STATISTICAL LEARNING PROJECT

Luca Sangiovanni – 33794A

ABSTRACT AND INDEX

The aim of the project is to apply different unsupervised and supervised learning techniques to real-world datasets. I used two different datasets for the two parts, and at the end of each part I tried to give an interpretation to the results.
In some cases I tried to use different techniques for the same task, in order to compare the result and interpret the data in different ways and with more accuracy.

INTRODUCTION

This project is divided into two parts, and two different datasets have been used.

UNSUPERVISED LEARNING

The data used in this part refers to the statistics of basketball players from the 2018 – 2019 season, across 5 different european leagues.

SUPERVISED LEARNING

The data used in this part refers to a survey made to the students of a portuguese high school, with questions regarding the consumption of alcohol with regard to their social and academic characteristics.

GOALS OF THE STUDY

What is the aim of this study?

DIMENSIONALITY REDUCTION

What's the relationship between all the characteristic of a player?

CLUSTERING

Is it possible to divide players in clusters based on their characteristics? How do we interpret the different clusters?

REGRESSION

Can we predict the final grade of a student? How do different variables contribute to the prediction?

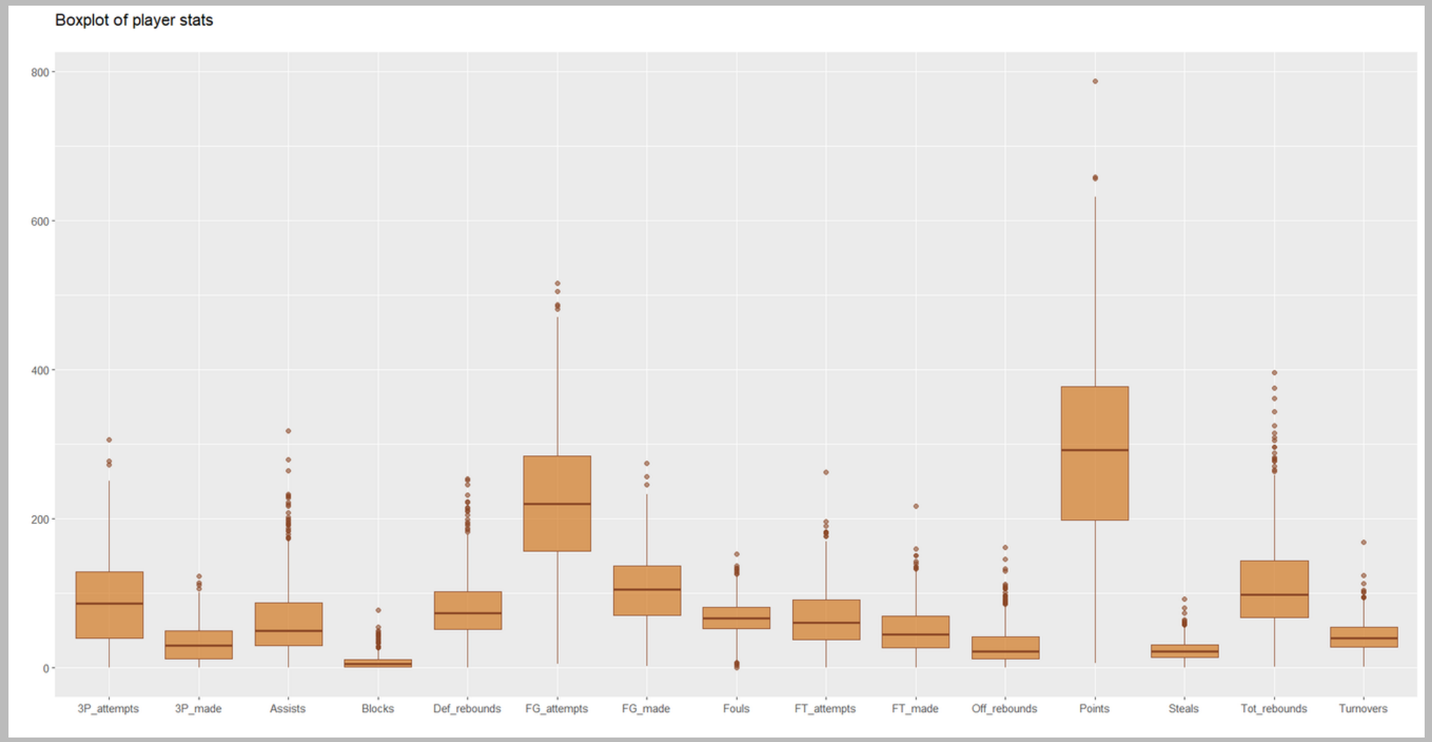
CLASSIFICATION AND DECISION TREES

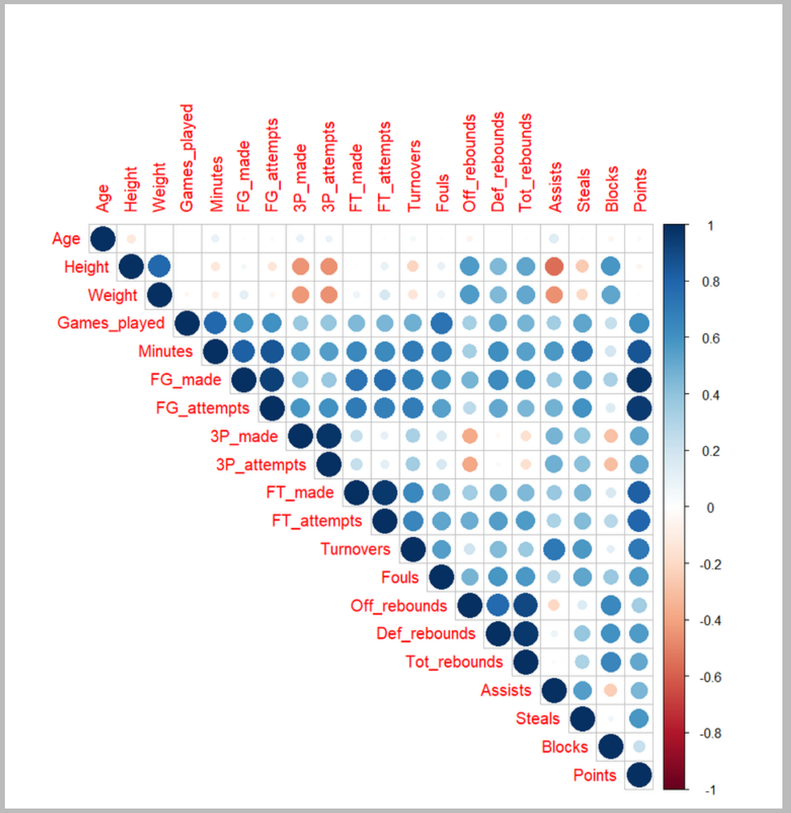
Can we determine, for each person and based on his characteristics, the type of alcohol consumer he is?

