Heart disease: what is my status?

Mireille P. Feudjio T.

Springboard

Outline

- **Context and Problem statement**
- **Exploratory data analysis**
- **Modelling**
- **Conclusion**

Context and Problem statement

Context

- Heart disease is the leading cause of death in the United States. In fact, Heart failure is a common event caused by Cardiovascular diseases (CVDs) which account for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age.

- The doctor said that she has a high risk of getting a heart attack because the exam showed that some values of the factors such as blood pressure, cholesterol, blood sugar are not good.

- Then patient said: I thought you were studying data, this is my lab exam result (.....), can you tell me what your models say? Do I potentially subject to heart risk? What are the factors that characterize people with high cardiovascular risk?

- In fact, people with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

The next presentation tries to answer Mom's question based on the heart failure prediction dataset taken from kaggle

Scope of solution space: the model development should take into account all the features with special attention on sex , blood pressure, blood sugar and cholesterol.

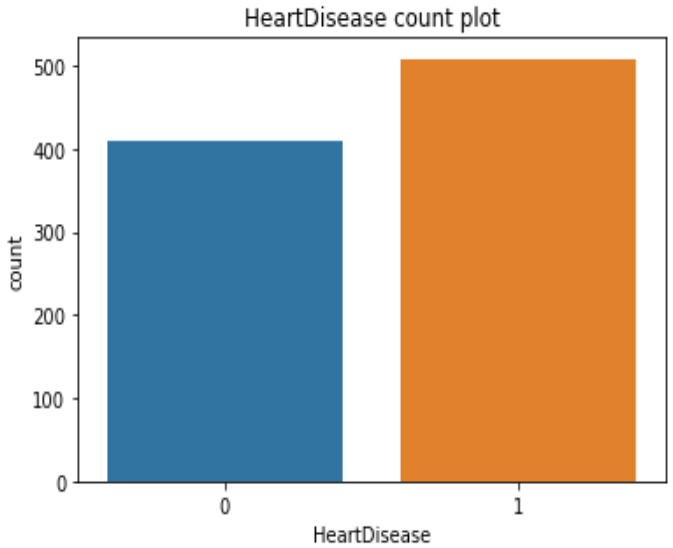
Constraints: the dataset miss information about weight, race , family history that I think could be important to the model.

Dataset informations

Attribute Information

data shape is (918, 12)

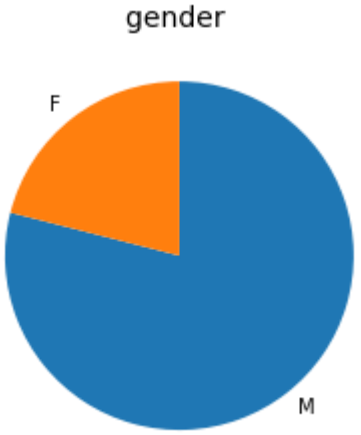
- **Age**: age of the patient [years]
- **Sex**: sex of the patient [M: Male, F: Female]
- **ChestPainType**: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- **RestingBP**: resting blood pressure [mm Hg]
- Cholesterol: serum cholesterol [mm/dl]
-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

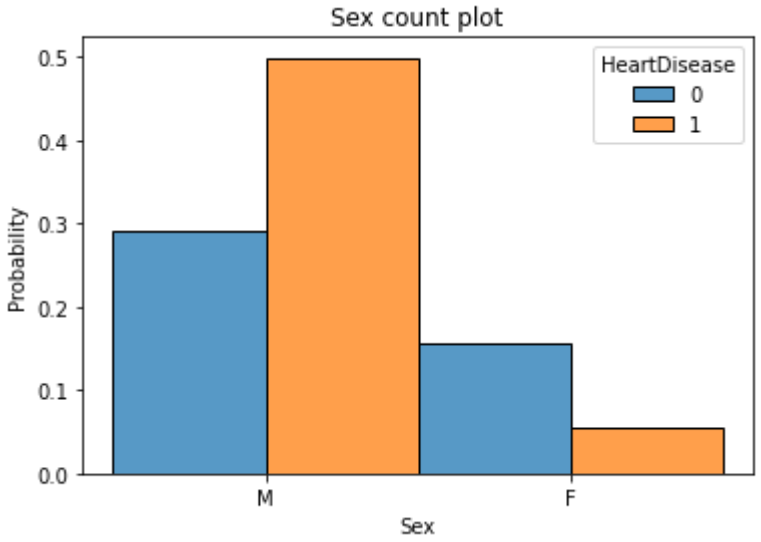


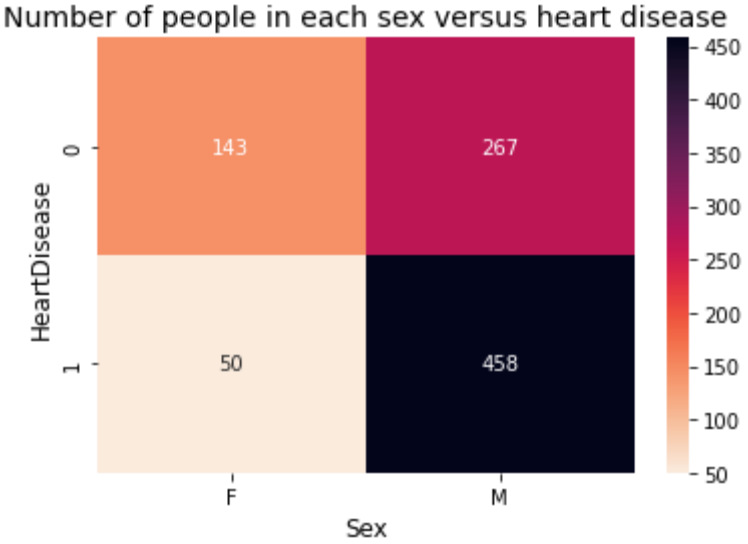
Exploratory data analysis

What are the factors that characterize people with high cardiovascular risk?

Gender analysis







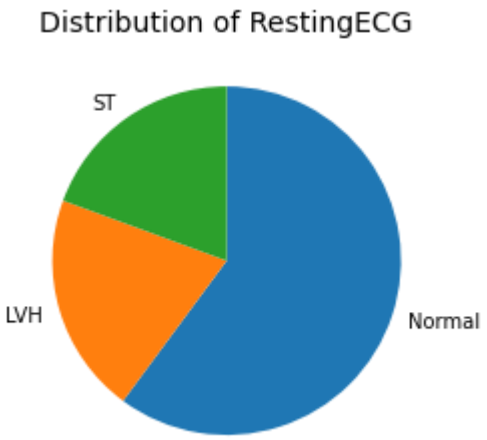
Keynote:

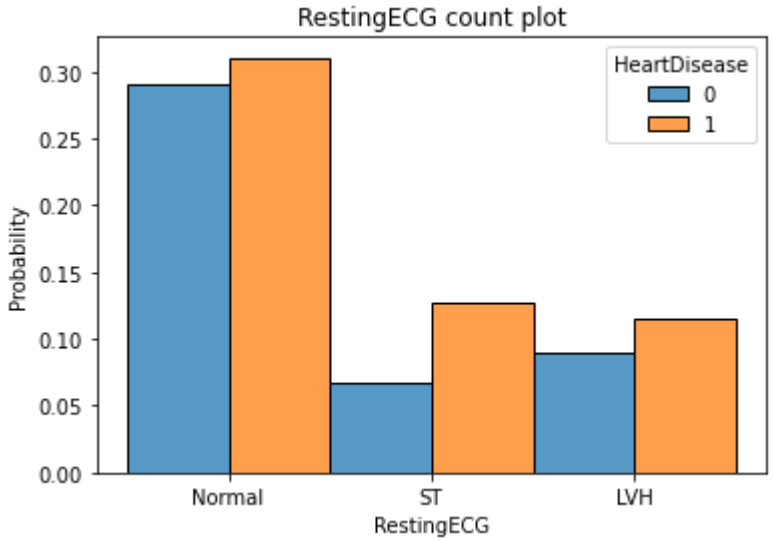
Person with heart disease is more likely to be Male more than female.
From all gender, about 50% of male with heart diseases.

