INTRODUCTION

The model described in this report was created to predict the use of drugs by future patients. The model is based on information about 600 patients divided in people who never used and used drugs throughout their lives.

The data set is composed of 13 variables. *Class* has two values: 0 for "never used" drug and 1 for "used at some point", and represents the response variable. The purpose of this analysis is to try to predict the consumption of drugs based on personality traits of future patients. The other variables are the explanatory variables and the first four provide information about age, level of education, country of origin and ethnicity. The other are more related to the personality of the subject: tendency to experience negative emotions such as nervousness, anxiety or depression (*Nscore*); being a talkative, active and cheerful person (*Escore*); being open to experiences (*Oscore*); type of interpersonal relations characterized by altruism, trust, kindness (*Ascore*); tendency to be organized, reliable and efficient (*Cscore*); *Impulsiveness* measured by Barratt Impulsiveness Scale and finally the sensation seeking (*SS*).

The second part of this report is the exploratory analysis which provides information about the general relationship within variables. Then there will be an analysis of different models created using different techniques, in this case are k-Nearest Neighbours, Classification Trees and Support Vector Machine. Finally, there will be discussion and conclusion about the best model predicting new patients whether they will be drugs user or not.

EXPLORATORY ANALYSIS

The data must have the right format; therefore, they need to be checked before being used and thus *Class* has been modified as factor. Furthermore, the column *X* has been removed as it only represents the row number of data, otherwise, the model would be led mainly by this variable using as cut off the number of row which divides the data that have *Class* value equal to 0 and those that have *Class* value equal to 1, as illustrated in figure 1. All the remaining variables have the same behaviour illustrated in figure 1.

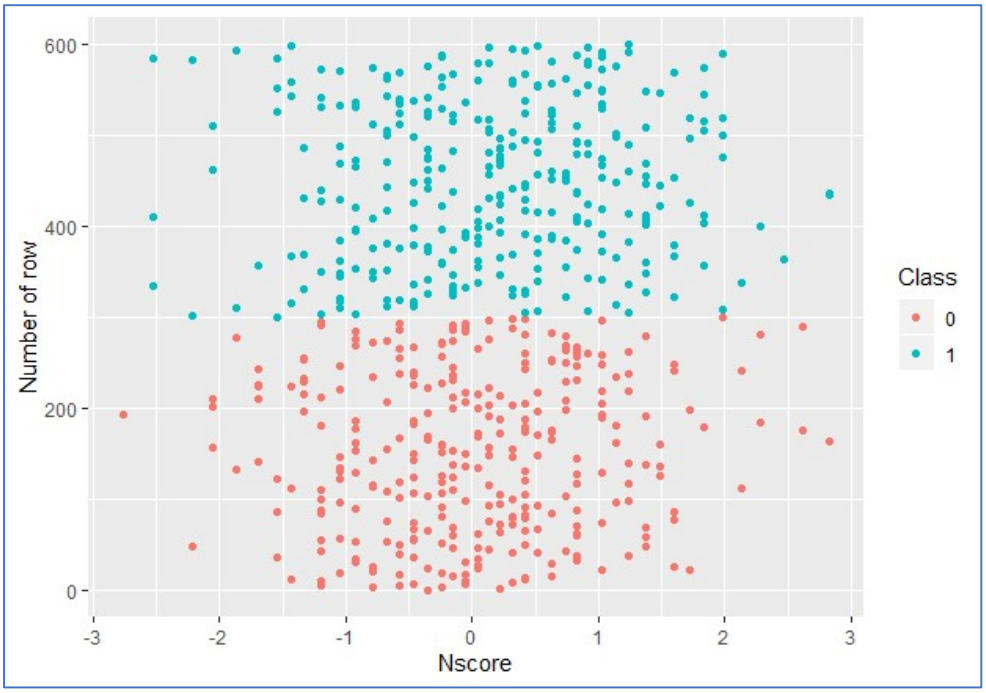

Figure 1

The whole data set has been divided in train test (50%) which is used to create a model; validation set (25%) is used for comparing different models and choose the best one based on the value of Correct Classification rate. Test set (25%) instead is used to make predictions and control how well the model chosen classifies future observations. Moreover, the train data set has been divided with the intention of obtaining plots of all variables versus all the other distinguished with regard to the two values of the variable *Class*.

