FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D
COIMBRA

Universidade de Coimbra

Departamento de Engenharia Informática

Pattern Recognition
Heart Disease prediction

Name:	Student Number:
Francisco Brilhante	**2018278239**
João Marcelino	**2018279700**

2021/2022

Contents

Chapter 1

Introduction

Heart disease is one of the leading causes of death in the USA population and is associated with risk factors such as high blood pressure, high cholesterol, smoking, etc.

The goal of the project is to use information about risk factors of patients to predict heart diseases. This report aims to explain how patient data was gathered, what tasks and techniques were conducted to detect patterns and what models were used to make predictions for each person health condition.

The project development can be divided into 3 main components:

- Pre-processing - each sample was relabeled according to Scenario A (all depended variables were dropped except the first - CoronaryHeartDisease);
- Feature Selection and Reduction - during this phase the correlation and redundancy of each feature was analysed. Dimensionality reduction was also performed here;
- Classification - 2 types of classification models were obtained in order to classify each data sample. Their performance is discussed in this report.

Chapter 2

Dataset

The dataset used in this project was collected by the CDC amongst U.S residents and available on Kaggle [1]. It comprises 318958 distinct samples with 4 dependent variables and 15 independent variables each.

Each dependent variable represents binary answers given by the patients regarding one of the following health issues:

- Coronary Heart Disease
- Myocardial Infarction
- Kidney Disease
- Skin Cancer

For the problem characterized in Scenario A, only the first dependent variable was kept. The final step in the pre-processing stage was the splitting of the data into training and test sets. 200000 samples were chosen randomly to compose the training set while the rest (118958) were assign to the testing. There are no duplicates in any of the sets and the splitting process took place multiple times with different RNG seeds in certain stages of this project in order to have representative measures of performance of the predictive models.

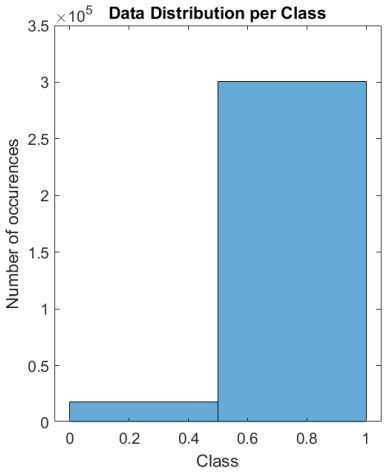


Figure 2.1: Coronary Heart Disease answer distribution. 0 means the patient has reported having such illness, 1 otherwise.

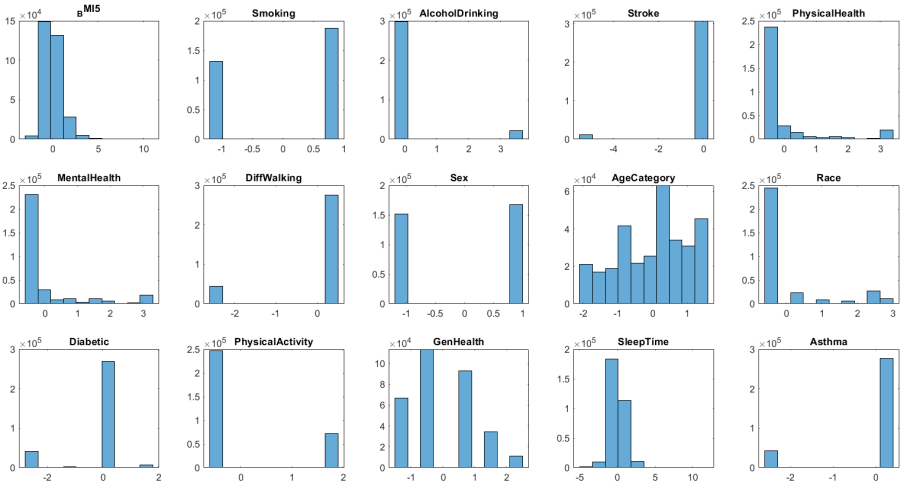


Figure 2.2: Independent variables distribution. Some features relate to Yes/No questions (smoking, asthma, difficulty walking,etc) while others are represented in discrete spectrums (age, sleep time in hours/day, etc).