Unsupervised Learning – Basketball data

EXPLORING THE DATA

Each player has characteristics about his person (age, height, weight) and his statistics on the court. The presence of a small amount of outliers doesn't influence the interpertations.

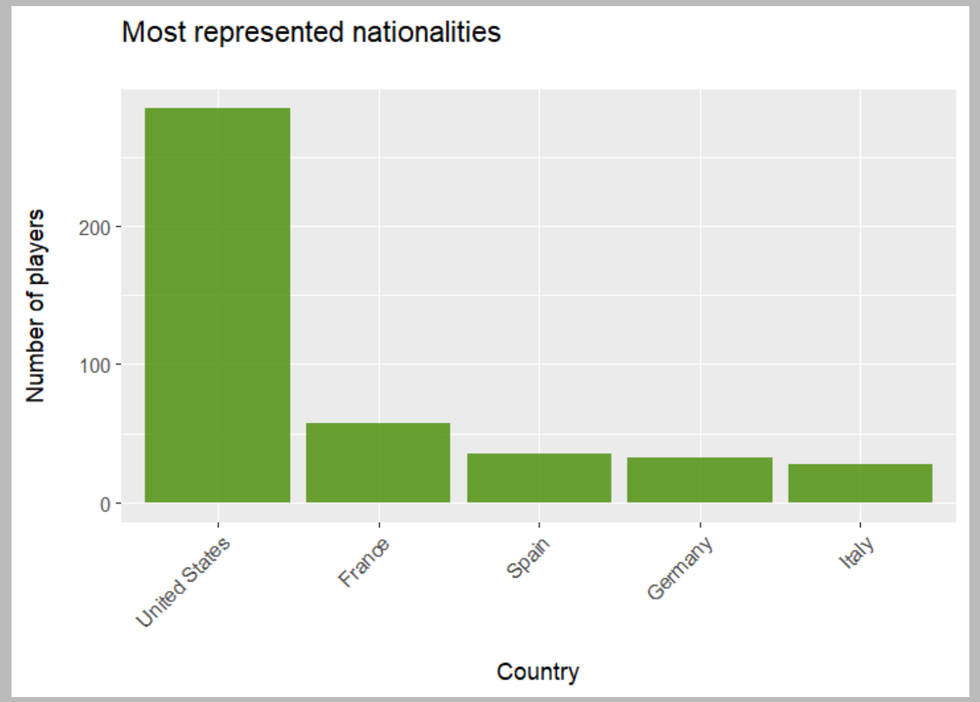


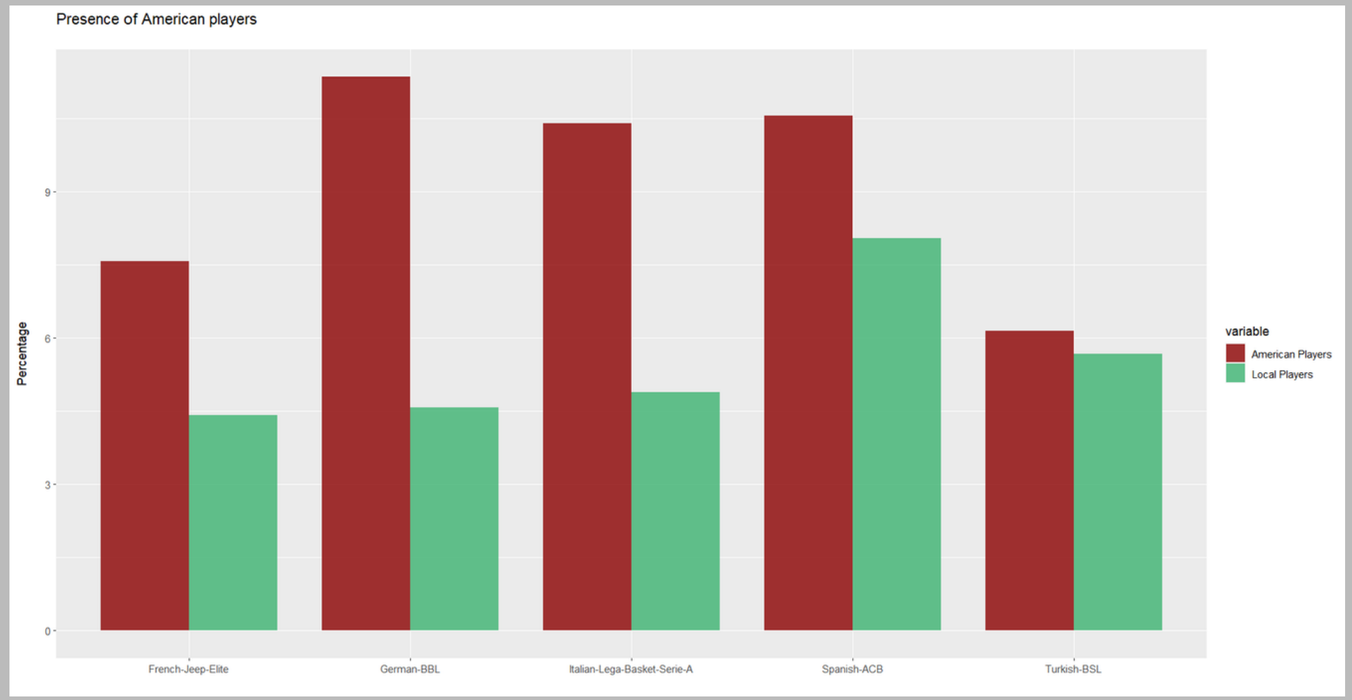


Unsupervised Learning – Basketball data

EXPLORING THE DATA

There is a strong presence of American players, that outnumber local players in all the leagues.