Figure 2 shows one of these plots about *Age* and it is possible to observe that in some graphs there is a noticeable separation between the two colours, with some overlap. For example, looking at the graph versus *SS* the blue points are mainly concentrated in the bottom right corner whereas the red ones are in the top left corner. This can be read as younger people with higher value of *SS* are more likely to use drugs. A similar behaviour is present also in graphs versus *Impulsive* and *Oscore*. From all graphs it seems that, in general, younger people are more likely to use drugs. For these reasons, the variable *Age* could be a strong predictor. On the other hand, the variable *Education* does not seem to be so important in predicting the use of drugs except maybe for *Age* and *SS*, as shown in figure 3. Same trend has *Xcountry* and *Ethnicity*.

Higher values of *Oscore* seem to be related with a high probability of drug use. In figure 4 it is perceptible, in almost all the plots, a quite clear separation of the data between people who never used drugs and those who used it at some point, making *Oscore* a strong predictor. Similar behaviour is present with other variables, like *Ascore, Cscore* and *SS*. The rest of the variables have a trend like *Education*, making them probably weaker predictors.

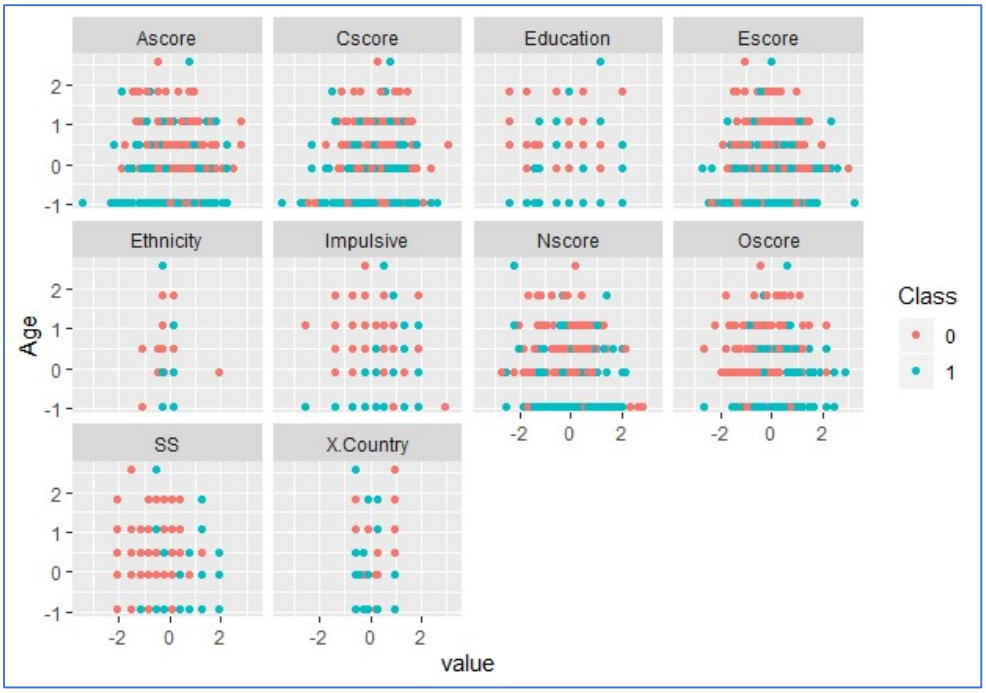

Figure 2

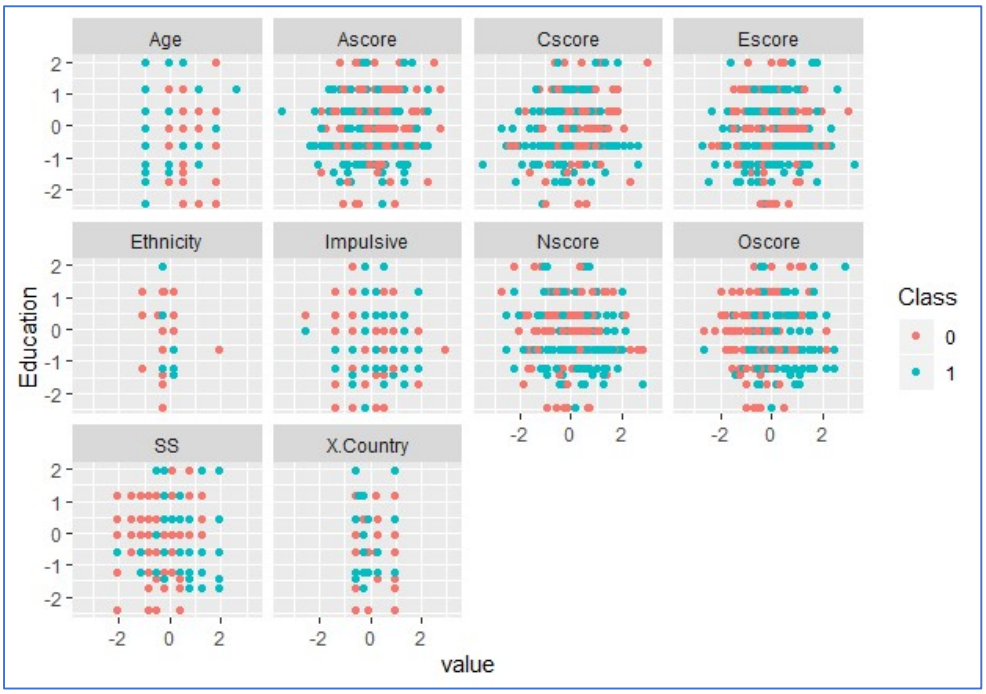

Figure 3

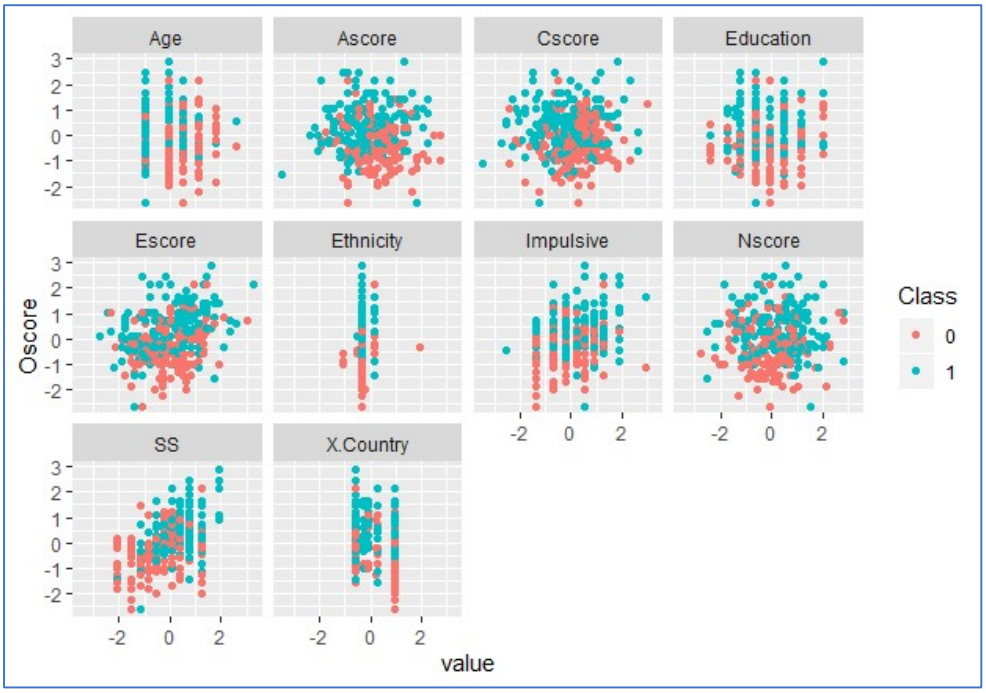

Figure 4

ANALYSIS

- **kNN**

The first model has been created using k-Nearest Neighbours (kNN) technique. This technique predicts the class of each point by calculating the majority of the k neighbouring points. For example, if k equals 5, the algorithm takes into consideration the 5 points closer to the point of reference. And if for example 3 of these 5 neighbours have class equal to A and 1 equal to B and the last equal to C, the point of reference will be of class A, because the majority of its neighbours have class equal to A. For these reasons, the choice of the value of k is important as different numbers can lead to different results, with different values of classification rate which indicates the performance of the model.

