Heart Disease Prediction Project

Kevin Rouille

MSCI 446

December 9th, 2020

Outline

Abstract 2

Introduction 3

Related Work 3

Data 3

Data Collection 3

Data Preprocessing 3

Data Description 4

Data Summary 6

Results 8

Model Evaluation Criteria 8

Supervised Learning Tasks 8

Decision Tree 9

Naïve Bayes Classifier 10

K-Nearest Neighbours Algorithm 11

Unsupervised Learning Task 12

K-Means Clustering Algorithm 12

Conclusions 14

References 15

Appendix 15

Abstract

In this project, I analyzed data from UCI's Machine Learning repository and then tested several supervised learning models that could help cardiologists to identify patients at risk of developing heart diseases. It seems that the Naïve Bayes classifier is the most performant in predicting occurrences of heart disease.

Introduction

Heart diseases are the 2nd leading cause of death in Canada. If these diseases are so deadly, it is mainly because they are difficult to predict and have many causes. The modern way of life (junk food, little physical activity, etc.) does not help to reduce the risk of cardiac arrests or heart disease. For this project, I analyzed data from UCI's Machine Learning repository and then tested several supervised learning models to determine which was the best at predicting occurrences of heart disease. Such classifiers could help cardiologists and other physicians to identify patients at risk of developing heart diseases.

Related Work

There has been extensive research in the prediction of heart disease using machine learning techniques, including classification, by the medical world and data science sector for many years. This project aims at discovering how algorithms like the decision tree, naïve Bayes classifier, k-nearest neighbours and k-means clustering can be applied to medical data and transform raw data into valuable information.

Data

Data Collection

The database from UCI's Machine Learning repository is composed of four datasets from medical institutes in different countries:

1. V.A. Medical Center, Long Beach, California, USA

2. Cleveland Clinic Foundation, Cleveland, Ohio, USA

3. Hungarian Institute of Cardiology, Budapest, Hungary

4. University Hospital, Zurich, Switzerland.

Data Preprocessing

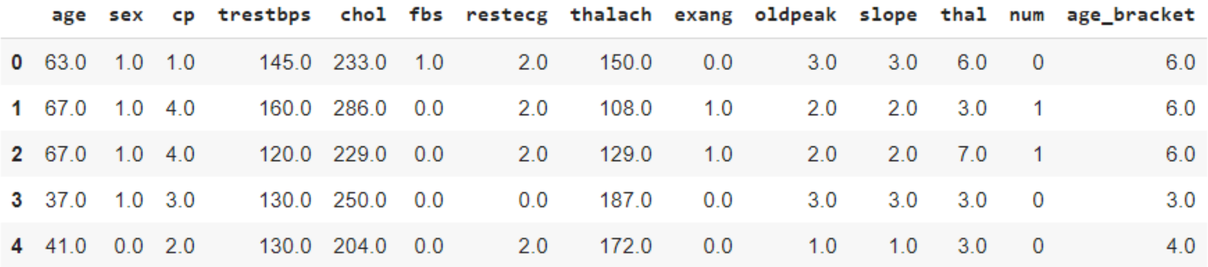
The datasets contain 200, 303, 294 and 123 lines respectively, with each line corresponding to one patient. Each dataset has the same 14 columns, all of which contain real number values.

First, I gathered the data of the four datasets to form a single one of 920 rows. I then deleted rows with out-of-range values for some features, as well as the 'ca' feature for which two thirds of values were missing.

I added an 'age_bracket' feature that assigns to each patient's age its corresponding 10-year bracket (i.e. 5.0 for a 54-year old). I used it to replace missing values for features 'exang', 'fbs', 'oldpeak', 'slope' and 'thal' with the median value for the patient's corresponding age bracket, as these features are highly correlated with age and have a high standard deviation. As for missing values in features 'chol' and 'trestbps', I replaced them with the average value for the patient's age bracket, since these features are also highly correlated with age but have a low standard deviation. Missing 'thalach' values were replaced with (220 – age) for men and (226 – age) for women.

Data Description

The final dataset contains 913 rows. Here are its first 5 rows:



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	thal	num	age_bracket
0	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	3.0	3.0	6.0	0	6.0
1	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	2.0	2.0	3.0	1	6.0
2	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.0	2.0	7.0	1	6.0
3	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.0	3.0	3.0	0	3.0
4	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.0	1.0	3.0	0	4.0

The description of the 14 columns is as follows:

1. age: age of the patient

2. sex:

- 0 indicates a woman

- 1 indicates a man

3. cp (chest pain): type of chest pain experienced by the patient

- 1 indicates typical angina

- 2 indicates atypical angina

- 3 indicates a different pain than angina

- 4 indicates no pain (asymptomatic).

Angina is a chest pain caused by a lack of oxygen in the blood in the heart.

4. trestbps: value of an individual's resting blood pressure in mmHg. Over time, high blood pressure can damage the arteries that supply the heart. Abnormally high blood pressure combined with above average cholesterol levels seriously increase the risk of heart attack and heart disease.

