Statistical and Machine Learning

Individual Assignment – Charlotte Gallet – 31/03/2022

1. Introduction

Statistical and Machine Learning is a part of Big Data Analytics that is in majority responsible for the outcomes that can be generated from the data received. It is necessary to denote the difference that exists between Statistical Learning and Machine Learning. The first one tries to find some insights from the data itself, it is a tool for understanding the data. On the other hand, we can see that Machine Learning models try to have an accurate and representative output. With time, there seems to be less and less difference between the two approaches. The general idea is to find the target variable by looking at the independent variables, also known as the predictors.

Having the best output per model depends on the data itself and what best fits the idea. Therefore, there are a wide variety of possible models as well as different methods to evaluate their performance. Once this has been done, it is possible to choose the best model and give the desired output.

In this paper, as we are talking about statistics and that machine learning has roots in the same field, a lot of term will be used to explain and describe the models. That is why some terms, which describe basic statistics will not be explained unless there can be a strong added value to the model's understanding.

The objective of this paper is to explain the different statistical and machine learning models. We will be looking at 5 different models such as Logistic Regression, Generalized Additive Models, Linear Discriminant analysis, Classification Tree and Support Vector Machines. Once the models have been presented, it will be interesting to perform a benchmark analysis on all of them to showcase the different elements that have been described through the theoretical part.

2. Machine Learning Predictive Models

K-nearest Neighbors

The idea revolving around the K-Nearest Neighbors (KNN) model is to capture the data points that are around the one point that interests us. Indeed, another way of calling this model is calling it Nearest Neighbors Averaging.

When predicting qualitative responses, the Bayes classifier is often used, as we will see in later models. However, finding the optimal solutions for this classifier is complicated, which is why attempts are done by estimating the probabilities. The KNN method tries to reproduce what the Bayes classifier would do in a transformed way. Instead of looking at the observation itself and estimating all the parameters, it looks at the other points that surround the observation.

The objective is to find the highest value for the probability that the observations is in a class that regroups all the other, closest points of that observation. The observation will then be categorized into the group where the probability is the highest. The equation to find the probability is as such:

$$P(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

Where K is the number of neighbors, x_0 is the observation also represented with $\mathcal{N}_0$ and j represents the class.

In this model it is possible to change the number of neighbors an observation has. This will increase the radius around which all closest points are considered as neighbors, which in turn will have an impact on the probability that an observation will be assigned to a certain neighborhood or not.

Therefore, a conclusion that could be made about this model is the fact that the more dimensions there are, the more features, the more it will be complicated to predict. Indeed, the sphere that represents the neighborhood will be bigger and bigger and in turn contain more and more neighbor points. This will lead to a greater bias and a larger chance of having errors.

This tells us that this model is better for datasets that are simpler and where there is small number of features.

Logistic Regression

In the same aspects as the KNN model, but with higher accuracy, the logistic regression can be used. Indeed, logistic regressions are used when looking at classification problems. That can be either for binary classification where 2 classes are created or multiple classification where we look at more than 2 classes.

The classification function can be written as follows:

$$Y = C(X) \epsilon\, C$$

In this case, we look at a qualitative feature X and a qualitative response Y in the set C. The output is a probability that X belongs to each category in C. Often, the default response is encoded as 0 if no, the vector probably doesn't belong in the category and 1 if yes, the vector probably belongs to the category. We are interested at the yes or 1 response.

To decide whether the probability is large enough to say if the predictor belongs to a category, a threshold is set and if the probability is above the threshold the output will be 1 otherwise it will be a 0. The objective is that this works better than random guessing, which is set at 50% probability of guessing right. However, some mistakes can still subsist which will then reduce the accuracy of the model.

In the case of a multiclass classification that sets more than 2 classes, it is not possible to use the linear regression. Therefore, we use the logistic regression. There are different approaches as to how to analyze a multiclass classification problem. The first way would be to look at a ONE vs ALL problem where we look at the first class versus the rest of the classes, then we move on to the second class versus the rest of the classes and so on. For each of the class that we look at, we create a linear regression. In the end, this outputs a probability for each regression and the selection is done on the class that has the higher probability. This method works best when there is a large number of classes as we only need the amount of model as there are classes.