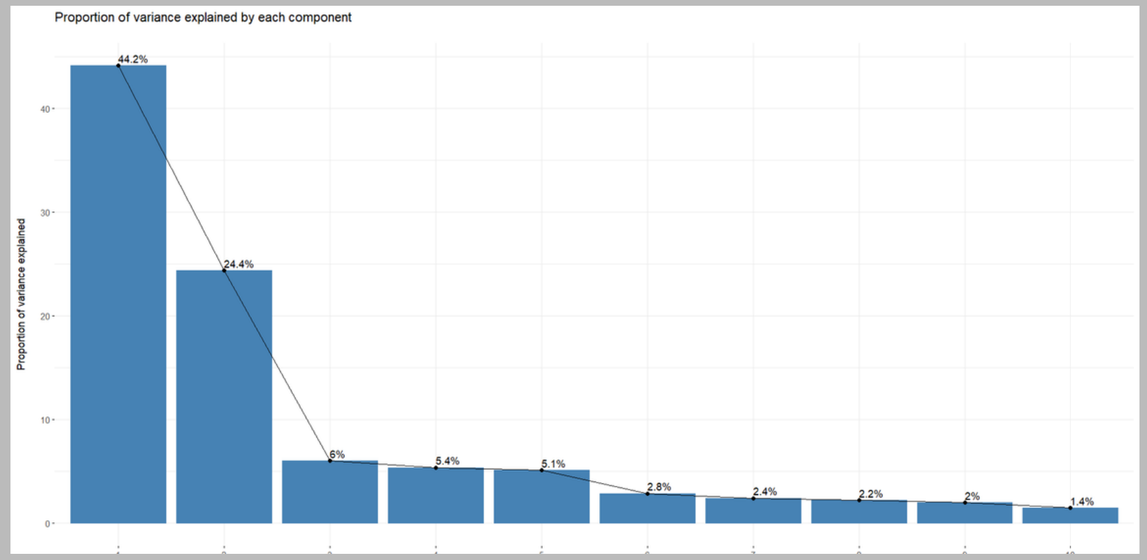




Unsupervised Learning – Basketball data

DIMENSIONALITY REDUCTION

```
Importance of components:
                          PC1    PC2     PC3    PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     2.9716 2.2073 1.09579 1.0354 1.01052 0.75381 0.69330 0.66226 0.62636 0.53822
Proportion of Variance 0.4415 0.2436 0.06004 0.0536 0.05106 0.02841 0.02403 0.02193 0.01962 0.01448
Cumulative Proportion  0.4415 0.6851 0.74517 0.7988 0.84983 0.87824 0.90227 0.92420 0.94382 0.95830
                          PC11    PC12    PC13    PC14    PC15   PC16    PC17    PC18      PC19      PC20
Standard deviation     0.49976 0.43145 0.36482 0.35578 0.25929 0.2145 0.13099 0.08945 1.149e-15 7.694e-16
Proportion of Variance 0.01249 0.00931 0.00665 0.00633 0.00336 0.0023 0.00086 0.00040 0.000e+00 0.000e+00
Cumulative Proportion  0.97079 0.98010 0.98675 0.99308 0.99644 0.9987 0.99960 1.00000 1.000e+00 1.000e+00
```



In order to reduce the dimensionality of the data I used the Principal Component Analysis (PCA). This method allows to better visualize and interpret the data. When choosing the optimal number of principal components, we look at the proportion of variance explained by each component.

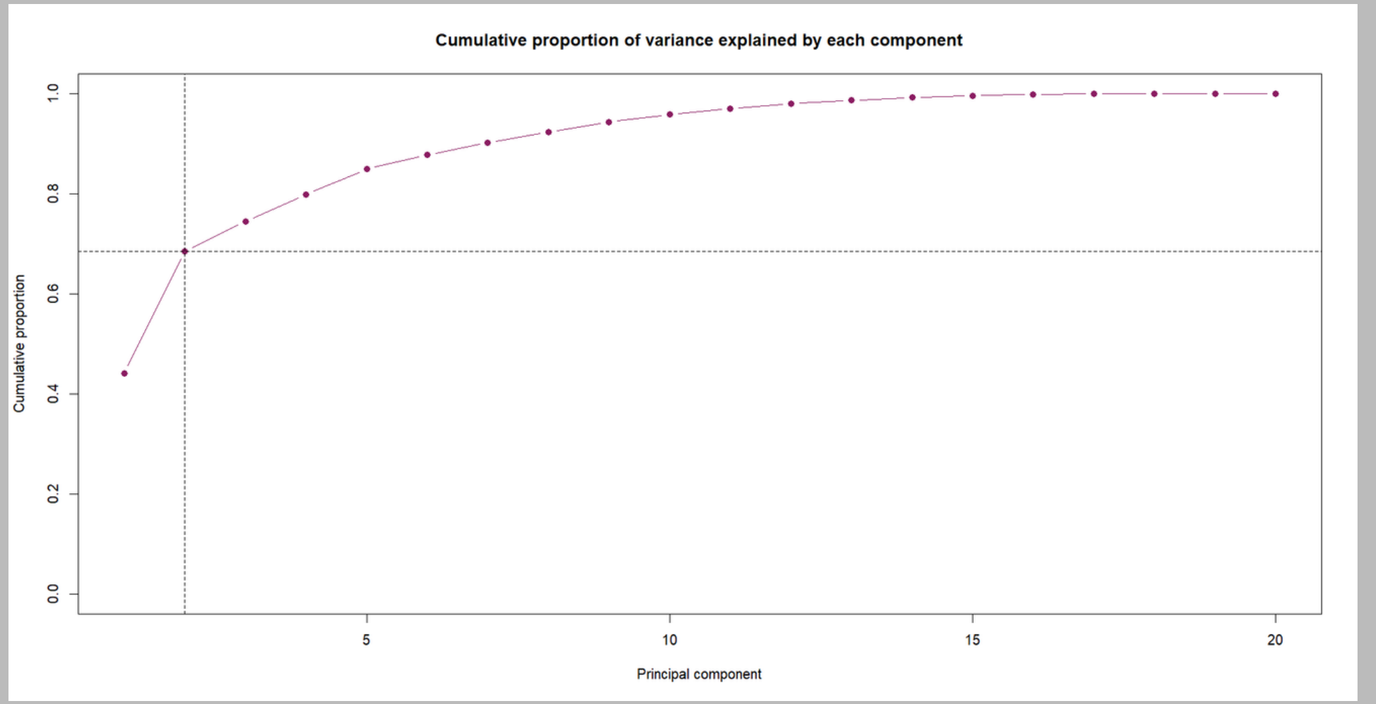
Unsupervised Learning – Basketball data

DIMENSIONALITY REDUCTION

By looking at the cumulative proportion of variance explained by each component, we conclude that the optimal number of components is 2.