5. chol: serum cholesterol level in mg/dl. There are however two main types of cholesterol (LDL and HDL). A high level of the former is detrimental for the heart, while the latter is beneficial.

6. fbs (fasting blood sugar): fasting blood sugar indicator

- 0 indicates a glycemia below 120 mg/dl

- 1 indicates a glycemia higher than 120 mg/dl.

Not producing enough of the hormone secreted by the pancreas (insulin) or not reacting properly to insulin leads to an increase in blood sugar, which increases the risk of heart attack. The value of 120 mg/dl is actually a threshold for defining the different types of diabetes.

7. restecg (resting electrocardiogram): result of the electrocardigram of the patient when being inactive.

- 0 indicates a normal result

- 1 indicates an abnormality in the ST wave of the electrocardiogram, i.e. the relaxation of the atrium of the heart. An abnormality in this relaxation can be a cause of cardio-vascular accident.

- 2 indicates a hypertrophy of the left ventricle of the heart. A left ventricular hypertrophy indicates a cardiac action characterized by an increase in the mass of the left ventricular muscle, which can also be a cause of a cardiovascular accident.

8. thalach: the maximum heart rate reached by the patient on the electrocardiogram. The increase in cardiovascular risk is related to the acceleration of the heart rate. Furthermore, it has been shown that an increase in heart rate of 10 beats per minute indicates a 20% increase in the risk of cardiac arrest, as does an increase in blood pressure of 10 mmHg.

9. exang (exercise-induced angina): indicates if physical exercise causes angina to the patient, i.e. a lack of oxygenation of the heart.

- 1 : yes

- 0 : no

10. oldpeak: value indicating a comparison of the ST segment of the patient's electrocardiogram during exercise with the segment when being inactive.

11. slope: the ST segment's direction during the physical exercise requiring the most effort from the patient.

- 1 indicates an upward slope

- 2 corresponds to a constant slope

- 3 indicates a descending slope

In general, a horizontal or descending ST-segment depression at a lower workload or heart rate indicates a higher likelihood of heart disease.

12. ca: number of major blood vessels detected by radioscopy (a radiology procedure that consists in acquiring instantaneous dynamic images of the interior of cardiac structures).

13. thal: indicator of thalassemia (qualitative or quantitative hereditary anomalies of the hemoglobin in red blood cells)

- 3: no anomaly

- 6: permanent anomaly

- 7: reversible anomaly

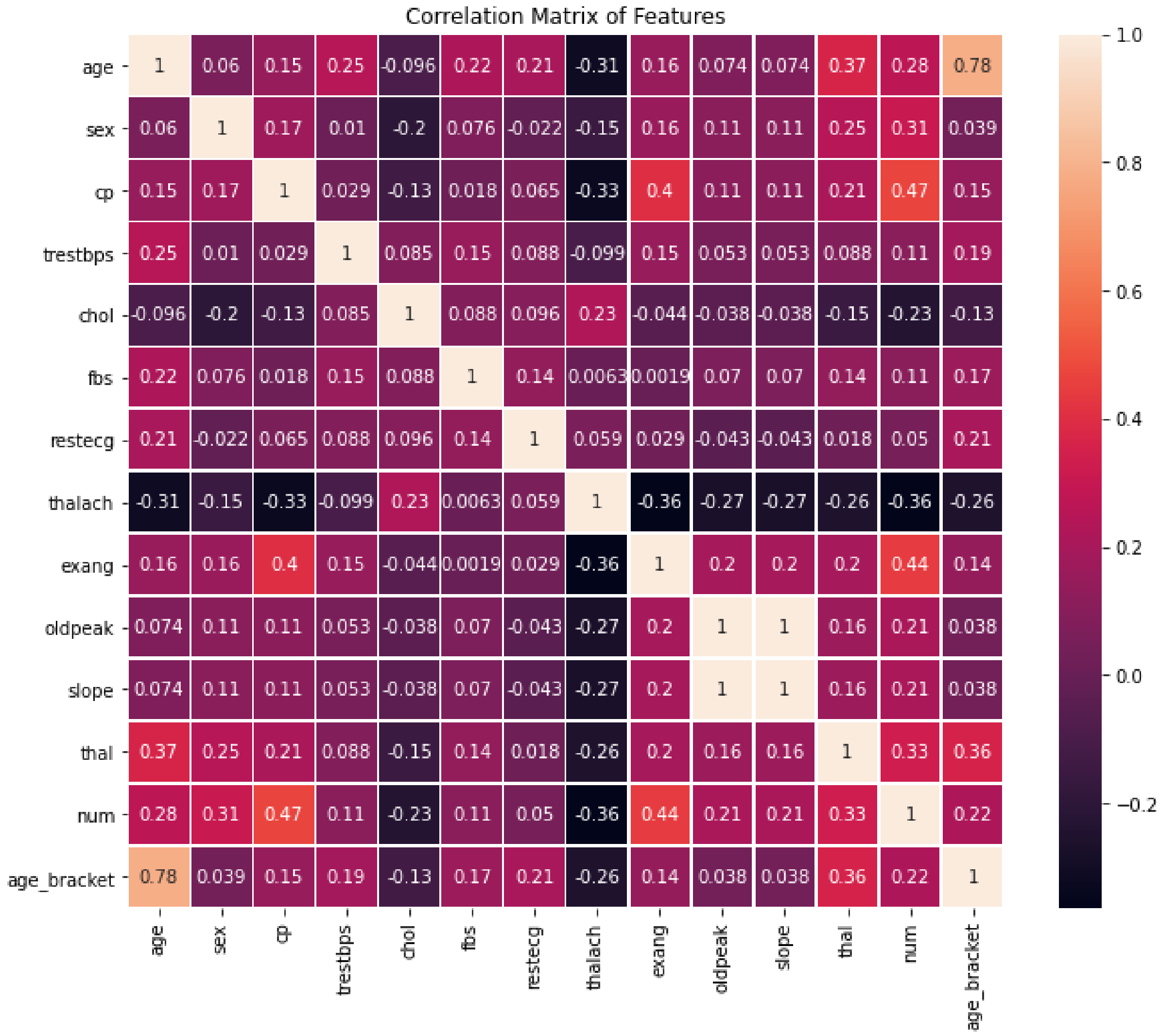
14. num: diagnosis of heart disease. I decided to restrict myself to supervised binary learning (two classes: positive and negative) because I consider that I do not have enough data for to have predictive tools powerful enough for four classes.