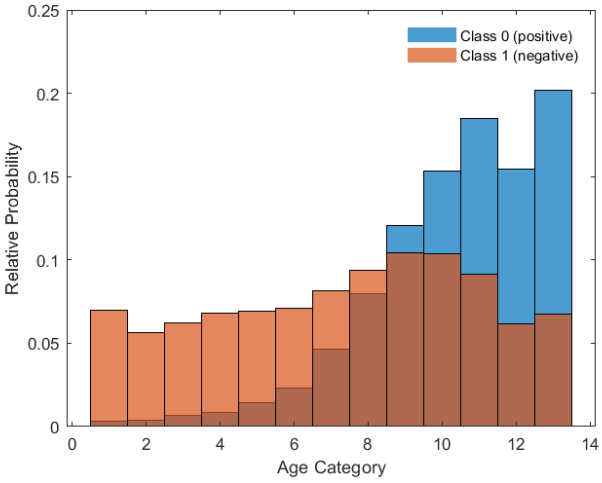


Figure 2.3: Example of one feature available in the dataset. Age category distribution per class. This project aims to take benefit of the differences in distributions to successfully identify Coronary Heart Disease in patients.

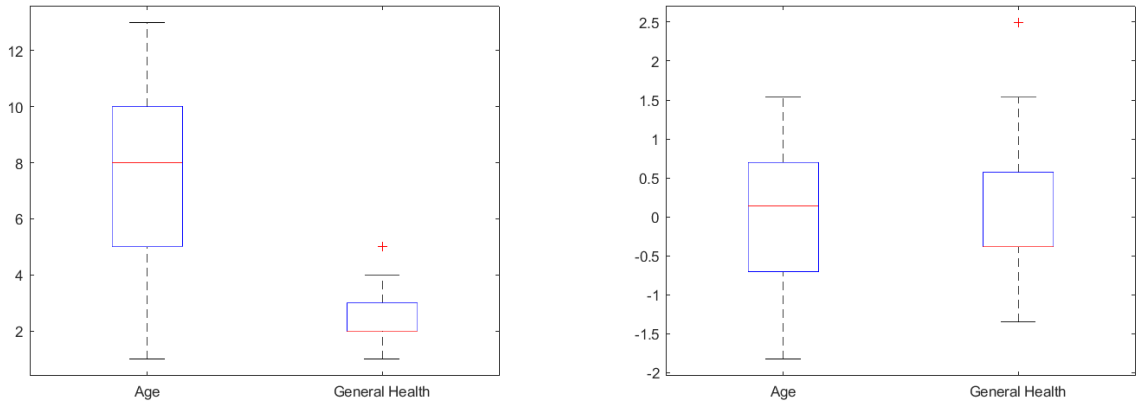
Chapter 3

Feature Selection and Reduction

As stated previously, each data point contains 15 different features. This number can be considered relatively high when the goal is to build predictive models with reasonably low time and space complexity and when other datasets, with substantially lower dimensionality, can be used to obtain almost equal levels of prediction accuracy using equivalent models. With this in mind, this stage aims to analyze the importance of each feature, produce new useful features and expose which ones should be selected to train the models of the next stage.

3.1 Principal Component Analysis

Before applying the PCA method to the data, it is important to normalize it to mean 0 and standard deviation 1. This is done because different features present different range of values and in order for PCA to work properly all features need to have the same amount of variance. PCA seeks to maximize variance of each component by projecting data into new directions.



(a) (b)

Figure 3.1: Exemplification of scaling impact in 2 features - Age category and General Health Category. On the right the dataset suffered scaling while on the left it did not.

After scaling all the data, PCA was applied, giving the results shown in 3.2. Unlike what was expected, the eingenvalues do not stabilize near 0 after a certain number of components as it happens with other datasets.

