Predicting Heart Disease Diagnosis

HarvardX Data Science Professional Certificate

Ismael Leal

2023-01-20

1 *Introduction*

Heart diseases are one of the leading causes of deaths. Just in the United States, one person dies every 34 seconds from a cardiovascular disease, making it the fifth cause of deaths in the US. Still, if diagnosed, with medications and healthy habits the majority of heart conditions can be stabilized. It is obtaining the diagnosis what poses a challenge.

Machine Learning has recently proved itself useful in many medical applications, including medical imaging and diagnosis. Thus, ML algorithms might be able to predict whether a patient has a heart disease or not, making it possible for them to manage their condition and improve survival rates.

The ECG stress tests first measure the heart rate at rest and the blood pressure of the patient. Then, they undergo some exercise (usually walking on a treadmill) that will become progressively more difficult. The heart and breathing rates and the blood pressure are monitored all throughout.

The Processed Cleveland data set is a subset of the Heart Disease Data Set from the UCI Machine Learning repository, a public data set that was donated on 1988. It is a multivariate set of labeled data that has got more than 2 million web hits from the UCI ML repository alone. It's been used in many published experiments with the goal of detecting the presence of a heart disease in a patient. This data set includes the results of ECG stress tests together with biological information on the patients.

This report will analyze the features, samples, and relationships between and within them of the Processed Cleveland data set, with the aim of developing and training various predictive algorithms, both supervised and unsupervised. Then, they will be compared, and the best one will be tested in a final test set. The limitations of the models and opportunities for future work will be discussed as well.

2 *Data set Exploration*

2.1 Data set summary

The Heart Disease Processed Cleveland data set can be accessed from the UCI Machine Learning repository. It is a data.frame consisting of 14 columns and 302 rows. The last column shows the diagnosis for each patient, where 0 indicates that the patient doesn't have a heart disease, while the remaining values (i.e. 1, 2, 3, 4) indicate a patient with a heart disease. The first 13 columns show different medical numeric data about each patient, described in Table 1:

Table 1: Feature description

Feature	Description
age	Age of the patient
sex	Biological sex (0 = female, 1 = male)
chest_pain_type	Chest pain type (1 = typical angina, 2 = atypical angina, 3 = non-anginal pain, 4 = asymptomatic)
rest_blood_pressure	Resting blood pressure (in mmHg)
cholesterol	Serum cholesterol level in mg/dl
fasting_blood_sugar	Fasting blood sugar > 120 mg/dl (0 = false, 1 = true)
rest_electrocardiographic	Resting electrocardiographic results (0 = normal, 1 = ST-T wave abnormality, 2 = left ventricular hypertrophy)
max_heart_rate	Maximum heart rate
exercise_induced_angina	Exercise-induced angina (0 = no, 1 = yes)
old_peak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment (1 = positive, 2 = flat, 3 = negative)
vessels_number	Number of major vessels (0, 1, 2, or 3) colored by fluoroscopy
thalassemia_type	Thalassemia type (3 = normal, 6 = fixed defect, 7 = reversable defect)
diagnosis	Diagnosis

Age is usually considered as a significant risk factor for heart diseases; the probability of having one triples with for every 10 years of life.

Regarding sex, men usually are in greater risk of a heart disease compared to women that have not yet experienced menopause. However, women with diabetes are in greater risk of a heart disease than men with diabetes.

The chest_pain_type refers to angina, a pressure-like discomfort that occurs when the heart does not receive enough oxygen. It may also manifest as pain in shoulders, arms, neck, jaw, or back.

A high level of blood pressure may damage arteries and veins near the heart, increasing the risk of a heart disease. Cholesterol levels shouldn't be too high as well, as it can narrow arteries and thus increase pressure. Another feature that, if taking high values, increases the risk of a heart attack, is the fasting_blood_sugar. The increase in sugar levels may happen because of a lack of insulin, of a defective response to it. Another dangerous increase is an increase in the heart rate. It has been shown that an increase in heart rate of 10 beats/min provokes a 20% increase in the risk of cardiac death.

The rest_electrocardiographic feature value of 1 (ST-T wave abnormality) indicates some type of heart arrhythmia, and together with the value of 2 (left ventricular hypertrophy) both are usually considered as risk factors.

Regarding the slope of the peak exercise ST segment, an upsloping curve show best heart rate with exercise, a flat curve indicates a typical healthy heart, while a negative slope show signs of an unhealthy heart.