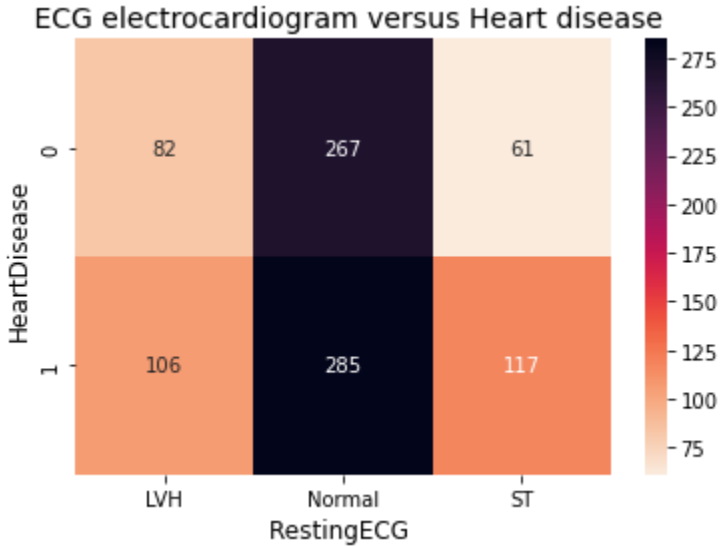
Exploratory data analysis

What are the factors that characterize people with high cardiovascular risk?

Resting ECG analysis







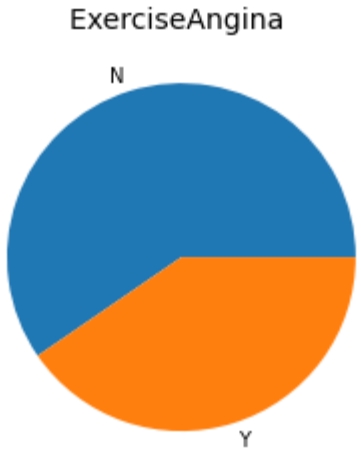
Keys notes:

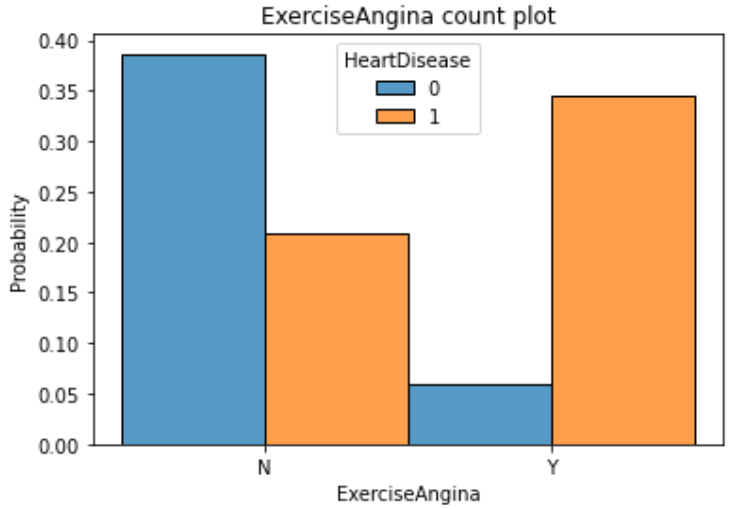
Person with heart disease is more likely to be person with normal resting electrocardiogram (about 30% of the entire observations).

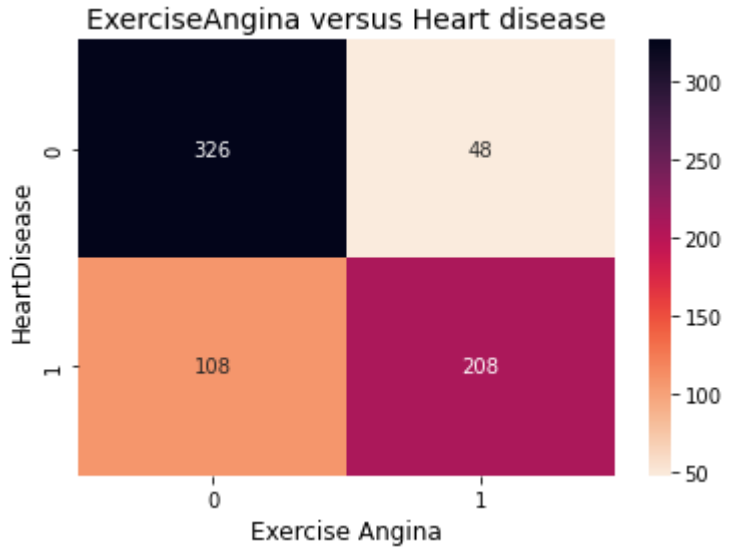
Exploratory data analysis

What are the factors that characterize people with high cardiovascular risk?

Exercise Angina analysis







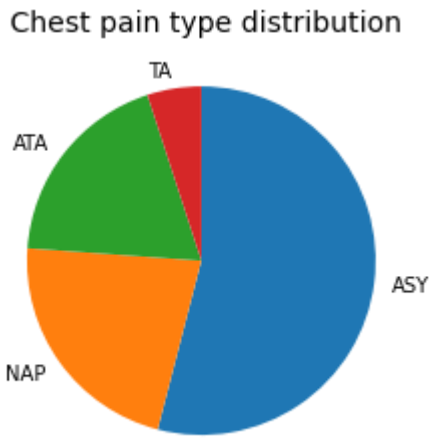
Keys notes:

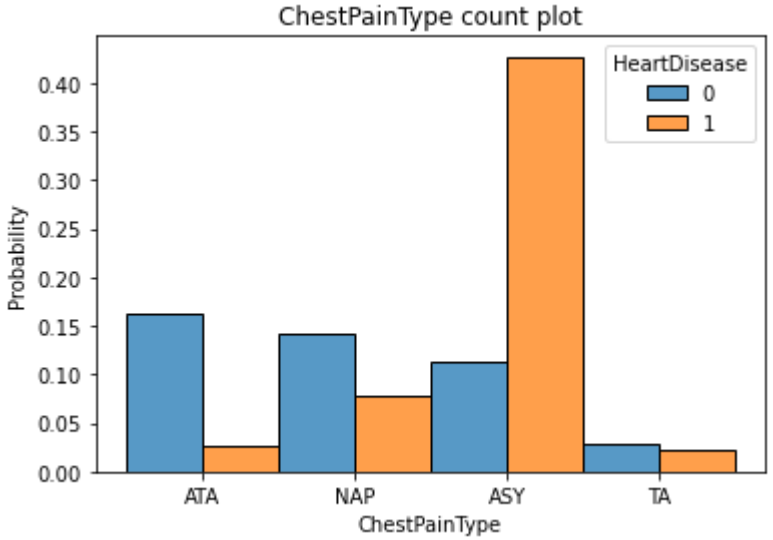
Person with heart disease is more likely to be person who exercise-induced angina (about 35% of the total observation).

