
  c. For continuous data, the Gaussian model is used (GaussianNB).
     - Features (Age, Max heart rate achieved, ST Depression, Resting blood pressure and Serum cholesterol) are normalized using .sqrt)
  d. The probabilities for outcomes from both models are added, averaged and binarized.

Method: Feature Engineering

Feature engineering:

1. Converted to dummy variables
2. Normalized data
3. Experimented with combining features

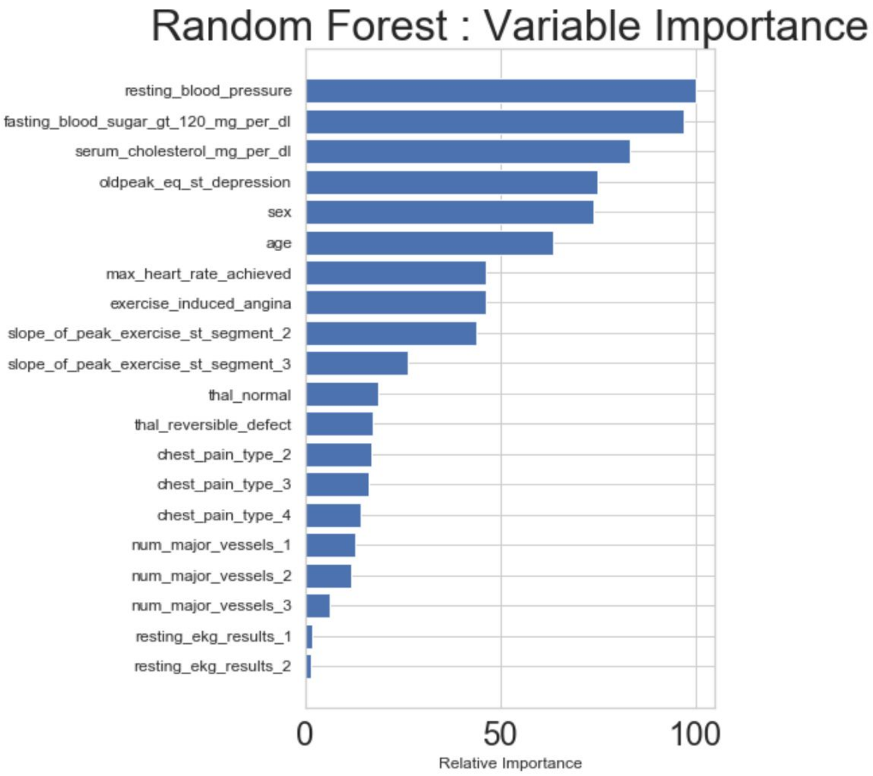
Feature Selection:

1. Feature selection tool from Gradient Boosting Model
2. KBest feature selection tool

Method :Feature Selection (Fandom Forest Model)

The outcome variable in this project is presence of heart disease. Relying on Feature Importance tool (Random Forest), the features that determine the outcome are:

