Machine Learning Project

Heart Failure prediction

Michele Cucchiaro

DTM

Business understanding

Business understanding

Context of the project

Measurements of patients' clinical features are taken. It is recorded whether they are affected by heart disease or not.

Business objective

Predicting the occurance of heart diseases in patients, based on the clinical features.

Data Mining goal

Correctly assigning the patients to a group (affected by heart dieses - not affected by heart diesease), based on the measurements taken.

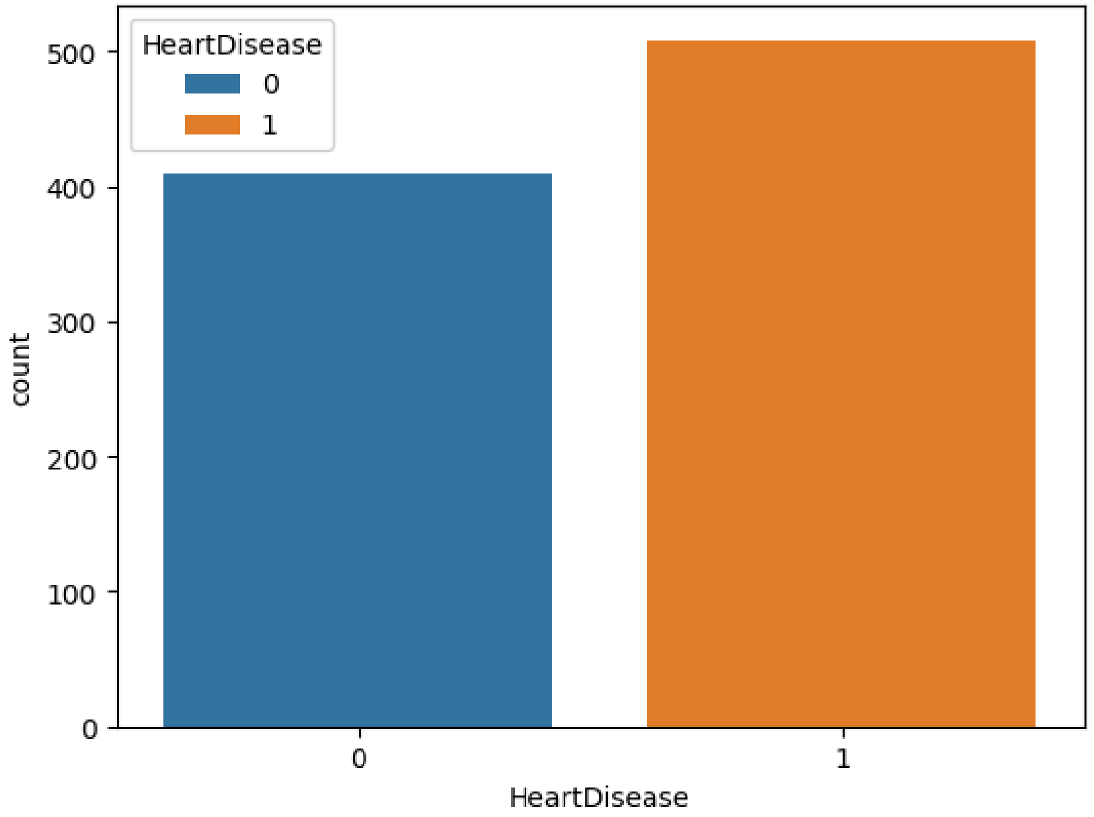
Data understanding

Data understanding

Key information: there are **918** total measurements, each with **12 features**. There are both numerical and non numerical values.

- **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 colesterol (mm/d)
- **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:** 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)