Finally, the more vessels a person shows, the more blood movement there is. Thus, patients with less major vessels are more likely to experience a heart disease.

There are 163 (53.97%) healthy observations and 139 (46.03%) observations with a heart disease present. This might mean that the false negative rate may be masked if not taken into consideration. Nonetheless there is not a massive imbalance between samples with different diagnosis, so it should not be a huge problem. Just in case, an analysis of the sensitivity through the confusion matrix will be done throughout the algorithm development.

Tables 2 & 3 show a summary of the data, demonstrating that every column's values are quite different from the others, as well as their ranges. Hence, the data could be normalized in order to get insightful visualizations that actually help comparing features.

Table 2: Feature summary (1)

age	sex	chest_pain_type	rest_blood_pressure	cholesterol	fasting_blood_sugar	rest_electrocardiographic
Min. :29.0	Min. :0.000	Min. :1.00	Min. : 94	Min. :126	Min. :0.000	Min. :0.000
1st Qu.:48.0	1st Qu.:0.000	1st Qu.:3.00	1st Qu.:120	1st Qu.:211	1st Qu.:0.000	1st Qu.:0.000
Median :55.5	Median :1.000	Median :3.00	Median :130	Median :242	Median :0.000	Median :0.500
Mean :54.4	Mean :0.679	Mean :3.17	Mean :132	Mean :247	Mean :0.146	Mean :0.987
3rd Qu.:61.0	3rd Qu.:1.000	3rd Qu.:4.00	3rd Qu.:140	3rd Qu.:275	3rd Qu.:0.000	3rd Qu.:2.000
Max. :77.0	Max. :1.000	Max. :4.00	Max. :200	Max. :564	Max. :1.000	Max. :2.000

Table 3: Feature summary (2)

max_heart_rate	exercise_induced_angina	old_peak	slope	vessels_number	thalassemia_type
Min. : 71	Min. :0.000	Min. :0.00	Min. :1.0	Min. :0.000	Min. :3.00
1st Qu.:133	1st Qu.:0.000	1st Qu.:0.00	1st Qu.:1.0	1st Qu.:0.000	1st Qu.:3.00
Median :153	Median :0.000	Median :0.80	Median :2.0	Median :0.000	Median :3.00
Mean :150	Mean :0.328	Mean :1.05	Mean :1.6	Mean :0.679	Mean :4.73
3rd Qu.:166	3rd Qu.:1.000	3rd Qu.:1.60	3rd Qu.:2.0	3rd Qu.:1.000	3rd Qu.:7.00
Max. :202	Max. :1.000	Max. :6.20	Max. :3.0	Max. :3.000	Max. :7.00

Here, vessels_number and thalassemia_type have been converted to integer class using the function as.integer(), as they were stored like characters.

Note how the features sex, chest_pain_type, fasting_blood_sugar, rest_electrocardiographic, exercise_induced_angina, slope, vessels_number, and thalassemia_type show discrete values, as the descriptions of Table 1 and the summaries of Tables 2 and 3 show. The kind of summary performed in tables 2 and 3 is usually more insightful for the continuous values (age, rest_blood_pressure, cholesterol, max_heart_rate, old_peak)

Table 4: Continuous features

List
rest_blood_pressure
cholesterol
max_heart_rate
old_peak

2.2 Data wrangling

The data shows numbers (both integers and rational numbers), and the class or data type of every feature can be seen in Table 5:

Table 5: Feature class

Column.name	Class
age	numeric
sex	numeric
chest_pain_type	numeric
rest_blood_pressure	numeric
cholesterol	numeric
fasting_blood_sugar	numeric
rest_electrocardiographic	numeric
max_heart_rate	numeric
exercise_induced_angina	numeric
old_peak	numeric
slope	numeric
vessels_number	integer
thalassemia_type	integer
diagnosis	integer

In the Data Set Summary section 2.1, the columns vessels_number and thalassemia_type were reclassified as integer for the creation of Tables 2 & 3. Also, the diagnosis column was reconverted to a factor using "as.factor()", where the levels are "H" for healthy when the diagnosis is 0, and "D" for disease otherwise (i.e. when the disease value is 1, 2, 3, or 4). While converting the columns vessels_number and thalassemia_type to integer, the NA values were removed from the data set "cleveland", obtaining a total of 296 observations.