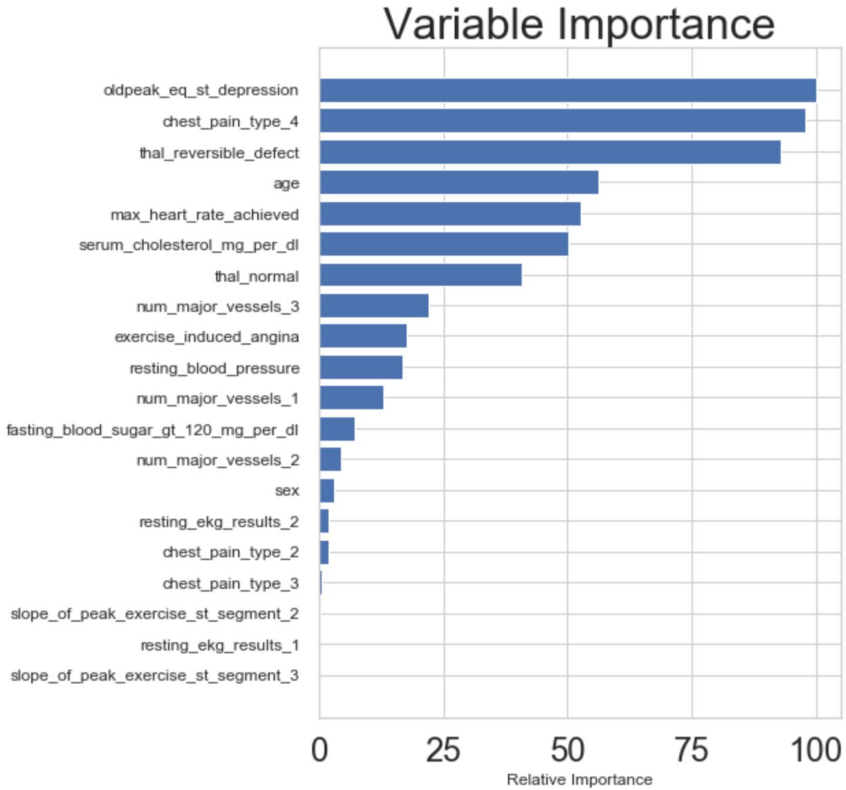
- Resting Blood Pressure
- Fasting Blood Sugar
- Serum Cholesterol
- ST Depression
- Sex
- Age



Method :Feature Selection (Gradient Boosting Model)

The outcome variable in this project is presence of heart disease. Relying on Feature Importance tool (Gradient Boosting Model), the features that determine the outcome are:

- ST Depression
- Chest Pain Type
- Thal
- Age
- Max Heart Rate Achieved



Method: Feature Selection (Select K best)

Relying on Feature Importance tool (Gradient Boosting Model), the features that determine the outcome are:

- Max Heart Rate Achieved
- ST Depression
- Exercise Induced Angina
- Thal Reversible Defect
- Chest Pain Type 4

Method: Model Tuning (Hyperparameters)

- LM:
  - C
  - Penalty
- KNN:
  - n= number of neighbors
  - Leaf_size
  - Weight
  - Algorithm

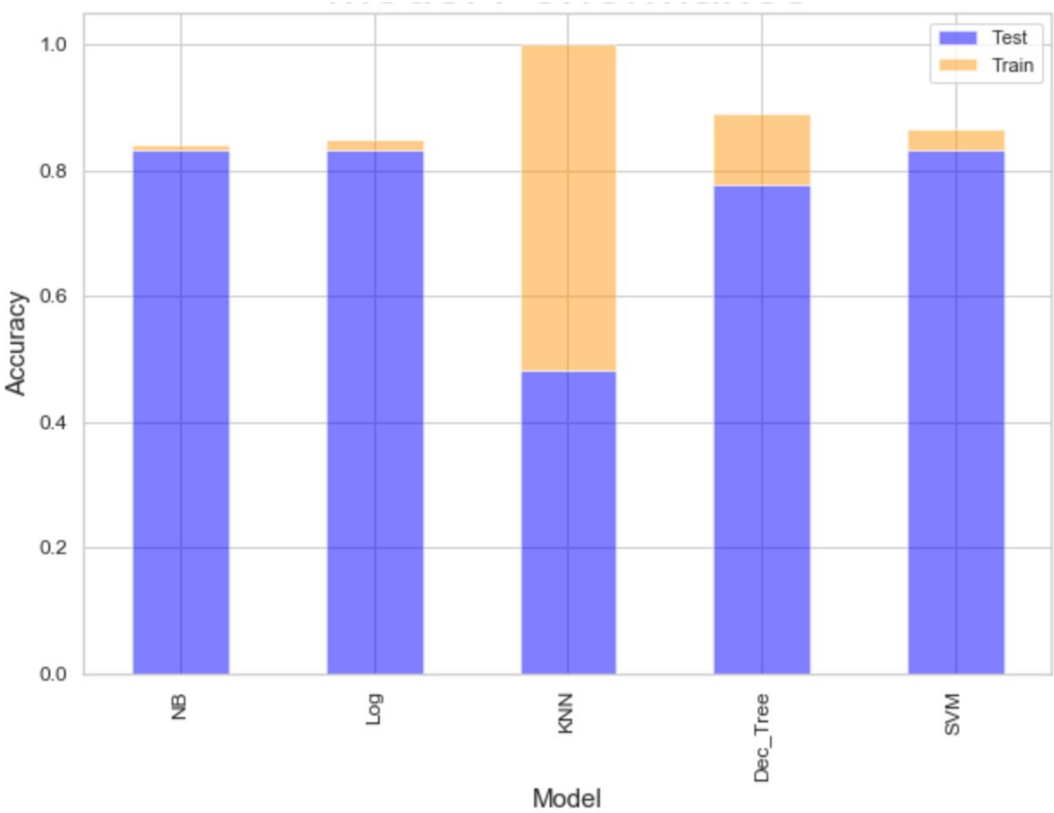
Method: Model Tuning (Parameters)

