Multivariate Analysis Report

Edgar Varela	Marisa Pereira	Ricardo Simões	Théo Di Piazza	Vicente Sobral
No. 86769	No. 88174	No. 93674	No. 101420	No. 102134

Course: Multivariate Analysis – *Instituto Superior Técnico* – MECD, MMA, MEBiom

January 2022

1 Introduction

Every year in the United States 200-300 thousand people die from acute myocardial infarction before arriving at the hospital. In the United States, every 29 seconds, one person becomes ill with MI, and every minute one patient with MI dies [1].
Considering the panoply of all available subjects of analysis in the field of Data Science, none is of more importance than a matter of life or death. Our work focuses on the understanding and consequent prediction of the complications that a Myocardial Infarction can pose to hospital patients - which are characterized by an array of continuous and categorical variables. In a silent and sudden disease such as the Myocardial Infarction, every instant counts. Every instant can change the life expectancy of a patient, and such thing as a prediction mechanism can eventually play a crucial role. Throughout the last half century, the occurrence of this disease has drastically increased, thus having become a dark and heavy name for medical doctors and staff, and one for which there is no easy resolution or closure. Therefore, our research lies on this topic, and in whether the patients that arrive at the hospital with the condition will eventually live or die.

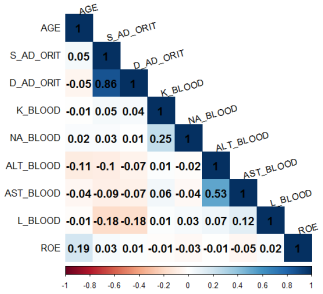
2 Dataset Description

The dataset that is central to our analysis [2] is very rich, featuring data for the patients in 4 different time frames: when the patient arrives at the hospital, and 24h, 48h and 72h after admission. It comprises 123 variables in total, where 89 are binary (variables that confirm the presence or absence of a condition in the patient, for example the presence of pulmonary edema in the patient), 12 are numeric (variables that are measured at the hospital by the doctors and nurses, such as the systolic blood pressure), 21 are ordinal (variables that inform about the time recurrence of a condition, e.g., the duration of the arterial hypertension) and 1 is categorical (the lethal outcome). This last one, which will be our target variable, is a very unbalanced one, where the individuals that have a lethal outcome represent 16% of of the data.

2.1 Descriptive statistics

2.1.1 Correlation and Association

In this section, we will use some descriptive statistics to extract some initial information about our data. We will begin by studying the correlation - a measure that indicates how closely two variables are related - of the numerical variables in the dataset and then the V of Cramer, that measures association between categorical variables. Let's consider the pictures from *Figure 1*.



(a) Correlation matrix

Feature 1	Feature 2	Cramer's V
MP_TP_POST	Ritm_ecg_p_2	0.81
Ritm_ecg_p_1	Ritm_ecg_p_7	0.76
Ritm_ecg_p_2	N_r_ecg_p_5	0.63
MP_TP_POST	N_r_ecg_p_5	0.57
FK_STTNOK	IBS_POST	0.53

(b) V of Cramer's

Figure 1: Correlation matrix and V of Cramer's

Few of the numerical variables appear to be highly correlated, which allows us to imagine that each numerical variable provides different (poorly correlated) information from the other variables. Note that the variables **S-AD-ORIT** and **D-AD-ORIT** (systolic and diastolic blood pressures, respectively) are highly correlated, which seems consistent given the description of the variables. Nonetheless, we decided not to exclude one of these two right away. In addition to this, we did not include the variables **KFK-BLOOD**, **S-AD-KBRIG** and **D1-AD-KBRIG** since they had too many missing values, which prevents the calculation of these correlations. We will address this issue again in section *3.1*.

As for Cramer's V, it measures the relation between two variables in categorical scale. The value returned is between 0 and 1, proportional to the association between the variables. The cramer V values for the categorical variables are grouped in a table and displayed above, for the values closest to 1. For some variables, we observe high values but not close enough to 1 to assert that the variables are completely related or even identical.

3 Data Processing

From analysing our data, we readily concluded that we needed to apply quite a few types of processing techniques: missing values imputation, transform the type of variables and outlier detection.

3.1 Missing Values

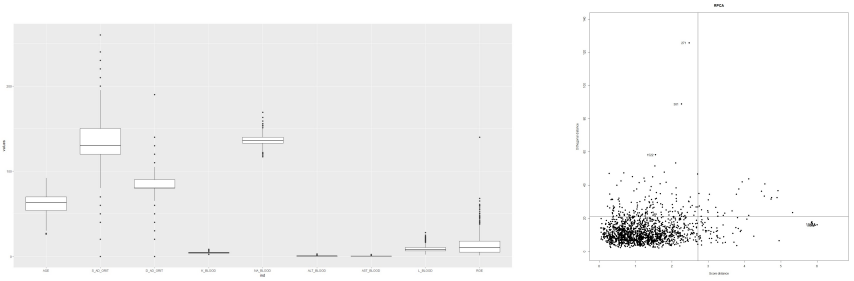
The first significant change we made to the dataset consisted in the removal of the 7 variables that had more than 25% of missing values, which were: KFK_BLOOD, IBS_NASL, S_AD_KBRIG, D_AD_KBRIG, NOT_NA_KB, LID_KB, NA_KB . To consider them we would need to introduce too much synthetic data.

In order to overcome the variables that still had missing values (after removind the already mentioned ones), we used the K-Nearest Neighbors with the Gower Distance as imputation procedure. We first assigned each variable to their respective data types, according to the descriptions provided. [3]. We decided to use this method since we have continuous, categorical and ordinal data, so it's appropriate for our case. We assumed 5 neighbours for the process, since it was the default value. We also decide to make the target variable binary, since the individuals with a lethal outcome are only a few and separating them by the cause of death would magnify the unbalance effect on the target variable.

3.2 Outlier Detection

Although we could have removed the unidimensional outliers, looking at every variable in an exhaustive way (see *Figure 2a*), we opted to remove only the multidimensional outliers, We made this decision since multidimensional outliers might not be one dimension outliers. This thought led us to discard an analysis that might have compromised our results, since univariate outlier detection is not correct when we are working with multivariate data.