Also, the data was reconverted to a list, given that many data sets have this form and are easier to work with. The list will have 2 elements:

- "x": a matrix with all the predictors (columns from 1 to 13 of the data set)
- "y": a list with all the diagnosis

2.3 Training and test sets creation

The data was separated into training and test sets (90% vs 10% of the original data), to create a final test set in which the final algorithm can be tested. This will prevent an over-fitting of the model developed. After the partition, the balance or prevalence of the disease is:

- 53.76% of healthy observations in the training set
- 53.33% of healthy observations in the test set.

Hence, the prevalence in the test and training sets is almost identical to the prevalence of the joint data in the "cleveland" data set.

2.4 Training set observations exploration

The normalization of the train set is important so that the different features (as Tables 2 & 3 show) can be compared together. The new data can be obtained by:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where μ is the feature mean and σ is the feature's standard deviation.

The dist() function was used to obtain the Euclidean distances between samples (i.e., between rows or patients), in a 266 dimensional space (the number of rows or observations in the training set). The average distance between all samples is 4.95. By calculating conditional averages, it can be inferred that healthy patients are closer to each other (4.84) than to patients with a disease (5.28). Patients with a disease are also closer to each other (5.07) than to healthy patients, though they are not as close between them as the healthy patients. Thus, there is greater variance for the features of patients with a disease.

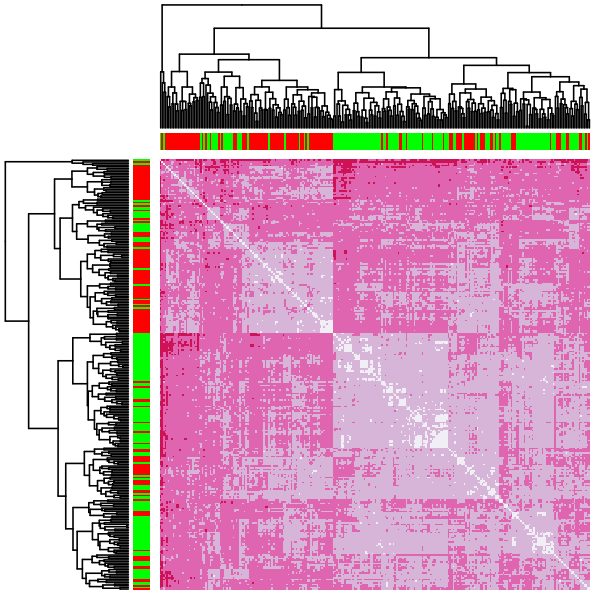

Figure 1: Distances between samples as a heatmap

The heatmap in Figure 1 shows the distances between samples with a color gradient, where darker colors are displayed for greater distances. Green samples are healthy patients while red samples are patients with a heart disease. It can be seen that observations are not perfectly clustered by diagnosis, although somehow, they are. In general, healthy samples are further apart from samples with a disease present (bottom left, top right) than among themselves (bottom right, central area). Also, in the top left quadrant a lower distance among non-healthy patients can be seen.

2.5 Train set feature exploration

Some unsupervised methods can be used to extract important features without the need of labeled outcomes. This reduces noise and over-fitting for the model, so it is an interesting option to consider.

