MVA - Project 7

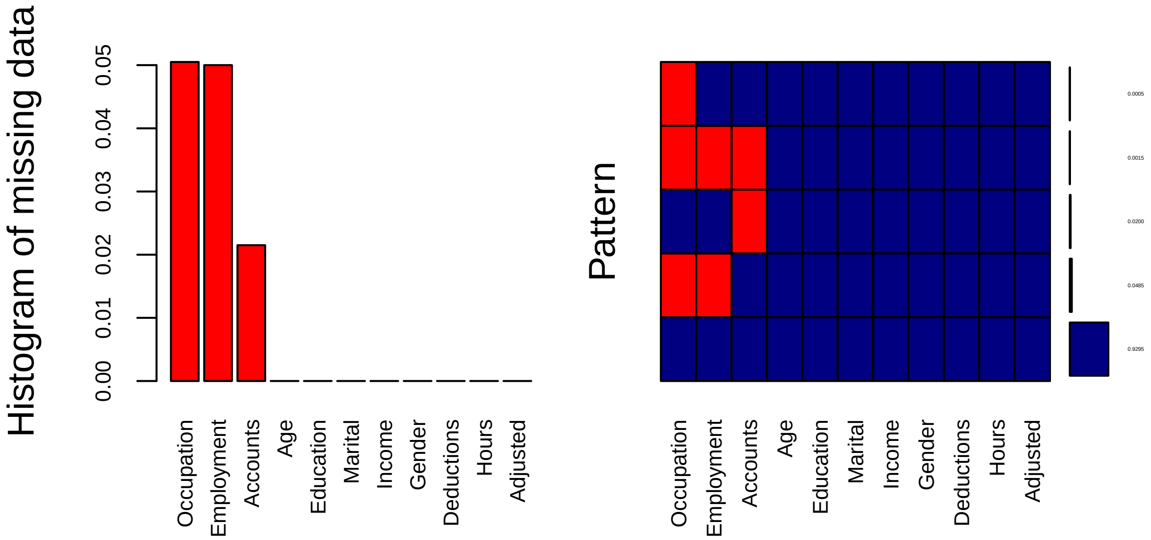
Carles Garriga Estrade i Balbina Virgili Rocosa

05/27/2018

All the code created for us to develop this assignment can be found on LAB07.R file. This file is located in the same folder of this document.

Exercise 1 and 2 First of all, the Audit dataset has been read from the file. As the main goal for this assignment is to use a decision tree to predict the binary *Adjusted* variable, in other words, if the individuals had made a correct financial statement or not, we have decided the predictors needed for it. After taking a look on the variables that it contains, we have discarded the *ID*, as it is just an identifier and does not provide any additional information for the analysis, and *Adjustment* because *Adjusted* variable has been calculated from it.

In order to be able to correctly manipulate it during the following exercises, we have prepossessed the variables. A missing values analysis has been performed.

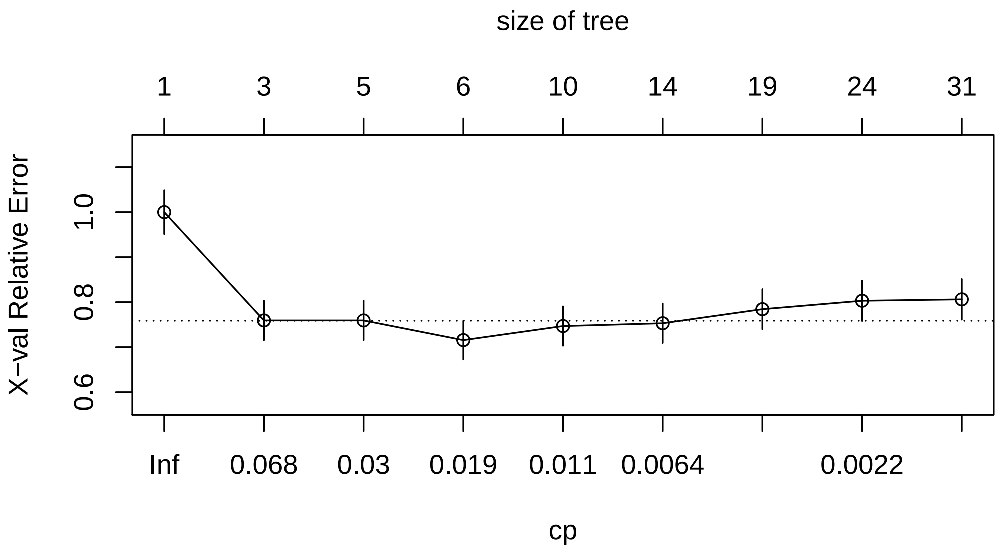


After checking the analysis, we have found that there are 243 missing values from variables *Occupation*, *Employment* and *Accounts*. As they not represent a significant quantity of the total data, the decision tree could be developed without performing them but random forest does not allow NA's values. That's why we have computed multiple imputation to fill them and we check that after its execution, there are no missing values left.

```
## integer(0)
```