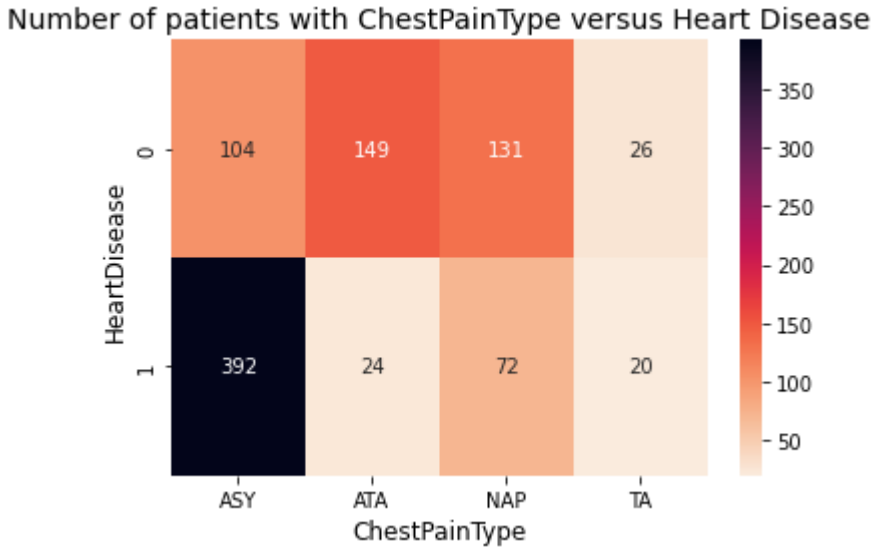
Exploratory data analysis

What are the factors that characterize people with high cardiovascular risk?

Chest Pain type analysis







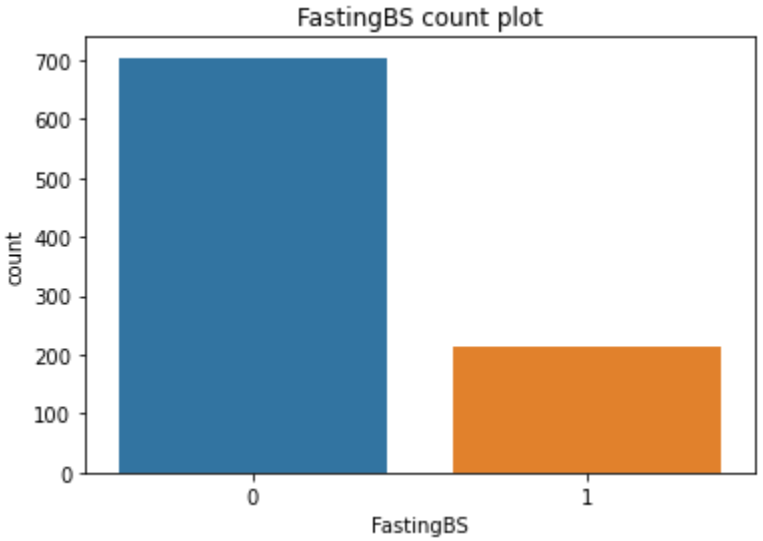
Keys notes:

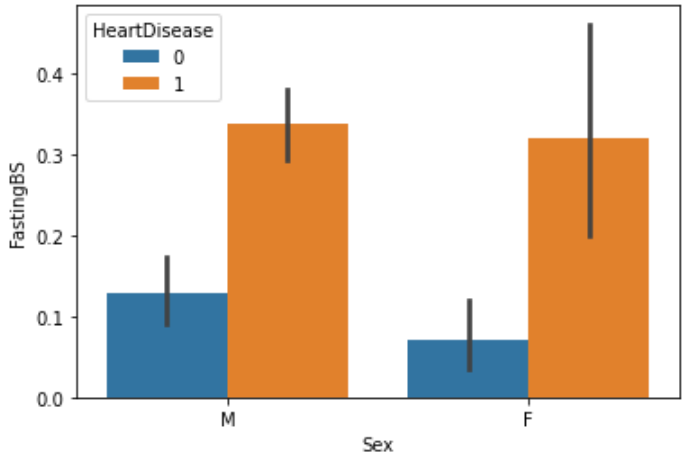
Person with heart disease is more likely to be person with asymptomatic chest pain type.

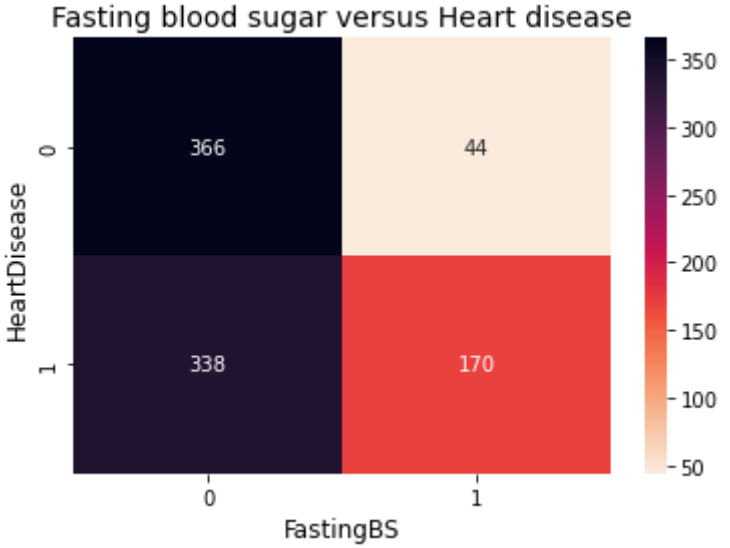
Exploratory data analysis

What are the factors that characterize people with high cardiovascular risk?

Fasting Blood Sugar analysis





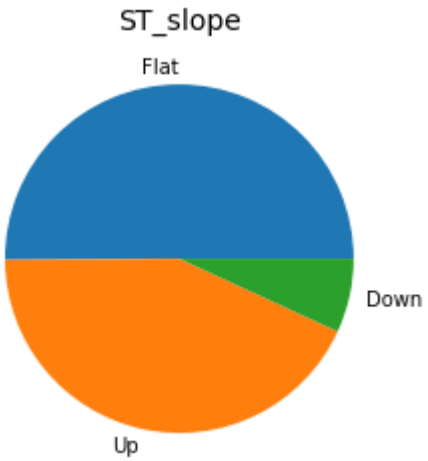


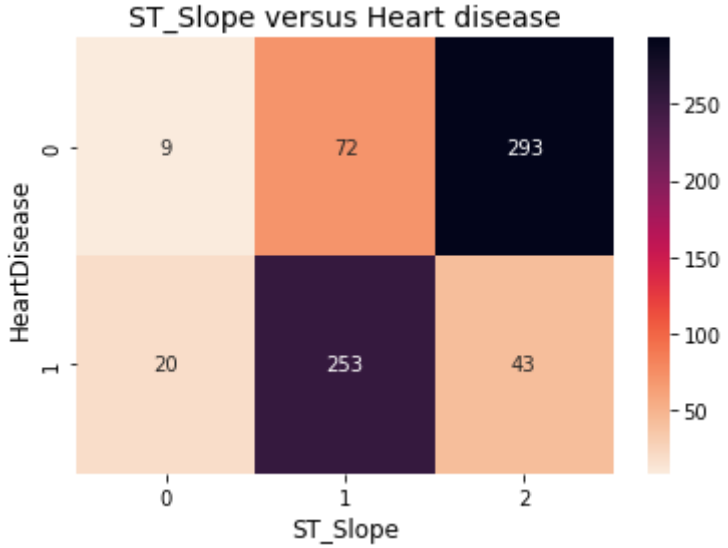
Keys notes:

Person with heart disease is more likely to be person with fasting blood sugar > 120 mg/dl (1).

Exploratory data analysis

What are the factors that characterize people with high cardiovascular risk?