2.5.1 Diagnosis dependence for each feature

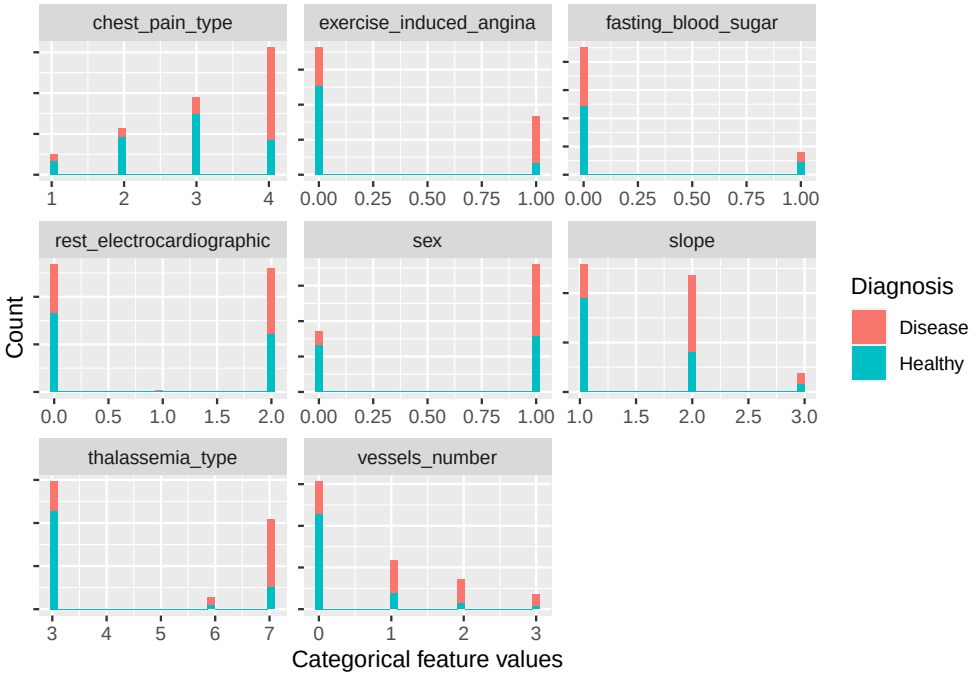

Figure 2: Histogram of values for every feature, colored by diagnosis

From Figure 2, it can be noted that patients experiencing exercise-induced angina are more likely to have a heart disease. Also, patients showing left ventricular hypertrophy (value of 2 in the rest_electrocardiographic column) are more likely to have a heart disease, as well as males. All of these conclusions match a naive intuition.

2.5.2 Feature variance

The caret package provides a function that allows an extensive analysis of the variance within features: nearZeroVar(). For the training set, it shows there is no near-zero or zero variance for none of the features, as Table 6 displays.

The mean frequency ratio is 2.21, with a 77% of the features showing a frequency ratio of less than 2.5. However, the percentage of unique values is not great, given that some features show discrete values, as explained in Section 2.1.

2.5.3 Hierarchical clustering

In order to calculate the Euclidean distance between features, the dist() function can be used again. However, as it operates between rows, the matrix with the data needs to be transposed so that the function works as we want it to. Then, hierarchical clustering can be performed, resulting in Figure 3.

Table 6: nearZeroVar outcomes

	freqRatio	percentUnique	zeroVar	nzv
age	1.12	15.038	FALSE	FALSE
sex	2.09	0.752	FALSE	FALSE
chest_pain_type	1.65	1.504	FALSE	FALSE
rest_blood_pressure	1.06	18.045	FALSE	FALSE
cholesterol	1.20	53.759	FALSE	FALSE
fasting_blood_sugar	5.65	0.752	FALSE	FALSE
rest_electrocardiographic	1.03	1.128	FALSE	FALSE
max_heart_rate	1.11	33.083	FALSE	FALSE
exercise_induced_angina	2.17	0.752	FALSE	FALSE
old_peak	6.50	14.662	FALSE	FALSE
slope	1.09	1.128	FALSE	FALSE
vessels_number	2.59	1.504	FALSE	FALSE
thalassemia_type	1.42	1.128	FALSE	FALSE