- 0: negative

- 1, 2, 3, 4: positive.

Data Summary

To determine which features have the largest impact on the outcome of the diagnosis of a heart disease, we plot a correlation matrix. It shows the correlation coefficient for each pair of features.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	thal	num	age_bracket
age	1	0.06	0.15	0.25	-0.096	0.22	0.21	-0.31	0.16	0.074	0.074	0.37	0.28	0.78
sex	0.06	1	0.17	0.01	-0.2	0.076	-0.022	-0.15	0.16	0.11	0.11	0.25	0.31	0.039
cp	0.15	0.17	1	0.029	-0.13	0.018	0.065	-0.33	0.4	0.11	0.11	0.21	0.47	0.15
trestbps	0.25	0.01	0.029	1	0.085	0.15	0.088	-0.099	0.15	0.053	0.053	0.088	0.11	0.19
chol	-0.096	-0.2	-0.13	0.085	1	0.088	0.096	0.23	-0.044	-0.038	-0.038	-0.15	-0.23	-0.13
fbs	0.22	0.076	0.018	0.15	0.088	1	0.14	0.0063	0.0019	0.07	0.07	0.14	0.11	0.17
restecg	0.21	-0.022	0.065	0.088	0.096	0.14	1	0.059	0.029	-0.043	-0.043	0.018	0.05	0.21
thalach	-0.31	-0.15	-0.33	-0.099	0.23	0.0063	0.059	1	-0.36	-0.27	-0.27	-0.26	-0.36	-0.26
exang	0.16	0.16	0.4	0.15	-0.044	0.0019	0.029	-0.36	1	0.2	0.2	0.2	0.44	0.14
oldpeak	0.074	0.11	0.11	0.053	-0.038	0.07	-0.043	-0.27	0.2	1	1	0.16	0.21	0.038
slope	0.074	0.11	0.11	0.053	-0.038	0.07	-0.043	-0.27	0.2	1	1	0.16	0.21	0.038
thal	0.37	0.25	0.21	0.088	-0.15	0.14	0.018	-0.26	0.2	0.16	0.16	1	0.33	0.36
num	0.28	0.31	0.47	0.11	-0.23	0.11	0.05	-0.36	0.44	0.21	0.21	0.33	1	0.22
age_bracket	0.78	0.039	0.15	0.19	-0.13	0.17	0.21	-0.26	0.14	0.038	0.038	0.36	0.22	1

The highest correlation coefficients in absolute value are for the correlations between ‘num’ and:

- Chest pain (r = 0.47)

- Exercise-induced angina (r = 0.44)

- Maximum heart rate (r = -0.36)

- Thalassemia (r = 0.33)

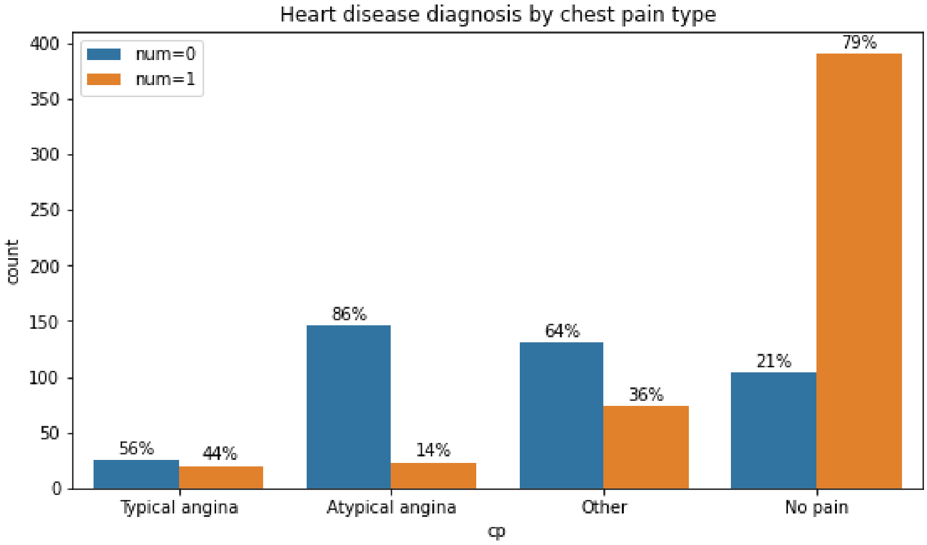
- Sex (r = 0.31)