- Decision Tree:
    - Number of max features
    - Depth
- RFC:
    - Criterion
    - Max_depth
    - Max_features
    - N_estimators
    - Class_weigth

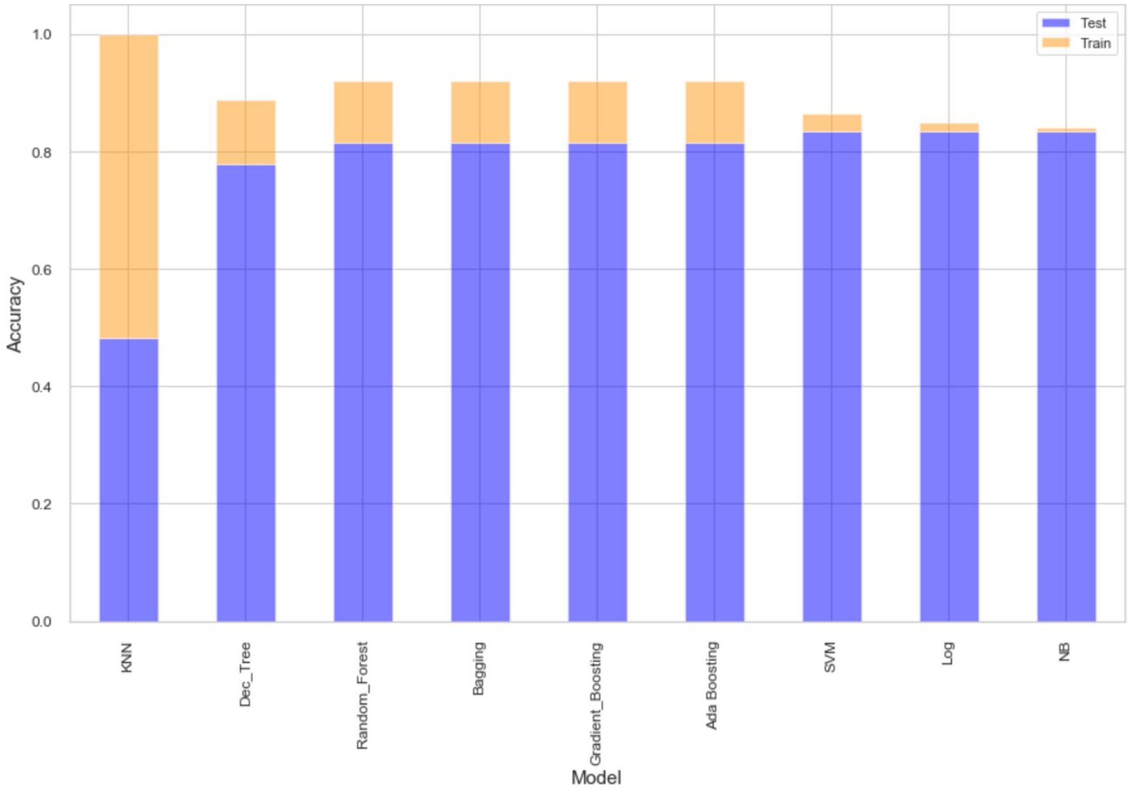
Method: Model Tuning (Parameters)