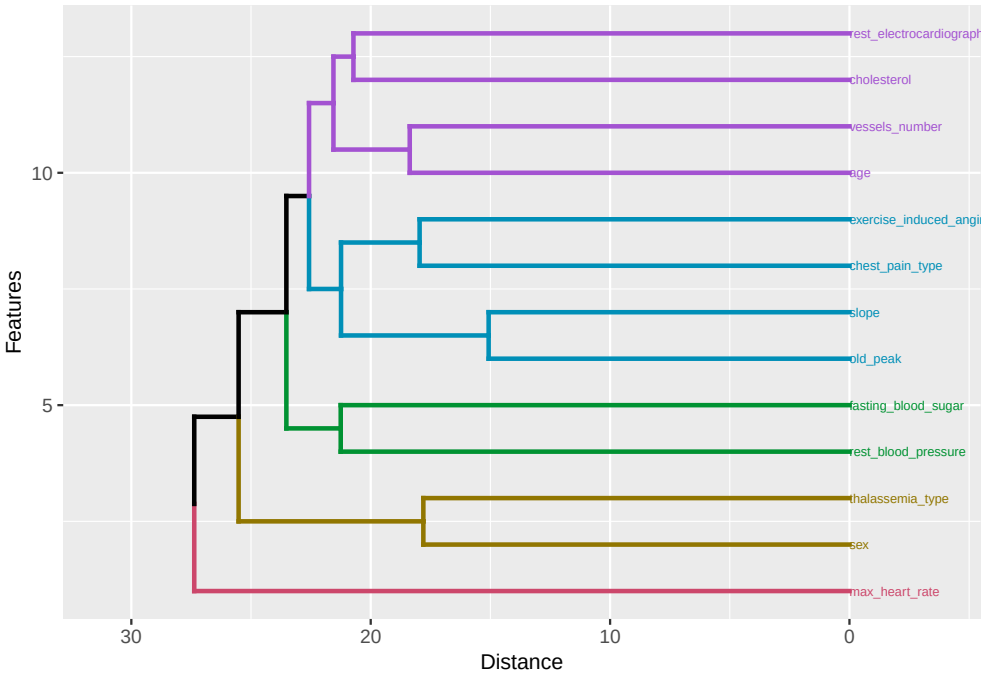

Figure 3: Dendrogram of features clustered

slope and old_peak are the ones that are closest together, and max_heart_rate is the feature that is furthest from the rest. However, except for these ones, the rest have similar distances, and none of the features are significantly close between them.

2.5.4 Correlation

The distances between and the variance within the features are insightful measures, as well as the correlation between them. The Pearson coefficient was used in this section to calculate the correlations between features and plot them as a heat map..

Figure 4: Heat map of correlations between features

The heatmap in Figure 4 shows that most features are not tightly related. Darker colors indicate stronger correlation. There is a significant correlation between old_peak and slope, as the dendrogram in Figure 3 showed. This could be due to the fact that both are measures related to the ST segment.

The age feature shows some correlation with rest_blood_pressure and max_heart_rate, as is intuitive to think.

The features that affect the max_heart_rate are age, thalassemia_type, chest_pain_type, exercise_induced_angina, and vessels_number. We can see that the heart rate is affected by the number of major vessels: the more vessels, the more blood flow. Also, the angina pain type and whether it is induced by exercise or not affect the heart rate, as expected.

It can be seen that max_heart_rate, exercise_induced_angina, and thalassemia_type have correlation with the values obtained in old_peak and slope.

However, the mean correlation is low: the mean is 0.23. This analysis supports the inclusion of all features in the algorithm.

The correlation of each feature with the final diagnosis might be more informative.

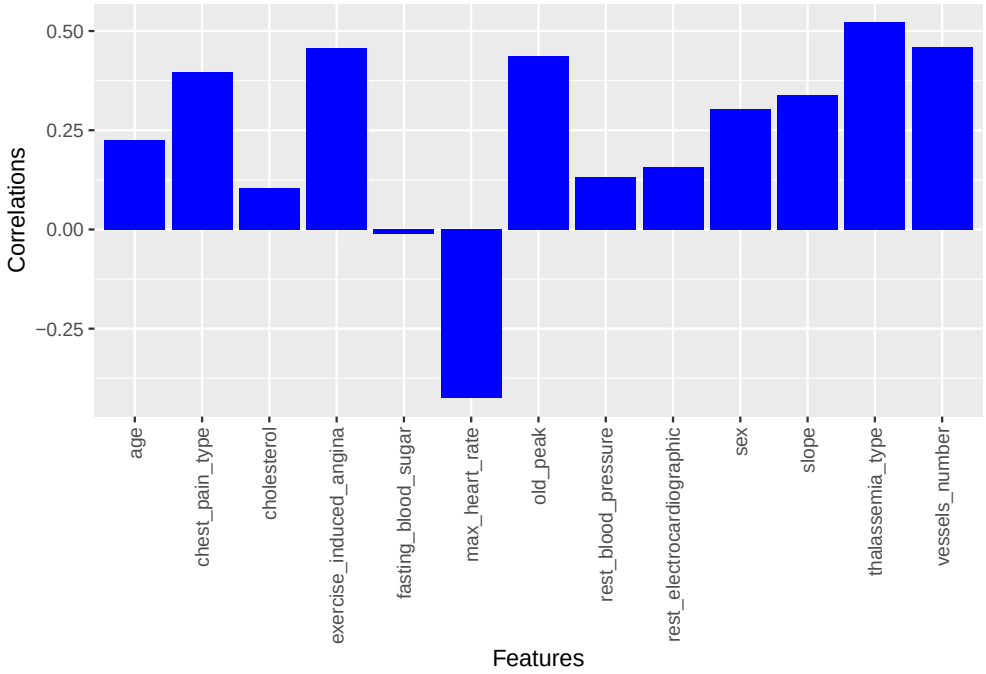

Figure 5: Correlation of each feature with the diagnosis

Figure 5 shows that the fasting_blood_sugar and the cholesterol are the least correlated with the final diagnosis. However, the rest of the features do show some correlation with the diagnosis, especially the chest_pain_type (the type of angina pain experienced), the exercise_induced_angina, the max_heart_rate, the old_peak, the thalassemia_type, and the vessels_number. Thus, predictions can probably be made with a decent accuracy.