- Age (r = 0.28)

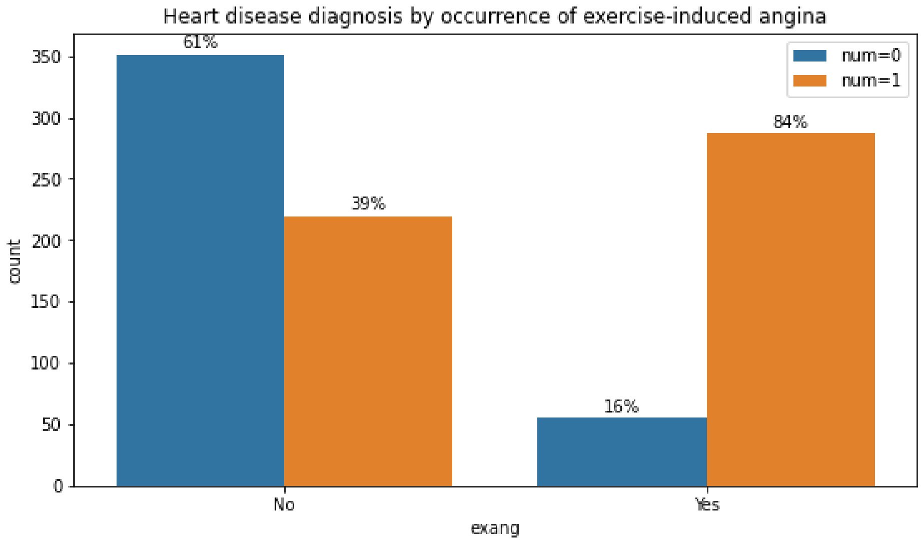
- Cholesterol level (r = 0.23).

We can note that age itself is less correlated with heart disease diagnosis than one would have expected.

With this new information, it seems useful to plot the repartition of positive and negative diagnoses by chest pain type and exercise-induced angina.



The presence of chest pain does not appear to indicate the occurrence of a heart disease, given that 79% of patients who have no chest pain have nevertheless been diagnosed with heart disease.



On the other hand, 84% of patients with exercise-induced angina have a heart disease. Therefore, it seems to be a strong indicator of heart disease.

Results

Model Evaluation Criteria

As we are in the case of binary supervised learning, we use the following statistical measures to evaluate our model:

- Accuracy: it is the proportion of correct predictions among the total number of patients examined
- Sensitivity (True Positive Rate): it is the physician's ability to detect all patients who have a heart disease, i.e. to have the least number of false negatives
- Precision (Positive Predictive Value): it is the probability that a positive diagnosis (i.e. the patient has a heart disease) is correct
- Specificity (True Negative Rate): it is a physician's ability to detect only those who have a heart disease, i.e. to have the fewest false positives
- F-Score: it is the harmonic mean of precision and sensitivity.

Supervised Learning Tasks

Before training different classifiers on the data, we need to normalize it. Indeed, normalizing a dataset is a common requirement for many machine learning estimators, as they could behave poorly if the features are not somewhat standardized and normally distributed.

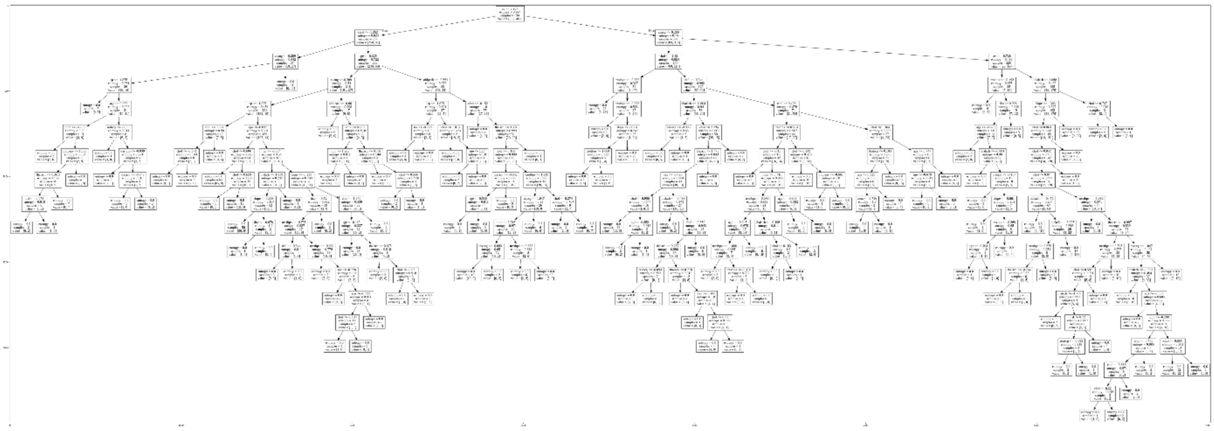
In addition, we need to split the data into two sets. The first set is used to train a model and the second to test the model. I decided to use 80% of the data for training and the other 20% for testing. Finally, before the split, the data is randomly mixed to ensure a balanced distribution of the data in both datasets.

