Report

L. Lavagna, E. Loru, A. Sgrigna

9th February 2022

Introduction

Heart Diseases and Cardiovascular Conditions are one of the leading causes of death in many countries of the world (cfr. [6]) but many of these conditions can be improved or even prevented by putting in place some changes in our daily lives (cfr. [7], [8]). In this context having early indication of possible future Cardiovascular Problems plays a crucial role, particularly when medical data are largely available.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Figure 1: Some features from the dataset against the actual diagnosis

Our aim is to use several indicators to predict whether a given patient is more likely to develop some heart disease in the future, making use of several techniques – such as feature scaling and dimensionality reduction – and well known models – such as Logistic Regression and K-Nearest Neighbors. We're not strictly interested into maximising the level of accuracy of the model, but rather its sensitivity: reducing to a minimum the mislabeling of sick patients is definitely the priority in this kind of classification problems. Finally, we will test the validity of our results by cross validation.

1 Dataset

We worked with a dataset freely available on Kaggle (cfr. [5]). This dataset contains 12 features of 918 patients selected from four different institutions around the world. Each of the features represents a particular medical information, with the 12th being the most pivotal: it indicates whether a given patient has indeed developed a heart disease or not.

The dataset is comprised of both categorical and numerical features. By inspecting their distribution, interesting correlations to the diagnosis of the patient become noticeable for some of them, for instance how the vast of majority of people who experienced

exercise-induced angina turned out to be sick, or how people who are older than average tend to be diagnosed more often; it's also curious how heart diseases seem to be more commonly diagnosed in males and in asymptomatic patients (Figure 2).

The data was almost complete, with no missing values or NaNs. There was a single out-of-the-ordinary value in the RestingBP feature and we decided to substitute it with the mean of the observations. In addition, about 20% of the values in the Cholesterol column were equal to zero (suggesting a missing measurement): we decided to replace those with values sampled around the mean of the non-zero measurements.

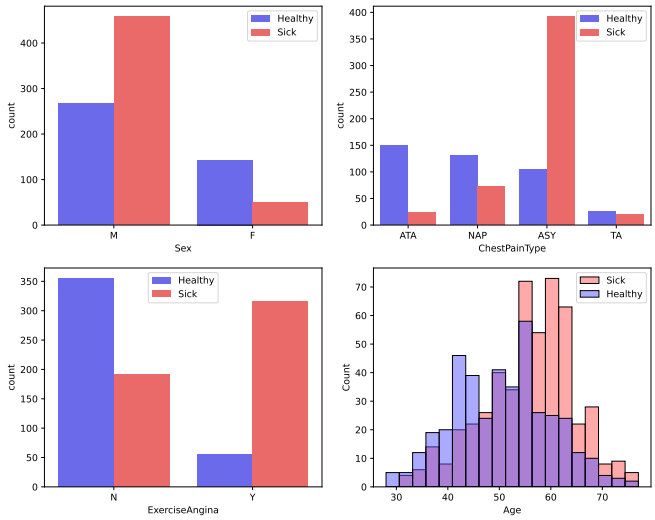

Figure 2: An extract of the dataset

2 Models and Techniques

We trained different models on the dataset and observed the results in terms of two metrics: the *accuracy* (i.e. the rate of correct predictions) and the *sensitivity* (i.e. 1 minus the rate of false negatives). In particular we tried different techniques, and the best results were given by:

- Decision Tree (DT)
- K-Nearest Neighbors (KNN)
- Logistic Regression (LR)
- Gaussian Naïve Bayes (GNB)

For each of the models we've investigated whether it was more efficient to work wit the whole dataset or with a reduced number of features (applying PCA). Finally, we checked the validity of the results with k-fold Cross Validation.

Decision Tree

At first we transformed the categorical features into dummy variables (one-hot-encoding). After that we trained the model on the 70% of the dataset and we used the remaining 30% for testing. We decided not to standardize the data because usually Decision Tree performs best in such a way. We obtained an accuracy of 80.8% and a sensitivity of 78.5%. We then tried to improve the model grouping together all symptomatic chest pain types (i.e. TA, ATA, NAP) which resulted in an astonishing accuracy of 100% and an astonishing sensitivity of 100%, namely all labels were correctly predicted. Also, k-fold cross validation gave excellent results: 99.4% of accuracy, and 99.4% of sensitivity. This astonishing result is probably due to the fact that the widest differences among healthy and sick patients can be found in categorical features.