68.5%

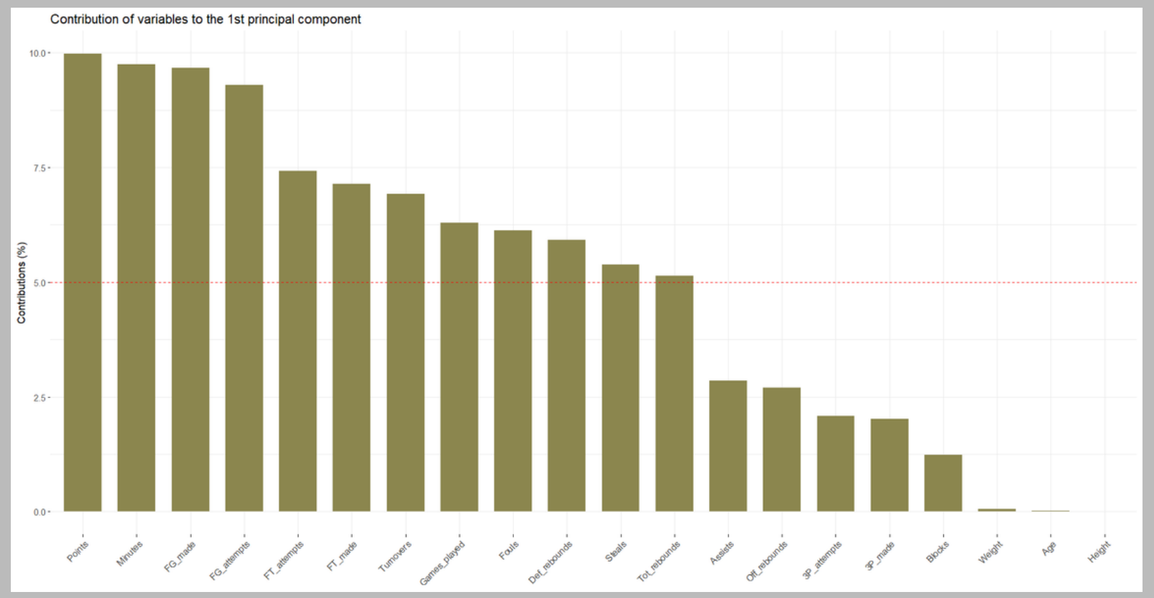
The variance explained by the first two principal components

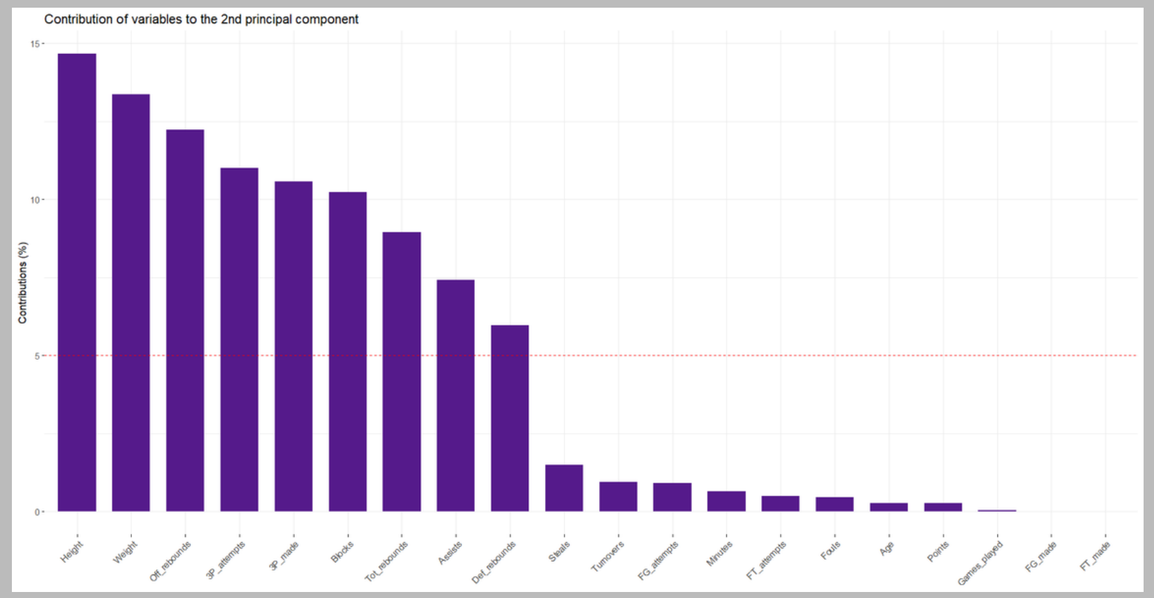


Unsupervised Learning – Basketball data

DIMENSIONALITY REDUCTION

By looking at the contribution of each variable to the two principal components, we can see that the first principal component is mainly composed by variables that influence the points scored, while the second principal component is composed by "secondary statistics" (height, weight, rebounds, etc).