Data understanding



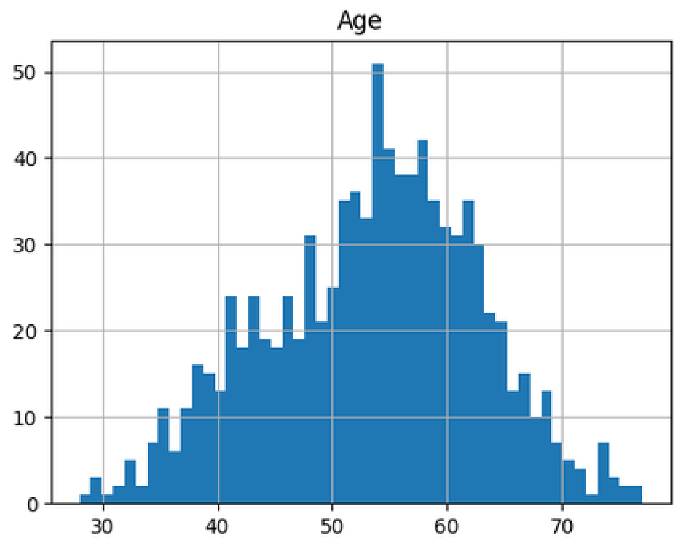
Out of the 918 observations, **508** resulted in patients being diagnosed with heart diesease, while **410** resulted in patients being healthy