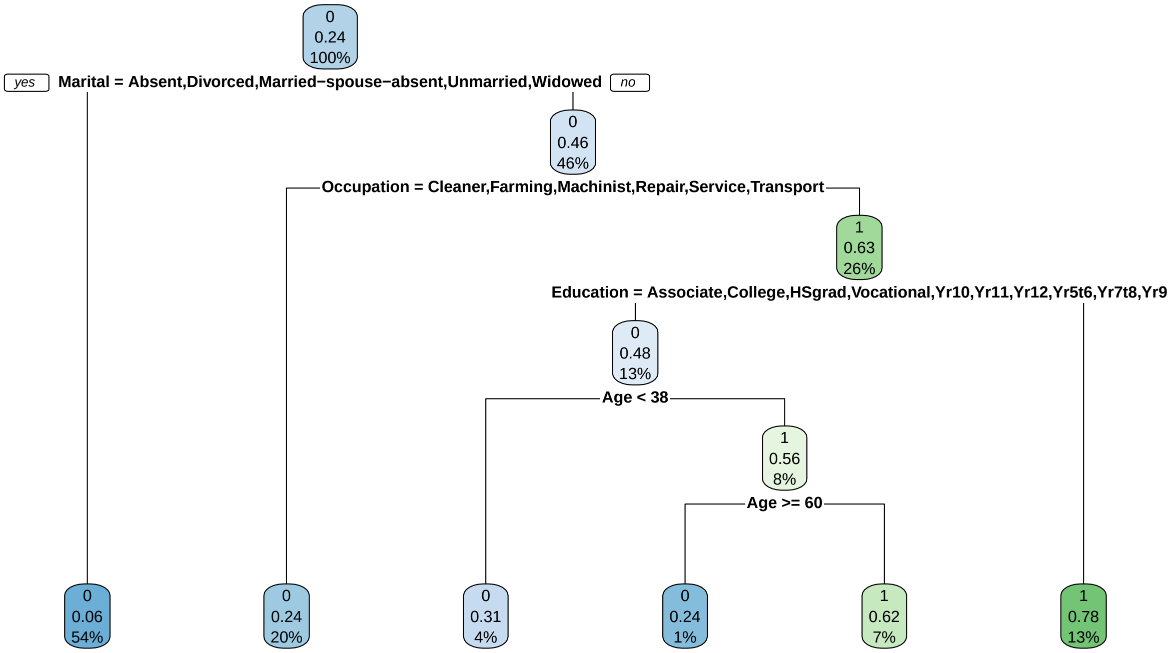
Exercise 3 Then, we have selected the last 1/3 of the observations as test data and the rest of the observations as training data.

Exercise 4 Once our data of test and training has been defined, we are able to obtain the decision tree to predict whether the variable Adjusted on the training data. To do so, the rpart function has been used by defining the appropriate formula for prediction the Adjusted variable taking into consideration all the other variables chosen for the analysis, passing the training data and also defining the complexity parameter and the number of cross-validations for our model. The visual representation of the cross-validation results obtained for our calculated decision tree is showed below. On it, we can see that the minimum relative error is obtained when the number of splits is 6.



We know that we obtain the optimal tree by pruning the maximal one up to the minimal cross-validation error. To decide the cutoff value for taking the decision more precisely, we have calculated the minimum error of our decision tree model. The optimal decision tree obtained with the values calculated is showed below, that as we can see there are 5 splits until the root point is reached.