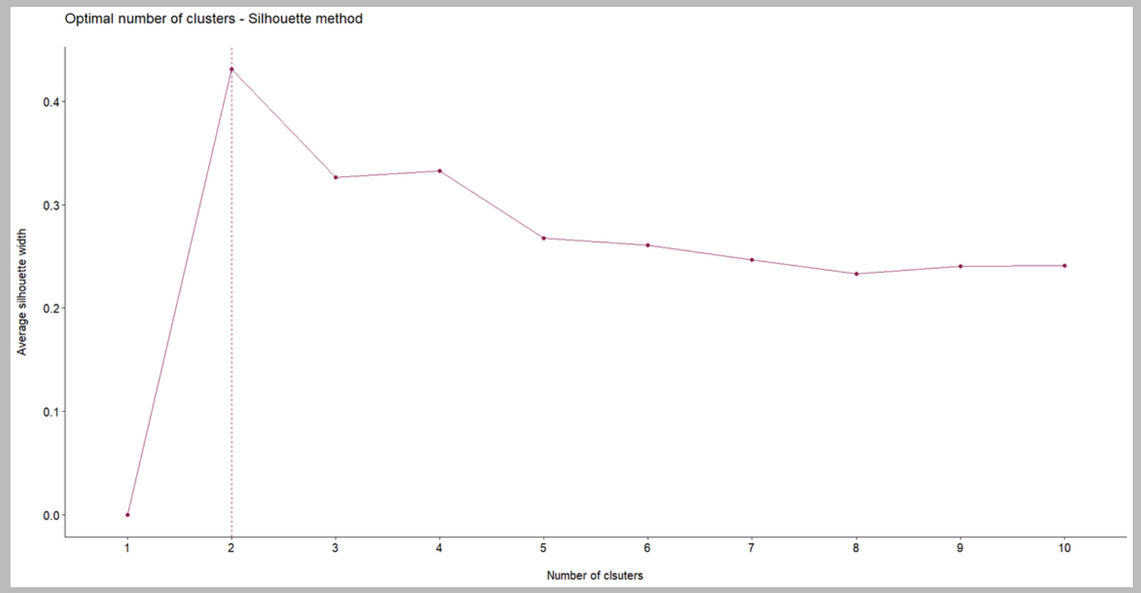
DIMENSIONALITY REDUCTION

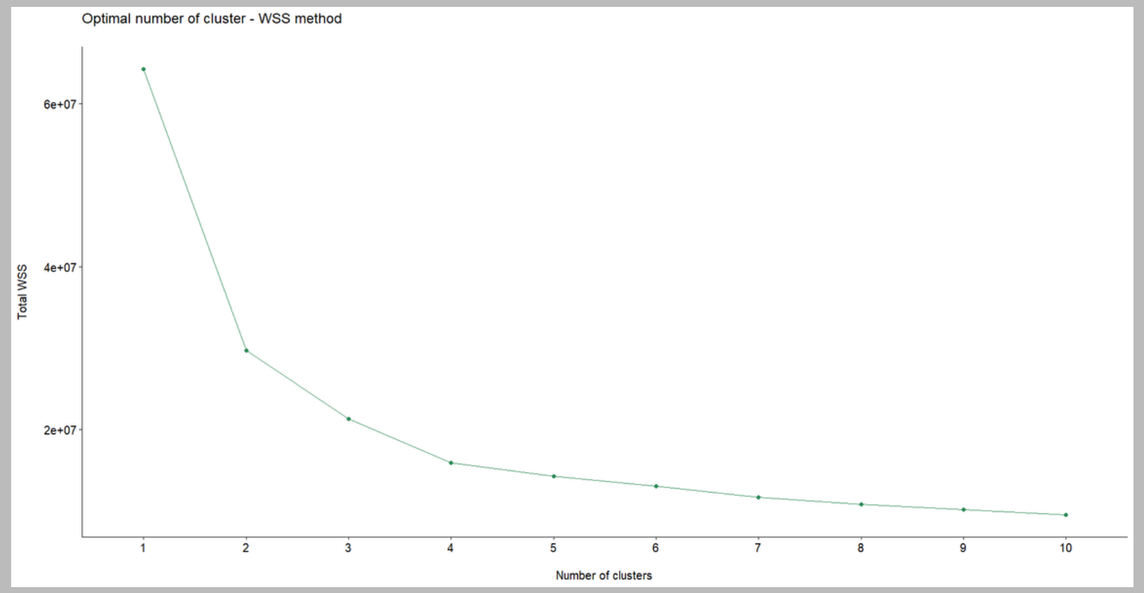
From the biplot we notice that there's no significant difference between american and non-american players. We notice instead some interesting correlations between variables.

Unsupervised Learning – Basketball data

CLUSTERING

When deciding the number of optimal clusters, we look at both the WSS plot and the silhouette plot. Following the “elbow method”, there is no clear evidence of a predominant value for the optimal number of clusters: we can see that a value of 2, 3 or 4 would be equally acceptable.

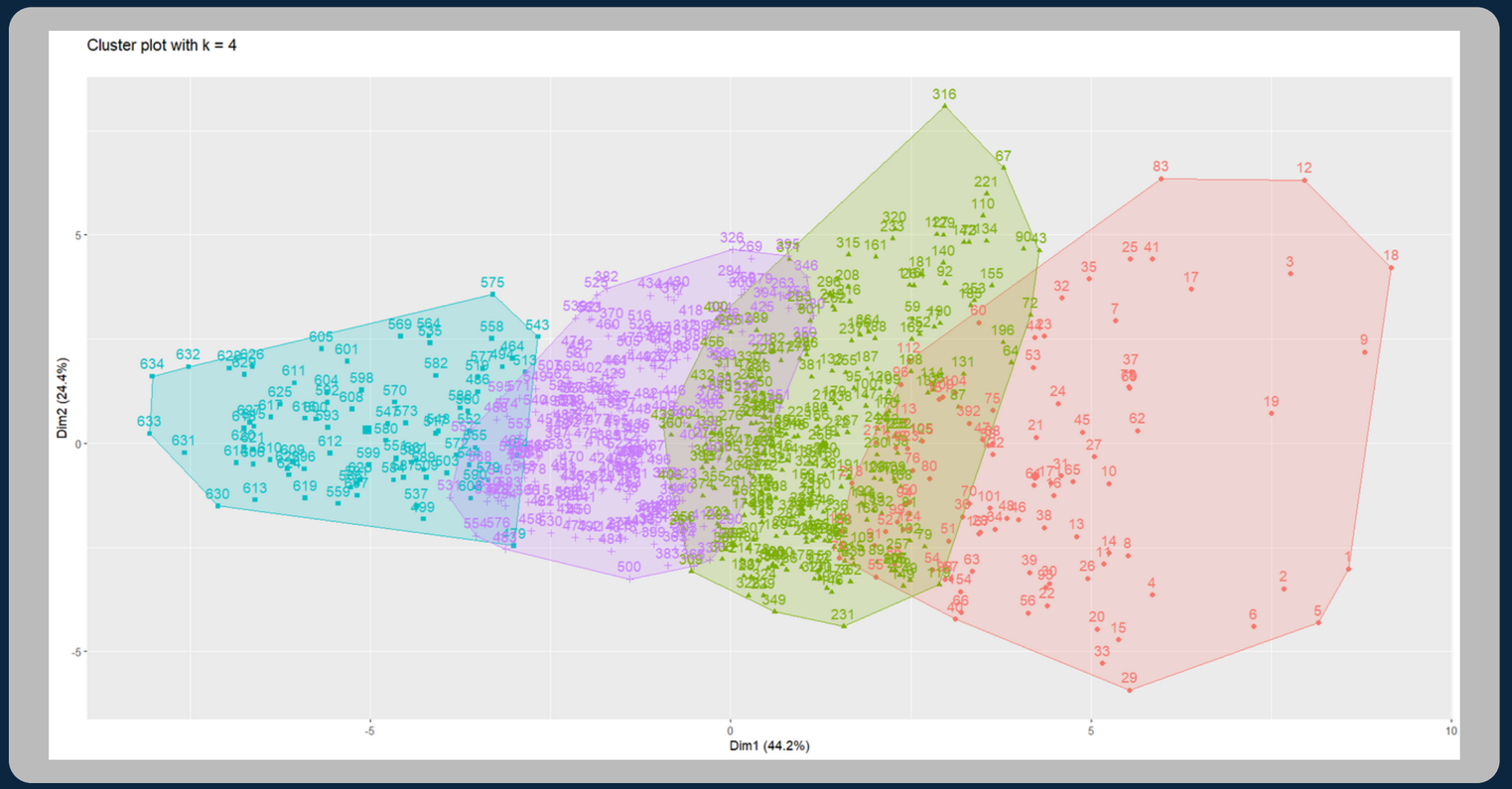




Unsupervised Learning – Basketball data

CLUSTERING

One of the possible clustering methods is the K – means. It creates a number of clusters based on the value preferred, and clusters the closest points, in terms of distance from the centroid. In this case, I chose k =4, so 4 clusters will be used:

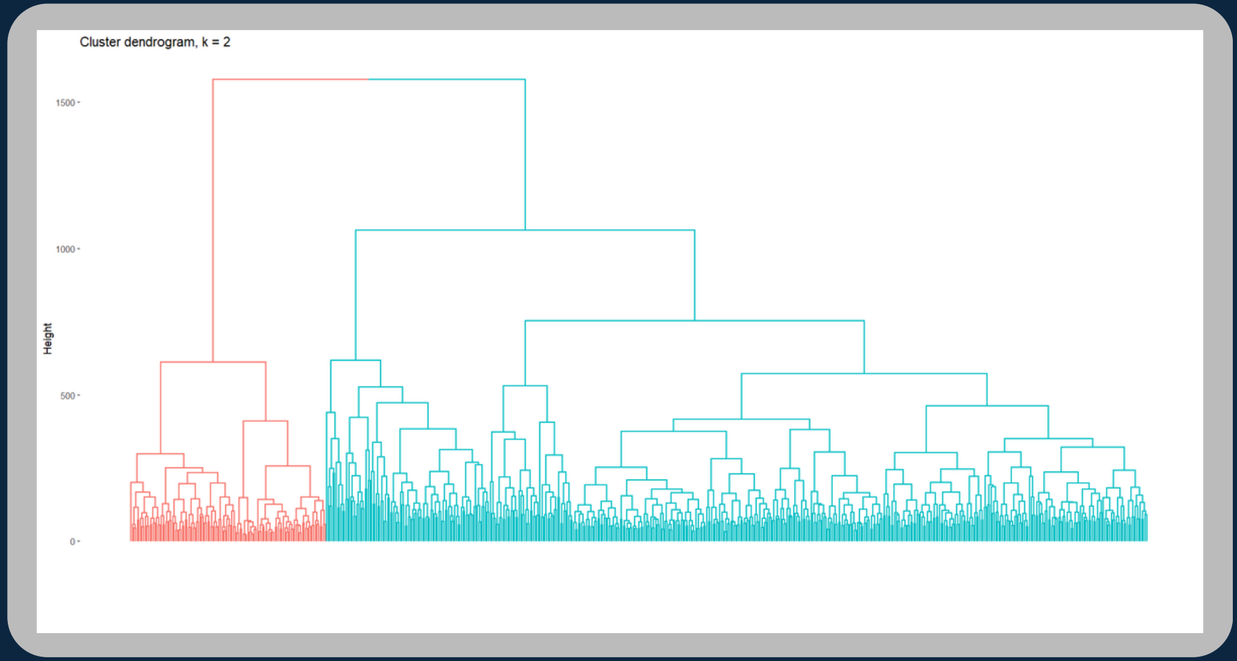


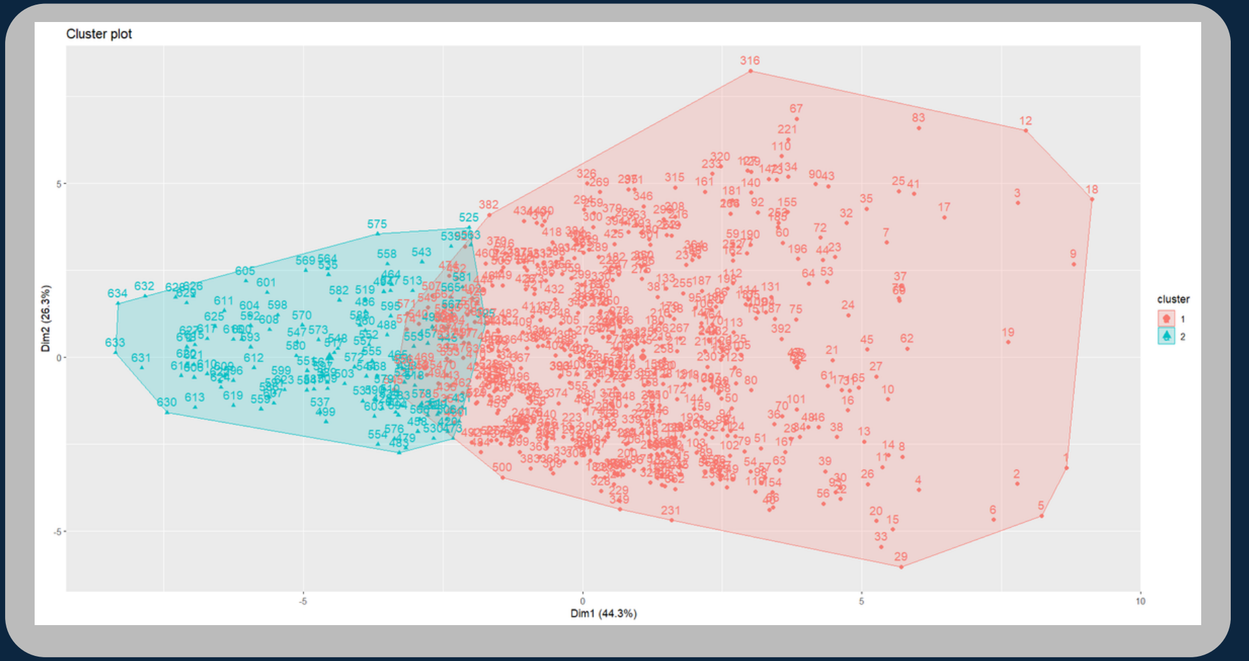
A centroid, in K – means clustering, is the center of each cluster. It corresponds to the arithmetic mean of data points assigned to the cluster.

Unsupervised Learning – Basketball data

CLUSTERING

An alternative clustering method is the hierarchical clustering. It creates groups so that objects within a group are similar to each other and different from objects in other groups. The custers can be visually represented in a hierarchical tree called dendrogram. In this case I chose to create 2 clusters.
On the left there is the dendrogram, and on the right the visual representation