For each supervised learning task, we compute the confusion matrix for both the training dataset and the testing dataset. It contains the numbers of patients grouped following the consistency between their actual heart disease diagnosis and their predicted diagnosis, as follows:

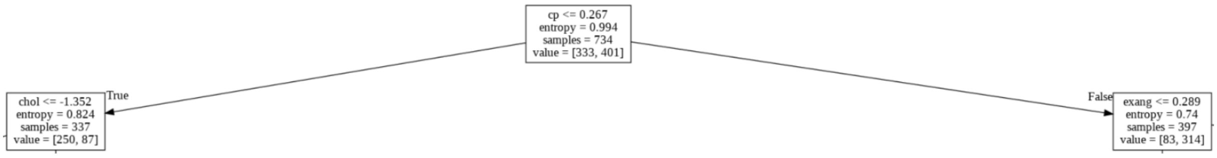
		Predicted	
		Negative	Positive
Actual	Negative	True Negatives	False Positives
	Positive	False Negatives	True Positives

Decision Tree

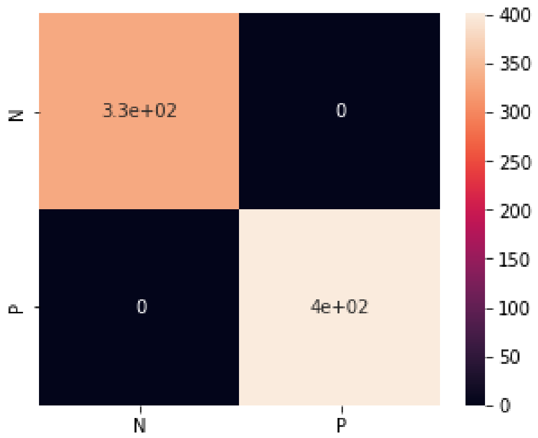
A decision tree classifier is trained on 80% of the data and tested on the remaining 20%. The resulting model is the following:



If we focus on the first two levels, we can see that chest pain is the feature with the lowest reduction in entropy, followed by cholesterol level and exercise-induced angina. This is in line with the findings from the correlation matrix.

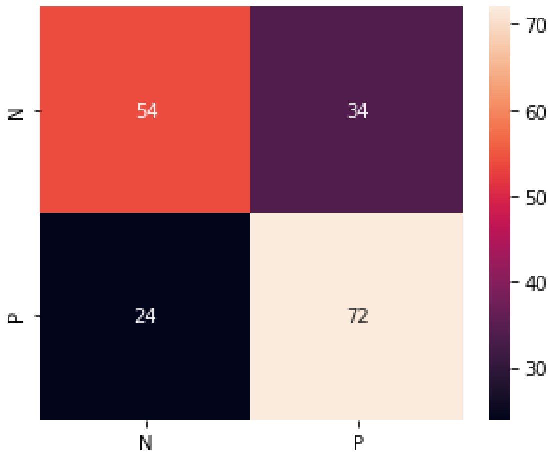


The confusion matrix for the predictions on the training dataset is:

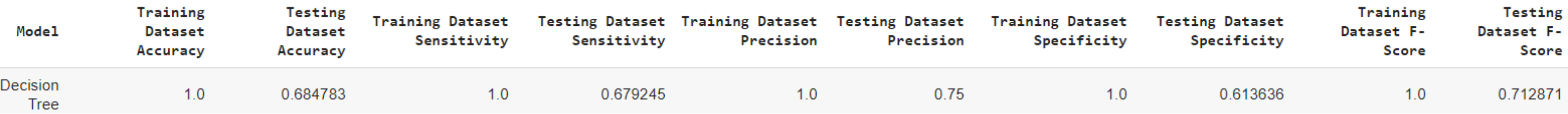


330 negative patients were correctly identified, 400 positive patients were correctly identified, no negative patients were identified as positive and no positive patients were identified as negative.

