Machine Learning Classification and Evaluation (2017-2018)
Universithé Catholique de Louvain

Final Project

Alejandro Simón Rodríguez

11 de mayo de 2018

Índice

Índice de figuras

1. Introduction

In this project I'm going to justify theoretically and empirically which model will be the best to predict the possibilities of Metastasis after a breast cancer.I have used R language,analyzed five possible machine learning algorithm and basic technique of essemble modeling to predict the results.

2. Exploration and Treatment of the Data

The idea of this section is to have an insight about how are the variables to analyze and to process the data before applying machine learning algorithms. The processing of the data is going to be realize by the treatment of NAs,outliers and normalization or standarization of the data while the analysis of the data is going to be realized by Univariate and Bivariate analysis.

2.1. Recognition and Processing of NA

At the beginning, I focus in locate the NA in the dataset. I have observed that they were are in the last variables of the set and one of the variable was completly NA, thus I discarded this variable. I have seen also one constant variable that I discarded too.I decided to fill the NAs by the mean and the median of the values of each variable building two set.

2.2. Types of variables

There are algorithms that only performs with numerical data or categorical data. I have studied the set to observe how were the variables. All variables were numercial less one after the two discarded. The categorical variable was binary then I have created a dummy variable.

2.3. Scaling the data

In general, there are two possibilities: normalization or standarization. As i did not know between what to choose, I perform the two scaled techniques to test which one would be better.

2.4. Outliers

At the beginning, I did not analyze the outliers of the variables because there were too much variables and were not possible to use boxplot and scatterplot to analyze them. One time the feature selection was done, I did the analysis. The problem to erase outliers is that you lose information of extreme case that maybe you have to predict in the test set. Thus, the correct way was analyze train and test set to observe when there were outliers only in the train set. After the analysis, I have tested the set with less outliers and the accuracy was lower, then I decided to keep the outliers because with them the models were able to better predict the test set.

2.5. Univariate and Bivariate Analysis

Univariate and Bivariate Analysis were not possible to do before feature selection for the same reason as Outliers' analysis. After that, I have realized both of them but the result were not too much interesting.

3. Feature Selection and Feature Engineering

3.1. Multicolineality

My first step to apply feature selection was to erase multicolineality variable.There was too much variable in the model, thus I thought that erase such variables that were correlated will decrease the noise of the model. Thus, I determined to erase variables with coefficient of correlation greater than 0.7 . After this, I had erased almos 2500 variables but they still being too much.

3.2. Feature Selection with Random Forest

Random Forest package in R has a function to analyze the importance of the variables in a model according to their Gini Index.There was too much variable to be well analyzed by a Random Forest, thus I built an algorithm to identify the most important variables.

The algorithm takes a sample of 1000 variables and apply random forest. Then, it took the variables according to:

$$\frac{GiniIndex}{\sum GiniIndex} > 0{,}003$$

After that, It saves in a vector the chosen variables if they were selected.The algorithm replicate this process 1000 times.Finally, it takes the variables with:

$$n^{o}appearances > 20$$

and

$$\frac{n^{o}timesselected}{n^{o}appearances} == 1$$

This give me a set of 66 variables to apply the models.

I have realized this method with and without have applied the multicolinealtity analisis and the result were more or less the same.

I also have realized this method with the measure of importance of GBM algorithm obtaining another set of variables,around 228 variables.I'm going to test each set of variable to search the best behaviour.

4. Model Selection

In this section, I have applied cross validation technique to tune the parameters of the model.For the final section I have discarded the models with less accuracy than 80 % after cross validation technique.To understand the graphics,the blue curve will always be the accuracy and the red the variance.

4.1. Random Forest

There are theree important parameters to tune in Random Forest algorithm according to accuracy, variance and overfitting. These parameters are: nodesize, mtry, replace and ntrees. I'm going to fix ntrees = 500 and replace = T to avoid overfitting to the train model and have more differents trees to predict the result.

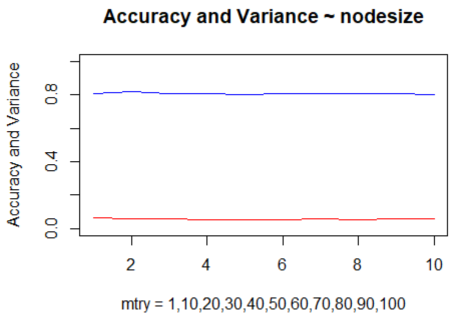

Figura 4.1: RandomForest tuning mtry value

The parameter mtry indicates the number of variables taken by each round of the algorithm. The default parameter is $\sqrt{n^{o} variables}$, in this case would be 15 and it is good enough as we can observe.There is not so much variation of the accuracy or variance and if we apply zoom the default value is one of the better.

