Universidad
Carlos III de Madrid


# STATISTICAL LEARNING

## Part 1: Statistical tools

Student: Ilán Francisco Carretero Juchnowicz
Professor: Francisco Javier Nogales Martín Academic course: 2020/2021

# Contents



# List of Tables



# List of Figures

# 1 Database, preprocessing and analysis

## 1.1 Context and motivation of the work

Cardiovascular diseases are one of the leading causes of death worldwide annually. Mainly myocardial infarcts and heart failure (HF) can be highlighted. In this way, the latter occur when the heart is not able to pump enough blood to meet the vital needs of the human body. Thus, in the following work, it is intended, from the electronic medical records of multiple patients, to find correlations and hidden patterns through statistical learning techniques in order to be able to classify the survival of patients based on their characteristics and also to extract which are really the most relevant characteristics when predicting the survival of a patient.

## 1.2 Database

In the present work, a dataset with a total of 299 patients with heart failure collected in 2015 has been analyzed. In this way, it is composed of the following features as shown in Table 1.

| Feature | Definition | Measurement | Range |
|---|---|---|---|
| Age | Age of the patient | Years | [40,95] |
| Anemia | Decrease of red blood cells or hemoglobin | Boolean | 0,1 |
| High blood pressure | If a patient has hypertension | Boolean | 0,1 |
| Creatine phosphokinase (CPK) | Level of the CPK enzyme in the blood | mcg/L | [23,7861] |
| Diabetes | If the patient has diabetes | Boolean | 0,1 |
| Ejection fraction | Percentage of blood leaving the heart at each contraction | Percentage | [14,80] |
| Sex | Woman or man | Binary | 0,1 |
| Platelets | Platelets in the blood | kiloplatelets/mL | [25.01,850] |
| Serum creatinine | Level of creatinine in the blood | mg/dL | [0.50,9.40] |
| Serum sodium | Level of sodium in the blood | mEq/L | [114,148] |
| Smoking | If the patient smokes | Boolean | 0,1 |
| Time | Follow-up period | Days | [4,285] |
| (target) death event | If the patient died during the follow-up period | Boolean | 0,1 |

**Table 1:** Meanings, measurement units, and intervals of each feature of the dataset

## 1.3 Preprocessing

First of all, it was checked if there were missing values in the database. In this way, there were none. However, if there were, the Predictive Mean Matching technique would have been used to replace them, where the variable where there are such missing data is adjusted to a distribution as similar to the existing data and replaced by observations that correspond to such distribution. In this way, the variability of the sample is preserved and possible correlations with other variables are not destroyed.

On the other hand, attempts have been made to detect outliers in order to be able to treat them. To do this, it was necessary to view the variables that we had to see their distribution. In this way, we have tried to carry out transformations to have distributions that approximately behave like a Gaussian and thus be able to apply detecting through the distance of mahalanobis outliers in order not to have observations that hide or destroy the possible correlations between variables. However, despite having tried to perform multiple transformations (e.g., logs, inverse, squaring) the distributions did not quite resemble a normal one. Likewise, the Shapiro-Wilk test was carried out to contrast normality in the data set and in no variable could the hypothesis that the data did not come from a normal distribution be rejected.

On the other hand, we also had the problem of having a relatively small database, and therefore, it is vitally important to keep as much data as possible. Consequently, it has been decided to work with all the observations taking into account that it is possible to have outliers in the data set to be analyzed. Thus, having discussed what to do with the possibility of missing values and outliers, a first analysis of the variables has been carried out.

## 1.4 Analysis of the variables and their respective correlations

To carry out this first analysis, a structured methodology has been carried out following these steps:

- Analysis of the distributions of the predictors as a function of the variable to be predicted (alive or dead). In the case of quantitative variables, the kernel density has been used, while in the case of qualitative variables a mosaic plot has been used (as it illustrates very well the frequency relationships between the qualitative predictors and the variable to be predicted).

- Analysis of the correlations between the predictors. For this, the pearson correlation and Kendall's tau have been used to study the linear and non-linear relationships between quantitative variables, the point biserial correlation to analyze the relationships between quantitative predictors and qualitative predictors, and the Cramer's V to analyze the degrees of association. among categorical variables.

