UNIVERSITÉ DE LORRAINE

STATISTICS FOR HIGH DIMENSIONAL DATA

PROJECT REPORT

Classification of cardiac arrhythmias

Mously DIAW – M2 IMSD

Academic year 2020-2021

Table of contents

Glossary

ECG: Electrocardiogram

KNN: K-Nearest Neighbors

SVM: Support Vector Machines

PLS: Partial Least Squares

PCA: Principal Component Analysis

LDA: Linear Discriminant Analysis

CART: Classification and Regression Trees

GBM: Gradient Boosting Machine

ROC: Receiver Operating Characteristic

AUC: Area under the curve

List of figures

List of tables

1. Introduction

Heart arrhythmias are characterized by irregular heartbeats, which could also be too slow or too fast. In order to diagnose cardiac arrhythmia, the heart activity is analyzed by recording an electrocardiogram (ECG). The parameters of the ECG combined with patient information allows to detect and categorize arrhythmia. Unfortunately, false arrhythmia alarm rates as high as 88.8% have been reported in Intensive care units [1]. This has a negative impact on both patients and clinical staff and can result in true alarms being ignored.

Guvenir et al created an arrhythmia dataset available in the UCl Machine Learning Repository [2]. They aimed at detecting the presence of cardiac arrhythmias and classify them in 16 groups.

In this project, I will use the same dataset to create a **binary classifier** that will distinguish between normal and anormal arrhythmias. This approach is due to the severe class imbalance noticed in the dataset. Before selecting the best classifiers and evaluating their performance, I conducted the following: data preprocessing, exploratory data analysis, feature selection and model tuning. The results obtained at each step are detailed in the next sections.

2. Data preprocessing and exploratory data analysis

The original dataset has 452 patients and 279 variables, 206 of which are quantitative. The population is composed of 203 men and 249 women across all ages (0-83). After converting the class attribute into a binary variable, there are 207 anormal cases and 245 normal cases. The class attribute is therefore fairly balanced.

Figures 1 and 2 shows the boxplots for the variable *height* and *weight* respectively. The variable *height* has two aberrant **outliers** (>600cm). These values were selected by looking for values with a z-score>3. They were then replaced by the median. The outliers in *weight* are realistic so I didn't replace them.

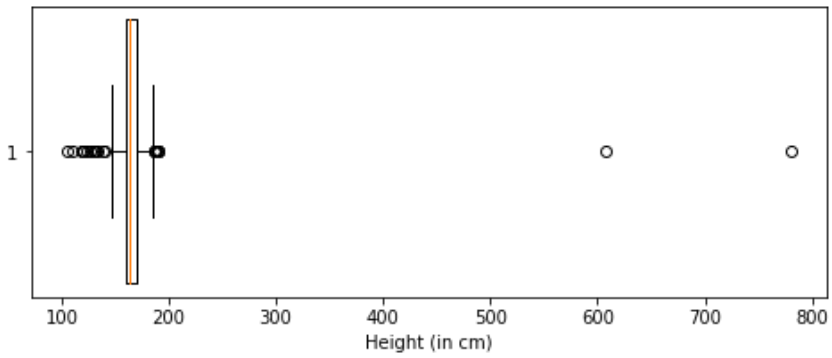

***Figure 1**: Boxplot showing the distribution of the height variable*

5 variables had **missing values**. I dropped the variable *J* where 83% of the values were missing. For the remaining variables that had a percentage below 5%, I performed KNN imputation. I also dropped the variables that had **zero-variance** or near-zero variance (when the ratio between the two most frequent values was above 20). Finally, the variables *II* and *IO* were dropped because of their **high correlation** (>0.9) with three other variables. **The final dataset has 142 predictors**.