Therefore, we applied RPCA [4] in a multidimensional space, comprised by the 9 numeric variables that we had left. This technique classifies the points as regular observations, good leverage points, orthogonal outliers and bad leverage points, taking as criteria the robust score distance and the orthogonal distance for every point, which then are compared with a cut-off value.



(a) Unidimensional outliers (b) RPCA analysis

Figure 2: Uni- and multidimensional outliers

We considered as outliers only the bad leverage points, which corresponded to 27 observations that can be found in *Figure 2b*. Instead of removing this outliers from the dataset, we decided to keep them and only remove them in the end to see how our models react without this datapoints present.

4 Supervised Learning

Supervised learning, which predicts an output variable (y) from the input variables (x) with the help of a training set, is one of the biggest tools for data scientists to apply in a dataset such as this one. Although it comprises several methods sometimes we could not use some of the most popular ones due to the limitations of our data, which proven to be the

biggest difficulty to overcome throughout the entire work. It is also relevant to mention that throughout this section there will be applied some techniques of data processing, because they were a consequence of the results we were obtaining in the different scenarios. Our main focus here will be in finding models that can as many deaths as possible, so we will look for specific evaluation metrics, which is something that will be addressed carefully in *4.2.*

4.1 Train and Test Set

In the following sections we ran different scenarios and algorithms, maintaining one thing constant: the same train-test split was performed for every one of them, with 80% of the data for training and 20% for test. Independently of the transformations applied, since we used the same seed to generate the data splits, the data of all these sets were the same, with exception for the scenario without outliers, since it has less 27 datapoints. For training, we resourced to parameter tuning. Therefore, the results for training data was obtained through a 5-fold stratified Cross Validation mechanism. This test worked as a validation of the selection of the best hyperparameters. For the sake of simplicity, we won't present the hyperparameters of all the models here; we advise the reader to carefully analyse the provided Notebooks that contain all the models, results and metrics obtained.

Summing up, we did a train and validation in a set to discover the best hyperparameters and then we applied these methods in a new test set, whose results are the ones we'll present.

4.2 Relevant Measures

Regarding important measures, since our main goal is to find a model that largest number of lives, **Recall** is the most important one to consider. We must also mention the **Precision**: a high precision signals than when the model detects an individual that will die, it is very probable (proportional to the recall value) that he will in fact die.

Taking the point of view of an hospital for a moment, it would not be sustainable to have a model with low precision, which is a model that deals with more false positives, representing false alarms. Therefore, it is interesting to understand that the importance of the metrics varies depending on who the final clients are. As a matter of fact, the importance of these two metrics can be expressed as one: the **F1 Score**, particularly important when the classes are unbalanced and precision and recall are both to care for.

Other metrics like the **Balanced Accuracy** are important, since it accounts for both the positive and negative outcome classes, taking into account an imbalanced dataset [5].

4.3 Initial Results

We started by considering only the 9 continuous variables, to see if you could obtain satisfying results with this small set of features. Discarding the non-continuous variables also allowed us to apply more Machine Learning models, since the most common ones can only deal with continuous variables.

These initial results were pretty bad (see Table 1). We verified that in general, every model had a pretty average balanced accuracy (BACC), low recall and the number of times that the models predicted that a patient would die were very scarce.

After that, we decided to apply the models for the complete dataset (the one with 115 descriptive variables), taking into account the limitations of having different data types. It is possible to observe that in general the results got better (see *Table 1*) for all the metrics presented. It's worth mentioning that for the ordinal variables, we decided to implement a cumulative form of one hot encoding, (present in the code provided). However, we verified that the machine learning models didn't produce good results, probably because the dimensionality increased a lot. In fact, for any variable with K labels it would create K-1 new variables. Therefore, we did not use it, maintaining the ordinal variables coded as integers.

Table 1: Supervised Learing results obtained in the first two datasets

	Numerical Variables Dataset				**All Variables Dataset**			
	BACC	**Precision**	**Recall**	**F1-Score**	**BACC**	**Precision**	**Recall**	**F1-Score**
Decision Tree	0.592	0.522	0.222	0.312	0.719	0.719	0.463	0.581
Random Forest	0.584	0.382	0.241	0.295	0.612	0.583	0.259	0.359
MLP	0.564	0.571	0.148	0.235	0.818	0.800	0.667	0.727
LDA	0.567	0.667	0.666	0.242	-	-	-	-
QDA	0.603	0.565	0.241	0.338	-	-	-	-

This lead us to conclude that the non-continuous variables are also important for the outcome we want to predict. More information about hyparparameter tunning or other metrics can be found in Notebook_1 and notebook_2 for these two scenarios, respectively.

4.4 Feature Selection

Since we had concluded that it was important to include some non-continuous variables in the analysis, the question of what were these relevant features soon emerged. For this reason we applied feature selection in order to gain some new insights about what were these relevant (or irrelevant) features.

4.4.1 Boruta

Therefore, we used Boruta [6], a relatively recent technique that helped us select features for a classification problem. Explaining it simply, it consists of a Random Forest algorithm which is fed with the initial variables and also with some shadow features, which represent copies of the initial variables that are randomized. The importance of the initial variables is then compared with a threshold, which is the highest feature recorded among the shadow features. The initial variable is only considered useful if it performs better than this shadow feature. This process is iterated a number of times (in our case, 200). In the end, it may reject some variables (in red), it might keep others (in green) and it may have an inconclusive answer about others (Tentative variables, in yellow). Blue boxplots correspond to minimal, average and maximum Z score of a shadow attribute

In the end, from the 115 descriptive variables that we had we only kept with 39 (4 continuous, 11 ordinal and 25 categorical), as seen in *Figure 3*.

