IBM Coursera Advance Data Science Capstone Project

Project: Heart Disease Prediction

Nicola Tombolan

Data set

Kaggel dataset : Heart disease UCI
https://www.kaggle.com/ronitf/heart-disease-uci

303 records with 13 attribute and target field refers to the presence of heart disease in the patient

Use case

The Heart disease UCI dataset contains 14 variables and 303 records along with a target field of having or not having heart disease.

The scope of this project is to find any correlation between the data in order to understand and calculate the event of an heart disease in the patient. The data will be used in different ML supervised and unsupervised models.

Solution

Different Machine learning models will be trained and tested to predict the event of an heart disease in the patient

A Jupyter Notebook in the IBM Watson Studio will be implemted to get the best model and data understanding

Data set

Dataset information

age: The person's age in years
sex: The person's sex (1 = male, 0 = female)
cp: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
chol: The person's cholesterol measurement in mg/dl
fbs: The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
thalach: The person's maximum heart rate achieved
exang: Exercise induced angina (1 = yes; 0 = no)
oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)
slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
ca: The number of major vessels (0-3)
thal: Thallium Stress Test (3 = normal; 6 = fixed defect; 7 = reversable defect)
target: Heart disease (0 = no, 1 = yes)

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

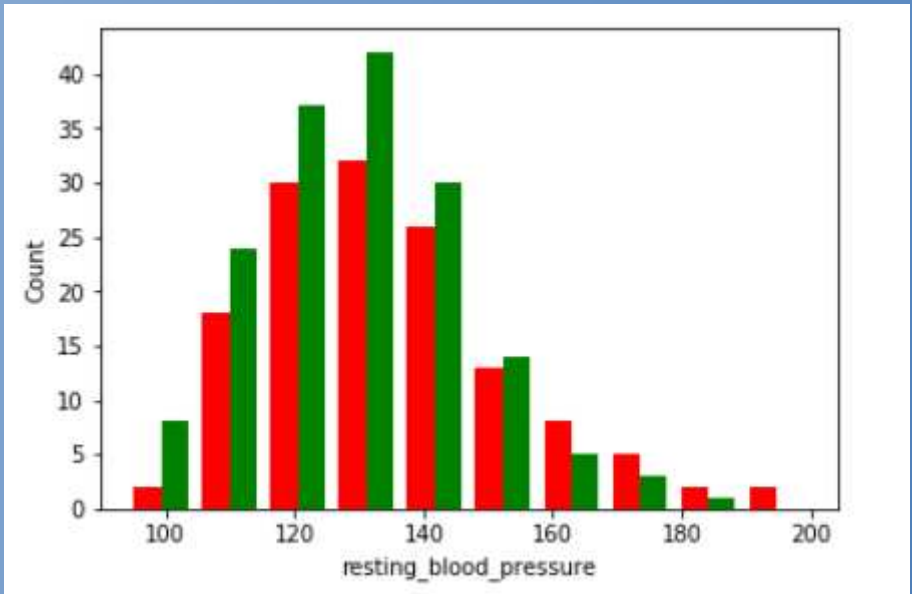
Data Quality Assessment

- The database doesn't contain null values
- All categorical information are indexed from the beginning (see dataset information)
- All records have the same format.
- Duplicate records have been removed

Data Analysis

Data visualization :

In the original dataset the categorical data have been indexed into numbers, this data have been then converted into categories to ease the data understanding



Data Analysis

Statistical information

	age	sex	chest_pain_type	resting_blood_pressure	cholesterol	fasting_blood_sugar	rest_ecg	max_heart_rate_achieved
count	302.00000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000	302.000000
mean	54.42053	0.682119	0.963576	131.602649	246.500000	0.149007	0.526490	149.569536
std	9.04797	0.466426	1.032044	17.563394	51.753489	0.356686	0.526027	22.903527
min	29.00000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000
25%	48.00000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.250000
50%	55.50000	1.000000	1.000000	130.000000	240.500000	0.000000	1.000000	152.500000
75%	61.00000	1.000000	2.000000	140.000000	274.750000	0.000000	1.000000	166.000000
max	77.00000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000