The parameter nodesize fixes the minimum size of terminal nodes. Setting this number too large causes smaller tres then underfitting and setting this number too small causes big trees but overfitting. Hence, we have to find the correct number to set.

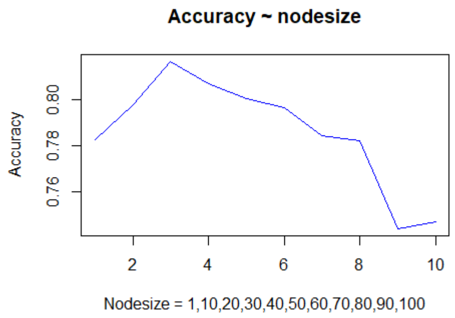

Figura 4.2: RandomForest tuning nodesize value according to accuracy

As we can observe,the accuracy is better with a nodesize value around 30 and too large or too small have worse results.

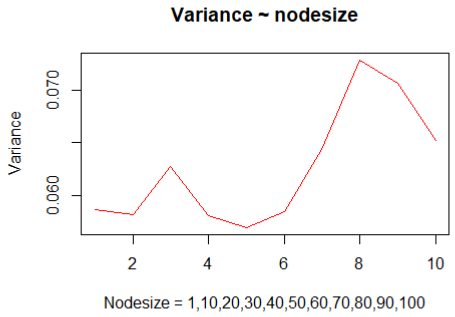

Figura 4.3: RandomForest tuning nodesize value according to variance

Furthermore, the variance has a best behavior with this nodesize value.

To conclude,before apply random Forest technique I expected a better performance of this model.Problably I did not find the best set of features for this algorithm or I did not tune well the parameters.

4.2. GBM

GBM is a esembling technique based in Gradient Boosting.It's a complex algorithm with a lot of parameters, thus I have tried to optimize the learning rate known as shrinkage and minimum number of observations in the trees terminal nodes.I have fixed the number of trees to fit in 500,the distribution to adaboost and the bag fraction to 0.5 .

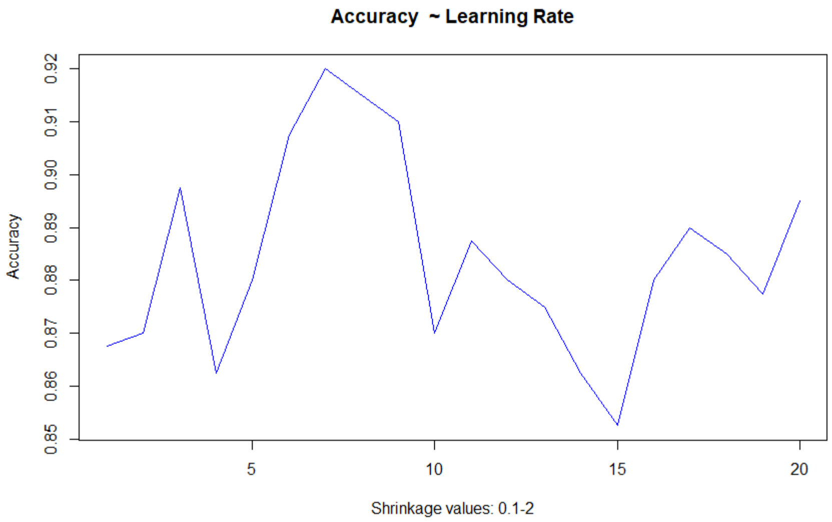

Figura 4.4: GBM tuning shrinkage value

The problem to tune the values in this algorithm is the variation depending of the number of patients chosen as training and test set.In this case,the best value was shrinkage=0.7 .