- Analysis of the correlations between the variable to be predicted and the predictors. For this, the point biserial correlation has been used in the case that the predictors were quantitative and the Crammer's V in the case that the predictors were qualitative.

Regard to the visualized distributions, it can be seen that most of them do not have a clear distinction between the living and dead populations. Consequently, this gives us a first guideline that these are reasonably overlapping populations and that therefore, it will be necessary to use statistical learning techniques to extract relevant information regarding the prediction and classification of patients. However, it can be seen as for a variable such as time, if a certain differentiation is perceived between the classes of the variable to be predicted, from which we could affirm a priori, that it may be a relevant variable. Regarding the mosaic plots, it has not been possible to extract relevant information from any of them to discern from the qualitative predictor variables between the classes of the variable to predict.

Thus, regarding the correlations between the predictors, it can be concluded that there is no reasonably strong correlation between them. This can be considered as something positive, since we ensure in this way that our model does not suffer from multicollinearity that can later negatively affect the use of the statistical models to be applied.

However, it can be seen as for a variable such as time, if a certain differentiation is perceived between the classes of the variable to be predicted, from which we could affirm a priori, that it may be a relevant variable. Regarding the mosaic plots, it has not been possible to extract relevant information from any of them to discern from the qualitative predictor variables between the classes of the variable to predict. Thus, the point biserial correlation is a measure of the difference between the sample means relative to the sample variances of the groups of observations defined by the binary variable. Consequently, if this measure has a high value (close to 1), the difference between means will be very high and therefore possibly it helps us to classify our samples, then considering that variable as relevant.

Extrapolating these conclusions to our data set, we can affirm after having carried out a first descriptive analysis that very probably the **age**, **ejection_fraction**, **serum_creatinine** and **time** variables are the most relevant because they have a point biserial correlation of -0.25, 0.27, -0.29 and 0.53 respectively. In this way, these values are reasonably high in comparison with the other variables and consequently, these variables can be very useful for the classification of the variable to predict.

# 2 Classification modeling using statistical tools

In this section the statistical tools used for the classification and extraction of relevant features from the data set to be analyzed are exposed. To do this, the development and application of the Bayesian classifiers are exposed first and then the logistic regression.

Thus, in order to train the models and then test them, a partition has been made where 80% of the data has been used to train the models while 20% of the data has been reserved for testing. Likewise, repeated cross-validation with 10 folds and 5 repetitions has been used in order to obtain robust training results. Thus, the training set is divided into 10 random groups and 9 are used to train the model and 1 to validate it. The process is then repeated, changing the validation set and, once all the folds have been part of the valuation set, this same process is repeated 5 times. It is worth mentioning that the K-fold cross validation has been selected over other validation methods such as leave one out cross validation (LOOCV) since a greater compromise between the bias and the variance is reached.

## 2.1 Linear Discriminant Analysis

Using this Bayesian classifier, it is assumed that all covariance matrices are equal and that the different observations come from a normal distribution. Thus, thanks to such assumptions, some bias is introduced at cost in order to reduce the variance. Now, before running the model and analyzing the results obtained, we could ask ourselves the following: Are all the variables relevant to the model? Is it possible that the best model is not with all the variables?

Currently, there are several algorithms in order to be able to select the optimal number of variables or, at least, a model with certain variables that has better results than the same model with other variables or is more parsimonious.

Therefore, in order to be illustrative in this work, different types of variable selection have been used. In the case of the LDA, the Recursive Feature Elimination (RFE) has been used, which is nothing more than a simple backward selection. In this type of selection, initially all the available variables are used and then variables are removed one at a time in an iterative way until the minimum number of possible variables is reached, i.e. 1. Various criteria are used to remove variables. In our case, the accuracy has been used. Consequently, the algorithm obtains, starting from all the variables (p), the best model with p-1, p-2,. . ., p- (p-1) variables.

Therefore, the best LDA model has been obtained using the following variables: **time, serum_creatinine, ejection_fraction, serum_sodium, age**.