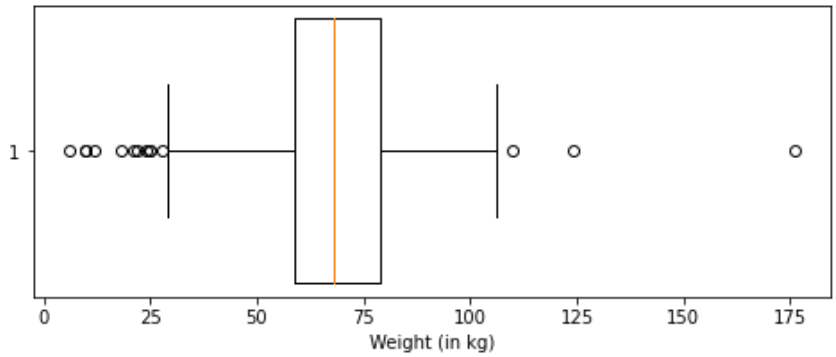

***Figure 2**: Boxplot showing the distribution of the weight variable*

We could expect the heart rate to be involved in the arrhythmia classifiers. Indeed, it is reported that, in some cases, the irregular heart rhythm could be either too slow (heart rate <60 bpm or beats per minute) or too fast (>100 bpm) [3]. Figure 3 shows the distribution of the heart rate in the cases of normal and anormal heart rhythms. As expected, for normal arrythmia the heart rate is roughly within the range 60-100 bpm. For anormal arrhythmia, the range goes beyond the normal range and the heart rate has several outliers. However, there is not an obvious rule to distinguish between anormal and normal rhythms by just looking at the heart rate.

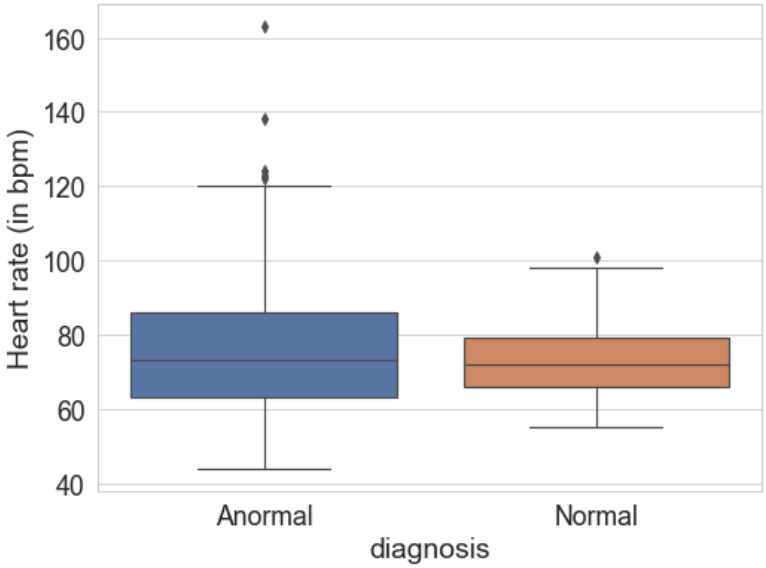

***Figure 3**: Comparison of the heart rate distribution between the normal and anormal heart rhythms*

The QRS duration is another important parameter in diagnosing heart failure. Figure 4 plots the heart rate against the QRS duration. We see that the two classes (*anormal* and *normal*) strongly overlap with the *normal* class contained within the *anormal* one. Therefore, it seems like classifiers such as linear SVMs or KNNs will not work well on this data. Radial SVMs could work but they will most likely have a lot of false negatives (*anormal* classified as *normal*).

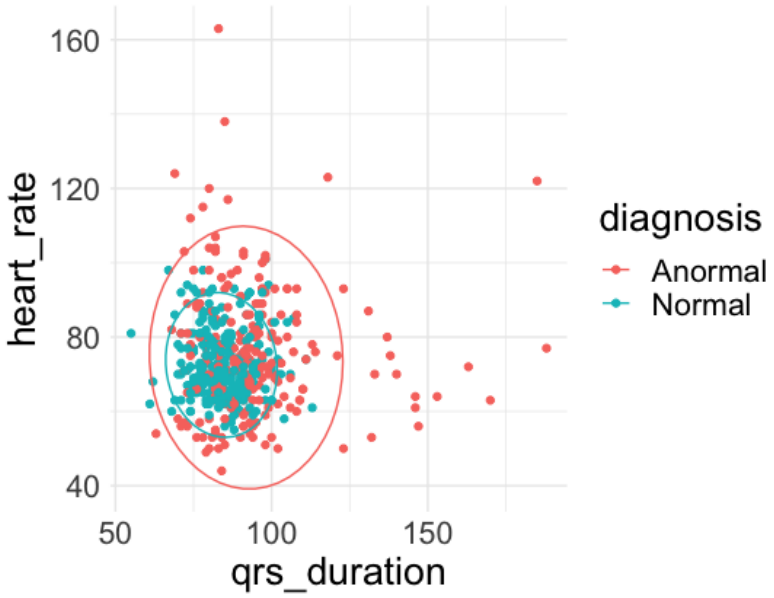

***Figure 4**: Heart rate against QRS duration by diagnosis outcome*

3. Data splitting and model strategy

The data has 452 samples available. 70% of them was used for training the algorithms while the rest was used to evaluate the final candidate models. Five repeats of 10-fold cross-validation were used to tune the models (50 resamples). AUC was chosen as the metric to maximize. Indeed, since the class attribute is balanced, AUC will optimize both classes, positive (*anormal*) and negative (*normal*).