2.5.5 Principal Component Analysis

Principal component analysis can reduce the dimensionality of the problem, which makes the visualizations more informative for data with multiple dimensions. It may also improve the quality of classification models.

Table 7: Principal Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	1.770	1.280	1.104	1.062	0.974	0.941	0.920	0.883	0.812	0.746	0.670	0.637	0.595
Proportion of Variance	0.241	0.126	0.094	0.087	0.073	0.068	0.065	0.060	0.051	0.043	0.035	0.031	0.027
Cumulative Proportion	0.241	0.367	0.461	0.547	0.620	0.688	0.753	0.813	0.864	0.907	0.942	0.973	1.000

According to Table 7, the first 4 principal components already account for almost 55% of the cumulative variance. The first 10 principal components account for more than a 90% of the variance.

Figure 6 shows two boxplots for each principal components, one for the healthy samples (colored blue) and other for the samples with a disease (colored red). The spread is similar for healthy and unhealthy samples for most principal components. However, the two boxplots of the first principal component are the only ones which don't have overlapping interquartile ranges.

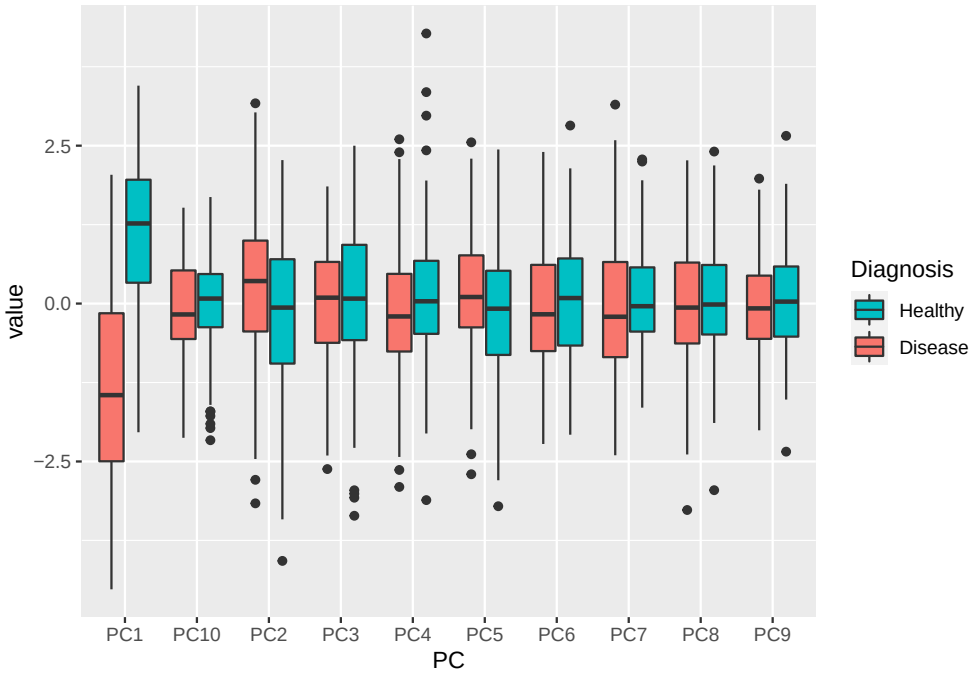

Figure 6: Boxplots by diagnosis

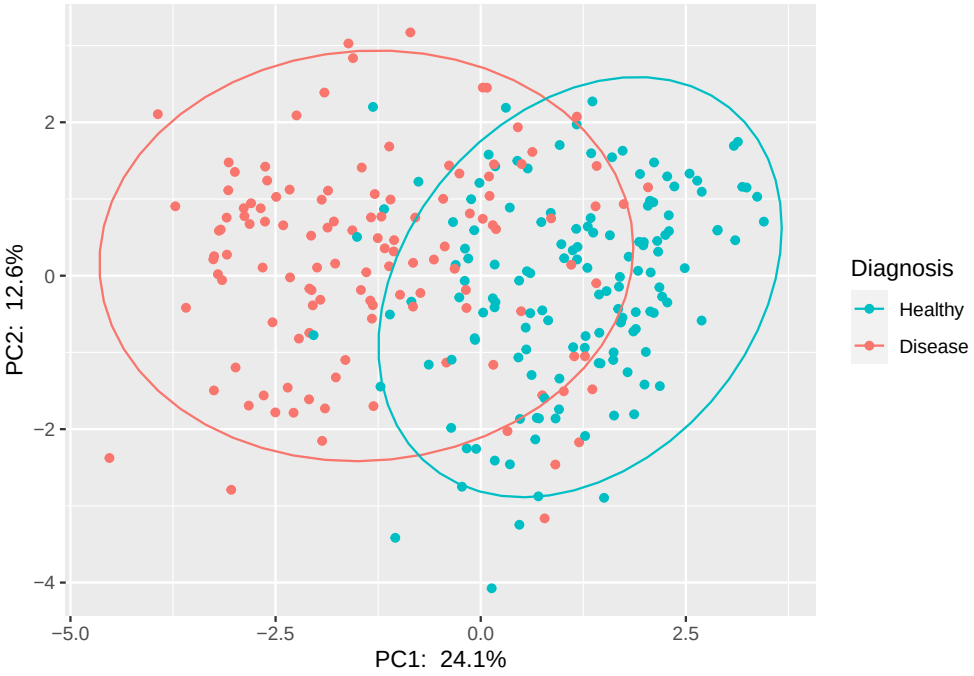

Figure 7: PC2 vs PC1

Figure 7 shows a scatterplot of PC2 vs PC1 colored by diagnosis with the same color code as that of Figure 6. Samples of patients with a disease are slightly more spread out. The ellipses drawn around both types of samples (healthy / unhealthy) show that there is an overlap between both diagnosis, with a significant

separation however.

Given the overlap in Figure 7 and the proportion of variance shown by the main principal components, it can be inferred that a Principal Component Analysis will not necessarily offer many more insights.

Summing up, there are patterns for both observations and features. Thus, the algorithm can be developed now.

3 *Methods*

3.1 Pre-processing

A data frame was created to store the performance of all the models. As specified in the Data set Exploration, an analysis of the confusion matrix is important. Thus, the performance will be measured with the accuracy, the sensitivity, specificity, and F1 score. This data frame was called "metrics"

Also, because of the normalization applied to the training set, the test set should be normalized. This is a safe step as it won't interfere with the result, preventing an over-fitting of the algorithm.