For the data set used in this analysis, as mentioned above, the model is initially created using the train data and the Correct Classification Rate is calculated for different values of k (in this case 55) using the validation data. It is possible to observe in figure 5 that the highest value of Correct Classification Rate is for k equivalent to 30, as indicated by the red line. Afterward, predictions have been made on this model using the validation data set giving a Correct Classification Rate of 80%.

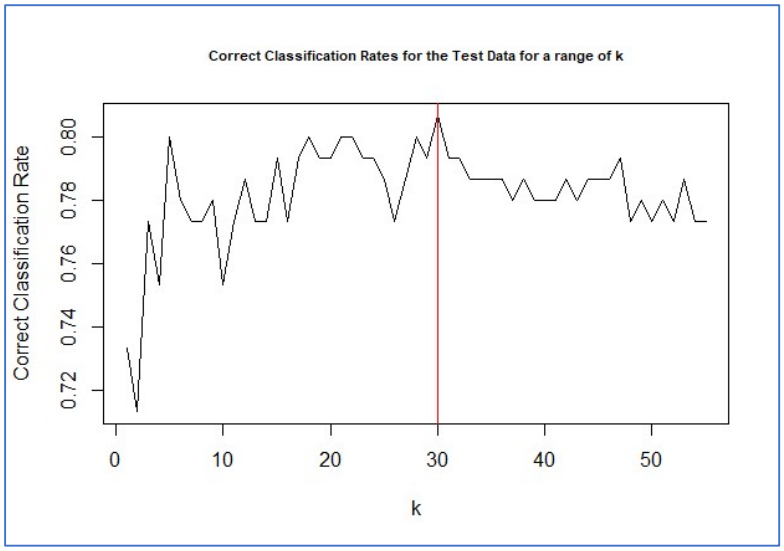

Figure 5

- **Classification Trees**

The second technique used is Classification Trees. This technique works by splitting the observations into a number of regions, and the prediction is made based on the mean or mode of all points that belong to a certain class falling in a specific area. Initially, a first general tree has been created and it is shown in figure 6.