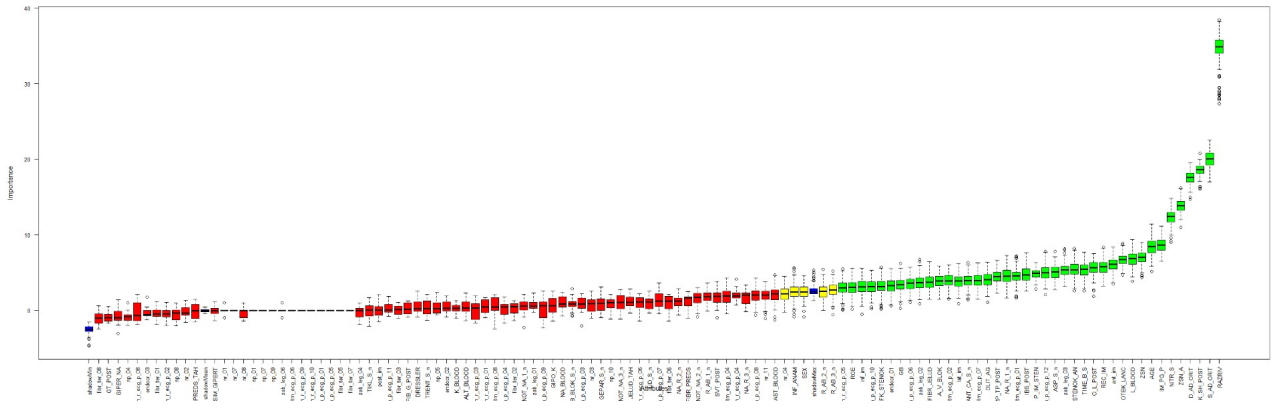


Figure 3: Boruta resulting variables in green

Considering the changes between the initial dataset and the Boruta one, we verify several improvements. Exemplifying, a Decision Tree Classifier, has BACC of 0.759 (4 percentual points higher than the one of the initial dataset) and the Recall is of 0.57 (+ 11 p.p. than the initial dataset). Even though the precision got worse, the F1 score is of 0.61 (3 percentual points higher than the one of the initial dataset). The model is now more capable in saving lives, while using much less information, which is a positive improvement. The top of the Tree can be seen in *Figure 4*.

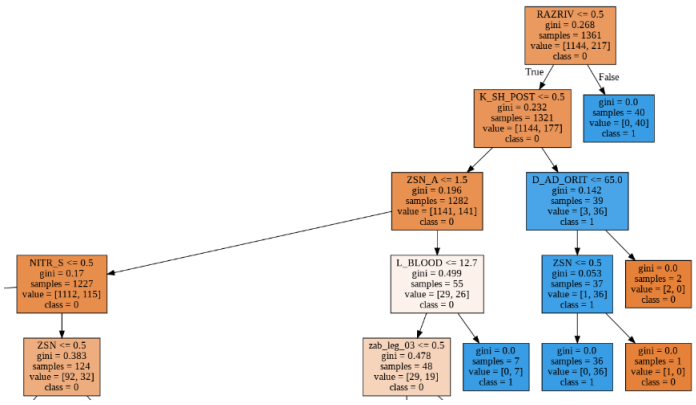


Figure 4: Detail of the truncated Decision Tree with the Boruta Dataset

The root node, right in the top of the tree, is *Razriv*. This feature leads to a pure node, that indicates us that a person will die if it has a value of 1 (signaling that the patient had had a myocardial rupture). For the individuals that didn't had this symptom, we can visually find the next important features such as *K_SH_POST*, *D_AD_ORIT*, *ZSN_A* (cardiogenic shock, diastolic pressure and heart failure, respectively). These are some of the features that very easily lead us to pure nodes. Naturally, these are factors that the common sense tells us that are important to predict if a person will die. However, since we don't have domain expertise about this subject, we will verify some of these insights with experts later in the report.

Other methods suitable for this dataset in particular were also applied and their results can be found on *Notebook_3*.

4.5 Dimensionality Reduction Techniques

In order to reach even more relevant results and predictions, we first applied the classical Multidimensional Scaling (MDS), using the Gower Distance, as a consequence of having categorical variables. The results we obtained with this method were not satisfactory. This was probably due to the fact that we only kept two dimensions with the objective of being capable to visualize the data in a 2D plot. We then decided to use Factor analysis of mixed data (FAMD) [7], a method that considers both quantitative and qualitative variables and thus would take in consideration all our variables.

4.5.1 FAMD

Roughly speaking, the FAMD algorithm can be seen as a mix between principal component analysis (for numerical variables) and multiple correspondence analysis (for categorical varibles). Not only does it allow to maintain a good portion of the information of the initial variables as it also generates new real dimensions, making the data more suitable to apply a bigger panoply of machine learning models.

Since a lot of the dimensionality was already reduced through the Boruta method, we decided to keep as many dimensions as the original variables that we had.

Consequently, the best results we obtained were produced by Linear Discriminant Analysis (LDA), a machine learning algorithm, while using the Boruta dataset with the FAMD method as dimensionality reduction technique, as seen in *Table 2*. As it can be seen, our dataset was very noisy in the start, and rather difficult to analyse due to the structure of its variables. As a Machine Learning algorithm, the LDA proved to be the most efficient method around, because despite being rather simple in its nature - and linear -, it provided a great generalization for the data. It is worth mentioning that QDA is the top classifier in terms of recall, but it has a lower Precision, so it provides more false alarms than the other methods. All the other non-linear models, most likely, overfitted to the training data, which translates in a poorer generalization power.

Table 2: Supervised Learing results obtained in the Boruta dataset with FAMD

	Boruta Dataset			
	BACC	**Precision**	**Recall**	**F1-Score**
Decision Tree	0.753	0.524	0.611	0.564
Random Forest	0.794	0.673	0.648	0.660
MLP	0.822	0.691	0.704	0.697
LDA	0.840	0.844	0.704	0.768
QDA	0.844	0.581	0.796	0.672

Since the LDA and QDA provided some really good results, we decided to implement a stacking classifier that combines the power of this two previous powerful models. We also used the default final estimator, which is a simple logistic regression model. Although it achieved good results (see *Table 3*), it didn't "combine" the good precision and recall from LDA and QDA, respectively. More information about these results can be found in *Notebook_4*.