ST_Slope	
Flat	460
Up	395
Down	63

ChestPainType	
ASY	496
NAP	203
ATA	173
TA	46

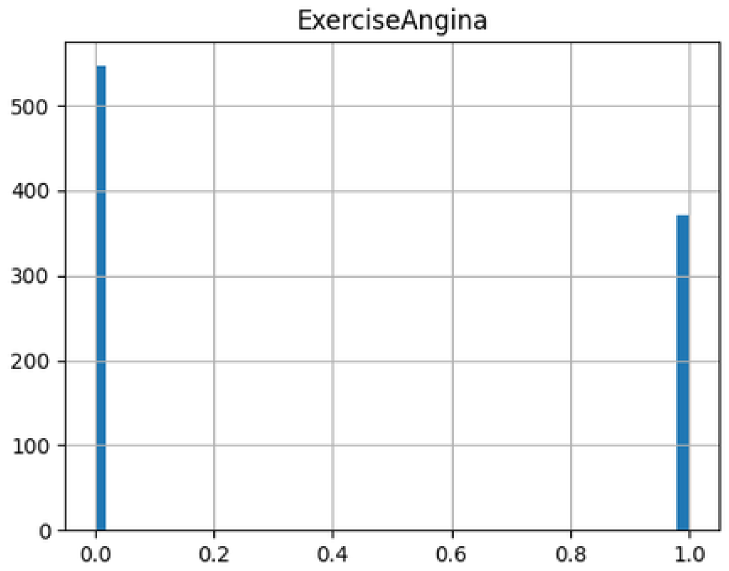
RestingECG	
Normal	552
LVH	188
ST	178

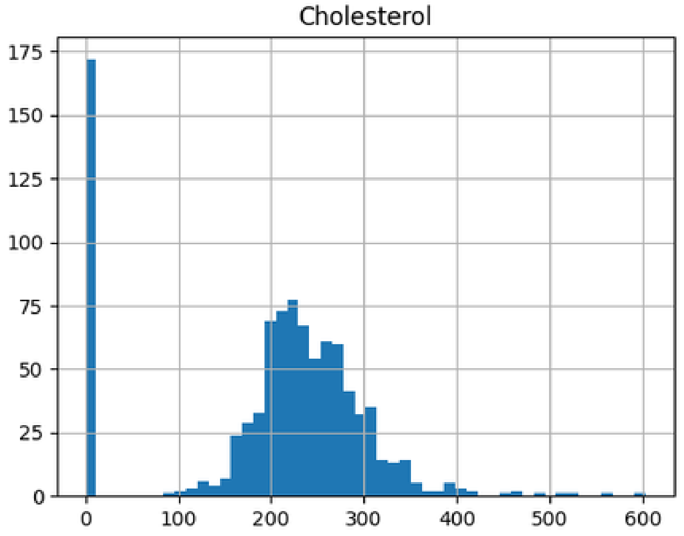
Data understanding



Distribution of **Age** feature

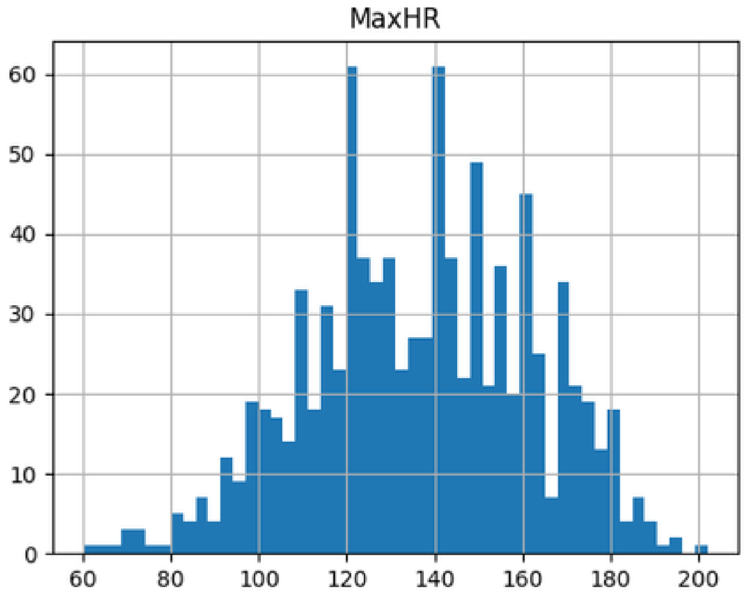
Distribution of **ExerciseAngina** feature