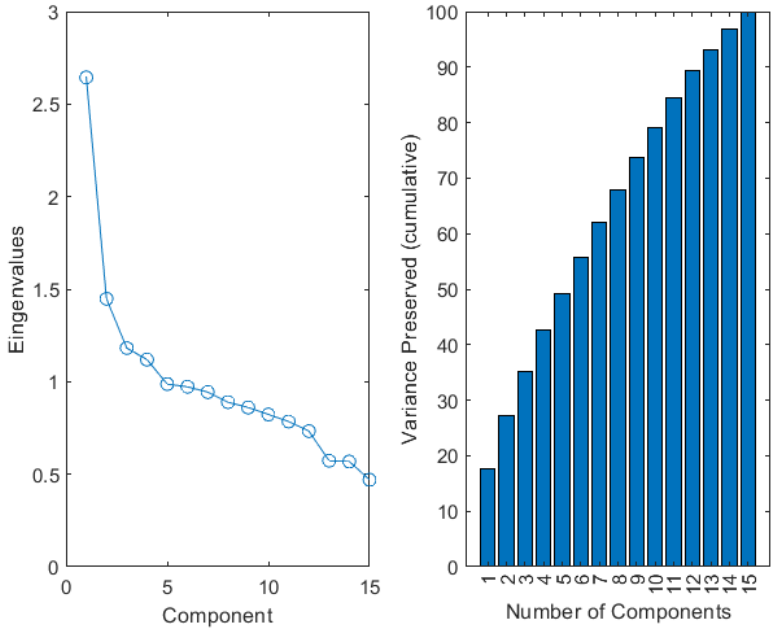


Figure 3.2: Eingenvalues (left) and percentage of variance preserved by number of components extracted (right).

Enforcing the Kaiser criterion in this context results in the extraction of 4 com-

ponents with values above 1 even though the variance preserved in that scenario is very low - 42.6%. On the other hand, selecting all components before the eingenvalues stabilize becomes quite a subjective task, leaving freedom to choose between 1, 4 or 12 features. Choosing the last one leaves a total variance of 89.3% while still maintaining a large number of features. which is undesirable, but that can be solved through other feature extraction methods addressed further in this report.

In spite of the high shift in variance according to the number of components chosen, the analysis of the prediction model performance reveals that such decision does not, in fact, impact the final results of the project in a substantial way. This is also investigated further in the next chapter.

3.2 Kruskal-Wallis Test and Correlation Matrix

Instead of projecting the initial data into new directions using PCA, one can also study their the redundancy and discriminative power.

The Kruskal-Wallis Test enables the sorting of each feature by its relevancy to the classification problem - a feature that renders a higher value in the test is more discriminative and shows higher dependency on the label values.

Feature	H value
AgeCategory	12988.028
GenHealth	12076.776
DiffWalking	9534.922
Stroke	7508.929
Diabetic	5855.576
PhysicalHealth	5165.189
Smoking	2284.627
PhysicalActivity	2068.247
Sex	1081.800
BMI5	863.678
Race	518.308
Asthma	435.784
AlcoholDrinking	249.975
SleepTime	6.945
MentalHealth	0.011

Table 3.1: Feature ranking by H values obtained in Kruskal-Wallis Tests.

From table 3.1 it is possible to observe well-known risk factors for Coronary Heart Disease - age, diabetes, smoking and stroke history - while other unexpected features, such as difficulty walking and general health report, also show high degrees of correlation.

With the support of these statistics, we can confidently extract 6 most important

attributes of patients, drastically reducing the amount of data to be processed by the models discussed later. Out of the first 6 attributes shown in table 3.1, we may need to discard even more elements depending on the values shown in the correlation matrix of figure 3.3.

The correlation matrix allows the identification of nearly identical features. Low values are preferred over pairs of features that present values near 1. In the context of our problem nearly all features show low levels of correlation. As expected, the highest value is shown between General Health(13) and Physical Health(5), which leads us to discard Physical Health from the set of 6 features selected earlier using Kruskal-Wallis Tests.