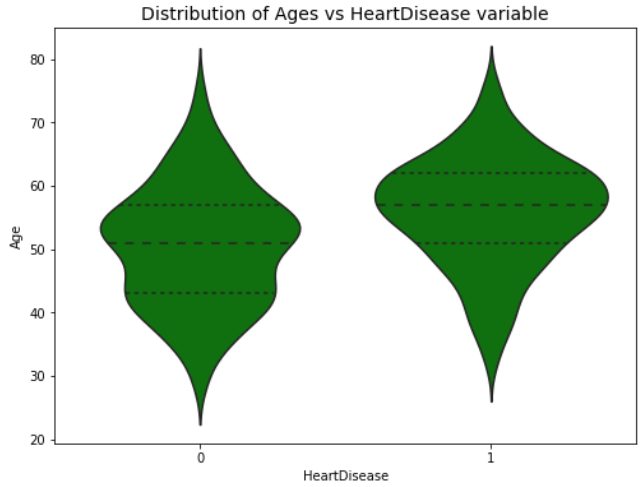


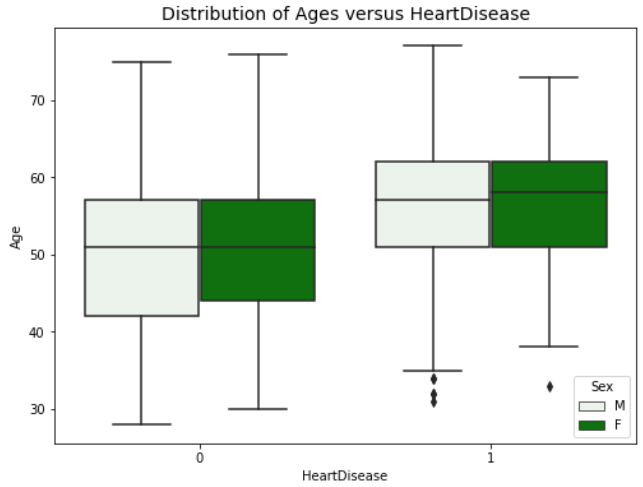
KeyNote:

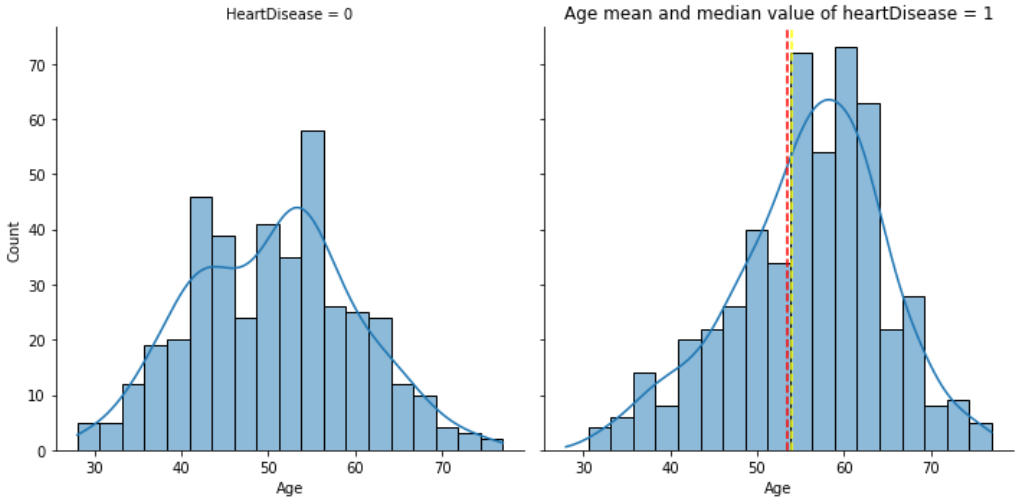
People with Flat ST_slope type are likely to have heart disease.

Exploratory data analysis

What are the factors that characterize people with high cardiovascular risk?







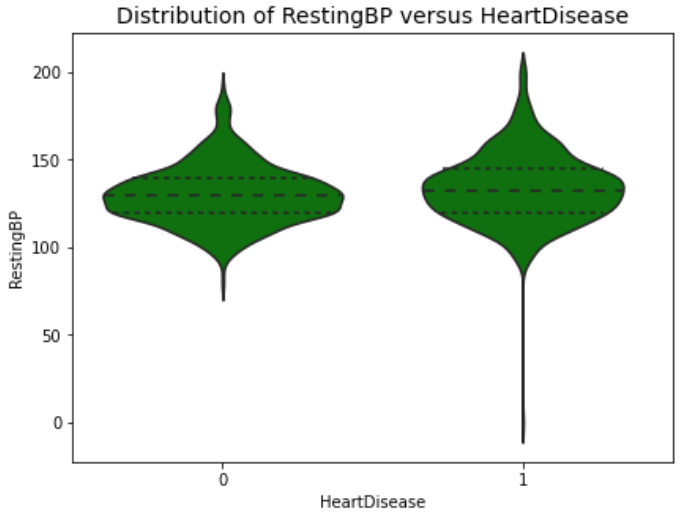
Keynote:

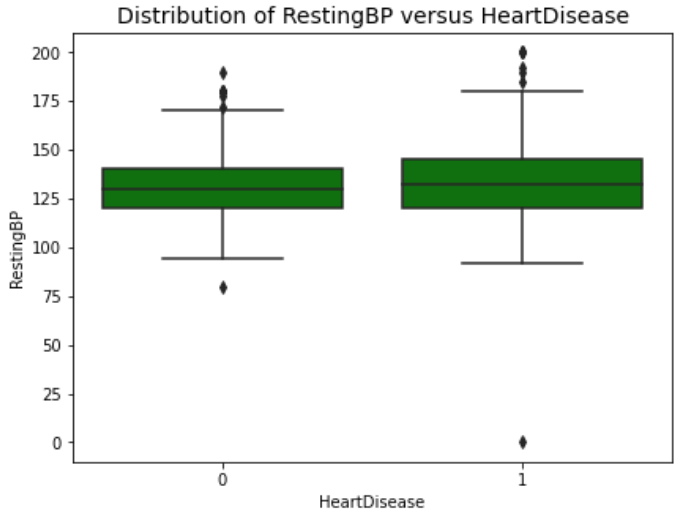
The average age of people who have heart disease is slightly higher than healthy people.

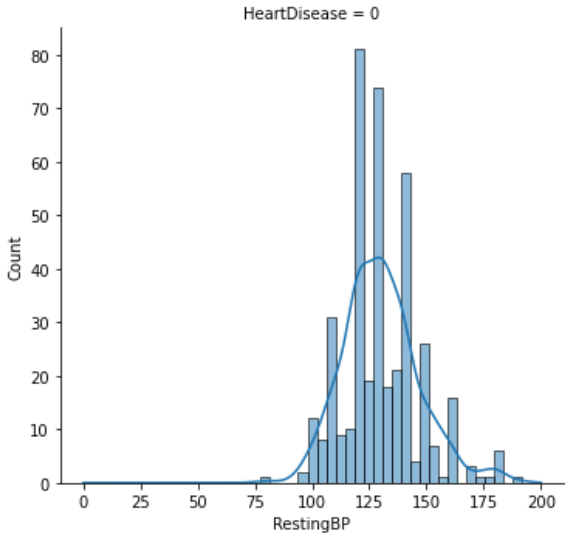
The mean value of aged person with heart disease is 55.58.