The other approach takes a look at a ONE vs One problem. The idea here is to look at the first class against the second, then look at the first class against the third, then look at the second class against the third and so on. Once again, a probability will be calculated for each interaction and the interaction that is the most frequent will be seen as the strongest and best one. The upside of using this approach is to have a more balanced data set and better understanding and evaluation of the interactions. However, it creates a lot of models, which is why it is preferred to be used if there are less features to be predicted on the target variable.

The idea being logistic regression bases itself on this understanding of a classification problem. Therefore, instead of modelling the response Y, which is categorical, directly, this model will compute the probability that the predictor fits into a class. In mathematical terms:

$$p(X) = P(Y = 1|X)$$

Which can be translated as the probability that Y belongs to class 1 knowing X, the predictor, or predictors.

The form of a simple logistic regression is based on the sigmoid function. It is not a linear function as the figure below shows.

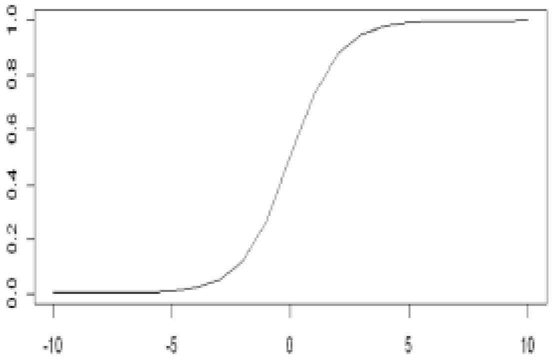


Figure 1: Graphic representation of the sigmoid function

The use of the sigmoid function is done while using the logistic function. We are using the form of the function to and its transformation to fit into the logistic model. This function allows the have a probability value between 0 and 1 for all the values of X. It can be written as follows:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Once we have this equation however, it is not sufficient which is why we will be looking at the log odds transformation of this $p(X)$ equation. This is the end will output the following equation:

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

This will be the equation used in the estimation of the parameters. In order to do so, the maximum likelihood will be used. The objective is to choose β_0 and $\beta_1$that will maximize the likelihood that the data is observed. This equation is also called the loss function.

$$L(\beta_0, \beta_1) = \prod_{i:y_{i=1}} p(x_i) \prod_{i:y_{i=1}} (1 - p(x_i))$$

It is called this way because if the likelihood is high than it means that the prediction is good, on the other hand, if it is low, the prediction is bad. For this reason, we always try to maximize the likelihood. This maximum likelihood function will output the cost function.

However, it is important to not that likelihood can sometimes be unstable is such situation. The way to overcome this weakness of the model is by using the log of the likelihood equation which must also be maximized. In order to do so, the function is derived to zero which is done through the gradient descent optimization algorithm, another algorithm that will not be explained here.

Generalized Additive Models

In going further than the logistic regression to produce a model for classification as well as for regression it is possible to use the Generalized Additive Models (GAMs). This type of model is in between linear regression and black box model, both in terms of flexibility and interpretability.

The idea that revolves around GAMs is to use various models on non-linear data, all while keeping some interpretability and simplicity in the model. Additive models get their names from the idea that one formula regroups individual models for each predictor in the dataset.

To illustrate what GAMs are about, this formula shows better in mathematical terms.

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i$$

Therefore, for each predictor in the dataset a function is associated in a way that best represents the distribution of this predictor. Once the best function for each predictor has been found, they are all added together into one global formula.

In the example of the Introduction to Statistical Analysis for R by Hastie and Tibshirany, three different predictors that are age, year and education are each plotted to show that they all have a different distribution.