Cross-validation allows a more extensive training of the algorithm without over-fitting it to the test set. It is also useful when trying to tune parameters, if needed. The control parameters were set for a 6-fold cross validation. A higher number was not chosen as the number of samples is not massive.

3.2 Random outcomes

3.2.1 Non-weighted random sampling

The most naive model would be one that predicted the diagnosis randomly with a 50-50 chance.

3.2.2 Weighted random sampling

The prevalence of the disease in the training set could be taken into account, as this would probably improve our random model.

3.3 Unsupervised methods

3.3.1 K-means clustering

The hierarchical clustering showed that unsupervised clustering might be able to cluster healthy observations from unhealthy observations. However, k-means clustering is a method for continuous data. Thus, the categorical or discrete features will be left out, with age, rest_blood_pressure, cholesterol, max_heart_rate, and old_peak remaining.

In this case, as there are only two classifications (i.e., "healthy" and "disease"), there will be 2 clusters. 2-means clustering will create 2 centers (one for each cluster) and assign every sample to one cluster, according to its Euclidean distance to the cluster's centers. A function can be created for this task. This function received the coordinates of the centers of both clusters by using the function kmeans().

3.4 Supervised methods

3.4.1 Generative modelling

These are supervised predictive methods that model how the data set is distributed and use probability distributions to predict the conditional probability of a specific outcome. The simplest model is the Naive Bayes, which uses the Bayes theorem and assumes equal importance for all features, which is not true in this case.

More advanced generative models might work better, such as Linear Discriminative Analysis. It assumes that the data are normally distributed.

Quadratic Discriminative Analysis assumes multivariate normal distributions, and is useful for classes that show different co-variances.

3.4.2 Discriminative modelling

3.4.2.1 Logistic regression This is a generalized linear model, assuming that predictors and outcomes follow a bivariate normal distribution, so that the outcome fits a regression line.