Table 3: Supervised Learing results for the Stacking Classifier, obtained in the Boruta dataset with FAMD

	Boruta Dataset			
	BACC	**Precision**	**Recall**	**F1-Score**
Stacking Classifier	0.832	0.860	0.685	0.763

To finalize this part, we decided to remove the outliers from this last dataset and repeat the analysis. For our surprise, the linear model and QDA were the only ones to obtain satisfying results, even though poorer than the previous ones (see *Table 4*).

Table 4: Supervised Learing results for LDA and QDA, obtained in the Boruta dataset with FAMD but without RPCA outliers

	Boruta Dataset			
	BACC	**Precision**	**Recall**	**F1-Score**
LDA	0.784	0.833	0.588	0.690
QDA	0.812	0.561	0.725	0.632

It is worth mentioning that 26% from this outliers were people that died, and they were all in the training set of all the previous results, so they probably had certain attributes on these continuous variables that were important for

the models, although they were outliers in the mathematically sense. More information, in particular about the other models, can be found in *Notebook_5*.

5 Unsupervised Learning

Unsupervised Learning has the goal to describe the structure of the data as well as its associations and patterns without *a priori* knowledge about its nature [8]. For this purpose, it may perform a variety of tasks with particular emphasis in cluster analysis. The cluster analysis task, in its turn, can be performed by several algorithms, being the hierarchical clustering methods and the partitioning methods the most widely studied ones, due to their simplicity and ease of implementation when compared to other clustering algorithms [9]. In this section, both hierarchical and partitioning methods will be explored, under different circumstances, in pursuit of the best possible results in terms of finding relevant patterns or coherent structures. All the applied methods, with a specific exception we will be explaining ahead, will benefit from the feature selection mentioned in previous sections. In terms of cluster validation, we will give high priority to external validation measures (rather than internal validation measures). Our main established objective here is to find clusters that match external information. In particular, we will calculate confusion matrices to observe how good are the different algorithms in finding patterns between individuals that live or die. We also calculated the Dunn Index as an internal validation measure but we will not give it too much importance in this analysis.

5.1 Hierarchical Clustering Methods

Hierarchical clustering algorithms address clustering problems by yielding a dendrogram, that can be split at different levels, according to the desired number of clusters. The dendrogram construction can happen in two ways, by starting to consider all the features as an individual cluster (agglomerative hierarchical clustering, AGNES) or by starting to consider them as a single big cluster (divisive hierarchical clustering, DIANA). The first approach reaches a final dendrogram by promoting the fusion of the individual clusters, iteratively, according to a proximity criterion, until there is only one cluster left, while the second approach, conversely, reaches a final dendrogram by splitting, iteratively, the initial single cluster until having a bunch of them. In this work, they will be applied only AGNES methods, namely single linkage, complete linkage, average linkage, ward's and diana.

5.1.1 Strategy driven by data-related concerns

Since the data contains both numerical and categorical variables, the distances had to be chosen carefully. We could have just performed clustering on the whole data using a compatible metric, such as gower distance, but we tried to take advantage of this mix data composition by performing clustering by parts (perform clustering not only on the data set as a whole but also on the two types of features - categorical and numerical - independently). Hence, we used both euclidean and manhattan distances to cluster only the numerical features and gower distance to cluster only the non-numerical features and also the dataset as a whole.

Exceptionally, clustering analysis on numerical-only and on the whole dataset were run using our original imputed dataset, that contains 115 descriptive variables, because it gave us better results in this case.

Due to its suitability for categorical variables, we also tried the ROCK method (Robust Clustering using Links) to cluster only the non-numerical features. The ROCK algorithm uses the L_p^3 metric or the Jaccard coefficient instead of using the distance measures to find the similarity between the data points. However, for non-numeric attributes, these distances will not be used because of their unsuitability. Thus, we will use the notion of links between the points of the dataset. For this, we will consider that a pair of points are neighbors if their similarity exceeds a certain threshold. The number of links between a pair of points is then the number of common neighbors for the points. An advantage of this method is that, unlike distances or similarities between a pair of points which are local properties involving only the two points in question, the link concept incorporates global information about the other points in the neighborhood of the two points. For the sake of understanding and simplification of the report, the algorithm will not be detailed.

5.1.2 Finding the ideal the number of clusters

First we determined the best number of clusters, using the function NbClust (see *Table 5*). This function does not work for Diana, Weighted, Gaverage, and Flexible methods, so for ones we used internal criteria imposes on each method, with k restricted between 2 and 20. Note that for these methods we used the original dataset, without missing values.

	Euclidean distance					**Manhattan distance**					**Gower distance**							
	S.	C.	A.	W.	D.	S.	C.	A.	W.	D.	S.	C.	A.	W.	D.	We.	G.	F.
Silhouette	2	2	2	2	2	2	2	2	2	2	2	8	2	2	2	2	2	2
Mcclain	2	2	2	2	20	2	2	2	2	20	2	2	2	2	20	2	5	16
Dunn	2	2	2	4	9	2	2	2	5	11	2	8	2	2	9	2	10	4
Cindex	5	2	5	10	20	4	2	5	2	20	10	5	10	6	20	9	20	19

Table 5: Results from the best number of k clusters to S-Single linkage; C- Complete Linkage; A-Average Linkage,; W-Ward method; D-Diana; We-Weighted; G-Gaverage;F-Flexible .

These results lead us to assume a k = 2 for the other methods that require a number of clusters as an input, since it's that most of the previous indexes suggest us. However, to consider k higher than 2 poses a interesting possibility, because it can express a different degree of state. Later, we will approach this issue.

5.1.3 Overall results

Numerical variables results: For the numerical variables (see *Table 6a*), when using the Single Linkage with Euclidean distance the results were bad, since almost every data point ended up in the same cluster. When the Average Linkage was used with euclidean distance, it had 92% of the people who died in the cluster where all the people that lived ended up. Furthermore, the second cluster only contains people that died, so this second cluster seems really good in finding key characteristics in people that died.