The confusion matrix for the predictions on the testing dataset is:



The model's statistical measures are summarized hereafter:



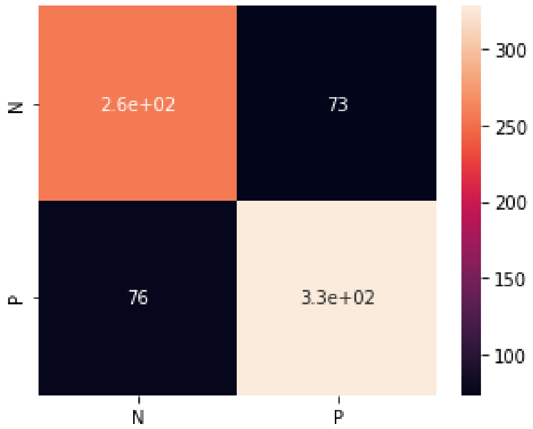
Model	Training Dataset Accuracy	Testing Dataset Accuracy	Training Dataset Sensitivity	Testing Dataset Sensitivity	Training Dataset Precision	Testing Dataset Precision	Training Dataset Specificity	Testing Dataset Specificity	Training Dataset F-Score	Testing Dataset F-Score
Decision Tree	1.0	0.684783	1.0	0.679245	1.0	0.75	1.0	0.613636	1.0	0.712871

The decision tree classifier shows perfect scores on the training dataset but significantly lower measurements for the testing set, which means that it is clearly overfitted for the training dataset. Therefore, so it is unlikely to chosen as the best model to determine heart disease occurrence.

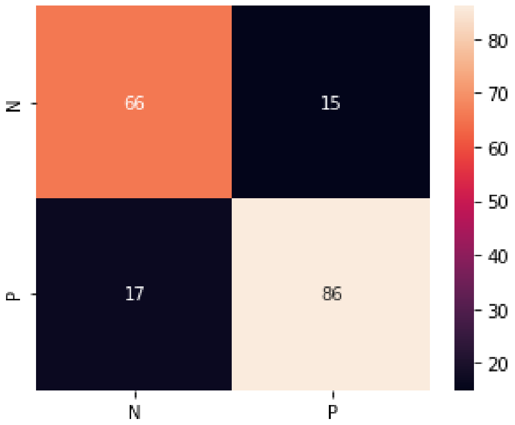
Naïve Bayes Classifier

We now train a Naïve Bayes classifier on a similar 80/20 split of the normalized data.

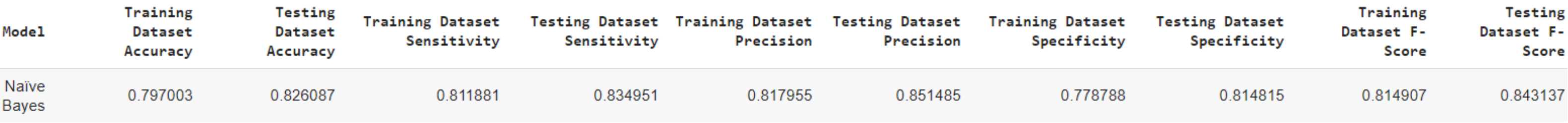
The confusion matrix for the predictions on the training dataset is:



The confusion matrix for the predictions on the testing dataset is:



The model's statistical measures are summarized hereafter:



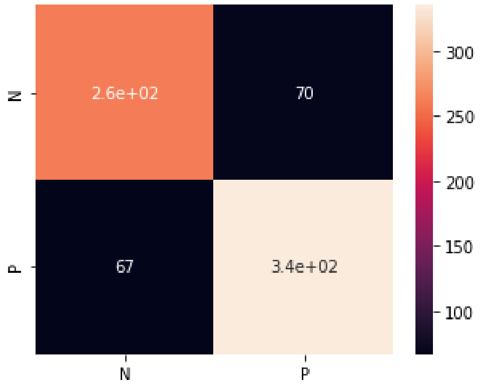
Model	Training Dataset Accuracy	Testing Dataset Accuracy	Training Dataset Sensitivity	Testing Dataset Sensitivity	Training Dataset Precision	Testing Dataset Precision	Training Dataset Specificity	Testing Dataset Specificity	Training Dataset F-Score	Testing Dataset F-Score
Naïve Bayes	0.797003	0.826087	0.811881	0.834951	0.817955	0.851485	0.778788	0.814815	0.814907	0.843137

The Naïve Bayes classifier gets high values for all measures for both the training dataset and the testing set, and the scores for the testing are consistently slightly above those of the training. Therefore, it is a strong candidate for the most performant classifier for heart disease occurrence.

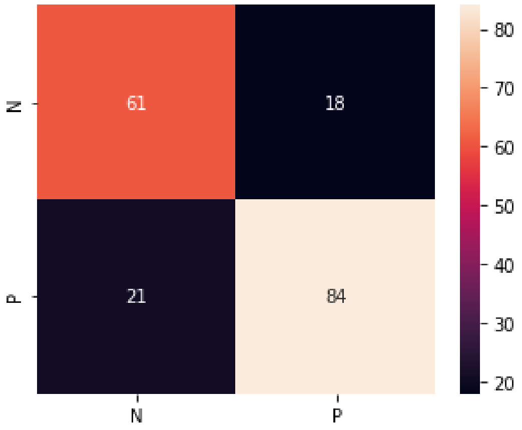
K-Nearest Neighbours Algorithm

In order to optimize the performance of the k-nearest neighbours classifier, we need to determine the number of neighbours for which the classification has the maximum accuracy. To avoid having to test each value manually, we use the GridSearchCV function from the Scikit-learn library on the k-NN classifier to test all values of k from 1 to 100. It outputs an optimal value of k = 33. Therefore, we train the classifier with a parameter of 33 for nearest neighbours and run the predictions.