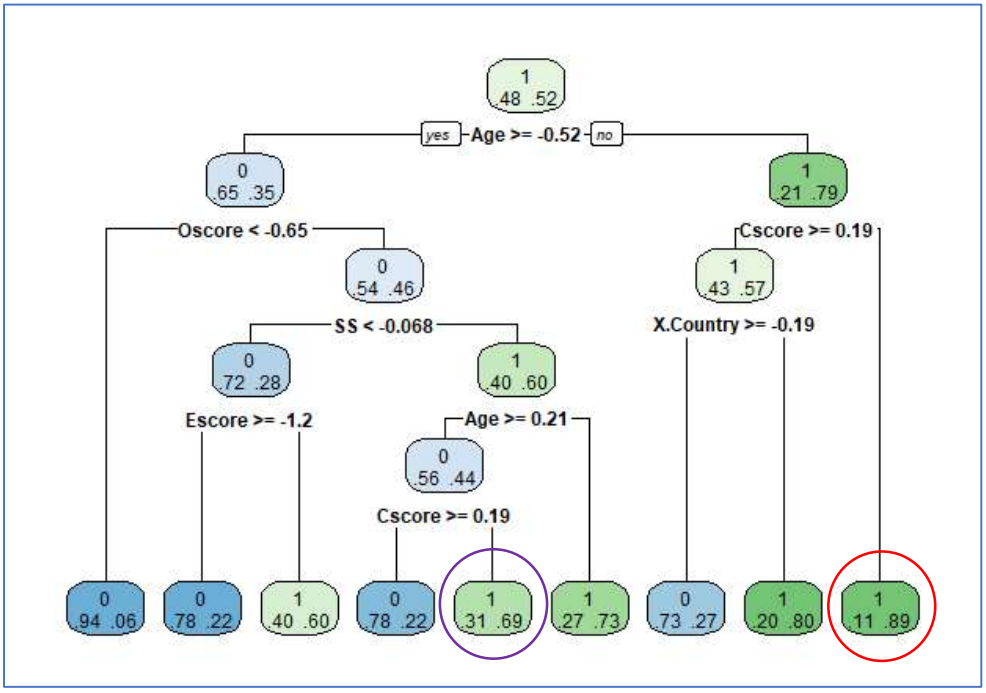

Figure 6

For this model, for example, a younger person (*Age* lower than -0.52) which is not organized, reliable and efficient (*Cscore* lower than 0.19) is more likely to use drugs (with a probability of misclassification equal to 0.11; red circle) than an older person who needs further more peculiar personality characteristics to be classified as a drug user (*Oscore* bigger than -0.65; *SS* bigger than -0.068; *Age* bigger than 0.21 and *Cscore* lower than 0.19, and a probability of misclassification equal to 0.31; purple circle).

Model trees are characterised by a high variance and they can be influenced by the way the data have been split. Bootstrap Aggregating (Bagging) helps with this issue and a second model have been created using Random Forest package and train data. The output of this model is the following:

```
Call:
 randomForest(formula = Class ~ ., data = train.data, mtry = 10, ntree = 200)
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 10

        OOB estimate of  error rate: 26.67%
Confusion matrix:
    0   1 class.error
0 103  40   0.2797203
1  40 117   0.2547771
```

The Out Of Bag (OOB) estimate of error is an error rate calculated on those data that are not included in the bagging process and thus used as test set. In the model above this error is 26.67% which is a good grade.

The third technique used is Random Forest which creates uncorrelated trees. This aspect of Random Forest deals with the fact that if one of the variables is a stronger predictor than the other variables, it will be used in all decision trees created with Bagging, obtaining an almost-identical series of trees. Random Forest randomly excludes some of the variables every time it makes a split. In this way it creates a series of uncorrelated trees and, on average, the predictions will be more reliable. The output of this model is as follow:

```
Call:
 randomForest(formula = Class ~ ., data = train.data, ntree = 200)
               Type of random forest: classification
                     Number of trees: 200
No. of variables tried at each split: 3

        OOB estimate of  error rate: 26.33%
Confusion matrix:
   0   1 class.error
0 99  44   0.3076923
1 35 122   0.2229299
```

The OOB estimate of error is 26.33, similar to the previous one and still a good rate.

As it is possible to observe in figure 7, the variable *Age* was the one leading the first split in Bagging whilst Random Forest brings it back to values similar to the other predictors in terms of Gini index which measures the impurity at a node of a tree.