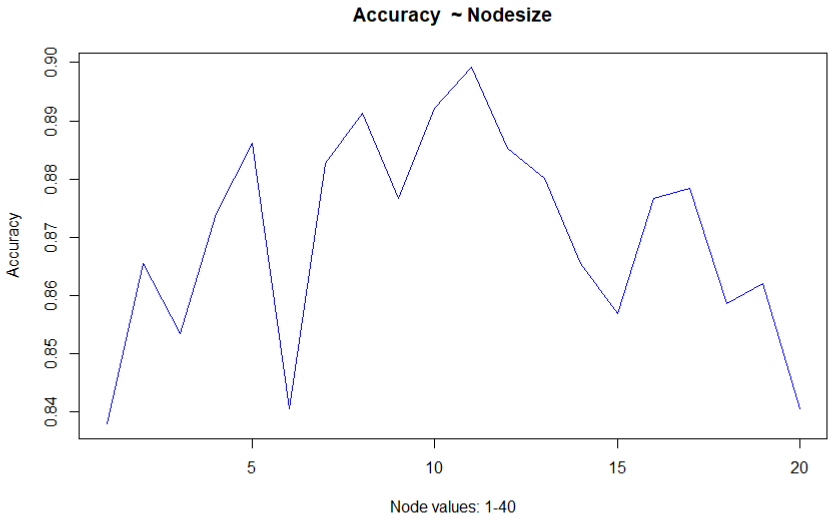

Figura 4.5: GBM tuning min node size value

This value and the learning rate are the values that we have to make the trade-off between a good prediction and overfitting.In this case the best value of minimun numbers of observation is between 15-20.

4.3. XGBOOST

XGBOOST is an advanced implementation of Gradient Boosting algorithm.It's a complex algorithm with a lot of parameters.I tried to do my best to tune this algorithm but I did not obtained a accuracy very higher,probably for the complexity of the model.

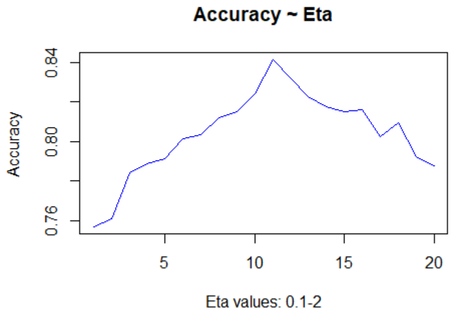

Figura 4.6: GBM tuning eta value

As we can see the eta value is very important for the accuracy and variance of the model.I have tested a lot this value and his performance because depends a lot of the quantity of training data and test data that we have.This test was made to predict 20 cases and the best values was $eta = 0.1$ but i have tested to predict 60 values and the best was $eta = 1$.

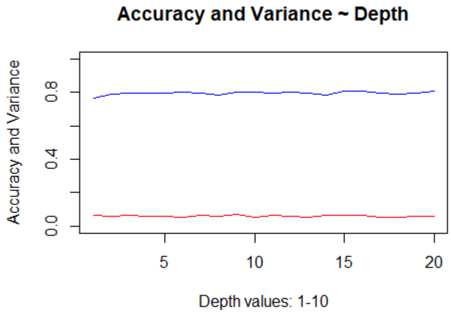

Figura 4.7: XGBOOST tuning depth value

High values of depth tends to overfit the model.There isn't exist a large difference between the values but the best is between 5-10.

To conclude, I have also discarded this algorithm due to his accuracy thas was lower than 83 % in average.

4.4. Naive Bayes

There is only a parameter to analyze in R with Naive Bayes and is the 'laplace' parameter.These are the results according to Laplace values.

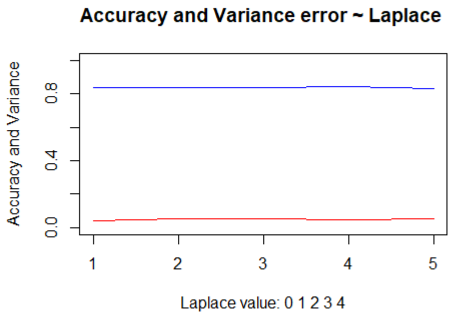

Figura 4.8: Naive Bayes tuning Laplace value

As we can see,the laplace value does not change the precision or the variance.I have tested with higher values for laplace parameter and the results were the same.

4.5. K-Nearest Neighbour

The main parameter in this algorithm is the number of neighbours.I have applied cross validation technique by running and testing the model with 100 times each neighbourg and this is the plot of precision according to neighbour.

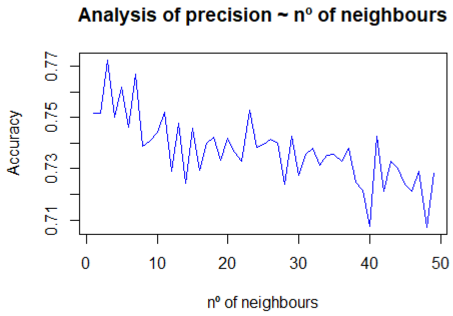

Figura 4.9: Knn tuning parameters

As we can see,the best election for the number of neighbours is $k = 3$.I have discarded this model because the accuracy is less than 80 %.