Several models were fit to the training set. When model tuning was relevant, the parameters providing the highest average AUC was chosen for the final model. Table 1 shows the models tested in this project.

***Table 1**: List of the tested models, preprocessing techniques and tuned parameters*

Models	Preprocessing	Tuning parameters
Stepwise Logistic Regression	Forward stepwise selection + Center-scale (CS)	-
PLS	CS	ncomp = 1:10
LDA 1	CS	-
LDA 2	CS + PCA	-
Sparse LDA	CS	-
Linear SVM	CS	tunelength = 10
Radial SVM	CS	tunelength = 10
Polynomial SVM	CS	tunelength = 4
kNN	CS	k= 1:2:101
Random Forest	-	mtry = 1:15

Single CART tree	-	tunelength = 10
GBM	-	interaction.depth = 1:2:7 n.trees = 100:50:1000 shrinkage = 0.01,0.1 n.minobsinnode = 10

4. Results

Figure 5 shows the boxplots of the resampling estimates of the mean AUC for each model. KNN and linear SVM did not perform well on this data. Feature selection improved the performance of the linear discriminant model. The performance of the SVMs were highly improved by using kernel methods (polynomial and radial).

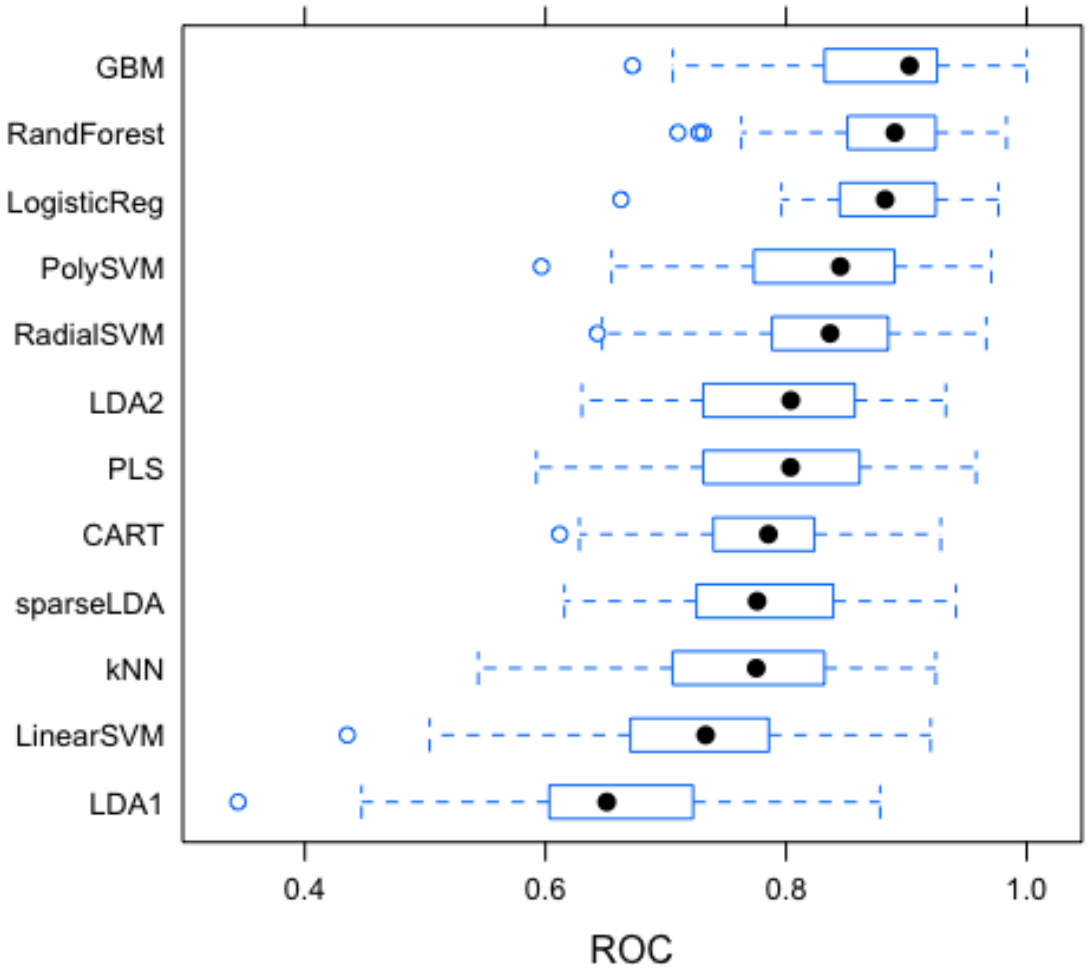

***Figure 5**: Average AUC resampling estimates for several models*

The top 4 models in terms of average AUC are GBM, random forest, logistic regression and polynomial SVM. A statistical test was conducted to see if the differences were significant. The results are shown in Table 2. For the top 3 models, the p-values for the model comparisons are at their maximum (p-value=1), which indicates that there is no significant difference between the 3 models. However, the polynomial SVM displays a significant difference in performance compared to the 3 others.

***Table 2**: p-values for the comparison of the top 4 models*

	LogisticReg	**RandForest**	**GBM**
RandForest	1		
GBM	1	1	
PolySVM	1.433e-06	1.404e-05	5.770e-05