And its results have been the following:

| **Linear Discriminant Analysis** | Accuracy | Kappa | Sensitivity | Specificity |
|---|---|---|---|---|
| | 0.797 | 0.534 | 0.684 | 0.85 |

**Table 2:** Linear Discriminant Analysis evaluation metrics.

Surprisingly, we found reasonably good values, despite knowing that the assumptions made by this model are not correct in this case. Likewise, it is worth mentioning that the model has a greater capacity to discriminate negatives (live) than positives (dead), as seen in the sensitivity and specificity values.

## 2.2 Quadratic Discriminant Analysis

In this model it is still assumed that the observations come from a normal distribution. However, unlike LDA, all covariance matrices are not assumed to be equal. In this way, this model is more flexible, so it generally reduces the bias at the cost of increasing the variance.

Since the models presented are different, it is logical to think that each model selects certain variables to obtain the best possible classification. In this case, we have chosen to run the model initially with all the variables, then to visualize which variables have had a greater importance in the classification and, finally, to carry out a new model only with the most relevant predictors.

Thus, in this case, the predictors with which the model has been run have been: **time, serum_creatinine, ejection_fraction, serum_sodium, age**.

While the results of the model trained with the test set have been the following:

| **Quadratic Discriminant Analysis** | Accuracy | Kappa | Sensitivity | Specificity |
|---|---|---|---|---|
| | 0.814 | 0.555 | 0.632 | 0.9 |

**Table 3:** Quadratic Discriminant Analysis evaluation metrics.

Since this model is more flexible, in this case it is reasonable to think that the results are better than in the LDA case. In this way, it is verified that the accuracy is better. However, this model classifies the positives worse (total death and as observed in the sensitivity).

### 2.3 Naïve Bayes

Regarding this model, it stands out mainly because it assumes that all the variables are independent. In our case, given the correlations, it may be an appropriate model, but its performance will have to be analyzed. As in the LDA, RFE has been used as a feature extractor.

In this way, once the backward selection has been applied, it has turned out that the most relevant variables for the Naïve Bayes model are: **time, serum_creatinine, ejection_fraction, serum_sodium**.

And likewise, the performance of the model with the set test has obtained the following metrics:

| **Naïve Bayes** | Accuracy | Kappa | Sensitivity | Specificity |
|---|---|---|---|---|
| | 0.814 | 0.555 | 0.632 | 0.9 |

**Table 4:** Naïve Bayes evaluation metrics.

The evaluation metrics of the Naïve Bayes model are identical to those of the QDA model. Thus, a fairly good accuracy is observed but relatively low sensitivity values.

### 2.4 Logistic Regression

Logistic regression is applied when the variable to be predicted is binary. Thus, instead of taking such a variable as if it were continuous, the model is built taking into account that it is dichotomous and therefore, instead of modeling the probabilities as such, a quantity commonly called the odds is modeled. In our case, we have imposed that the log-odds or logit (p) be linear. In this way, logistic regression allows us to classify the different observations of the variables to predict from qualitative and quantitative predictors.

Furthermore, in order to obtain the best possible model, the best subset selection has been used as a feature extractor. In this algorithm, all possible number s of variables are tested and the best model is left with 1 to p variables. Although it is true that computationally it is much more demanding than for example backward or forward selection, it allows us to obtain the optimal model according to the relationship of variables imposed (in our case linear).

In this way, the model obtained through logistic regression and the application of the best subset selection has considered the following variables relevant: **ejection_fraction, serum_creatinine, time, age, sex**.

And it has obtained the following results:

| **Logistic Regression** | Accuracy | Kappa | Sensitivity | Specificity |
|---|---|---|---|---|
| | 0.831 | 0.612 | 0.737 | 0.875 |

**Table 5:** Logistic Regression evaluation metrics.

Logistic regression has some pretty good evaluation metrics. It is observed that the sensitivity and the specificity are more balanced in this case and also, although by a little, it is the model with the best accuracy.

## 2.5 Model comparison and selection

Once all the results have been obtained for all the models, it has been considered appropriate to compare and analyze the results obtained for each of them both in the training set and in the test set.