4.6. SVM

The main parameters in this algorithm are the kernel function,the gamma coefficient and the cost function.The two coefficients controls the trade off between smooth decision boundary and classifying the training points correctly. First, I'm going to analyze the kernel function's analysis.

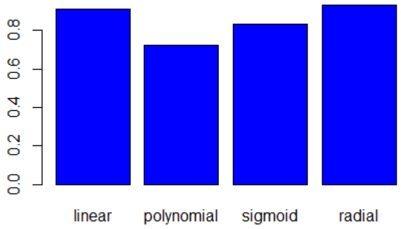

Figura 4.10: Election of Kernel function

After the analysis, we can observe that there are two kernel functions,the linear and the radial kernel functions that has an accuracy above to 90 %.

To choose between this kernel function I have seen the variance os each model by the standard deviation of the errors and was almost the same so I decided to use radial function due to it has better accuracy.

The general observation tuning the model parameters is: the models have better behaviour in terms of bias and variance errors if you take gamma $\in$ 0,001, 0,0001 and cost $\in$ 1000, 10000, 10000.

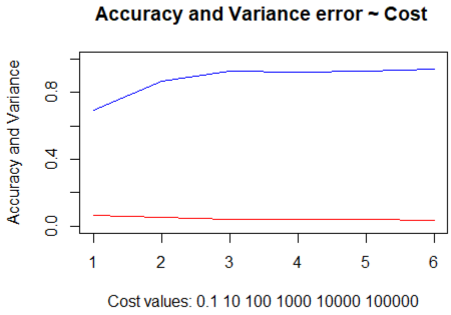

Figura 4.11: Kernel function tuning cost parameter

We can observe that as much increase this value the variance is reduced and the accuracy improved. The gamma value was 0.001 . The overfitting of the model increases if the gamma value is very small or the cost value too large, thus I have let this value in 1000 .

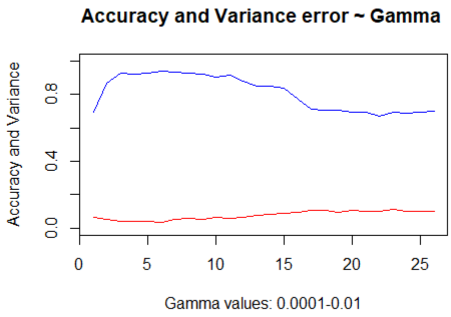

Figura 4.12: Kernel function tuning gamma parameter

We can note that the accuracy improves and the variance decreases until gamma = 0.001 and them the accuracy decreases and the variance increases. The cost value was 10000. Hence,I have decided a value of $gamma = 0{,}001$.

Finally,I have chosen the next parameters for my model: $(kernel = radial, gamma = 0{,}001, cost = 10000)$

5. Essemble Modeling

The main idea of essembling modeling is combining diverse set of learners (individual models) together to improvise on the stability and predictive power of the mode.I have tryed to perform two techniques.To combine these models we assign weights to the prediction of each model to make the final prediction combining all the others.

5.1. Neural Network

As we can expect assign the same weight to every model is not the most optimal way to realize this method. Thus, I have tryed to use a neural network without hidden layers, where the input would be the decision of the model and the output the combination of the decision.The 'problem' of this technique is that you have to train the neural network to find the optimal weights. To train the neural network you have to had trained the individual models and then use them to make predictions to train the neural network.

The problem is if you don't have enough data I think that is difficult to the neural network find the optimal weights and finally it has a similar performance as the simple majority vote.

I have studied this technique taking the five algorithms and then erasing XGBOOST and randomForest.The best results were with GBM,NaiveBayes and SVM.

For the next graphic I have created a test and validation set with 70 patients and the training set with 118 patients.