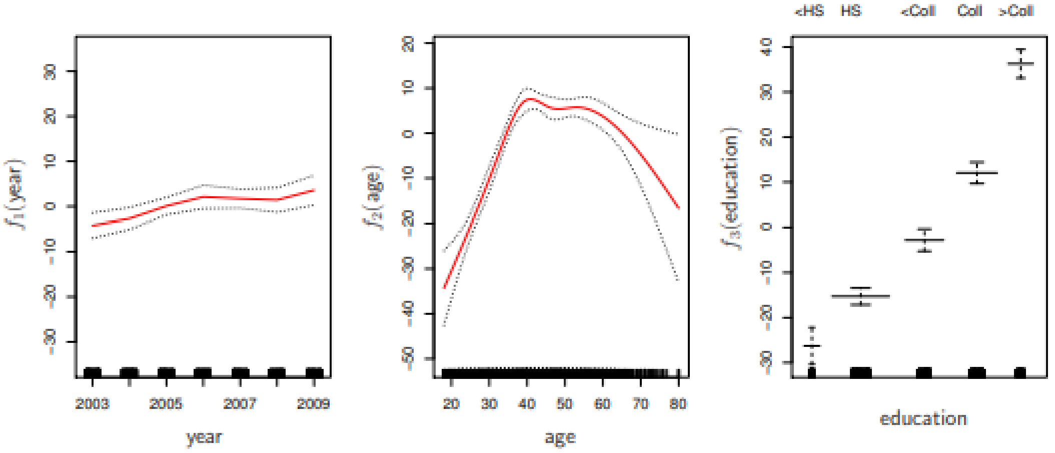


Figure 2: distribution of three different predictors in Wage data set

As they all have a different distribution, it seems complicated to allow only on type of model that would better fit linear data take its place. Hence, it is important to remember that each predictor can have a different type of function that will suit better the data. One type of function that is used in this case but also as a model itself is the smoothing spline model.

We will not go into details about what splines are and what they do as it is another model entirely. However, we need to understand that they try to fit part of the data as best as they can. Hence, the reason why it is often used for GAMs as they will try to perfectly fit with the data that was provided.

GAMs can be used both for regression and for classification. When doing regression multiple linear regression will be applied, a linear regression is applied to each of the model. In the case of classification, the functions are replaced by flexible functions, often it means splines which then allows for non-linear relationships.

The objective is then to use the least squares function as a way to fit all the regression and models into one model. It is also necessary to use backfitting when using splines. That means that we are fitting each predictor to a model while holding all other predictors fixed. The difference between using backfitting or least squares is usually small, so either one can be used for the model's fitting.

This model is useful when looking for non-linear relationships while combining a high interpretability. It is also important to use the GAMs model when there are no a priori about the relationships between variable as it is possible to change the regression associated to each variable and choose the one that fits best.

Finally, this model has been used and experienced with for more than 30 years which implies a great reliability with its results and outcomes.

There exists some advantages and inconveniences in the use of this model. The advantages can be listed as the fitting of non-linear relationships to each variable in an automatic way which in turn creates a better relationship between the desired outcome and the predictors. This leads to the second advantage which can lead to better accuracy of the dependent variable when the data fits this model well. Another advantage that this model offer is the examination of one variable and its impact on the target variable, while the other variables are fixed. This means that we can look at each predictor individually and combine them into one accurate and impactful model. Finally, when using the smoothing splines, degrees of freedom are used. These degrees of freedom are different for each predictor, which means that there is a large possibility to use the right degree of freedom for each variable individually, therefore administrating the best possible result.

On the other hand, some inconveniences can occur. The first being the restriction for the model to be additive. It is not possible to look at multiplication or other type of relationship between the variables. This leads to the second inconvenience which is the missed interactions between predictors as we are looking at each one of them independently. Indeed, in some cases, it is possible to have a stronger impact when the predictors are combined into one predictor that better explains the target variable.

In that last case, there is a solution. That is the creation of interaction variables by hand. Instead of letting the model try to find the best interactions, the objective is to combine two predictors together and fitting a model to that newly created variable. This adds a new, low-dimensional interaction function where we have one function for two variables. Each variable can be used separately and then in an interactive function. This allows to add some dynamism into the model.

Linear Discriminant Analysis

In the same idea as the logistic regression but with adjustments, some classification problems will require more than two classes. The use of logistic regression in this case seems to be quite unstable and will not provide the right predictions. Therefore, if the number of observations we have is small and the distribution of the predictors tends to be normal, the linear discriminant model (LDA) is more stable.

The idea that is behind the LDA is that instead of looking at a set of parameters on which to set the response, the distribution of each estimator is calculated, and this distribution is then transformed into an estimation of the probability.

There exist two approaches to achieve that goal. The first approach, statistical approach, bases itself on Bayes' Theorem and is used when there is only one predictor (p=1). Assuming that the function is normally distributed, we can calculate the probabilities for each class thanks to the equation where each observation is replaced by its density function.