Unsupervised Learning – Basketball data

CLUSTERING

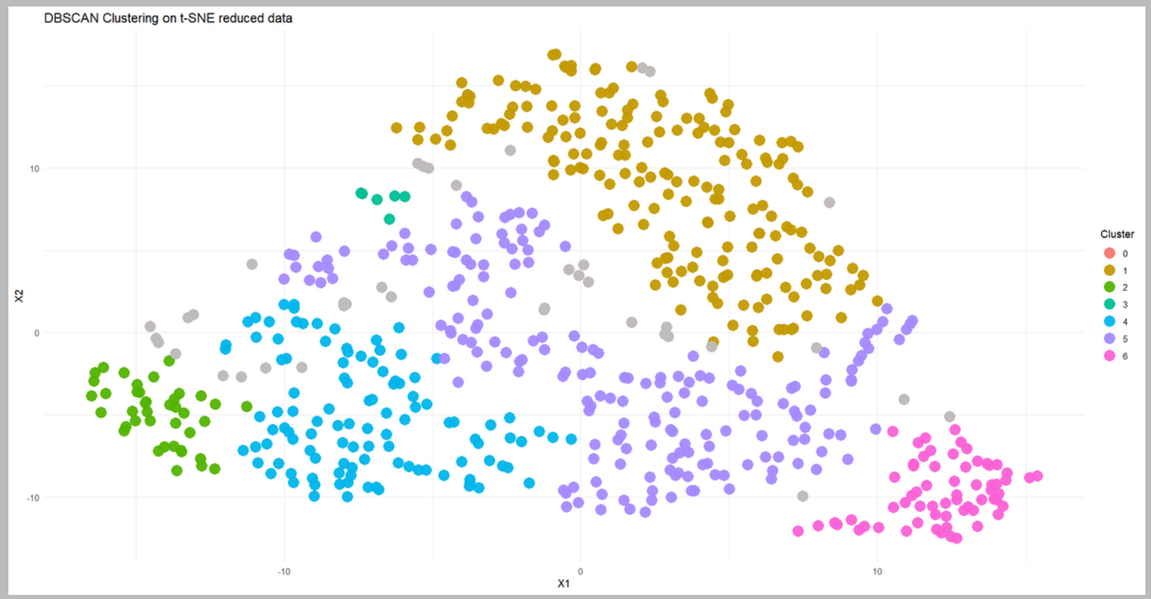
An alternative of reducing the dimensionality and creating clusters is the t – SNE. I then represented the clustering using the DBSCAN clustering method.

t – SNE

Alternative, non–linear dimensionality reduction technique. It focuses on preserving local structures and clusters in the data, meaning nearby points in the high–dimensional space are likely to remain nearby in the lower–dimensional.

DBSCAN

Density–based clustering algorithm. The key fact of this algorithm is that the neighbourhood of each point in a cluster which is within a given radius must have a minimum number of points.



INTERPRETING THE RESULTS OF THE UNSUPERVISED LEARNING

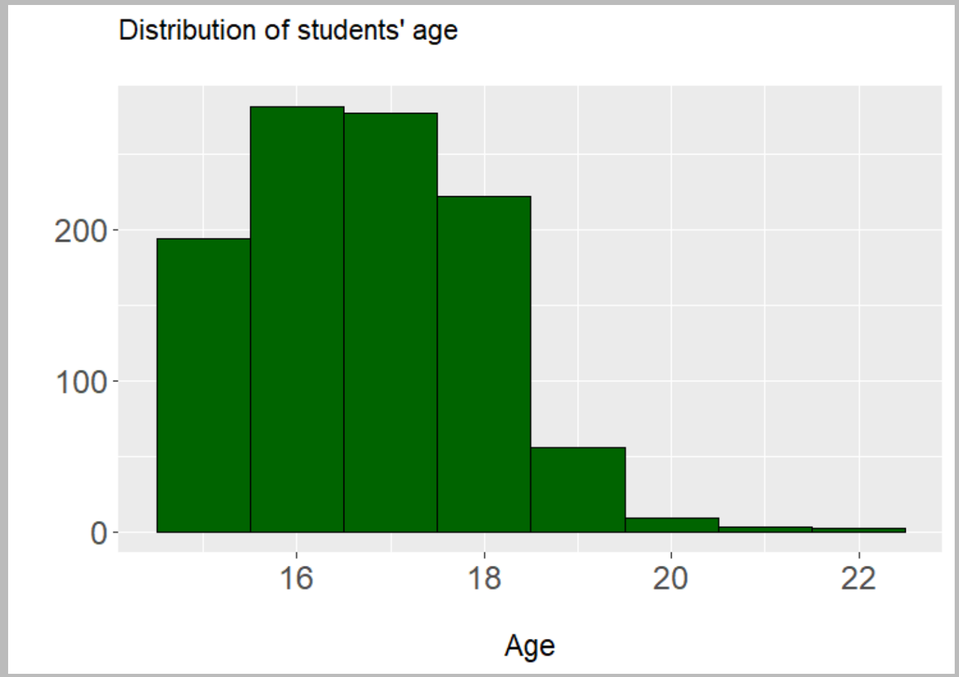
Using Principal Component Analysis was very useful in interpreting and analyzing the different variables present in the dataset. I was able to understand the correlations between variables, and understand how each variable is present in the component.

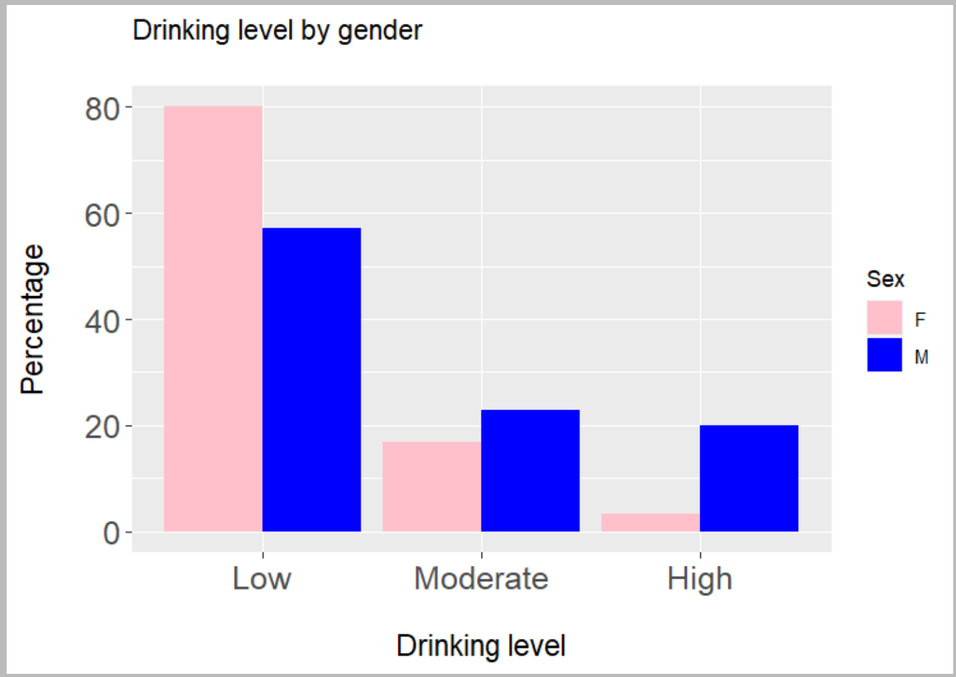
Clustering results can be interpreted in different ways: different clusters may refer to different levels of “strength” of the players, or even the role of the player. Another interpretation could be that each cluster refers to a different league. However, this last interpretation could be wrong, as most of the players are similar depsite the league they play in. So, the first two interpretations could be correct.