KNN

Before applying the model we've standardized the data and we've divided it into 70% for training and 30% for testing as we did in Decision Tree. We've obtained an accuracy of 84.8% and a sensitivity of 87%. We then decided to improve the results reducing the number of features using PCA (Principal Component Analysis). In this way we've obtained 81.9% of accuracy and 84% of sensitivity. Good results, but not as precise as those given without applying PCA. Finally, we performed k-fold cross validation on the non-reduced dataset, that confirmed the results we got earlier.

Logistic Regression

We reserved to the data the same treatment as KNN: standardize, convert categorical data into dummy variables, and group the chest pain types. Then we applied the model on the whole dataset and we obtained an accuracy of 83.7% and a sensitivity of 87.7%. These results seem very comparable to KNN's. We then tried to improve them by reducing the number of features using PCA, but we obtained the same accuracy and the same sensitivity. Finally, using k-fold cross validation we saw that the results were confirmed.

Gaussian Naïve Bayes

As we did before we one-hot encoded all categorical features, grouped the chest pain types in asymptomatic and symptomatic and finally standardized all data. With this model we obtained an accuracy of 84.1% and a sensitivity of 86.5%. Unlike previous models, GNB seems to work better with the reduced dataset via PCA, in fact it gave 83.7% of accuracy and 88.3% of sensitivity. Finally, we checked the results with k-fold cross validation and we saw that the sensitivity was slightly lower but the overall accuracy remained good.

3 Results

By plotting the different ROC curves o we see that all the models performed quite well, having an AUC (area under curve) greater than 0.9. It is astonishing how well Decision

Tree predicted the labels: its curve is basically a step function. Among the other models, KNN is visibly better than the other two, which on the other hand are very comparable.

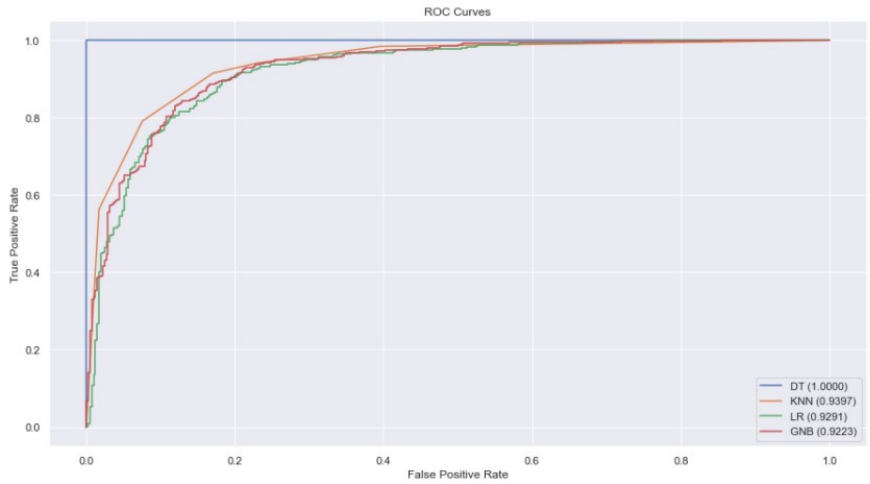

Figure 3: ROC curve comparison

We summarize in a table the results we obtained previously by applying the various models to our dataset and the k-cross validation.

Model	**Accuracy**	**Sensitivity**	**Area ROC Curve**
Decision Tree	99%	99%	1.00
K-Nearest Neighbors	84%	87%	0,94
Logistic Regression	84%	86%	0.93
Gaussian Naïve Bayes	85%	85%	0.92

References

[**1**] Frank L.J. Visseren, et al: *2021 ESC Guidelines on cardiovascular disease prevention in clinical practice*. European Heart Journal (2021) 42, 32273337. Available also online.

[**2**] Institute of Medicine (US) Committee on Preventing the Global Epidemic of Cardiovascular Disease: *Meeting the Challenges in Developing Countries* (Fuster V, Kelly BB, editors). Available also online www.ncbi.nlm.nih.gov/books/NBK45693/pdf/Bookshelf_NBK45693.pdf.

[**3**] S. M. Ross: *Probability and statistics for engineers and scientists*. Academic Press 2014.

[**4**] Christopher M. Bishop: *Pattern Recognition and Machine Learning*. Springer 2006.

[5] www.kaggle.com/fedesoriano/heart-failure-prediction
[6] www.cdc.gov/heartdisease/facts.htm
[7] www.cdc.gov/heartdisease/about.htm
[8] www.medicinenet.com/heart_disease_pictures_slideshow_visual_guide/article.htm
[9] github.com/g-shreekant/Heart-Disease-Prediction-using-Machine-Learning