In this way, the results in the training set have been the following:

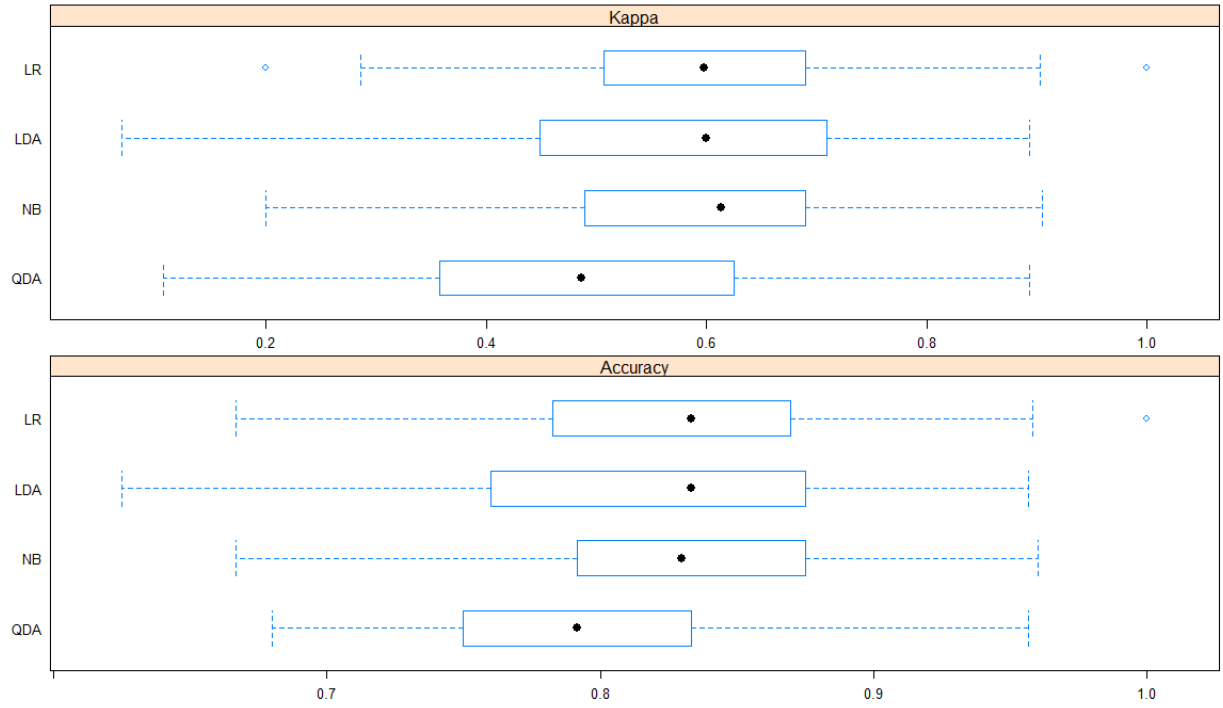


**Figure 1:** Results of accuracy and kappa for all the models for the training set.

As can be seen in Figure 1, the performance of the models in the training set is quite similar in terms of accuracy and kappa. Although it is true that the worst model in the training set is QDA, the rest of the models have similar values. However, if we take into account the variability of the results, according to the results of the a priori training set, we would be left with the logistic regression, as it is a model with good metrics and less variability in its results.

However, there are always variations (even if they are slight) between the predictions in the training set and the test set, since the models can be too inflexible, having too much bias, or on the contrary, be too flexible, generating overfitting. Thus, the results in the test set have been the following:

| **Models / Evaluation Metrics** | Accuracy | Kappa | Sensitivity | Specificity |
|---|---|---|---|---|
| Logistic Regression | 0.831 | 0.612 | 0.737 | 0.875 |
| Naive Bayes | 0.814 | 0.555 | 0.632 | 0.9 |
| Linear Discriminant Analysis | 0.797 | 0.534 | 0.684 | 0.85 |
| Quadratic Discriminant Analysis | 0.814 | 0.555 | 0.632 | 0.9 |