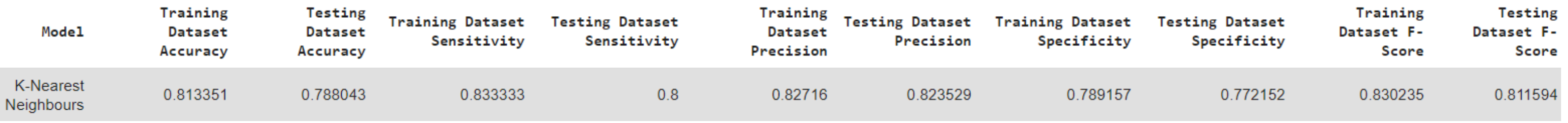
The confusion matrix for the predictions on the training dataset is:



The confusion matrix for the predictions on the testing dataset is:



The model's statistical measures are summarized hereafter:



Model	Training Dataset Accuracy	Testing Dataset Accuracy	Training Dataset Sensitivity	Testing Dataset Sensitivity	Training Dataset Precision	Testing Dataset Precision	Training Dataset Specificity	Testing Dataset Specificity	Training Dataset F-Score	Testing Dataset F-Score
K-Nearest Neighbours	0.813351	0.788043	0.833333	0.8	0.82716	0.823529	0.789157	0.772152	0.830235	0.811594

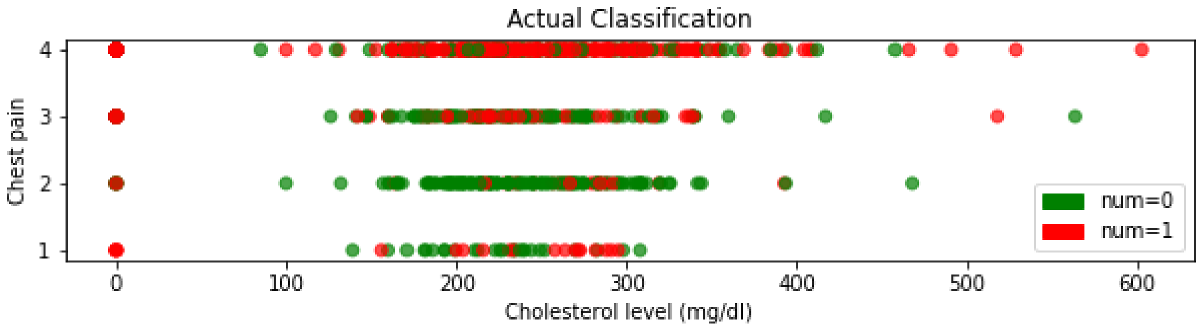
The k-nearest neighbours classifier also shows high scores across the board, with slightly better results on the training dataset than for the testing set. Its performance is almost as satisfactory as the one of the Naïve Bayes classifier.

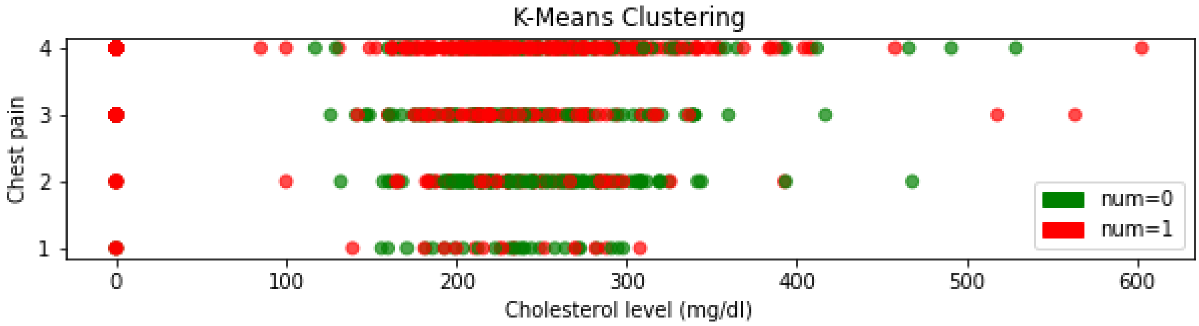
Unsupervised Learning Task

Even though our data is already labeled with the class variable 'num', we are going to check if an unsupervised learning algorithm like the k-means clustering algorithm can group the data points into two clusters that correspond to patients diagnosed with heart disease (num = 0) and healthy patients (num = 1) without having access to the class variable.