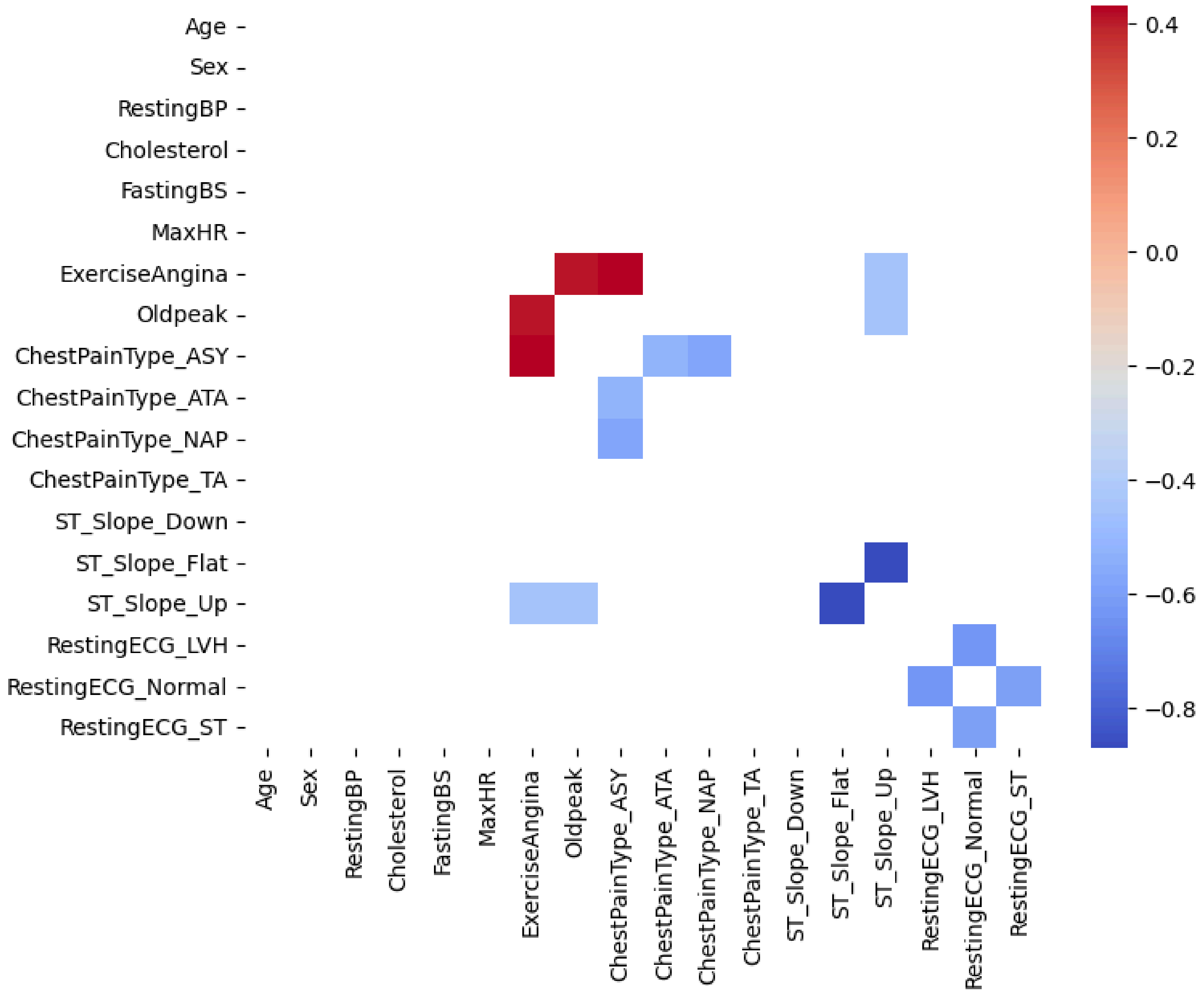
Distribution of **Cholesterol** feature

Distribution of **MaxHR** feature



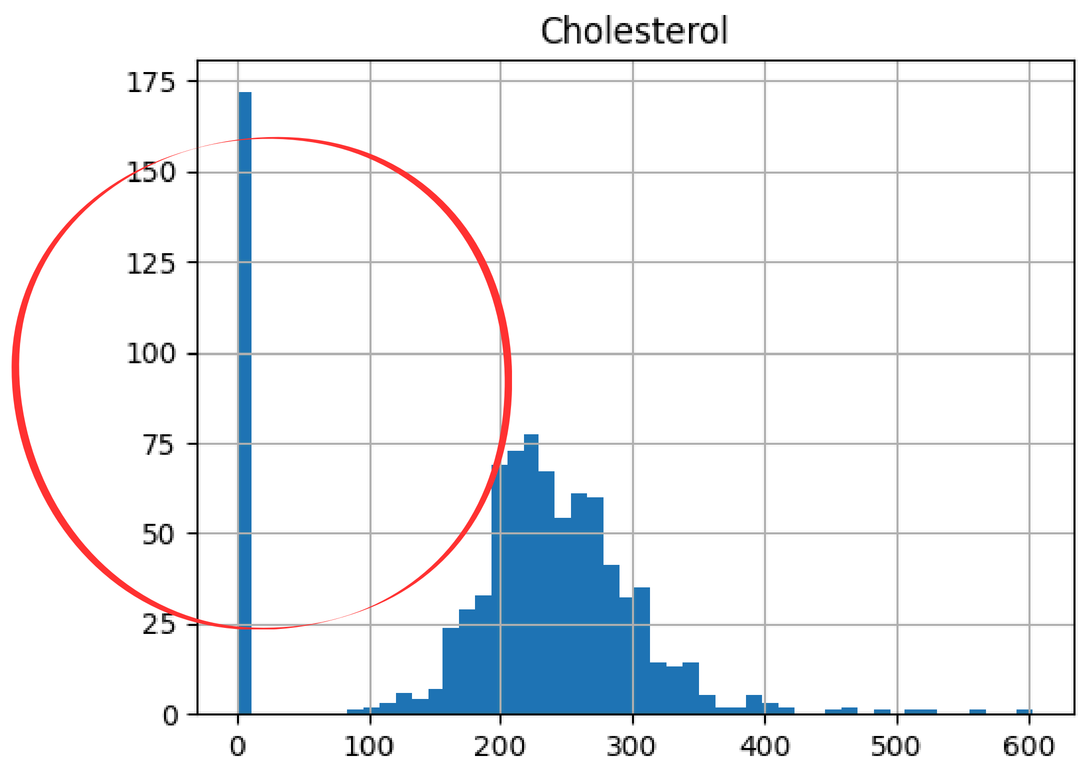
Data understanding

The **correlation** between features is genrally low, therefore should can not drop any of them.
The plot shows only the values with a correlation **>4.**