Non-Numerical variables results: As it can be seen in *Table 6b*, the results for the Rock algorithm weren't satisfying. In opposite, the results for Complete Linkage were a little bit better.

Whole dataset results: Analysing the whole dataset results (see *Table 7*), we can see that the single, average and weighted linkage place every individual in same cluster, which is not the most relevant result. The method that yields better results is *Gaverage Linkage*, which was able to outperform the Complete Linkage applied to categorical features alone (see *Table 8*).That seems to agree with what we saw in *4.3*: *"it is important to consider both continuous and non-continuous variables in the analysis"*, because the Gaverage method with all dataset has better results than the categorical alone.

NOTE: Each clustering method generates 2 clusters and we don't know which one corresponds to those who live or die to perfom cluster validation. Thus, from now on we assume as the alives cluster the one that shows a higher proportion of living samples.

Table 6: Resume of the best results, based on Recall, for hierarchical methods using...

(a) ... only the continuous variable

		Clusters			
		Single		Average	
		①	②	①	②
True Prognosis	Alive	1428	1	1429	0
	Death	271	0	250	21

(b) ... only the categorical feactures

		Clusters			
		Complete		Rock	
		①	②	①	②
True Prognosis	Alive	1167	262	1429	0
	Death	168	103	270	1

Table 7: Resume of the overall results for hierarchical methods focusing in the whole dataset

		Clusters													
		Single		Complete		Average		Ward		Weighted		Gaverage		Flexible	
		①	②	①	②	①	②	①	②	①	②	①	②	①	②
True Prognosis	Alive	1429	0	794	635	1428	1	1124	305	1429	0	1062	367	911	518
	Death	270	1	184	87	270	1	135	136	270	1	99	172	91	180

Table 8: Resume of the cluster validation for the best hierarchical clustering methods (using external indexes and Dunn)

	Recall	F1 Score	Accuracy	Precision	Dunn
Complete (for categoricals only)	0.38	0.32	0.75	0.28	0.103
Gaberage Link (for whole dataset)	0.63	0.42	0.54	0.32	0.094

5.2 Partitioning Clustering Methods

Partitioning clustering algorithms address clustering problems by optimizing a specific objective function and iteratively improving the quality of the partitions [9]. Unlike the hierarchical methods, they require the number k of clusters as input to know how many clusters they should generate, which implies *a priori* knowledge about the data structure. Since we are trying to distinguish death and alive cases, it would be logical to make k=2. Even though, the value of k will be confirmed using average silhouette index.

5.2.1 Strategy driven by data-related concerns

Once again we tried to take advantage from the mix composition in numerical and non-numerical features of the data, to also perform clustering by parts in addition to do it at once, hoping to find interesting results.

Focusing on numerical features alone: k-means and kernel k-means. On the 9 continuous variables of the complete dataset (the one with 115 descriptive variables), we applied k-means algorithm, that uses the euclidean distance as its metric, and because of that, it is suitable do deal with numerical variables alone. We also tried the kernel k-means, which works like the usual k-means algorithm, but has an extension for identifying also non-linearly separable clusters. Normally, it has better results than k-means, however it was not the case in most metrics. We used the *rbfdot* kernel and the *laplacedot* kernel, and we had better results with the *rbfdot* kernel.

Focusing on non-numerical features alone: k-modes. K-modes is a variant of k-means that uses a different matching dissimilarity measure (where the distance between 2 data points X and Y is the number of observations in X and Y whose values are different) and modes instead of means, which makes it suitable to deal with categorical variables instead of numerical ones [10]. Apart these differences, the k-modes works similarly to k-means, which could be interesting from the analysis point of view. The results were quiet similar to the Kernel k-means' ones.

Focusing on data as a whole: k-prototype and k-medoids. K-prototype is also a variant of k-means which is able to handle the whole data, regardless the type of features (numerical or categorical), because it uses a dissimilarity measure that takes into account both the squared euclidean distance (dissimilarity measure for numerical features), s^r, and the number of mismatches of categories between two objects (dissimilarity measure for categorical features), s^c, using a weight,γ, to balance the two dissimilarities, according to the expression: $s^r \pm \gamma s^c$. To apply the k-prototype algorithm, we used the function *kproto()* avaiable in the package *clustMixType* [11].

Apart from these differences, as it happened with k-modes, the k-prototype also works similarly to k-means. Thus, in order to have at least one algorithm that does not work like k-means, we decided to apply also k-medoids (with Gower distance as the metric), that besides being a k-means variant top, it works differently by choosing data points as the prototypes instead of using virtual ones to make the clustering task, which makes it less sensible to outliers [9]. The k-medoids was applied using the Partioning Around Medoids (PAM) algorithm and to support in the visualization, it was applied the t-SNE method.

5.2.2 Partitioning Clustering Methods: Overall Results

All the partitioning methods have shown consensus on 2 as the ideal number of clusters (determined by average silhouette index). Thus, that is the chosen K for all the algorithm (which makes sense given we are trying to distinguish between live or death).

The overall results were not bad at all (see *Table 9* and *Table 10*). In in this partitioning clustering context, putting the focus only on continuous features with kernel k-means, got the best results and proved that performing clustering "by parts" was a wining strategy in this case. Focusing on the non-numeric features alone with k-modes was not so good as kernel and k-means but showed a tiny better cluster cohesion according to its Dunn index. Looking to the dataset as a whole looked to be the worst strategy here. One possible explanation for the success of kernel k-means could have been the fact that the numeric features, beyond their probable relevance, represent a much lower proportion of the data, being less prone to contain irrelevant data.

Recapping the results from hierarchical methods (in *Table 8*), we see that one of the best partitioning method (kernel k-means) has a similar performance compared to the best hierarchical method (Gaverage Linkage). In particular, the kernel k-means' recall (the most relevant metric) was better. Despite of the fact that the hierarchical methods have also some good results, like the Gaverage, they needed all the categorical features on the whole dataset to achieve their best results - the partition methods with the 9 continuous variables achieved the best recall.