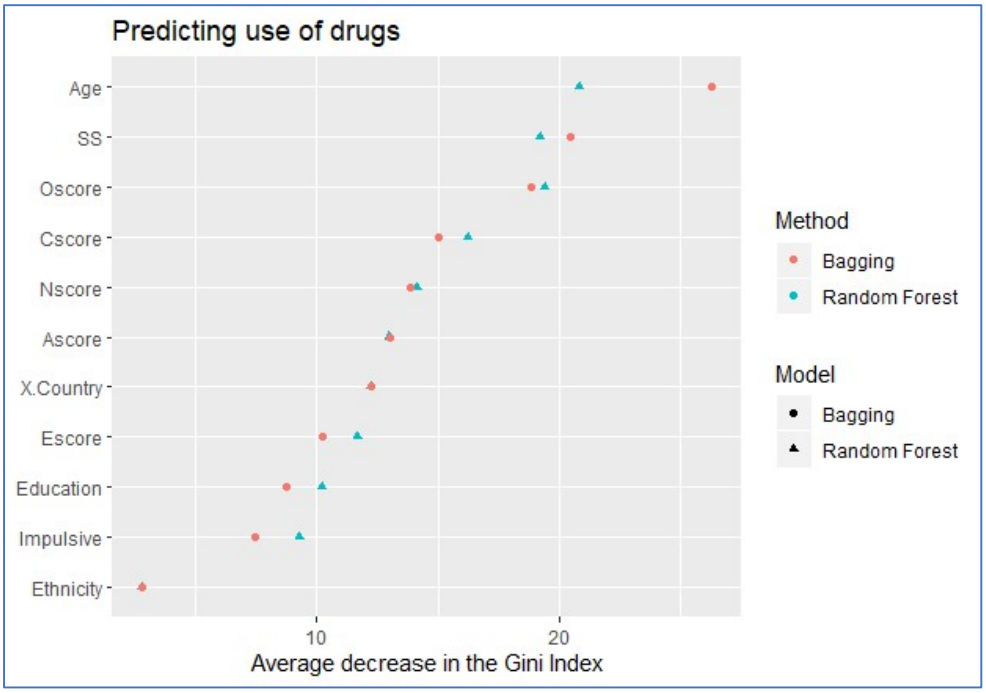

Figure 7

The three models presented above, have been created using the train data set. Subsequently, they have been validated using validation data set and the Correct Classification rates have been calculated.

- **Support Vector Machine**

The third and last method used is Support Vector Machine (SVM) which creates a surface to separate the data belonging to the two classes, maximising the distance between the closest points of the two classes and the surface itself. A new observation will be simply classified based on which side of the surface will fall. To make the technique more flexible a small misclassification is allowed. This comes with a cost, and the factor C, that must be specified in the function *tune* of the library *e1071* used in R, indicates how much "error" it is admissible.

The first model created with SVM technique is the linear classification. This model has been created testing different value of the factor C and the best performing model is the one with C equal to 0.01, as it is possible to observe below at the voice *cost:*

```
Call:
best.svm(x = Class ~ ., data = train.data, cost = c(0.001, 0.05, 0.01, 0.1, 1, 10, 100),
    type = "C-classification", kernel = "linear")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  0.01

Number of Support Vectors:  209
```

The number of *Support Vector* indicates how many observations (vectors) have been used to create (support) the surface.

The second model created with SVM is the polynomial classification. Since it is needed to consider other factors in addition to C, two different models have been created in order to find the best model considering different values of the factors C, Gamma and Coef0. The best parameters of the two polynomial models are indicated in the two tables below:

Best Parameters Polynomial model 1		
Gamma	**Coef0**	**Cost**
1	2	0.001

Best Parameters Polynomial model 2		
Gamma	**Coef0**	**Cost**
0.01	3	100

Both models, of second degree, have been validated using validation data set and the corresponding Correct Classification rates have been calculated.

The last model has been created using the radial classification. Again, different values of the parameters C, Gamma and Coef0 have been checked and the best ones are indicated in the table below:

Best Parameters Radial Model		
Gamma	**Coef0**	**Cost**
1	0	10

The model has been afterward validated using validation data set and the corresponding Correct Classification rate has been calculated.

Form the table below it is possible to observe that amongst the three models created using Support Vector Machine technique, the linear model has the highest Correct Classification rate, therefore, the linear SVM model will be used to create a plot in order to have a visualisation of the separation surface.