Let's dive a little more into what these terms mean. The density function bases itself on the Gaussian density function and is a combination with Bayes' equation. The density function for a categorical variable can be written as such: $f_k(X) = P(X|Y = k)$. This equation makes it possible to see the density of the estimator from a class k to have a probability of x. Therefore, the higher the result of this function, the higher the probability of this observation to have a probability of $X \approx x$.

This function is then introduced inside Bayes' equation which outputs Bayes' Theorem as follows.

$$P(Y = k|X = x) = \frac{\pi_k f_k}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

Where π_k represents the probability for the observation to be in the kth class and there are more than 2 classes.

To estimate this probability, the first element we need to estimate is the density function. As said before, it is important to have some assumption which says that the density is normally distributed. By using this assumption, we can transform Bayes' equation to add in the mean and the variance, that is already known due to the distribution's nature. This transformed equation is called the Bayes classifier as it will assign the observation to one of the k classes. Therefore, the observation will always be placed in the class where the equation was the highest.

The idea behind the LDA is to approximate the Bayes Classifier through the approximation of the discriminant score: $\delta_k(x) = x.\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$. To do so, it needs to estimate some of the parameters notably as follows:

$$\hat{\pi}_k = \frac{n_k}{n} \qquad \hat{\mu}_k = \frac{1}{n_k}\sum_{i:\, y_i=k} x_i \qquad \hat{\sigma}_k = \sum_{k=1}^{K} \frac{n_k - 1}{n-K}\hat{\sigma}_k^2$$

These terms are then replaced inside the discriminant score and the LDA goes through the process of estimating the parameters while knowing the number of observations in total, number of observations per class and the number of classes.

The second approach is through Fisher's LDA where we look at more than one predictor. When using this approach, we try to change the perspective from high dimensions to a vector and analyze the results from that perspective. The idea is to look for a new and lower dimensions coordinator that maximizes the separability among the categories.

To look at multiple predictors we use the multivariate Gaussian. This distribution type tends to look at the distribution of each predictor and adds some correlation between the predictors. Looking at two different predictors, we use the vector representation of these predictors and find the distribution of these vectors. In doing so, we create a new orthogonal projection with the data aligned into one vector that fits the distribution of both vectors. The same is done when we increase the number of predictors.

The process is the same as for the statistical approach of LDA, however the equations change according to the multivariate Gaussian distribution as can be seen in the following equation:

$$f(x) = \frac{1}{(2\pi)^{p/2}|\sum|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\sum^{-1}(x-\mu)\right)$$

Where the mean or μ represents the expected value of x and $\sum$ represents the correlation between the predictors. Once we have this function, it is possible to once again estimate the discriminant score that has been modified according to the new parameters.

$$\delta_k(x) = x^T\sum^{-1}\mu_k - \frac{1}{2}\mu_K^T\sum^{-1}\mu_k + \log(\pi_k)$$

Finally, the LDA will maximize the objective function. That is the division between the in-class variance and the variance between each class, in mathematical terms:

$$w = \arg\max \frac{w^T S_B w}{w^T S_w w}$$

Where S_B represents this between class variance, S_w the in-class variance and w^T the linear transformation matrix. The main objective is that each class is as far away as possible from each other all while keeping the variance within each class, as small as possible.

This model helps when trying to classify the observation into two or more classes. However, it is sometimes possible to see that mistakes are made, especially by assigning the wrong class to the observations. That is why the confusion matrix is calculated as a way to understand what the error rate is. The objective is therefore to minimize this error rate, which can be done by improving the parameters of the model or by switching to a model that better fits the data and the desired output.

Decision Trees

Once again, we look at a model that can be applied both for a regression and a classification problem. However, in this case the idea is to segment and stratify the different classes into regions and then make a prediction about the class that the observation should be situated in.

When first taking a look at regression trees the idea is to separate the space where all predictions are possible, that is all the possible values of each observation, into smaller regions. For each region we want to observe what the probability is of the observation ending up in one of the regions.

This model is called a decision tree due to its looks as well as the terms used to describe the model it represents. The visual representation helps to better understand.