Table 9: Resume of the overall results for partitioning clustering methods

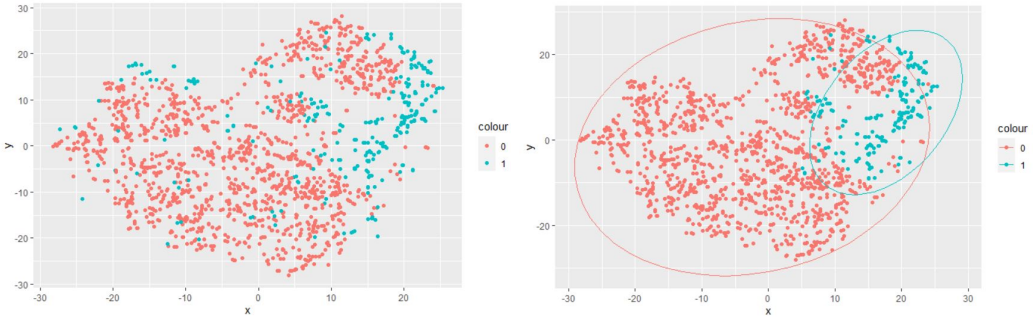
		Clusters									
		K-means		Kernel k-means		K-modes		K-prototype		K-medoids	
		①	②	①	②	①	②	①	②	①	②
True Prognosis	Alive	1209	220	733	696	766	663	739	690	722	707
	Death	152	119	74	197	104	167	156	115	166	105

Table 10: Resume of the overall results of partitioning clustering methods (using external indexes and Dunn)

	K-means	Kernel k-means	K-modes	K-prototype	K-medoids
Recall	0.44	0.73	0.62	0.42	0.39
F1 Score	0.39	0.34	0.30	0.21	0.19
Accuracy	0.78	0.55	0.55	0.50	0.49
Precision	0.35	0.22	0.20	0.14	0.13
Dunn Index	0.0298	0.0111	0.0709	0.022	0.0084

5.3 Manual clustering with t-SNE

In this part, the objective is to find a way to split the data "manually" thanks to t-SNE. The t-SNE method allows to project the points of the dataset in a low dimensional space. In this procedure, the issue is to know if the fact of projecting the data in a 2-dimensional space makes it easier to perform clustering (see *Figure 5*). As a first step, let's perform a projection of the data in a 2-dimensional space using the t-SNE method and observe the distribution of individuals in each class.



(a) Projection of the dataset into 2D space (t-SNE).

(b) Projection of the dataset into 2D space (t-SNE), after filtering on class 1 individuals.

Figure 5: Scatter plot, projection of data with t-SNE

The figure on the left is the projection of the initial dataset obtained with t-SNE. We clearly observe a cloud of blue points (class 1 individuals) at the top right. The rest of the class 1 individuals are scattered over the remaining space.

Thus on the right, we have the same projection after filtering out the blue points (class 1) that do not seem to belong to the identified point cloud. By adding ellipses, we clearly see a cluster with a high concentration of blue points.

Thus, the next step is to identify the parameters of this ellipse to be able to identify if a point belongs or not to the ellipse, and thus to the identified cluster (of class 1).

After determining the parameters of the ellipse, it performs a manual cluster method: if the point is in this ellipse, then it will be added in the cluster of class 1. If it is not in the ellipse, then it will be added in the cluster of class 0. The results can be sen in *Table 11*. We also obtained a Dunn Index of (0.0042).

Table 11: Confusion matrix and overall results for Manual Clustering with t-SNE

		Clusters	
		①	②
True Prognosis	Alive	1210	219
	Death	123	148

	Recall	F1	Accuracy	Precision	Dunn
t-SNE	0.55	0.46	0.80	0.40	0.0042

6 Supervised Learning on Clustering Results

In the previous sections we verified that the clusters didn't separate labels of the variable of interest in the best way. Nevertheless, we decided to apply supervised learning methods to predict cluster labels in order to check if there is potential in this technique. It can be seen has motivation for future works where a good partition of a variable of interest was achieved.

Since we generated several partitions using different datasets, here we decided to use two that we particularly liked. The first one is the cluster labels from Hierarchical Clustering using Gaverage and the other one is the cluster labels from TSNE results.

When compared to the results obtained in the supervised learning section, these results are extremely better. In particular, for the t-SNE scenario, it's easy to understand why non-linear models work and linear models fail. The visualization of Figure 5 shows us that the data is not linearly-separable, but for most non-linear models it's possible to obtain a decision boundary in such a way that the classification is really good for both the ones that live and die. The linear models has, with no surprises, average results.

This circumstance is due to the fact that clustering techniques are able to find patterns and structure in the data - but this exactly what traditional machine learning models search for when defining their classification boundaries. For this reason, the good results we obtained are actually not that surprising. We advise the reader to check the results in a more detailed way on notebook_6 and notebook_7.

Table 12: Supervised Learning results obtained on cluster labels prediction

	Gaverage Hierarchical Clustering				t-SNE			
	BACC	**Precision**	**Recall**	**F1-Score**	**BACC**	**Precision**	**Recall**	**F1-Score**
Decision Tree	0.903	0.854	0.863	0.859	0.963	0.985	0.930	0.957
Random Forest	0.889	0.859	0.832	0.845	0.977	0.986	0.958	0.972
MLP	0.934	0.955	0.884	0.918	0.974	0.958	0.958	0.958
LDA	-	-	-	-	0.692	0.537	0.500	0.518
QDA	-	-	-	-	0.784	0.750	0.625	0.682

7 Discussion

In order to understand whether the variables that were selected through dimensionality reduction made sense, and also to verify the relevance of the decision tree built from our dataset, we decided to promote a discussion with a group of students from Nova Medical School, the Faculty of Medical Sciences from Universidade Nova de Lisboa.

Working closely with these students, we concluded that from the 39 variables that composed our final dataset, 37 were very relevant for the final outcome of the patient, while the other 2 were only moderately relevant.