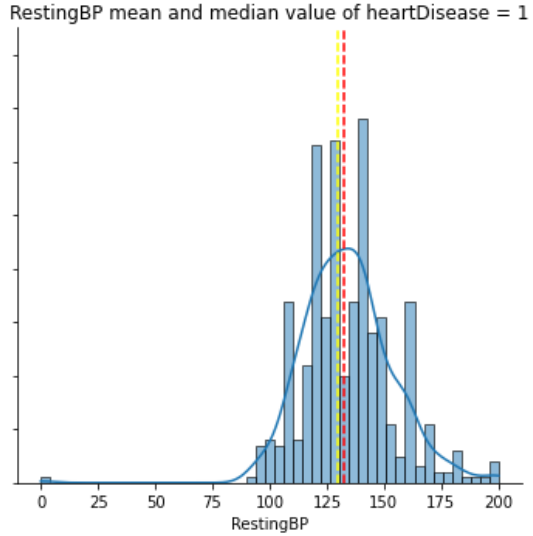
Exploratory data analysis

What are the factors that characterize people with high cardiovascular risk?









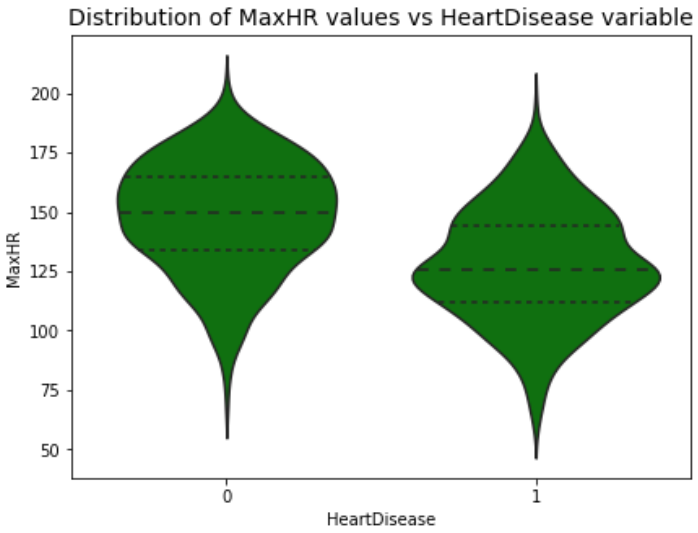
Keys notes:

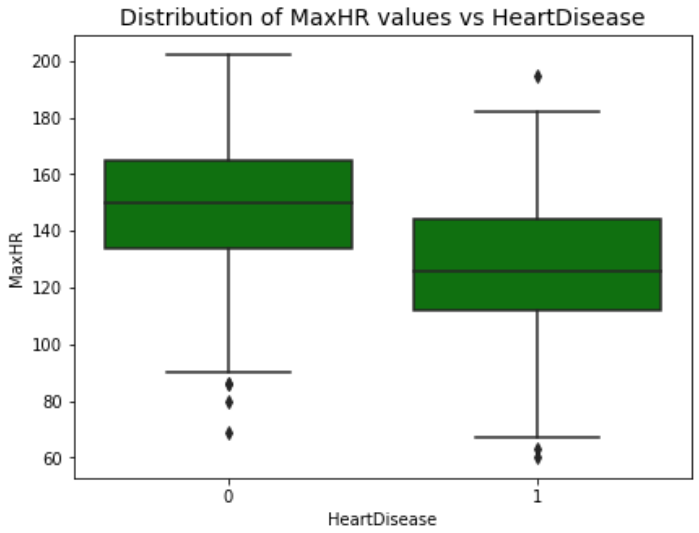
The distribution of Resting Blood Pressure seem to be similar in both healthy people and person with heart disease.

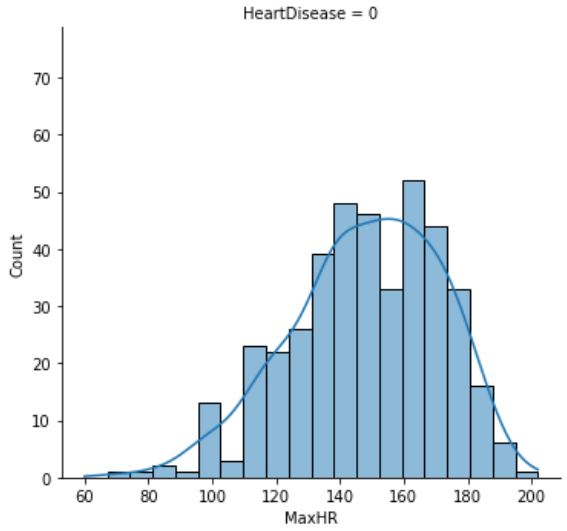
The mean value of resting **b**lood pressure of person with heart disease is 134.18.

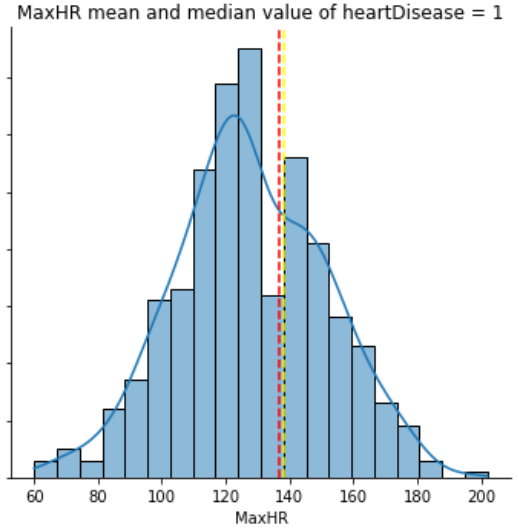
Exploratory data analysis

What are the factors that characterize people with high cardiovascular risk?







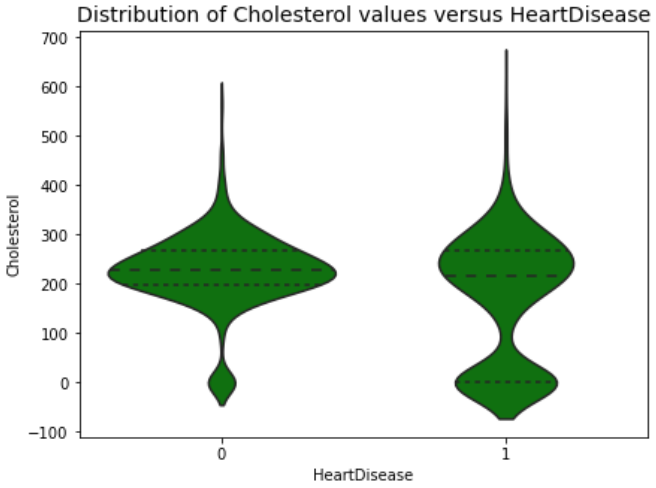


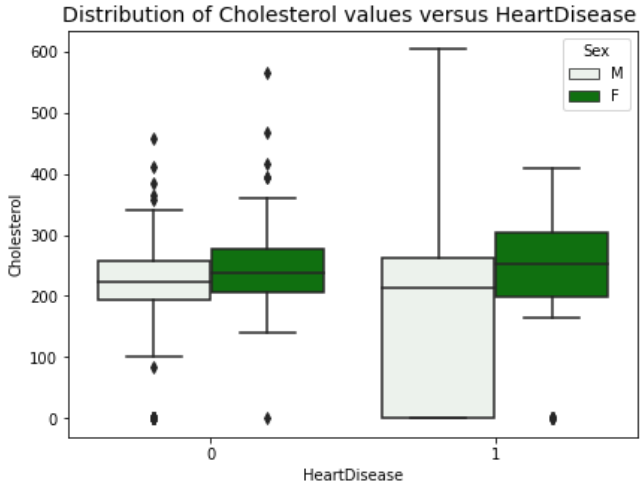
Keys notes:

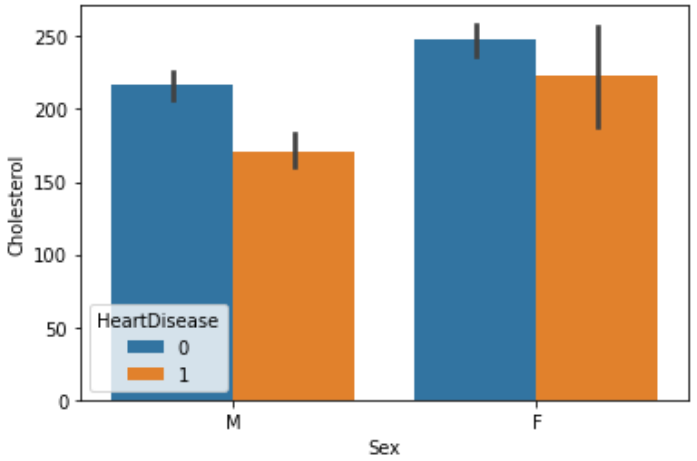
The average value of MaxHR for sick people lower than that for healthy.
The distribution of Max Heart rate for healthy people in center around 130 and 170.
The mean value of Max heart rate of person with heart disease is 127.65.