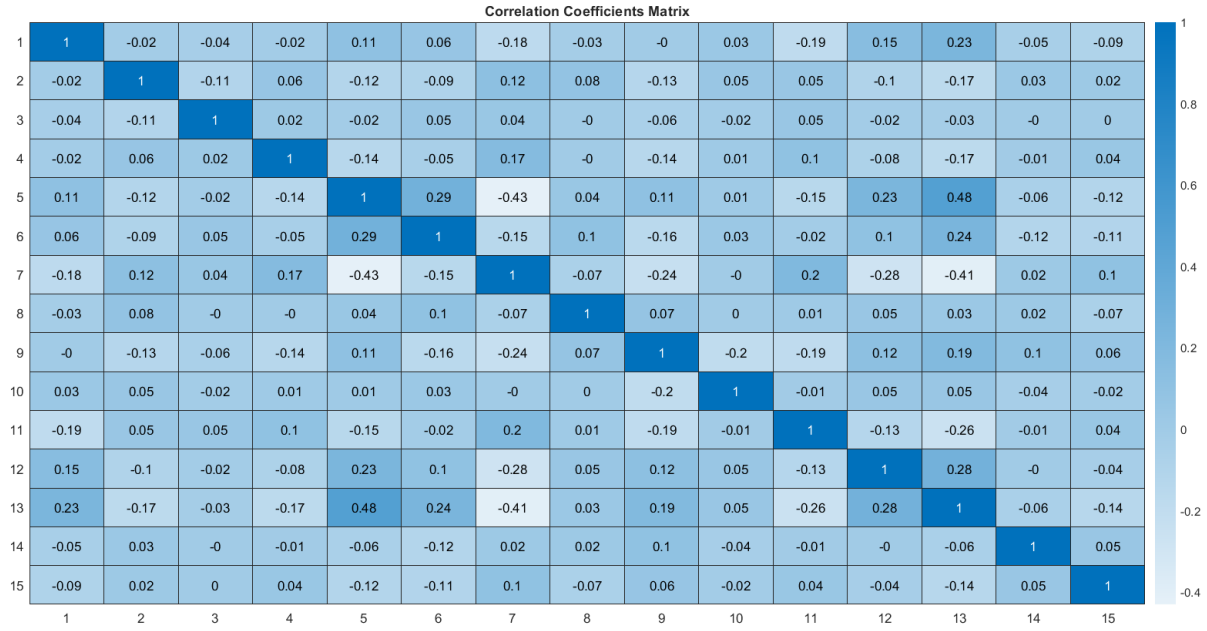


Figure 3.3: Correlation coefficients matrix computed for the entire dataset.

3.3 Fisher's Linear Discriminant

The third way to reduce the number of dimensions addressed in the project was to use Fisher's Linear Discriminant. FLD works similarly to PCA in the sense that it projects data in different directions. The difference between the two techniques is that, in FLD case, data is projected onto directions that maximize inter-class separability and minimize intra-class variability. The number of dimensions left from this process is independent of the initial dimensionality and is always equal to c-1, where c is the number of classes.

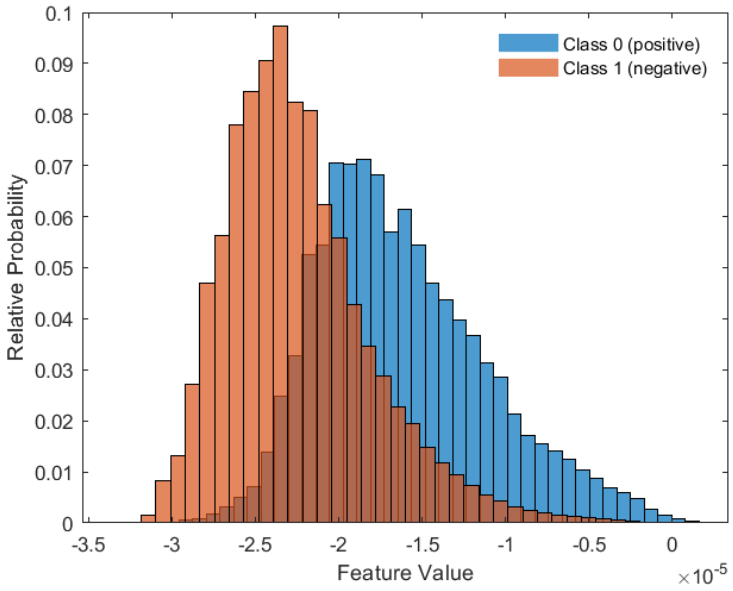


Figure 3.4: Histogram of feature extracted from Fisher's LDA using initial raw data.

Comparing figures 2.3 and 3.4, the data obtained from this technique displays a higher level of separation between labels, which hopefully will translate into equivalent or even better classification results.

Chapter 4

Classification Models and Results

In order to predict the class of each test data sample, 3 distinct models were used: Euclidean Minimum Distance Classifier, Mahalanobis Minimum Distance Classifier and Fisher's Linear Discriminant.

The dataset was divided into train and test sets. The training set is composed of 200000 samples and is used to compute the centroids for the minimum distance classifiers, for instance. The remaining samples constitute the test set and were used to compute the accuracy, sensibility, specificity and confusion matrix of each model.

Given that the splitting process applied to the dataset has stochastic properties (as described in Chapter 2), each run was replicated 30 times using different RNG seeds. Then, the mean and standard deviation were computed for each performance metric.