Regarding the decision tree that we used for classification, the root node regarded the presence of a myocardial rupture, which, according to the tree, would cause certainly the death if it was present. Here, our colleagues were peremptory in their answer: "the Myocardial rupture is a very serious and life-threatening condition in which there is an acute rupture of, usually, the atria and/or ventricles (heart cavities), causing a massive bleeding and sudden death. It can occur as a complication of stroke, usually in the following 24 to 72 hours" [12]. The other relevant nodes (cardiogenic shock, diastolic pressure and heart failure) made also a lot of sense, according to our colleagues.

In the Appendix, we present the full version of the remarks and relevance of every variable from the dataset.

8 Conclusion

Since the subject of myocardial infarction is an active area of research, due to its relevance in modern times, we end our analysis with some positive results that can have practical application for this urgent issue and be used as a benchmark for future research projects. We were able to find some reasonably good machine learning models that can serve as support decision tools for medical staff. At the same time, we got some insights about this domain and were able to validate this analysis with domain experts that reviewed some of the results we were obtaining.

There are multiple ways researchers can improve these results. For that, and in particular, we have suggestions that we will mention. We advise researchers to find a better way to make ordinal encoding, since this can have a high impact in the model's results. We also advice researchers to try and use techniques like SMOTE, that are useful when training models with unbalanced target variables; we did not use it because we would be adding even more synthetic data to the analysis which can be a controversial thing to do, specially in such a delicate subject as this one. Also, Cluster Analysis can be heavily improved. There are other techniques worth applying here like Multidimensional Scaling (MDS), which can provide visual representation of the data and allow us to visualize clusters and investigate what variables are important in the cluster. Due to the good results on supervised learning over cluster labels, since we didn't obtain satisfying results with two clusters, maybe it is worth investing time in trying to find if there are more clusters. Maybe it is natural to have similarities between individuals the ended up living or dying - and thus more clusters can be found, helping to define medical staff's priorities and somewhat revolutionizing the efficiency of the hospital unit, saving time and lives.

References

[1] B. P. Griffin, *Manual of Cardiovascular Medicine*. Lippincott Williams amp; Wilkins, 2008.

[2] S. Golovenkin, J. Bac, A. Chervov, E. Mirkes, Y. Orlova, E. Barillot, A. Gorban, and A. Zinovyev, "Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data.," 2020. Available at https://archive.ics.uci.edu/ml/datasets/Myocardial+infarction+complications.

[3] "Myocardial infarction complications database (documentation)." Available at https://leicester.figshare.com/articles/dataset/Myocardial$_i nfarction_c omplications_D atabase$/12045261, (*accessed* : 24.01.2022).

[4] M. Hubert, P. Rousseeuw, and K. Branden, "Robpca: A new approach to robust principal component analysis," *Technometrics*, vol. 47, pp. 64–79, 02 2005.

[5] K. P. Shung, "Accuracy, precision, recall or f1?." Available at https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9, (accessed: 24.01.2022).

[6] S. Mazzanti, "Boruta explained exactly how you wished someone explained to you." Available at https://towardsdatascience.com/boruta-explained-the-way-i-wish-someoneexplained-it-to-me-4489d70e154a?gi=f2c6ff81471f, (accessed: 22.01.2022).

[7] J. Pagès, *Multiple Factor Analysis by Example Using R.* 2015.

[8] T. Hastie, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 01 2009.

[9] C. Reddy and B. Vinzamuri, *A Survey of Partitional and Hierarchical Clustering Algorithms*, pp. 87–110. 09 2018.

[10] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Min. Knowl. Discov.*, vol. 2, pp. 283–304, 09 1998.

[11] G. Szepannek, "clustmixtype: User-friendly clustering of mixed-type data in r," *The R Journal*, vol. 10, p. 200, 01 2019.

[12] L. M. N. G. Hutchins KD, Skurnick J, *Cardiac rupture in acute myocardial infarction: a reassessment.* 03 2002.

[13] M. J. Mulder, S. Ergezen, H. F. Lingsma, O. A. Berkhemer, P. S. Fransen, D. Beumer, L. A. Van Den Berg, G. Lycklama à Nijeholt, B. J. Emmer, H. B. van der Worp, *et al.*, "Baseline blood pressure effect on the benefit and safety of intra-arterial treatment in mr clean (multicenter randomized clinical trial of endovascular treatment of acute ischemic stroke in the netherlands)," *Stroke*, vol. 48, no. 7, pp. 1869–1876, 2017.

[14] D. G. Hackam and J. D. Spence, "Antiplatelet therapy in ischemic stroke and transient ischemic attack: An overview of major trials and meta-analyses," *Stroke*, vol. 50, no. 3, pp. 773–778, 2019.

[15] K. D. Hutchins, J. Skurnick, M. Lavenhar, and G. A. Natarajan, "Cardiac rupture in acute myocardial infarction: a reassessment," *The American journal of forensic medicine and pathology*, vol. 23, no. 1, pp. 78–82, 2002.

Appendices

Individual analysis of the variables of the dataset, from a medical standpoint - with Ana Sofia Faria, Inês Falcão, Leonor Carvalho, Luís Fialho, Pilar Pegas

Age: if someone is older, they will have less regenerative capacity and less flexibility to maintain homeostasis (balance of all the physiological patterns in the organism), making it harder to recovery from such a lesion to the myocardium

White blood cell count: a high WBC count is associated with inflammatory and/or infectious states. This might reflect a worst response from the organism, such as a reperfusion injury (state of oxidative stress and inflammation after revascularization by solving the obstruction of the occluded artery) or, eventually, post-infarct complications, such as hospital-acquired infections.

Exertional angina pectoris: angina that appears with exercise. It is the somatic manifestation of myocardial ischemia, reflecting a "bigger" ischemia, which is symptomatic

FC AP in the last year: someone with a previous history of AP is more likely to have a stroke. AP in the previous year reflects a prolonged state of ischemia to the myocardium, causing lesions previous to the stroke in the muscle, and worsening the recovery and prognosis

CHD: coronary disease that worsens the stroke

Essential hypertension: hypertension that does not have an identifiable origin, or cause; the one most people have. In the long term, it causes lesions to arterial walls, promotes atherosclerosis and enhances overall mortality. Opposed to secondary hypertension, which is the direct consequence of another condition or disease.