Correlation matrix

	age	sex	chest_pain_type	resting_blood_pressure	cholesterol	fasting_blood_sugar	rest_ecg	max_heart_rate_ach
age	1	-0.09	-0.06	0.28	0.21	0.12	-0.11	-0.4
sex	-0.09	1	-0.05	-0.06	-0.2	0.05	-0.06	-0.05
chest_pain_type	-0.06	-0.05	1	0.05	-0.07	0.1	0.04	0.29
resting_blood_pressure	0.28	-0.06	0.05	1	0.13	0.18	-0.12	-0.05
cholesterol	0.21	-0.2	-0.07	0.13	1	0.01	-0.15	-0.01
fasting_blood_sugar	0.12	0.05	0.1	0.18	0.01	1	-0.08	-0.01
rest_ecg	-0.11	-0.06	0.04	-0.12	-0.15	-0.08	1	0.04
max_heart_rate_achieved	-0.4	-0.05	0.29	-0.05	-0.01	-0.01	0.04	1
exercise_induced_angina	0.09	0.14	-0.39	0.07	0.06	0.02	-0.07	-0.38
st_depression	0.21	0.1	-0.15	0.19	0.05	0	-0.06	-0.34
st_slope	-0.16	-0.03	0.12	-0.12	0	-0.06	0.09	0.38
num_major_vessels	0.3	0.11	-0.2	0.1	0.09	0.14	-0.08	-0.23
thallium	0.07	0.21	-0.16	0.06	0.1	-0.03	-0.01	-0.09
target	-0.22	-0.28	0.43	-0.15	-0.08	-0.03	0.13	0.42

ML Algorithms

Binary classification model to predict the event of an heart disease.

Models :

Supervised:

- Linear model
  - Logistic Regression
- Ensabled Models
  - Decision Tree
  - Random Forest
  - Gradient Boosted Trees
- Deep learning
  - Feed Forward Neural network

Unsupervised:

- K-means model

Technology

Spark Mlib for ML algorithms

Keras for neural network

Feautures

- Clean data set (no null or missing values)
- Data indexing, vectorization, normalization
- New feature: age field aggregated per decade
- Split data set for training and test

Model evaluation

- Target class is well balanced

```
Total records number  = 302
Heart Disease class   = 54 %
No heart Disease class  = 45 %
```

- Evaluation metric : accuracy and f1 score used to evaluate perfomance
- review evaluation metrics of train/test based on model paramete
- check target class balancing for train and test dataset
- model overfitting were considered to evaluate the best performance

Model performance

LogisticRegression

- Train Accuracy = 0.8380 Train Error = 0.1619
- Train f1 = 0.8366 Train Error = 0.1633
- Test Accuracy = 0.8545 Test Error = 0.1454
- Test f1 = 0.8545 Test Error = 0.1454

DecisionTreeClassifier

- Train Accuracy = 0.9554 Train Error = 0.0445
- Train f1 = 0.9554 Train Error = 0.0445
- Test Accuracy = 0.7636 Test Error = 0.2363
- Test f1 = 0.7665 Test Error = 0.2334

Model performance

RandomForestClassifier

- Train Accuracy = 0.9716 Train Error = 0.0283
  Train f1 = 0.9716 Train Error = 0.0283
- Test Accuracy = 0.8363 Test Error = 0.1636
- Test f1 = 0.8371 Test Error = 0.1628

GBTClassifier

- Train Accuracy = 0.9958 Train Error = 0.0041
- Train f1 = 0.9958 Train Error = 0.0041
- Test Accuracy = 0.8196 Test Error = 0.1803
- Test f1 = 0.8176 Test Error = 0.1823

Model performance

K-means

- Accuracy 0.7847

FF Neural Network

- Test Accuracy: 0.8999
- Test loss: 0.4385

Deployment

Use Jupyter Notebook with IBM Watson Studio: supervised and unsupervised algorithms implemented and evaluated

Conclusion

Best performance with:

RandomForestClassifier and GBT Classifier perfomance changes with different train and test samples.

A bigger and balanced dataset could help to get better performances.