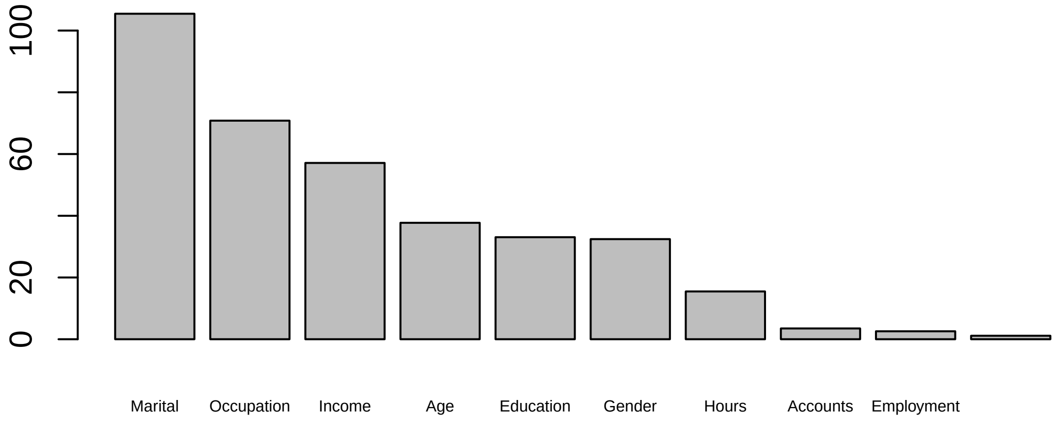
```
##   optimal_CP optimal_splits
## 1     0.0125              5
```


0
0.24
100%
yes
Marital = Absent,Divorced,Married−spouse−absent,Unmarried,Widowed
no
0
0.46
46%
Occupation = Cleaner,Farming,Machinist,Repair,Service,Transport
1
0.63
26%
Education = Associate,College,HSgrad,Vocational,Yr10,Yr11,Yr12,Yr5t6,Yr7t8,Yr9
0
0.48
13%
Age < 38
1
0.56
8%
Age >= 60
0
0.06
54%
0
0.24
20%
0
0.31
4%
0
0.24
1%
1
0.62
7%
1
0.78
13%

Taking a look at the pruned tree, we can observe that 54% people unmarried, widowed or with no spouse tend to have their finalcial statement adjusted. For those still being married and being working on a low income job (cleaner, farming, service, transport or machinist) 20% of those have had their finalcial statements adjusted. For those who have not any of the previous occupations, but they have had a minimal education, a 13% of them have had their financial statement adjusted. 13% of those who hadn't received a minimal education also had their financial statement adjusted.

Exercise 5

The importance of each variables on our obtained optimal tree are showed with the plot below. As we can see, they are ordered from more to less importance, so Marital, Occupation, Income and Age variables have a deep impact on the prediction of Adjusted variable. Then, Education, Gender and Hours variables also have an impact but not that much. While the other variables, Accounts, Employment and Deductions have such a poor impact on our predictions, in comparative to the other ones, that we could even consider to remove them as predictors.



Exercise 6

To be able to obtain the accuracy, precision, recall and AUC on the test individuals, first we have performed a prediction with the pruned model obtained during the exercise 4 and the data defined as test. For it, we have used the predict function of the library stats and we have needed to state the type of predicted data as class because the Adjusted data is an integer but limited to 0 or 1. Afterwards, the confusion matrix has been calculated with caret library in order to be able to compare this predicted classes with the observed ones. Thanks to the results retrieved with this confusion matrix we have been able to calculate all the qualities of our solution asked. Confusion matrix calculates the true positive, false positive, true negative and false negative values obtained with our prediction model.

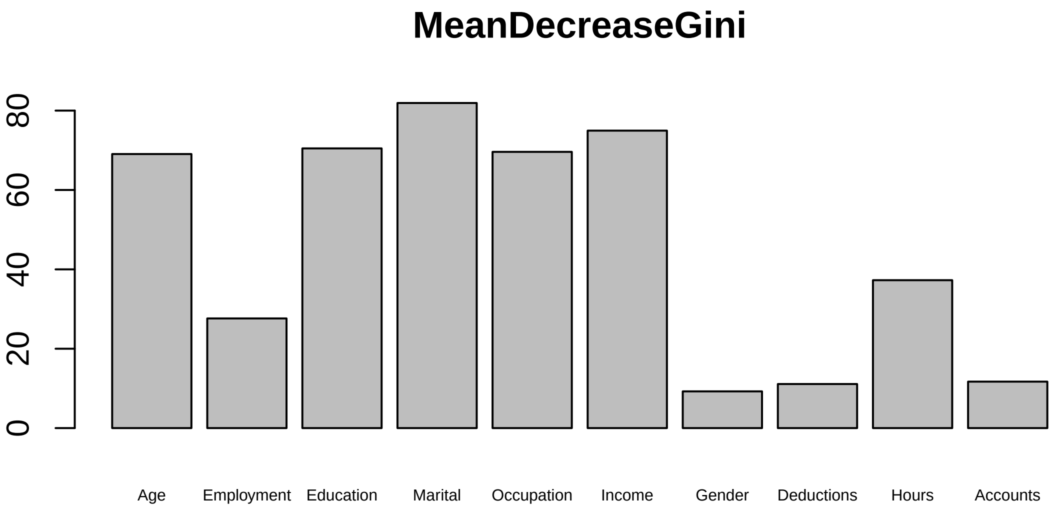
```
##   accuracy precision   recall       AUC
## 1 84.25787  74.21862 88.58195 0.8207108
```