Exploratory data analysis

What are the factors that characterize people with high cardiovascular risk?





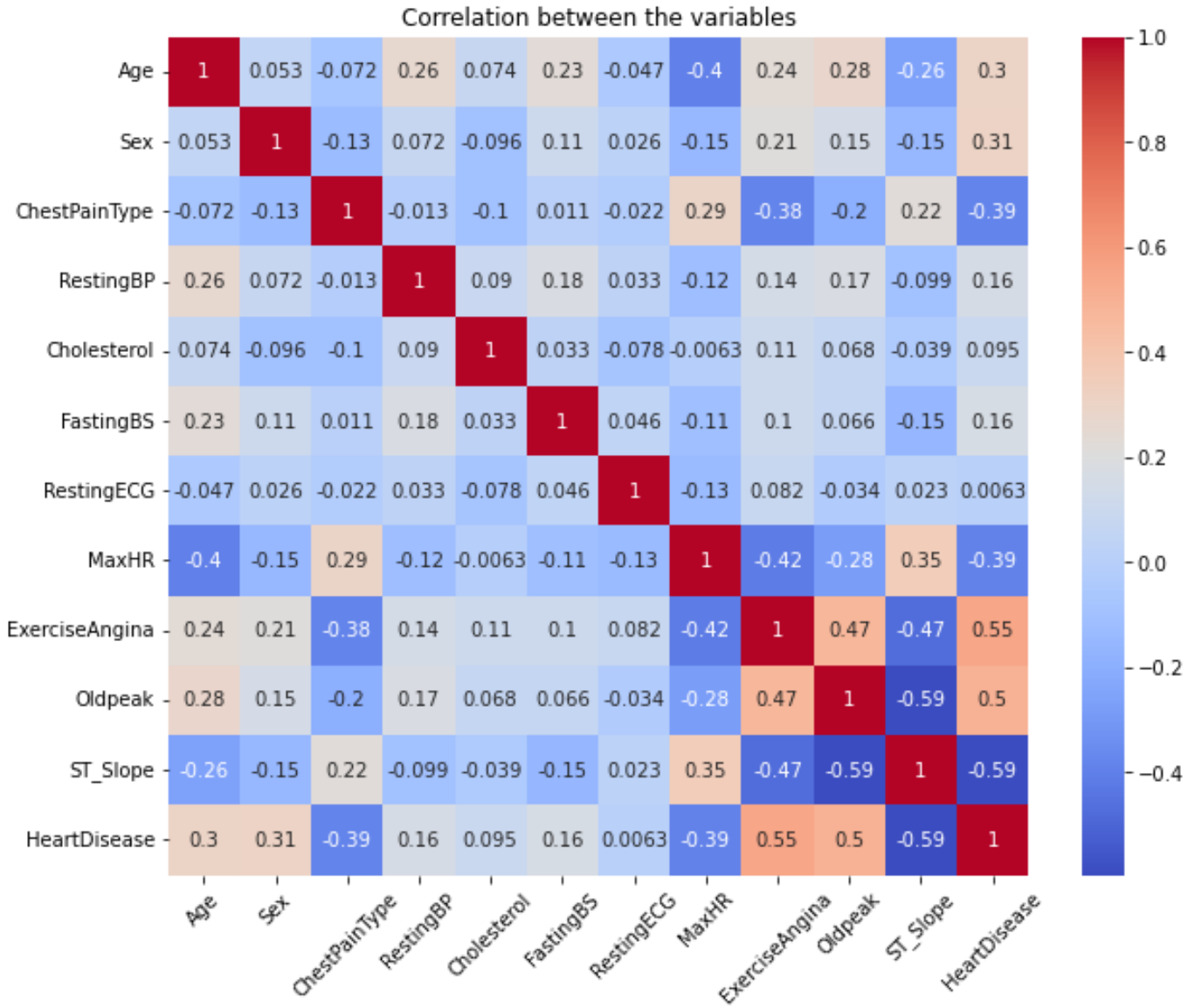


Keynote:

Female with heart disease tend to have Hight cholesterol than male.

Exploratory data analysis

How do sex , blood pressure, blood sugar and cholesterol are related to heart Disease?



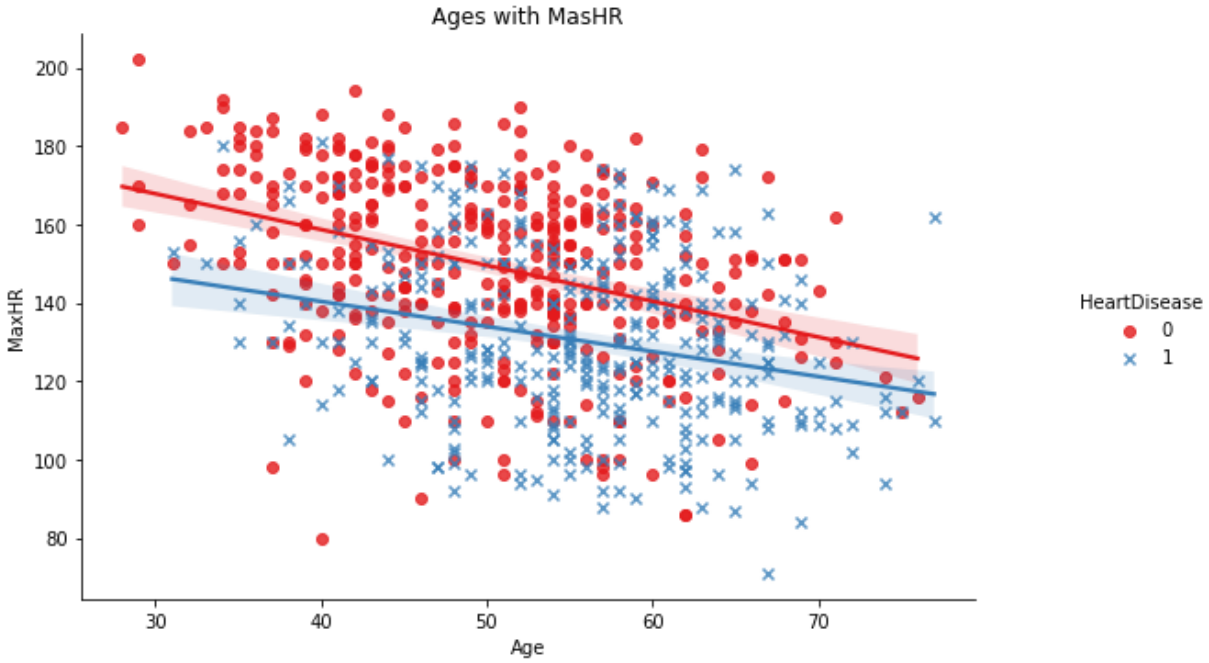
Keynote:

The features related to heart disease are :

- age and sex of the patient (weak: 0.3).
- Chestpaintype (moderate , 0.4).
- exercise-induced angina (moderate, 0.5).
- Oldpeak (0.5).
- ST_slope: the slope of the peak exercise (-0.59).