**Table 6:** Evaluation metrics for all the models in the test set.

In this way, it can be observed that in the test set the model with the best results is the logistic regression. Thus, although it is true that the accuracy of the four models is quite similar, the logistic regression stands out above the others for having the best kappa, for which it is the one that has learned the most to classify, and for its sensitivity , while in the specificity it has practically the same value as Naïve Bayes and QDA. It should be noted that, as discussed in the next section, the sensitivity of our model in this case is of great importance, since the false negative rate is the error that we want to prioritize in this case. It is also worth mentioning that, by relating our first hypothesis about the relevant variables with those selected by each model, we observe that in general all the models have selected them and in turn correspond to a large extent with those initially proposed. Therefore, this demonstrates the importance of carrying out the pre-modeling analyzes in order to understand why the models select some variables or others.

### 2.6 Cost-sensitive learning and final model

While it is true that the Bayesian classifier is the best of all possible classifiers, i.e. It is from which the best results are obtained in terms of accuracy, there are multiple situations in the real world where, perhaps, we are interested in having a worse accuracy but in return to be able to better classify a certain class.

In our case, interpreting the problem in question, the possible errors in this problem do not have the same cost. By this we mean that, if we find ourselves in a hypothetical situation where this model is used in a hospital as a clinical decision-making aid system, it is convenient for us that the model is more conservative and has a greater predictive capacity on those individuals at risk of dying than those with a higher probability of survival. Therefore, it is convenient for us that the error that the model may have regarding classifying a person as a survivor, when he or she can actually die, is minimal.

However, here it is necessary to reach a compromise situation, since, if we see the problem from an economic perspective, where a greater number of resources are added to each patient that the model classifies as dead, that the error of the model is High of predicting that someone will die when not really is also worrying.

Therefore, a cost-sensitive learning has been carried out where both possible errors have been penalized, although to a greater extent the error concerning classifying a person as a survivor when it really should be classified as dead (thus prioritizing the survival of those affected ).

In this way, cost-sensitive learning has given us the following result:

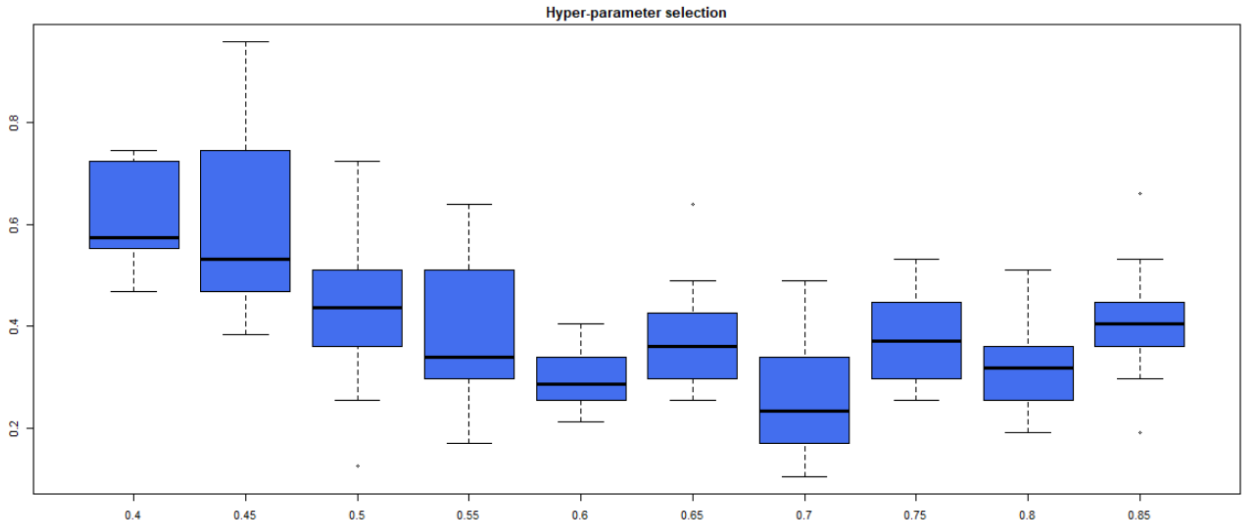


**Figure 2:** Cost-sensitive analysis.

Where the lowest cost threshold is 0.7. Consequently, the threshold for the selected logistic regression model has been modified and the following result has been obtained:

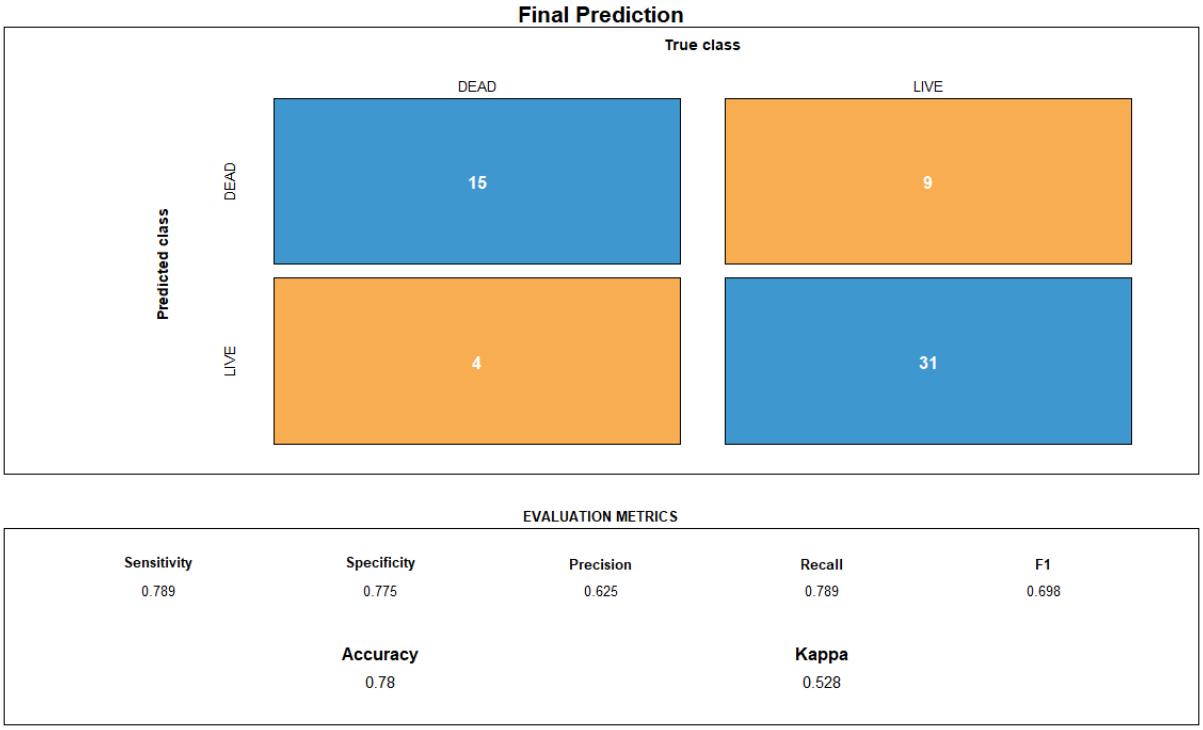


**Figure 3:** Final metrics of the logistic regression model.

Thus, both in the evaluation metrics and in the confusion matrix of our final logistic regression model, we observe that as the threshold has changed, the accuracy, kappa and specificity of the model have decreased slightly. However, we have slightly increased sensitivity thanks to the fact that our model has been designed to further reduce the rate of false negatives, which is the error associated with predicting that a patient will live when they actually die.

Finally, it is worth mentioning that by carrying out this work, what was learned has been put into practice both during the theoretical sessions and during the laboratories, where, based on a problem, statistical learning has been used to draw conclusions and modify the models to Get the best possible results based on what you want to prioritize.

## Bibliography

1. Chicco, D. & Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making* **20,** 16 (2020).
2. James, G., Witten, D., Hastie, T. & Tibshirani, R. *An introduction to statistical learning* (Springer, 2013).