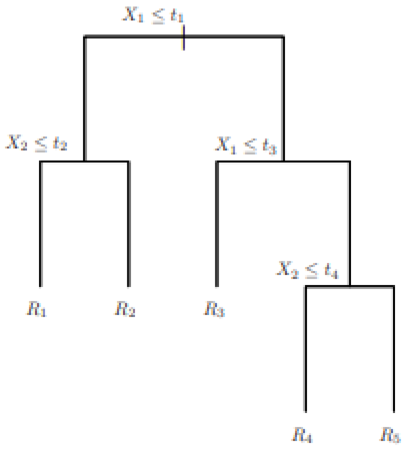


Figure 3: Graphic representation of a Decision Tree

In this case each region, which can be found at the bottom of the graph is called a terminal node or leaves. The points where there is a probability are called the internal nodes. Finally, each line that links terminal nodes to internal nodes are called branches.

When constructing the decision tree, it is important to keep in mind that all the regions are separated. This means that while creating a regression we first divide the space of prediction into distinct regions. The predictor space as explained before is then comprised of *J* regions noted $R_1, \ldots, R_J$.

There are infinite ways to split the data into different regions. The assumption that the split can only be done on one dimension is used to simplify the problem. To know where to split the data, the residual sum of squares (RSS) is calculated. A split is being done only if it possible to improve the model, otherwise there is no need to split. To see if there is an improvement, the RSS is used. The split is being done at the place where the RSS is the smaller.

For the first split in the data, the mean of the observations is taken, and the split is made it the middle, then the RSS is calculated for each region. If one of the regions has a small RSS, this is where the split will be. To make a second split, we assume that we do a top down or recursive, which means that it is only possible to do a split the new regions that were created before. This ensures that each region is distinct and non-overlapping.

With this method comes along some risks. The decision trees have a high variance which means that they are prone to overfitting. To reduce this flexibility, the idea is to cut the lower leaves and to remove the branch where there is not a lot of information. That way we reduce the complexity through pruning. Pruning can be done either before creating the tree or also after letting the tree grow as big as possible and then reducing the number of leaves and branches that there are.

In the case of a classification tree however, it is not possible to use the RSS, instead other metrics such as the classification error rate, the Gini index or the cross-entropy are used. The same disadvantage as for regression trees can be found for classification trees, that is overfitting and high variance. In this case the solution would be bootstrap sampling, also known as bagging.

While doing bagging or bootstrap sampling, the idea is to average a set of observation. That where we reduce the variance. The main issue, however, is that is it possible that some predictors are stronger and therefore the bagging will have less effect because the same bagging tree will be created at every step. This can be counteracted by randomizing the trees.

Once all these regions have been created, we calculate the means response values for each observation in that region. Therefore, for each region there will only be one prediction, which is then used as a probability for a new observation to be a part of a certain region.

The way this is used in practice is that for each observation there is a variable value. This variable value is then compared to what are at the internal and external nodes. Based on the value of at the internal nodes and the values of the observations, we can predict in which region will this observation be.

Some advantages come with the use of this model. This type of model is used both for classification and regression, which can be helpful in many situations. It is also capable of handling dummy variables. Indeed, as it looks at the probability for each variable, it will independently look at their values and place the prediction in the right category for each dummy variable. Finally, it is possible to say that this model is easily interpretable by non-technical people. Indeed, from just a glance at the tree, it is possible to understand how it works and how the predictions are made. There is no need to have a background in statistics or machine learning to grasps the concepts associated with the decision tree modeling.

On the other hand, some disadvantages subsist. As discussed earlier, the structure of the trees is quite unstable. Indeed, it can change every time a new model is created, with the same features. As the first split of the tree will have an impact on the rest of the splits. It is difficult to decide where the first split will be with certainty, which is why some more advanced version of the decision trees are also used. Another disadvantage is the fact that it could easily overfit the data. By knowing and estimating so many aspects of the data, it will always try to have the best probability to create the best desired outcomes.

This is the reason why there are many types of models, as each data set has its particularities and each model will suit, or not, the desires of the data and the output.