- Bagging:
  - Number of max features
  - Depth
- SVC:
  - Kernel
  - C
  - Gamma

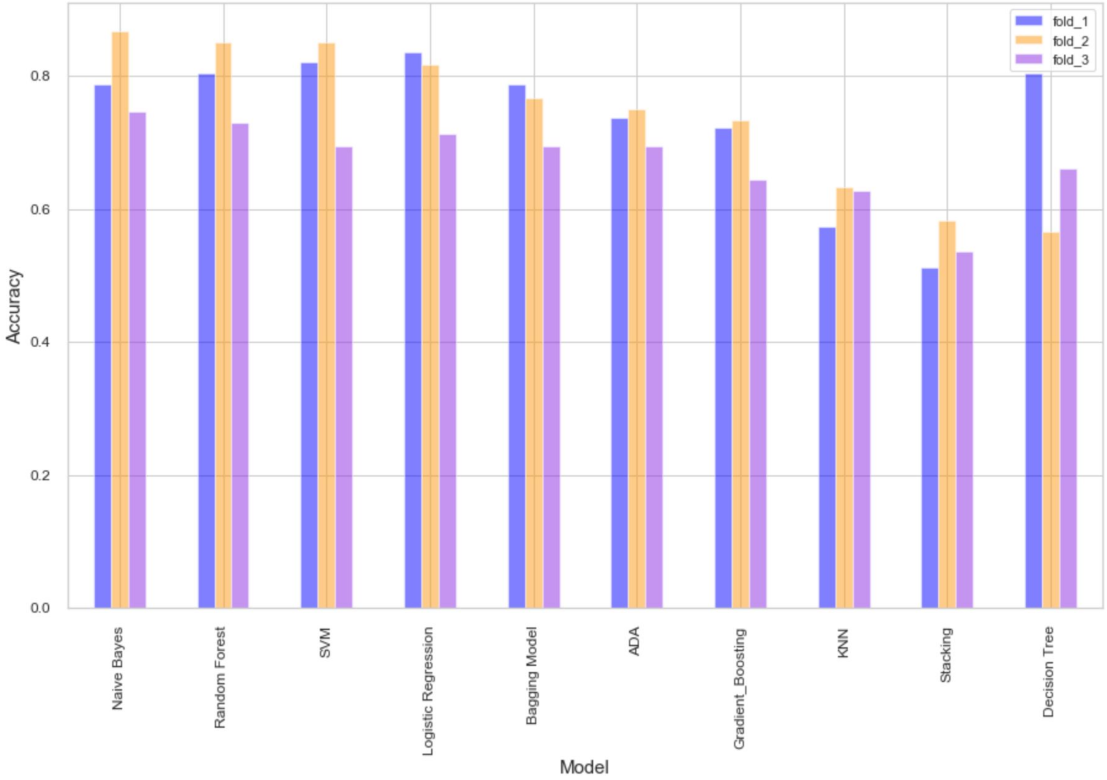
Evaluation: Individual Predictors Performance



Evaluation: Ensemble Predictors Performance



Evaluation: Ensemble Predictors Performance



Conclusion

- Relying on accuracy scores, NB performed the best in terms of generalization gap.
- SVM performed the best out of ensemble models when using accuracy score and cross evaluation.

Next Steps and Missed Opportunities

- Invest more time on feature selection.
- Invest more time on tuning (Ensemble Models).
- Use GridSearchCV for model selection
- Pipeline method to optimize code, tune parameters and try multiple functions.
- Create two different models for males and females.
- Visualize differences in Classification Reports results
- Use confusion matrices to identify specificity

Limitations

- Time
- Dataset publicly available
- Superficial knowledge of subject matter

Thanks!!

Questions?

NB

- Correlations between key features are relatively low (>0.45).
- Relationship between Outcome variable and key features is linear.

Rationale for choosing the models