Since the logistic regression, random forest and GBM models displayed a similar performance, I chose as **final candidate models logistic regression and random forest**. Indeed, GBMs are harder to tune (more parameters compared to random forests), they take longer to train and are more sensitive to overfitting.

5. Final models

The 2 final candidate models, logistic regression and random forest, were then applied to the test set. The AUC for random forest was 0.86 while the logistic regression had an AUC of 0.77. Table 3 shows the two confusion matrices for the logistic regression model and the random forest model respectively. Overall, the differences between the 2 models are not large although the random forest model is better at identifying anormal cases. The advantages of a logistic regression model are that it is more interpretable and faster compared to random forest models so it could be a good choice for the purpose of our application.

***Table 3**: Test set confusion matrices for the logistic regression and random forest models*

Logistic regression

	Reference	
Prediction	Anormal	Normal
Anormal	39	15
Normal	23	58

Random Forest

	Reference	
Prediction	Anormal	Normal
Anormal	45	14
Normal	17	59

The logistic regression model has 21 variables that were selected by forward stepwise selection. On Figure 6, the variables are ranked by order of importance. It is interesting to note that the 20 most important variables in the random forest model (Figure 7) differ from those in the logistic regression model. They only have 3 variables in common (*DD*, *HT*, *qrs_duration*) but these variables do not have the same importance except for the *DD* variable, which is fairly important in both models.

The most important variables in the random forest model, which are the heart rate and the QRS duration, are more comprehensible and have been reported as important ECG parameters to diagnose arrhythmia [4]. However, the interpretation of the variables in the 2 final models is beyond my expertise and would require the help of a specialist.