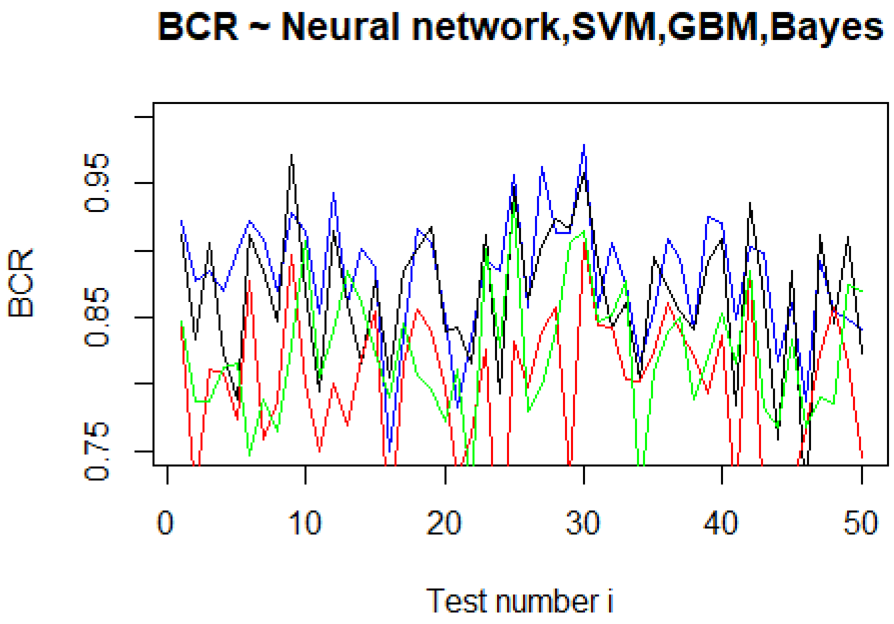

Figura 5.1: BCR in function of Nerual Network, GBM, SVM and Naive Bayes

The red lines is naive bayes, the green is GBM,the black is Neural Network and the blue is SVM.As we can see, naive bayes and GBM model are outperformed by the others.In the graphics is difficult to see that the best in average was the SVM model.It also was the one with less variance.

5.2. Majority Voting

The technique of majority voting assign the same weights to each model and the final decision is decided by the majority decision taken for the models.The problem of this technique is that treat by equal every model and does not take into account the precision of each model.

I have used theree models to apply this technique: SVM, GBM and Naive Bayes.I have taken a test set with 70 patients and a train set with 188 patients to build this models.

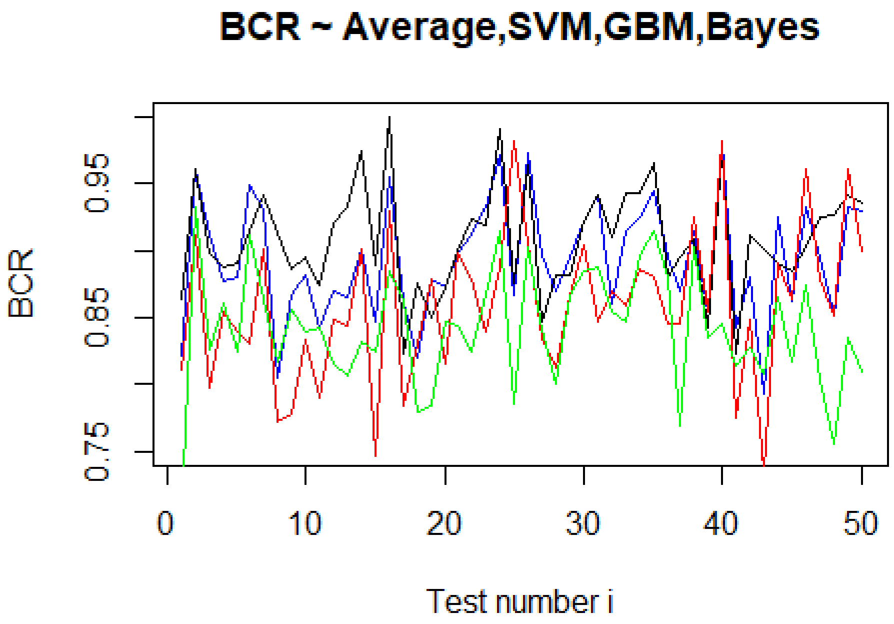

Figura 5.2: BCR in function of Majority Voting, GBM, SVM and Naive Bayes

The colors are assigned in the same way as the last graphic.As we can see Naive Bayes and GBM are outperformed again and the best was SVM.It was also the one with less variance.

6. Conclusion

I have chosen SVM as my model for the competition with the variables taken by the GBM model.There was not difference among to choose the one scaled with normalization or standarization. This model was the better in terms of accuracy,bcr and variance.

7. Annexe

I have apply all the cross validation technique measuring the accuracy and the variance. I know that probably would have been better to study the parameters according to BCR value.I had performed all the cross validation technique before to start to measure the BCR value and cross validation requires a lot of time. Thus, I have tested a little with

BCR value, and the value of BCR was very near to the Accuracy value and also their variance, thus I have decided to believe in the cross validation already done.

Furthermore, I know how to put legends but they did the graphics more difficult to understand due to the space that legends need. I also know that i can put more graphics but maybe this will boring the reader.

I have enjoyed a lot this project. Thank you.