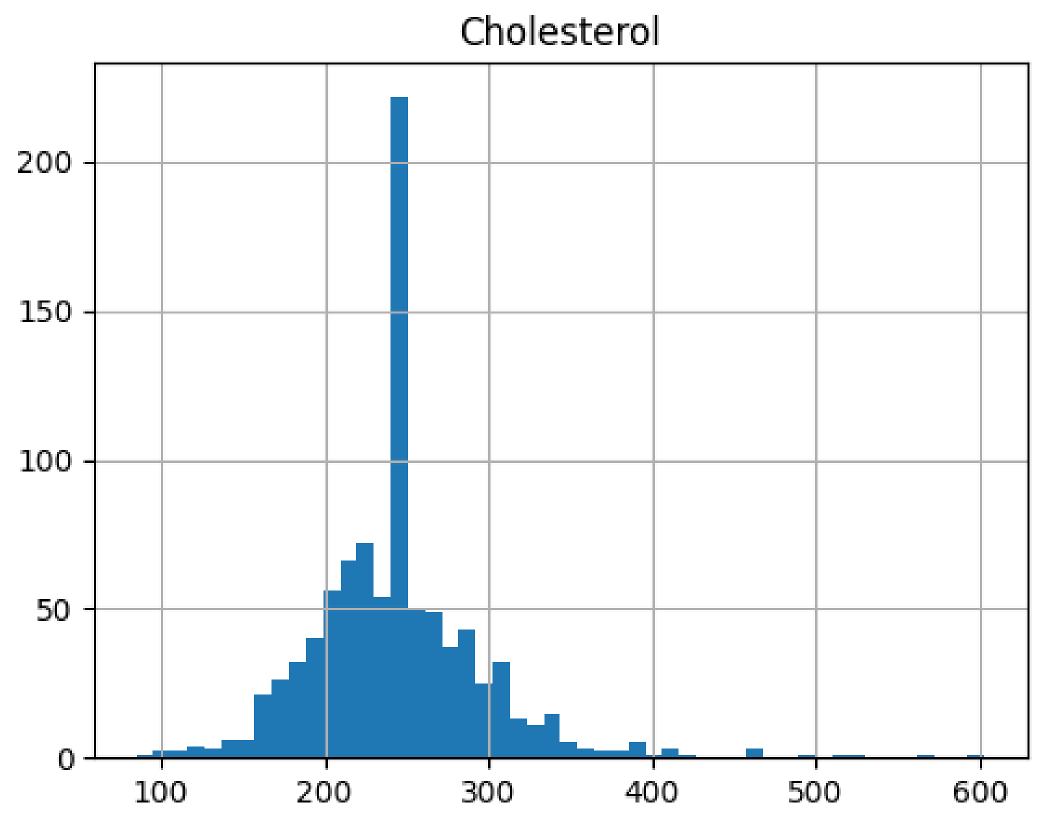
Data preparation

Data preparation



There are around 170 measurements which represent a level of **cholesterol = 0**, which is not possible.

The solution is to replace these values with the **mean value** of cholesterol.



Data preparation

Encoding values: the main encoding technique used was **one-hot encoding,** for the categorical features: **ChestPainType**, **RestingECG** and **ST_Slope**

Data preparation

ST_Slope_Down	ST_Slope_Flat	ST_Slope_Up
FALSE	FALSE	TRUE

One-hot encoding for **ST_Slope**

ChestPainType_ASY	ChestPainType_ATA	ChestPainType_NAP	ChestPainType_TA
FALSE	FALSE	TRUE	FALSE

One-hot encoding for **ChestPainType**

RestingECG_LVH	RestingECG_Normal	RestingECG_ST
FALSE	FALSE	TRUE

One-hot encoding for **RestingECG**

Data preparation

Additional encoding has been carried out for the features **Sex** and **ExerciseAngina**, in order to turn them from categorical to numerical. In this case, **ordinal encoding** has been used.

ExerciseAngina		ExerciseAngina
Y	→	1
N		0

Sex		Sex
M	→	1
F		0

Modeling

Modeling

Three different models have been used:

- Logistic Regression
- Random Forest
- K-Nearest Neighbours

Modeling

Logistic Regression

```
from sklearn.linear_model import LogisticRegression

logisticRegr = LogisticRegression(random_state=seed)
logisticRegr.fit(X_train, y_train)
y_pred = logisticRegr.predict(X_test)
metrics.accuracy_score(y_test, y_pred)
```

```
0.8586956521739131
```