3.4.2.2 k-Nearest neighbours (knn) Here, k is the tuning parameter (unlike for k-means clustering) that can be determined using cross-validation. The value of k will be the number of neighbors for the data points taken into consideration when training the algorithm. A greater k results is smoother estimates. To find the best value for k, the tuneGrid argument of the train() function from the caret package was used. The values were all integeres from 1 to 50.

3.4.2.3 Random Forest Random forest models use decision trees to partition the data, allowing final predictions to be made with a smaller subset of predictors. However, decision trees are prone to over-train models. Thus, an ensemble of multiple decision trees will ease this problem. Then, an average of the predictions is calculated to yield the final prediction. In this case, it is the number of randomly chosen predictors included in each decision tree that has to be tuned. The values tunes were the odd numbers between 3 and 15.

3.4.2.4 Neural networks These models are useful for both multidimensional data and for non-linear data. However, these are computationally expensive, being rather complex algorithms. The most basic neural network models, sigle-layer neural networks, deal with linear data. They are models that apply a weighting to the multidimensional inputs and sum them to classify the model. The method "nnet" from the caret package was used.

3.4.3 Ensemble

The ensembles are combinations of different predictive models. They effectively improve stability and accuracy. The selected models to include in this final ensemble were the ones with higher accuracy. Also, only supervised models were included, given that they are clearly better for labeled data, which is our case.

4 *Results*

4.1 Performance

Table 8 shows the performance of every model tested in this report. It can be confirmed that specificity was higher than sensitivity for most models. This means that there were more false negative predictions than

Table 8: Performance of models used

Model	Accuracy	Sensitivity	Specificity	F1
Random sampling	0.48	0.48	0.48	0.46
Weighted random sampling	0.52	0.48	0.55	0.48
K-means clustering	0.85	0.72	0.97	0.82
Naive Bayes	0.83	0.76	0.90	0.81
Linear Discriminant Analysis	0.80	0.68	0.90	0.76
Quadratic Discriminant Analysis	0.80	0.72	0.86	0.77
Logistic regression	0.83	0.76	0.90	0.81
K Nearest Neighbour	0.80	0.72	0.86	0.77
Random Forest	0.89	0.84	0.93	0.87
Neural Network	0.78	0.68	0.86	0.74
Ensemble	0.81	0.72	0.90	0.78

false positives.

Random sampling, as expected, is the least accurate method. There are 25 patients with a heart disease (true positives) and 29 healthy patients (true negatives) in the test set. Looking at Table 8, a sensitivity of 0.48 means that there were 27 false negatives, so 27 patients would be incorrectly diagnosed as healthy. The specificity for this model is also 0.48, meaning 31 patients would be incorrectly diagnosed as having a heart disease.

The weighted random sampling accounted for the prevalence in the data set and improved the specificity to 0.55, meaning now 23 patients would be incorrectly diagnosed as having a heart disease. The sensitivity remained the same as without the weighting, thus improving the overall accuracy.

The first model significantly improving the accuracy of the predictions was the k-means clustering. Its accuracy increased up to an 85% and the specificity was at 0.97. This entails that just 1 person was incorrectly diagnosed as having a heart disease. The sensitivity also improved to a 0.72, meaning 10 patients were classified as negative while actually having the disease.

Supervised models were expected to work better for the labeled data, however only the random forest improved the accuracy of k-means clustering, reaching an 89%.

#Discussion

Both the unsupervised and supervised models developed improved the performance compared to the random and weighted random sampling models. Thus, it can be said that the creators of this data set did choose relevant features that could be used as predictors. Even k-means clustering (unsupervised) worked well assuming the data were unlabeled.

Supervised models performed well, all having an accuracy above 80%, except for the neural network model. These are not perfect predictions, they are not even near. Moreover, the F1-score was even lower than the accuracy for all models. Specificity in general was good (around 0.90 for most models), meaning the number of false positives (incorrectly diagnosed as healthy) was low, which is good. In spite of this, the sensitivity had a much lower value (around 0.70 for most models), meaning there are some patients diagnosed as having a heart disease while being healthy. This, of course, poses a risk of over-diagnosis and unnecessary treatments. Usually discriminative models are preferred over generative models. The best model overall, the random forest one, is a discriminative model.

However, the ensemble model might be better, given that it lowers the risk of over-fitting with a single model,

instead of considering various.