Model	Correct Classification Rate (%)
SVM Linear	79
SVM Polynomial 1	78
SVM Polynomial 2	78
SVM radial	61

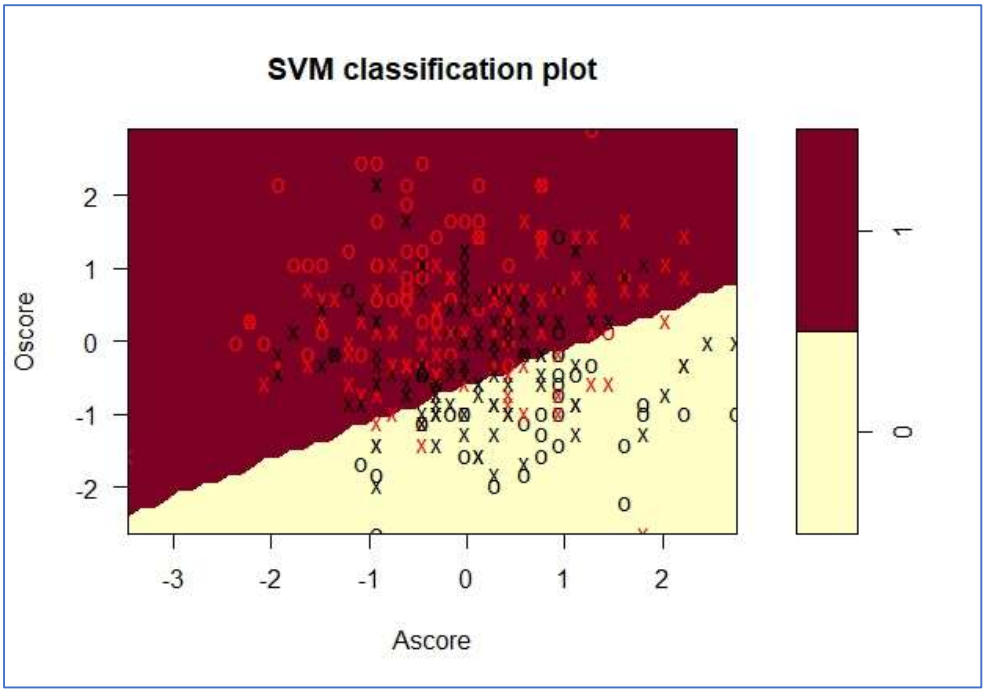

Figure 8

Figure 8 is the plot of *Oscore* and *Ascore*, two of the strongest predictors that have been recognize from the exploratory analysis, and shows the separation surface that classifies data as belongings to *Class* 0 or *Class* 1, i.e. people as drug user or not. In the figure it possible to distinguish "o" from "x", the latter indicates vectors that have been used to create the separation surface. Another characteristic noticeable from the plot above is a slightly higher amount of black points in the red area than red points in the yellow area. That could be explained by a higher error in classifying observations belongings to *Class* 0 as *Class* 1 than classifying observations of *Class* 1 as *Class* 0. This is confirmed from the calculation of the misclassification rate shown if the following table.

	CLASSIFICATION	
REAL VALUES	0	1
0	0.7215	0.2785
1	0.1408	0.8592

The table above shows the Correct Classification rates (on the diagonal) and the misclassification rates. It is possible to see that the model has a slightly higher error in classifying observations that have real value equal to zero as observations belonging to Class 1.

CONCLUSIONS

All the models described above have been trained using train data set, then validated using the validation data set. The Correct Classification rates have been calculated for all models and they are shown in the following table.

MODEL	CORRECT CLASSIFICATION RATE (%)
kNN	80
Single tree	74
Bagging	77
Random Forest	77
SVM Linear	79
SVM Polynomial 1	78
SVM Polynomial 2	78
SVM radial	61

The model that best performs is the one created with k-Nearest Neighbours showing a Correct Classification rate of 80%, therefore, predictions will be made on this model using test data set.

Summarising, in the table below are reported the accuracy, sensitivity and specificity of the chosen model.

Accuracy	Sensitivity	Specificity
79%	82%	77%

All the three parameters are high, indicating that the model performs well when predicting both used and never used drugs. However, the sensitivity indicates that amongst all of the people who actually used drugs, 82% of them are predicted as drugs users and 18% are classified as non-drug users when in reality they have used it, or will use it. On the other hand, specificity indicates that amongst all of the people who never used drugs, 77% are classified as non-drug users and 23% are classified as drug users when in reality they have never had or will never use drug throughout their lives. Nonetheless, the model classifies well both classes with a little higher misclassification error for those who never used drugs.