Duration of HTA: longer corresponds to more lesion to blood vessels and tissues, particularly myocardium

Chronic HF: reflects a myocardium that, by its own, is not capable of pumping the blood well enough to the rest of the body. If present in the anamnesis, difficults the stroke recovery.

Anterior, lateral or inferior stroke: all left ventricular. Left heart (left ventricle) is the part of the blood system responsible to pump blood to the body (whereas the right heart pumps blood to the lungs, requiring less pressure). Left heart strokes are worse than right heart strokes, so the presence of this variable makes a lot of sense.

Time elapsed from the beginning of the attack of CHD to the hospital: more time corresponds to a bigger damage. Essential measure.

Use of opioid drugs in the ICU in the first hours of the hospital period: while with a slight relation, it is a disposable measure. Opioids: a group of potent analgesic drugs used to reduce very strong pain. They might be used to reduce chest pain in the event of a stroke. When used as an abuse substance, and injected, they might promote the appearance of infectious endocarditis by repeated injections.

Diabetes mellitus in the anamnesis: diabetes favors aterosclerosis. Worse injury recovery for heart problems.

Obstructive chronic bronchitis in the anamnesis / Bronchial asthma in the anamnesis : disease that results, for example, from chronic smoking. Associated with chronic and systemic inflammation, which worsens prognosis. Overall, patients with COPD have other risk factors, such as hypertension and atherosclerosis.

Pulmonary edema at the time of admission to intensive care unit: pulmonary edema is the reflexion of heart failure, which, in turn, is a reflexion of a more severe stroke, with a hemodynamic impact and without ability to preserve heart function.

Cardiogenic shock at the time of admission to intensive care unit: Shock is a state of very low blood pressure with compensatory tachycardia (heart fastens trying to compensate). Because of a stroke, shock appears when there is heart failure, which means that the heart is unable to keep pumping blood to the tissues. This reflects a much worse stroke. Shock by itself is a very serious condition with a high rate of mortality.

Paroxysms of atrial fibrillation at the time of admission to intensive care unit, (or at a pre-hospital stage): atrial fibrillation is a state in which there is not an organized contraction of the atria. This leads to a turbulent flux of blood, which, in turn, is one of the main factors contributing to thrombosis. This enhances the risk of another ischemic event in the recovery of the stroke, worsening the prognosis.

Presence of a right ventricular myocardial infarction: while left heart strokes are worse than right ones, the fact that it is ventricular poses a special problem here.

ECG rhythm at the time of admission to hospital – sinus (with a heart rate 60-90): normal state ECG rhythm.

ECG rhythm at the time of admission to hospital – atrial fibrillation: similar to the one above.

ECG rhythm at the time of admission to hospital – sinus with a heart rate above 90 -tachycardia: a tachycardic state reflects an uncompensated heart, probably tending to enter or in shock, which, as we saw, is associated with overall higher mortality.

Paroxysms of atrial fibrillation on ECG at the time of admission to hospital: Paroxysms are not constant, so they do not pose a risk as big as the one of tachycardia. However, this is still a condition to take in account.

Third-degree AV block on ECG at the time of admission to hospital; Complete LBBB on ECG at the time of admission to hospital; Complete RBBB on ECG at the time of admission to hospital: RBBB: right bundle branch block LBBB: left bundle branch block AV: atrioventricular Third degree AV: complete block All these conditions represent problems in the electric cardiac flow, preventing the heart from contracting. They worsen the results significatively.

Ventricular fibrillation: the same as atrial fibrillation, but in the ventricles, which are responsible to pump the blood to the lungs and body. When an individual is in ventricular fibrillation, the heart is not working, causing cardiorespiratory arrest and, potentially, death.

Liquid nitrates: nitrates are used in the event of a stroke to enhance vasodilation, both in the body (systemic) and coronary circulations, decreasing the amount of "effort" that the heart needs to do to effectively pump the blood. This way, it decreases the amount of oxygen the myocardium needs to work properly. As a side effect, it might present arrythmia (irregular cardiac rhythm), causing secondary problems to its use. It might also make some organs, such as the spleen and gastrointestinal tract, receive less blood (by reflex vasoconstriction).

Calcium channel blockers (CCB): drugs that are used to control arterial hypertension and to prevent primary and secondary (stroke that occurs a short time after a primary stroke) stroke.

Regarding nitrates and CCB: if used in the ICU (intensive care unit), they are used to lower blood pressure. Lowering BP in the event of an acute stroke is known to be safe, however, studies have shown a neutral or negative effect in patients treated with BP-lowering drugs. [13]

Systolic blood pressure: 3 in 4 stroke patients have increased blood pressure by an unknown mechanism, which is known as post-stroke hypertension. Patients with high BP (134,59 mmHg is slightly above upper limit of normal – 120 mmHg) are known to have a worse outcome. https://www.ahajournals.org/doi/full/10.1161/strokeaha.117.017228

Use of acetylsalicylic acid in the ICU: Acetylsalicylic acid (or aspirine) is used in low doses to prevent the formation of blood clots that could potentially cause a thrombotic event. It is used to prevent stroke.

However, it should not be used in the event or treatment of a stroke (in this case, the ICU). Despite the majority of the strokes being caused by atherosclerosis, a stroke could also be caused by a ruptures vessel (hemorrhagic stroke). In this case, it could worsen the clinical situation. It might be used to prevent a secondary stroke after the recovery of the first. [14]

Myocardial rupture is a very serious and life-threatening condition in which there is an acute rupture of, usually, the atria and/or ventricles (heart cavities), causing a massive bleeding and sudden death. It can occur as a complication of stroke, usually in the following 24 to 72 hours. [15]

Relapse of the myocardial infarction: recurrent stroke, defined by a second episode shortly after the primary stroke. It is associated with a increased morbility and mortality, and overall worse prognosis.

Post-infarction angina: chest pain caused by reduced oxygen supply to the cardiac muscle, occurring in the context of a stroke. It reveals a defective blood supply in the coronary circulation, even after the primary event is solved, associated with increased morbility and mortality.