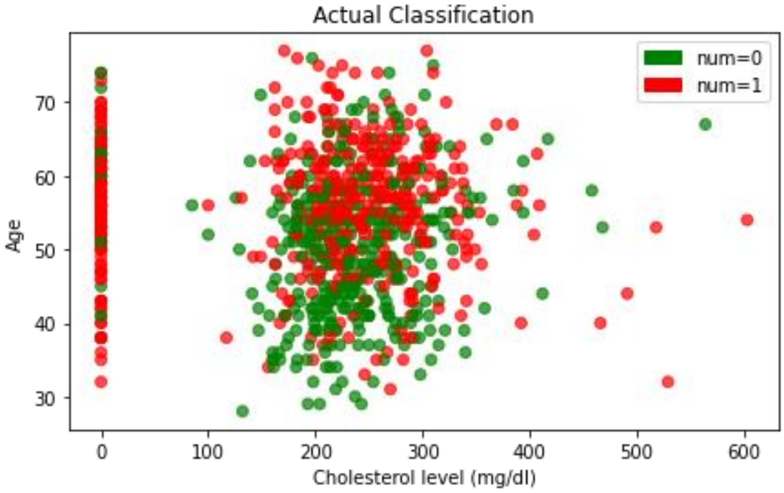
K-Means Clustering Algorithm

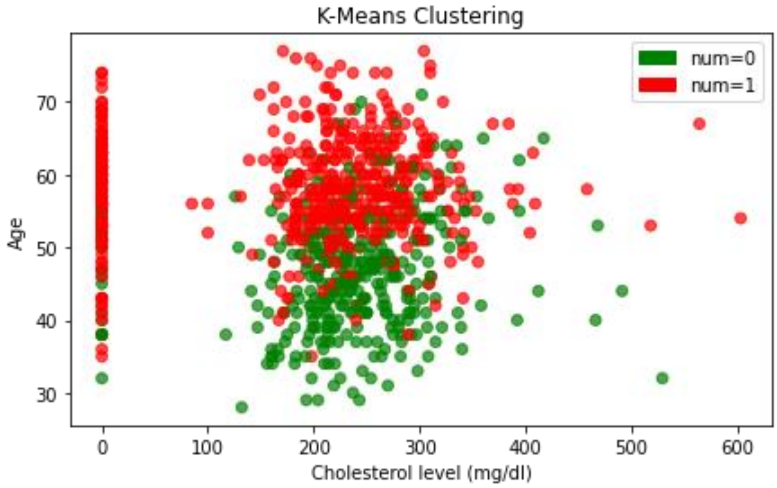
After splitting the 'num' class variable from the features and normalizing the data, we perform k-means clustering on the features and compare the clustering made by the algorithm to the actual classification for three of the most indicative features as determined by the correlation matrix: chest pain, cholesterol level and age.





The algorithm seems to have correctly distinguished ill patients (who generally have a chest pain of type 4) from the others.





The same somewhat holds true when we replace chest pain with age. However, in this case the graph for k-means clustering shows a clear-cut divide between the upper and the lower halves, whereas patients with heart diseases are more vertically and horizontally spread out in the graph of the actual distribution. The clustering algorithm therefore appears to overestimate the importance of age in outcome of the diagnosis.

Conclusions

The most important statistical measure for a physician conducting heart disease diagnoses is sensitivity, which quantifies the detection of all patients who have a heart disease and minimizes the number of false negatives. Accuracy comes in the second place, as it is the proportion of correct predictions among the total number of patients examined.

Out of the three unsupervised learning models we have trained, the Naïve Bayes classifier is the most performant. Indeed, it has both the highest sensitivity (0.83) and accuracy (0.83) for the testing dataset, and it does not overfit nor underfit.

Therefore, the Naïve Bayes model could be used by cardiologists as a tool to conduct a very first diagnosis on a new patient, and from which the physician will determine if the patient needs to undergo further and more thorough diagnosis. It is obviously very perfectible, and other machine learning techniques could prove to be more adapted for this task.

References

Amita Malav et al., Prediction Of Heart Disease Using K-Means And Artificial Neural Network As Hybrid Approach To Improve Accuracy, International Journal of Engineering and Technology (IJET), Vol 9 No 4 Aug-Sep 2017

S. Ekız and P. Erdoğmuş, "Comparative study of heart disease classification," *2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, Istanbul, 2017, pp. 1-4.

Randa El-Bialy, Mostafa A. Salamay, Omar H. Karam, M. Essam Khalifa, Feature Analysis of Coronary Artery Heart Disease Data Sets, Procedia Computer Science, Volume 65, 2015, Pages 459-468.

Fatima, M. and Pasha, M. (2017) Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications*, **9**, 1-16.

M. C. Tu, D. Shin and D. Shin, "Effective Diagnosis of Heart Disease through Bagging Approach," *2009 2nd International Conference on Biomedical Engineering and Informatics*, Tianjin, 2009, pp. 1-4, doi: 10.1109/BMEI.2009.5301650.

Jesmin Nahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen, Computational intelligence for heart disease diagnosis: A medical knowledge driven approach, Expert Systems with Applications, Volume 40, Issue 1, 2013, Pages 96-104.

Resul Das, Ibrahim Turkoglu, Abdulkadir Sengur, Effective diagnosis of heart disease through neural networks ensembles, Expert Systems with Applications, Volume 36, Issue 4, 2009, Pages 7675-7680.

Fahd Saleh Alotaibi, Implementation of Machine Learning Model to Predict Heart Failure Disease, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019

Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS international conference on computer systems and applications* (pp. 108-115). IEEE.

Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. *Procedia Technology*, *10*, 85-94.

Appendix

The code for this project can be found in a separate zip folder.