Exploratory data analysis

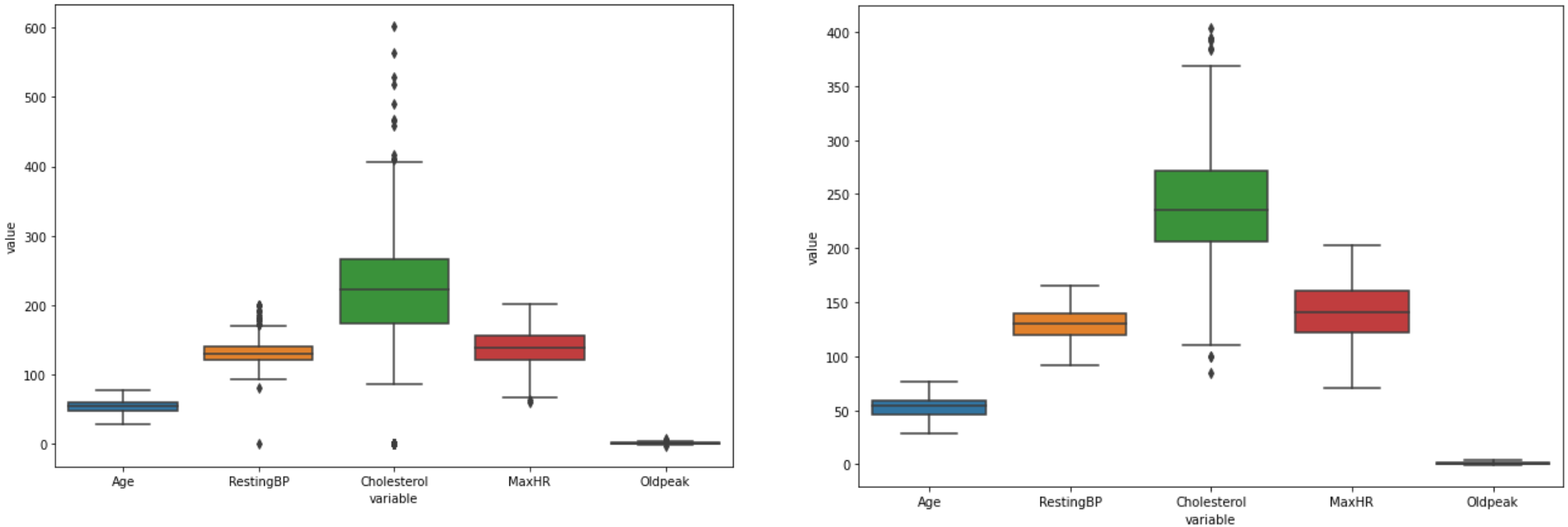
How do sex , blood pressure, blood sugar and cholesterol are related to heart disease?



Keynote:

Max heart rate in negatively correlated with age. The correlation is weak and there is no clear distinction between healthy and sick people.

Data preprocessing



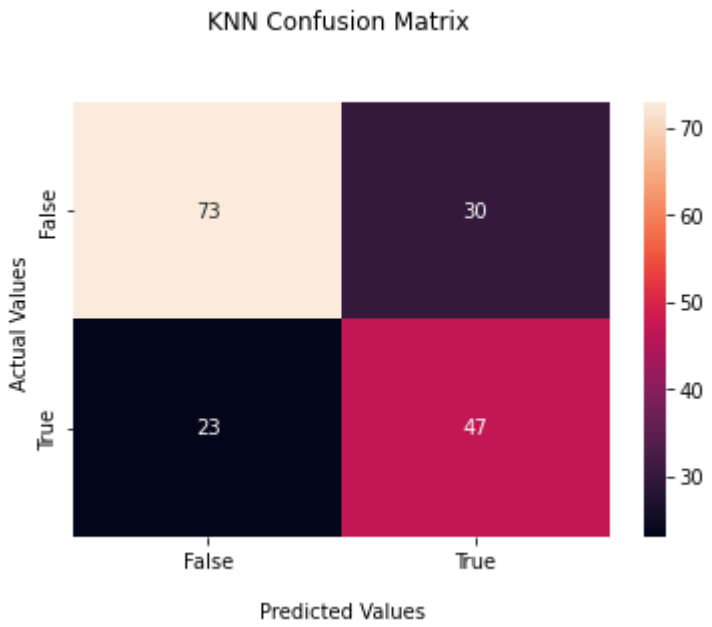
Outlier treatment with interquartile range (before and after)

- Outlier treatment with interquartile range.
- Label Encoder of categerical variable.
- Dataset divided in test set and train set (30% , 70%).
- No missing value in the dataset .

Modelling

Can you tell me what your models say? Do I potentially subject to heart risk?

KNN model



Classification_report with test set and predicted value

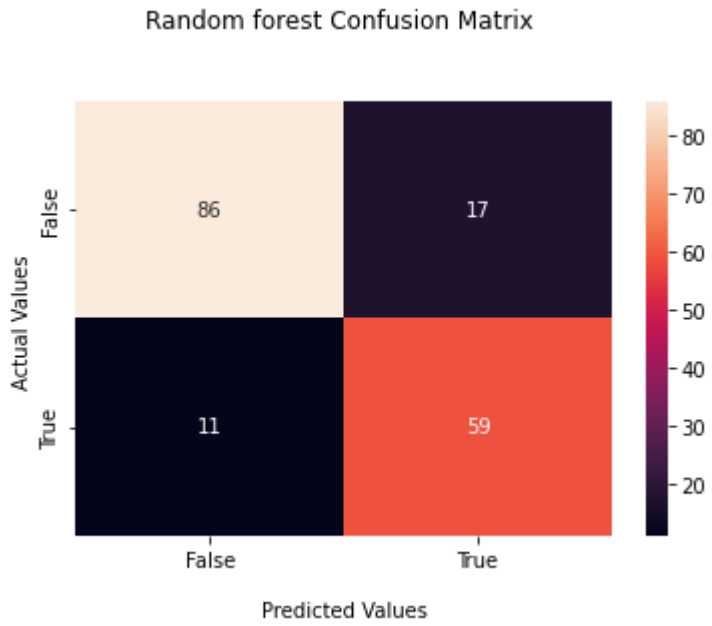
	precision	recall	f1-score	support
Heart Diseases	0.76	0.71	0.73	103
Normal	0.61	0.67	0.64	70
accuracy			0.69	173
macro avg	0.69	0.69	0.69	173
weighted avg	0.70	0.69	0.70	173