3. Benchmark

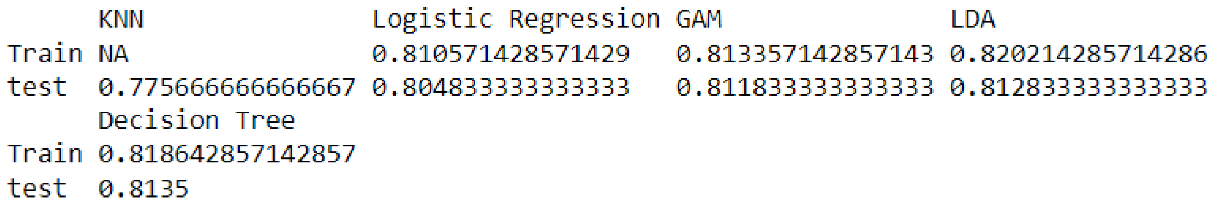
While it is interesting to have the elements of the theory associated with each model, it is also necessary to see what is happening in practice when using a dataset.

In this part, we will take a look at the result of each model on a dataset that contains the default of a credit from a bank. Several features are contained in the dataset. These features correspond to some categorical variables such as the gender, the marital situation, the education, and the age group. There are also some numerical features such as the billing amount and the payment amount.

The primary objective, with each of the model, is to be able to predict whether a client from this bank will default on its credit or not. For that, we have used all of the above models.

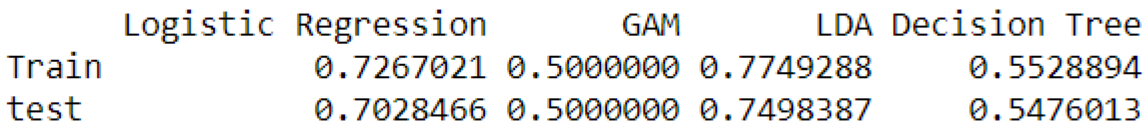
After some pre-processing, cleaning of the data as well as separating a train and test data set, to ensure that no information is leaked in the model, we have measured the accuracy as well as the AUC, a comparison metric for each of the model.

Here are the results for accuracy and AUC for each of the model.



```
      KNN               Logistic Regression GAM               LDA
Train NA                0.810571428571429   0.813357142857143 0.820214285714286
test  0.775666666666667 0.804833333333333   0.811833333333333 0.812833333333333
      Decision Tree
Train 0.818642857142857
test  0.8135
```

Figure 4: Accuracy results for both train and test set of the models



```
      Logistic Regression       GAM       LDA Decision Tree
Train           0.7267021 0.5000000 0.7749288     0.5528894
test            0.7028466 0.5000000 0.7498387     0.5476013
```

Figure 5: AUC results for both train and test set of the models

The KNN model as we can see, is less interpretable than the other models as it is not possible to get the AUC nor the train accuracy. It is only possible to rely on the accuracy of the test set. However, to make sure that the best possible accuracy is found, the KNN model has been fitted in such a way that produces the optimal number of neighbors to look at. In the case of this model, this number amounted to 41 neighbors.

Moving on to the logistic regression, it is possible to see that this model is very accurate. By looking at the difference between the train and the test set we can see that there is no overfitting. Indeed, this difference is quite small, therefore not significant. To ensure that this is the best model possible, the idea was to run a forward stepwise selection of the features. This way, we use in the logistic regression only the most important features that will have the maximum impact on the target variable. Therefore, we went from 32 features to the 21 most important.

Concerning the Generalized additive models, we can see that the accuracy is quite strong both for the train and the test set. This would mean that the model fits well the data. However, when we look into the AUC of the model, it is clear that it is performing poorly and does not, in fact, fit the data as well as thought. Once again, to create the best model possible a stepwise regression was done to select the most important features out of the numerical features.

We then move on the Linear Discriminant Analysis model. This model seems to be the best performing model. Its accuracy is close to the accuracy of the other models and no overfitting can be observed. The AUC of both the train and the test are, however, very high in comparison to the other models. The Linear Discriminant Analysis has also had a stepwise forward regression applied, to select the best features. In this case, only 5 features have been selected. Thanks to the choosing of these 5 variables, the model seems to perform very well.

Finally, the decision tree has been applied. While this model seems to be quite accurate, even more accurate than the LDA for the test set, it is very poor when looking at its AUC. This means that compared to other models, it is not worth using. This is the case even though some pruning has been done to select the best tree possible and improve its performance.

In conclusion, from this paper as well as from the benchmark, it is possible to say that every model has its advantages and drawbacks. It is therefore necessary to try to use many different models on the data that we have to select the one that performs the best. In the case of this case study, it is possible to conclude that the Machine Learning Model that will be suit the data is the Linear Discriminant Analysis.