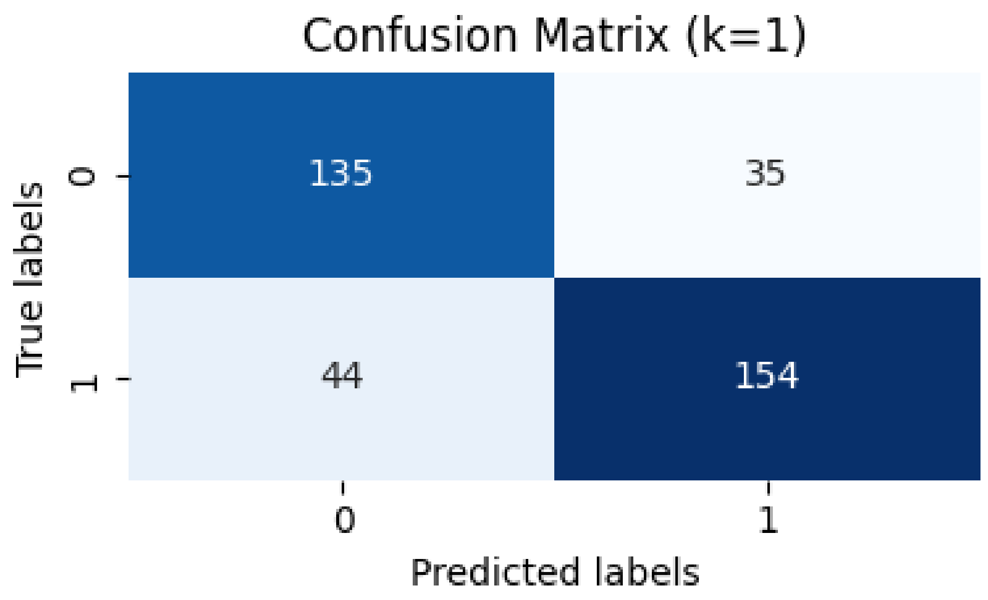
Logistic Regression should work well with predicting catergorical variables, like in this case (HeartDisease Y/N). The accuracy of this model was just below **86%**.

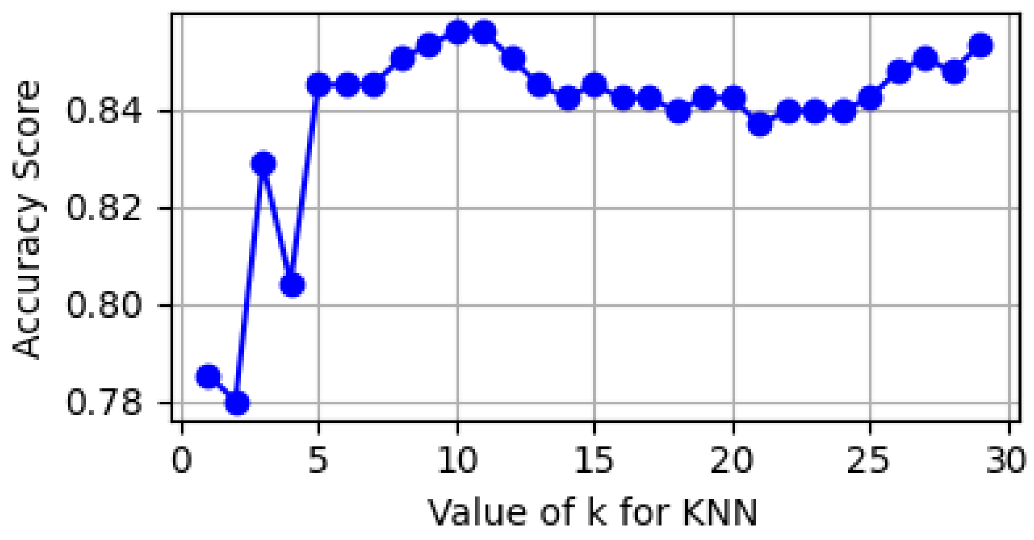
Modeling

K-Nearest Neighbours

Also in this case, the maximum accuracy achieved was just below **86%.**

With the results cealing at a value of **k = 10.**





Modeling

Random Forest

The results achieved with the random forest model were sligly better, at **87,5%.**
The most relevant features for the model are in line with what the research on heart disease tells us.

```
Accuracy: 0.875

Features sorted by descending importance:
ST_Slope_Up           0.153458
Cholesterol           0.140666
Oldpeak               0.125167
Age                   0.117781
ChestPainType_ASY     0.114956
ST_Slope_Flat         0.083500
ExerciseAngina        0.064620
Sex                   0.042073
FastingBS             0.039789
ChestPainType_ATA     0.029083
RestingECG_LVH        0.021064
ChestPainType_NAP     0.017649
RestingECG_ST         0.017454
RestingECG_Normal     0.016690
ST_Slope_Down         0.008133
ChestPainType_TA      0.007916
```

Conclusions