Features	**Model**	**N**	**Accuracy**	**Sensibility**	**Specificity**
Raw(15)	Linear MDC	30	0.8291 ± 0.0012	0.3865 ± 0.0060	0.8557 ± 0.0014
Raw(15)	Mahal. MDC	30	0.7912 ± 0.0016	0.6962 ± 0.0062	0.7970 ± 0.0018
Raw(15)	Fisher's LD	30	0.7912 ± 0.0016	0.6962 ± 0.0062	0.7970 ± 0.0018
PCA(1)	Linear MDC	30	0.7794 ± 0.0005	0.5836 ± 0.0056	0.7911 ± 0.0006
PCA(4)	Linear MDC	30	0.7827 ± 0.0006	0.6333 ± 0.0050	0.7917 ± 0.0008
PCA(12)	Linear MDC	30	0.7911 ± 0.0007	0.6290 ± 0.0050	0.8009 ± 0.0008
PCA(1)	Mahal. MDC	30	0.7794 ± 0.0005	0.5836 ± 0.0056	0.7911 ± 0.0006
PCA(4)	Mahal. MDC	30	0.7767 ± 0.0008	0.6690 ± 0.0048	0.7832 ± 0.0009
PCA(12)	Mahal. MDC	30	0.7984 ± 0.0011	0.6455 ± 0.0053	0.8076 ± 0.0012
Kruskal-Wallis(5)	Linear MDC	30	0.6344 ± 0.0008	0.8012 ± 0.0041	0.6244 ± 0.0009
Kruskal-Wallis(5)	Fisher's LD	30	0.7774 ± 0.0019	0.6909 ± 0.0054	0.7826 ± 0.0020

Table 4.1: Performance results of different models and different feature reduction/extraction techniques.

As expected, the highest accuracy value (82%) was achieved with the initial features using Linear MDC. On the other hand, this was the combination with the lowest sensibility value (38%), which is unacceptable in a medical application such as this one. This low sensibility value can be, in part, attributed to the imbalance of number of samples per class. Taking data displayed in figure 2.3 as an example, the centroid of the dominant class - class 1 - tends to overlap a big portion of class 0 samples, leading to their miss-classification.

Overall all combinations performed reasonably well when compared with the base model and, at the same time, most of them required less than a third of the initial amount of information, highlighting the power of feature reduction techniques such as the ones used in this project.

From all the combinations presented in table 4.1 apart from the base model, PCA (12) + Mahal. MDC had the best accuracy(79.84%) while Kruskal-Wallis(5) + Linear MDC produced the best sensibility value (80%), making these two combinations good alternatives to the initial model.

As referenced before, although the amount of variance preserved after PCA varied drastically depending on the number of components selected, this choice did not affect in any significant way the final results presented in table 4.1.

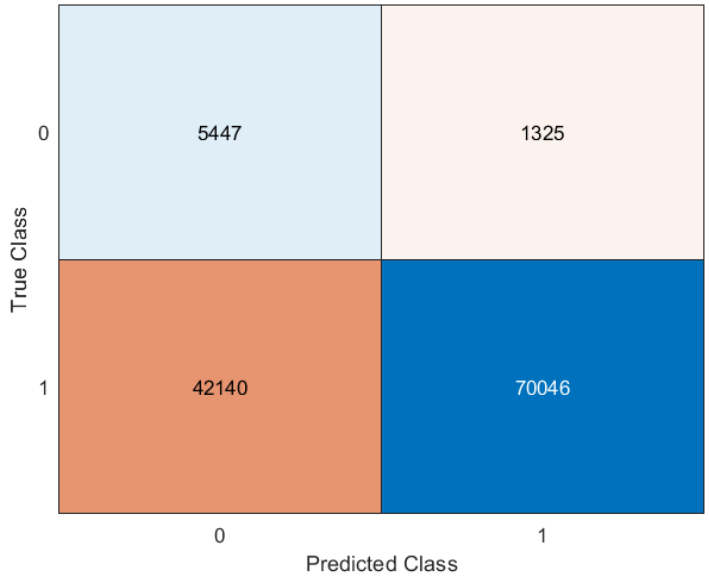


Figure 4.1: Confusion matrix of a run with the Kruskal-Wallis(5) + Linear MDC combination.

Chapter 5

Conclusions

With the objective to predict heart diseases, it was implemented and tested different techniques to process, select and reduce the data while maintaining a high percentage of information, as well as different classification models.

The results displayed in Chapter 4 emphasize the power of feature reduction and selection processes in diminishing the computational power required to accomplished prediction results equivalent to the ones of the raw dataset.

It is also crucial to highlight the importance of analysing various metrics when choosing the best models as sometimes good accuracies can hide low sensibility values, which are inadmissible in critical applications such as those found in medical contexts.

In future work, more robust models could be developed and compared against the linear models presented in this report.

Bibliography

[1] Kamil Pytlak. Personal key indicators of heart disease, Feb 2022.