- The accuracy measures a ratio of correctly predicted observation to the total observations. As our result obtained is 84%, then we can state that our model just miss 16% of its predictions.
- The precision is the ratio of correctly predicted positive observations to the total predicted positive observations. We have obtained a 74% which is pretty good.
- The recall, also known as sensitivity, is the ratio of correctly predicted positive observations to the all observations in the actual class. Our result obtained is 88
- Finally, the AUC (Area Under the Curve) is the average value of sensitivity for all possible true negative values. As our result obtained is 0.82, we can consider it as good because it is much close to the perfect accuracy that is represented by 1.

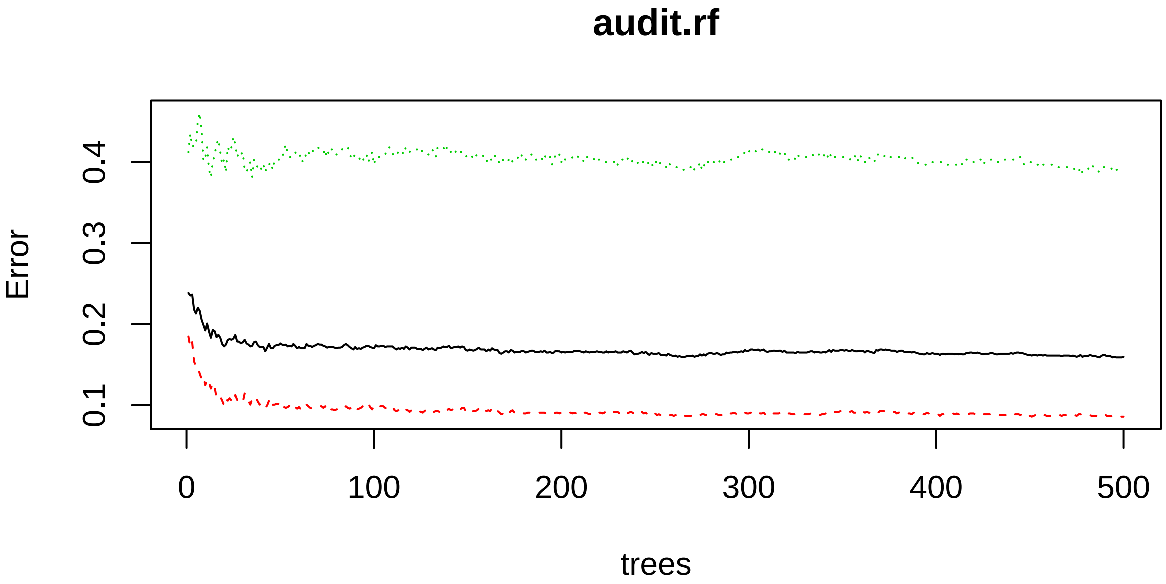
Exercise 7 Now, we want to perform a Random Forest on the same data. So, we have trained a random forest using the randomForest package with the same formula for predicting the Adjusted variable than the one used during the exercise 4. With the results obtained, the confusion matrix has also been calculated in order to be able to compare this predicted classes with the observed ones and we have calculated again the qualities of our solution obtained.

In order to compute the average precision on the testing, we have averaged the class precision computed from the confusion matrix. The average precision has turned out to be 74%.

```
## [1] 74.19167
```



Taking a look at the MeanDecreaseGini, we find that the most important variables, according to the resulting Gini index, are Age, Education, Marital, Ocupation and Income. Even the 5 most important variables are the same to the ones observed in exercise 5, we can observe that there is a remarkable difference regarding their magnitude.



Then, the accuracy of the classifier can be computed by subtracting 1 to the average error rate OOB (grey line in the previous plot) for all trees. The final accuracy has been 83%.

```
## [1] 83.18435
```

Finally, we can see that the accuracy and precision obtained for both predictors are almost the same. So we can see that Random Forest obtains the almost as good solution as Decision Tree for this dataset given, but no prunning to obtain an optimal solution is needed.