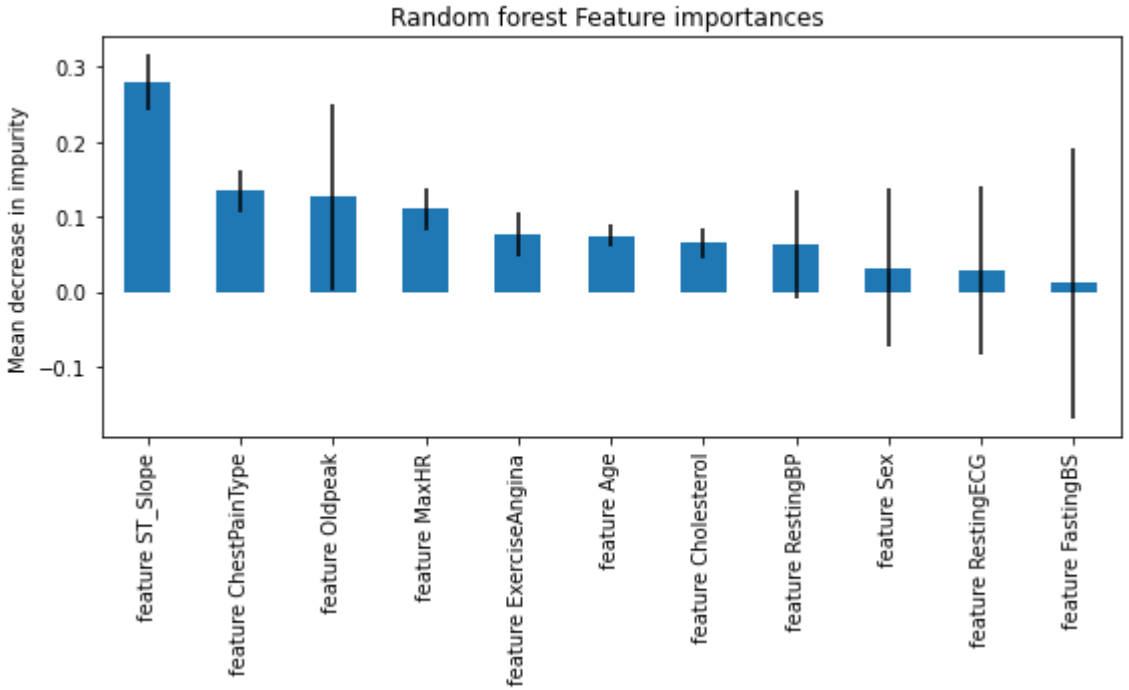
```
# lab Exam result and prediction
Exam = [{'Age': 70, 'Sex': 0, 'ChestPainType':3, 'RestingBP':165, 'Cholesterol':395, 'FastingBS':1,
      'RestingECG': 2, 'MaxHR': 192, 'ExerciseAngina':1, 'Oldpeak':3.3, 'ST_Slope':2}]
exam = pd.DataFrame(Exam)
mom_pred = knn3.predict(exam)
mom_pred : 0
```

Modelling

Can you tell me what your models say? Do I potentially subject to heart risk?

Random Forest model





Classification_report with test set and predicted value

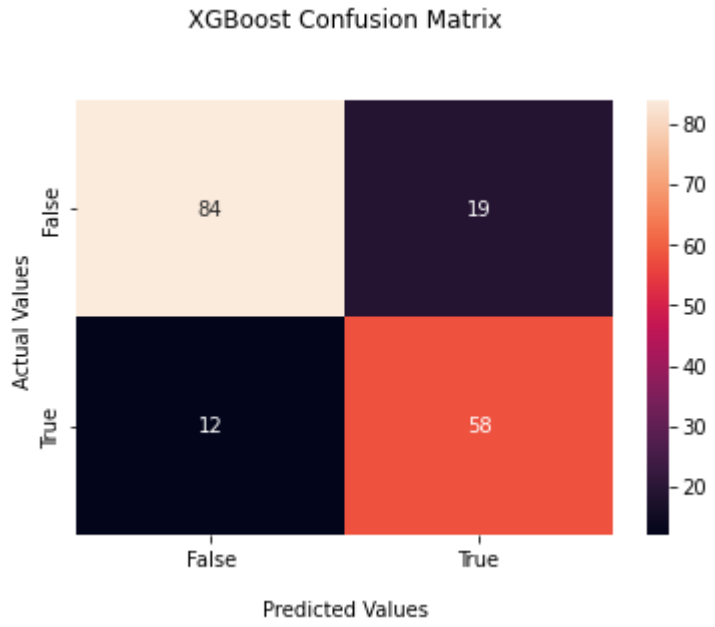
	precision	recall	f1-score	support
Heart Diseases	0.89	0.83	0.86	103
Normal	0.78	0.84	0.81	70
accuracy			0.84	173
macro avg	0.83	0.84	0.83	173
weighted avg	0.84	0.84	0.84	173

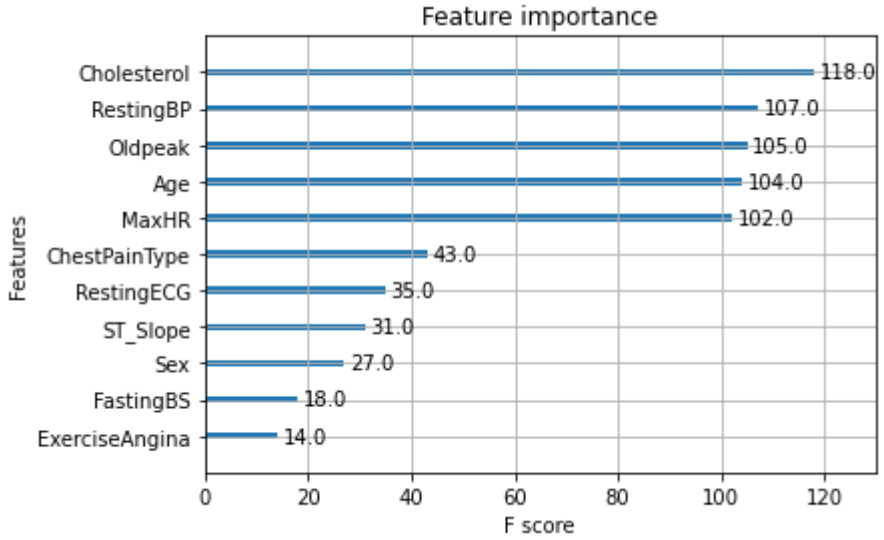
```
# lab Exam result and prediction
Exam = [{'Age': 70, 'Sex': 0, 'ChestPainType':3, 'RestingBP':165, 'Cholesterol':395,
'FastingBS':1,
     'RestingECG': 2, 'MaxHR': 192, 'ExerciseAngina':1, 'Oldpeak':3.3, 'ST_Slope':2}]
exam = pd.DataFrame(Exam)
mom_pred = rfOpt.predict(exam)
mom_pred : 0
```

Modelling

Can you tell me what your models say? Do I potentially subject to heart risk?

XGBoost model





Classification_report with test set and predicted value

	precision	recall	f1-score	support
Heart Diseases	0.88	0.82	0.84	103
Normal	0.75	0.83	0.79	70
accuracy			0.82	173
macro avg	0.81	0.82	0.82	173
weighted avg	0.83	0.82	0.82	173

```
# lab Exam result  and prediction
Exam = [{'Age': 70, 'Sex': 0, 'ChestPainType':3, 'RestingBP':165, 'Cholesterol':395,
'FastingBS':1,
      'RestingECG': 2, 'MaxHR': 192, 'ExerciseAngina':1, 'Oldpeak':3.3, 'ST_Slope':2}]
exam = pd.DataFrame(Exam)
mom_pred = rfOpt.predict(exam)
mom_pred : 0
```

Conclusion

Factors that characterize people with high cardiovascular risk

Person with heart disease is more likely to be:

- male more than female;
- be person with normal resting electrocardiogram;
- person who exercise-induced angina;
- person with Asymptomatic chest pain type;
- be person with fasting blood sugar > 120 mg/dl (1) ;
- Female with heart disease tend to have Hight cholesterol than male.

- mean value of aged person with heart disease is 55.58.
- mean value of resting blood pressure of person with heart disease is 134.18.
- mean value of Max heart rate of person with heart disease is 127.65.

Model choice

3 models developed

- KNN;
- Random Forest;
- XGBOOST;

- all the model predict from lab exam that mom have no risk to heart disease;
- we cared about correct prediction, and we have unbalanced dataset;
- we considered F1_score to select the best model;
- the best Model is Random Forest;

THANKS