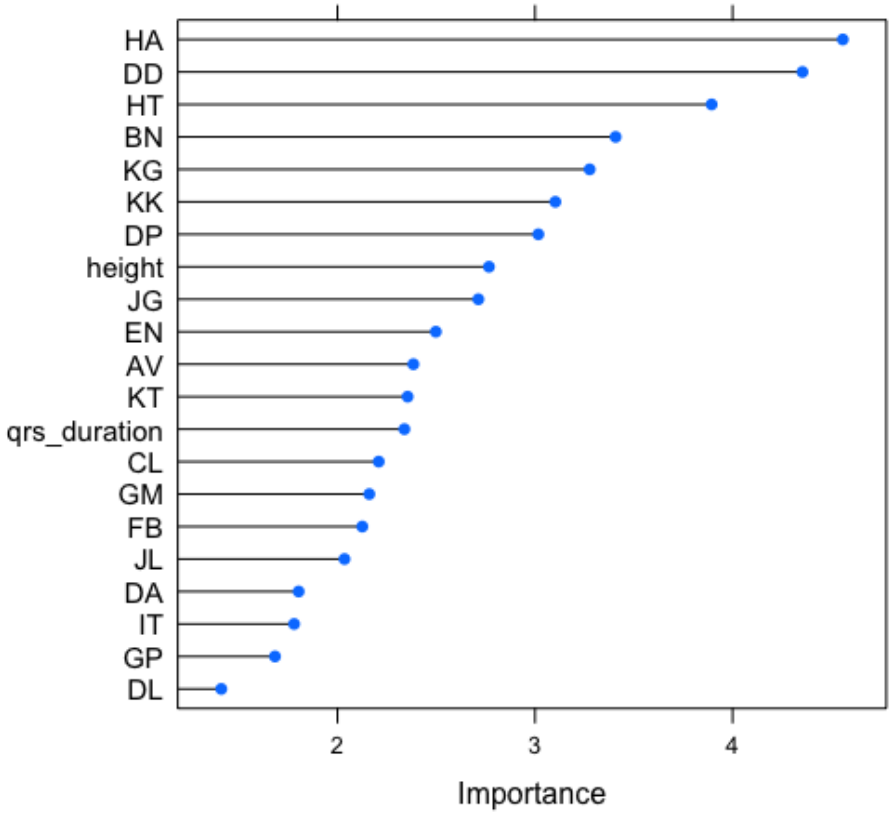

***Figure 6**: The 21 variables in the logistic regression model by order of importance*

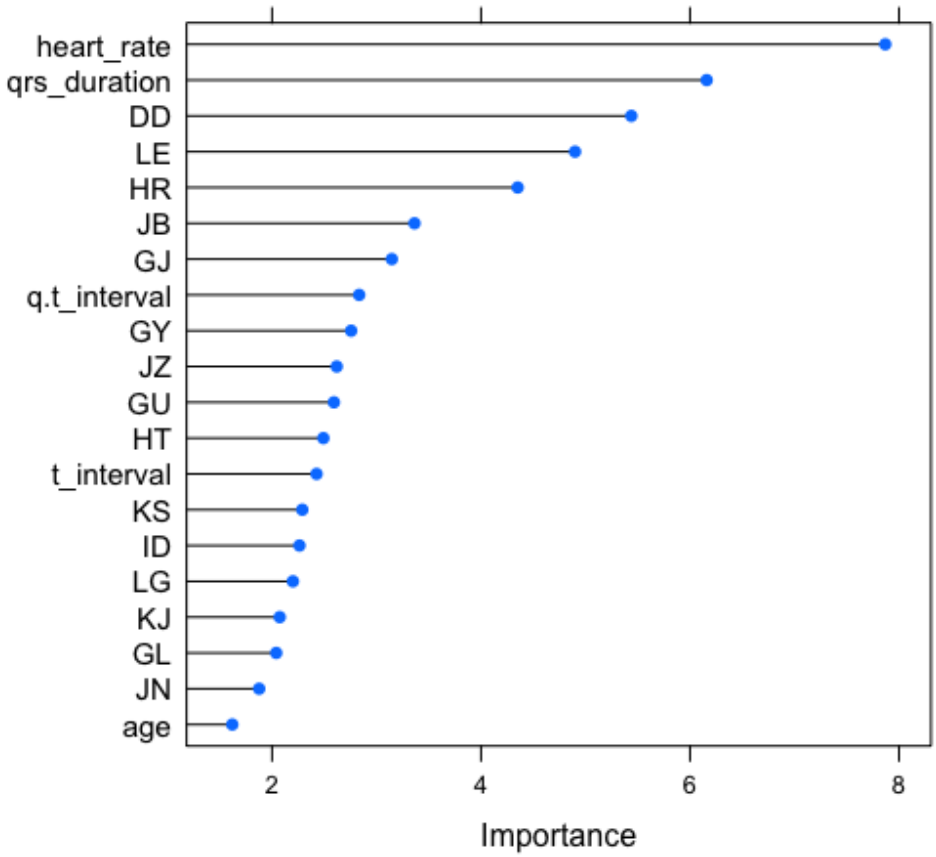

***Figure 7**: 20 most important variables in the random forest model*

6. References

[1] Drew, B. J., et al. "Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients." *PloS one* 9.10 (2014): e110274.

[2] Guvenir, H. A., et al. "Arrhythmia data set in UCI machine learning repository." *UC Irvine* (1998).

[3] Fu, D. G. “Cardiac Arrhythmias: Diagnosis, Symptoms, and Treatments.” *Cell biochemistry and biophysics* vol. 73,2 (2015): 291-296.

[4] Mandala, S., and Di, T. C. “ECG Parameters for Malignant Ventricular Arrhythmias: A Comprehensive Review.” *Journal of medical and biological engineering* vol. 37,4 (2017): 441-453.