Supervised Learning – Alcohol consumption data

EXPLORING THE DATA

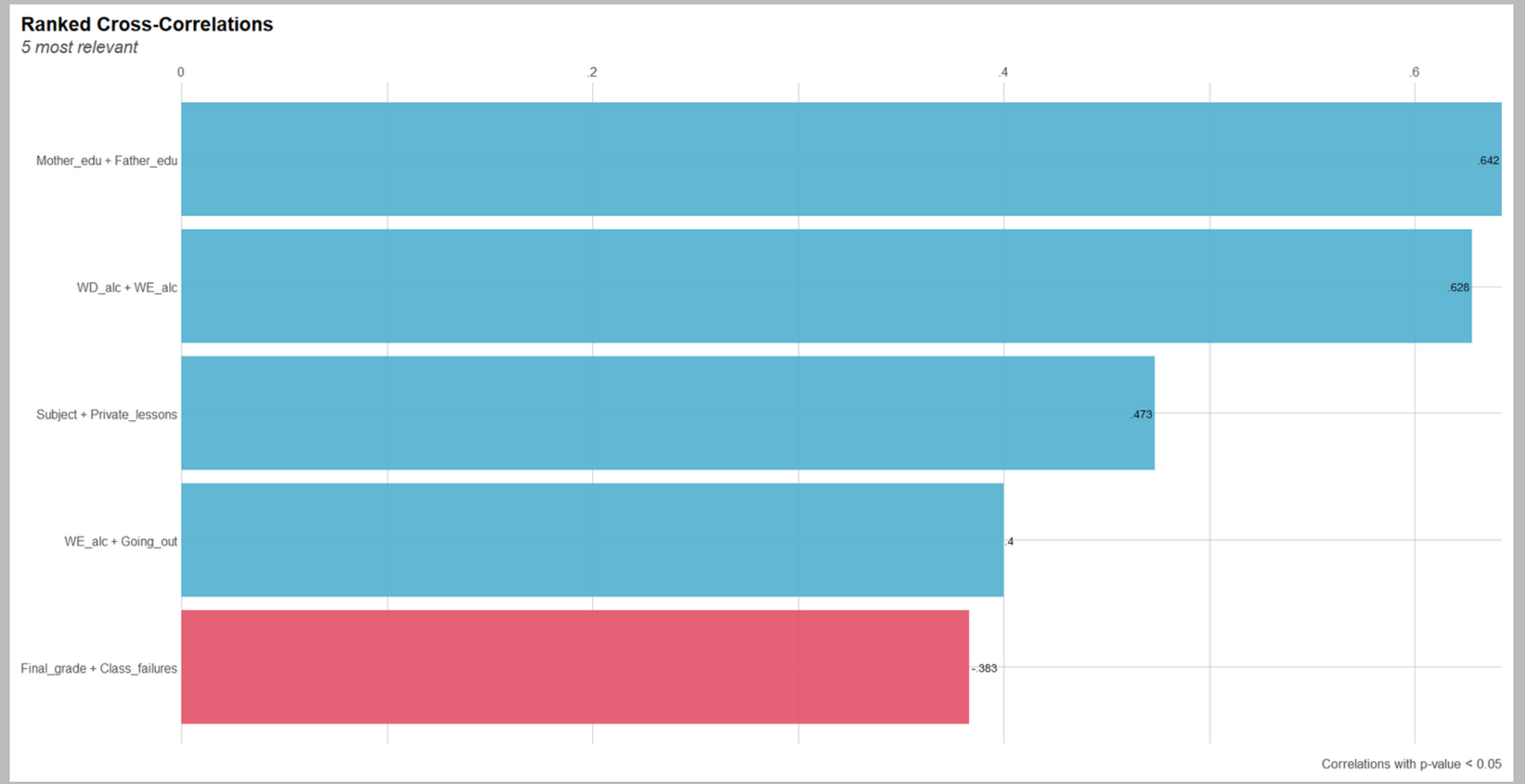
Passing on to the second dataset, we can notice how most of the people interviewed are betwee 15 and 18 years old, and male students tend to drink more alcohol compared to females.





Supervised Learning – Alcohol consumption data

EXPLORING THE DATA



Some variables are more correlated than others. Here we can see the 5 most correlated to each other.

Supervised Learning – Alcohol consumption data

REGRESSION

QUESTION

CAN WE PREDICT THE FINAL GRADE OF A STUDENT? HOW DO DIFFERENT VARIABLES CONTRIBUTE TO THE PREDICTION?

I used regression to create a model that would predict the final grade of the students, based on the informations provided.

75% Training data

25% Testing data

1.934

The value of the Durbin – Watson test. We can deduce that there's no autocorrelation.

1.047 - 2.047

The min and max values of the Generalized Variance Inflation Factor (GVIF). We can say that there's no multicollinearity.

Supervised Learning – Alcohol consumption data

REGRESSION

VERY SIGNIFICANT

Age

Subject

N° of class failures

Weekday alcohol consumption

Weekend alcohol consumption

School

Absences

Grades at the end of the first trimester

Grades at the end of the second trimester

SLIGHTLY SIGNIFICANT

Continue studies

Going out

Family size

In a relationship

Father job

Health status

Mother job

COULD BE OMITTED

Sex

Father education

Mother education

Travel time

Familty relationship

Extra activities

Parents status

Internet

Private lessons

Free time

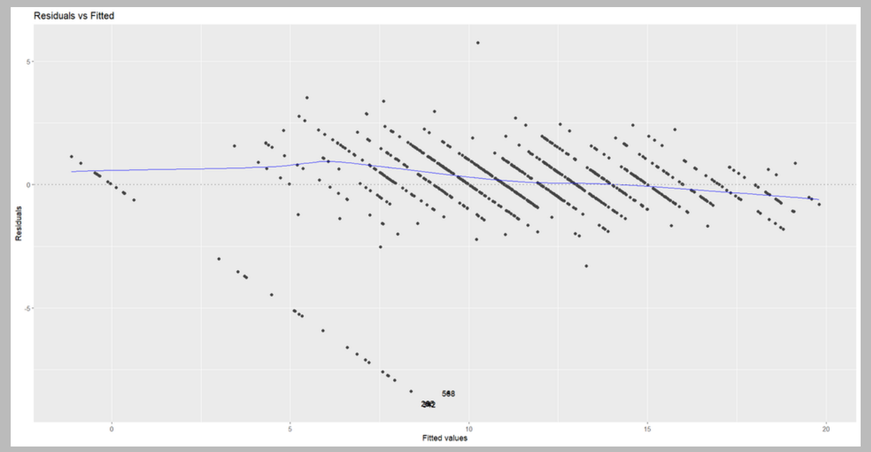
Something to notice:

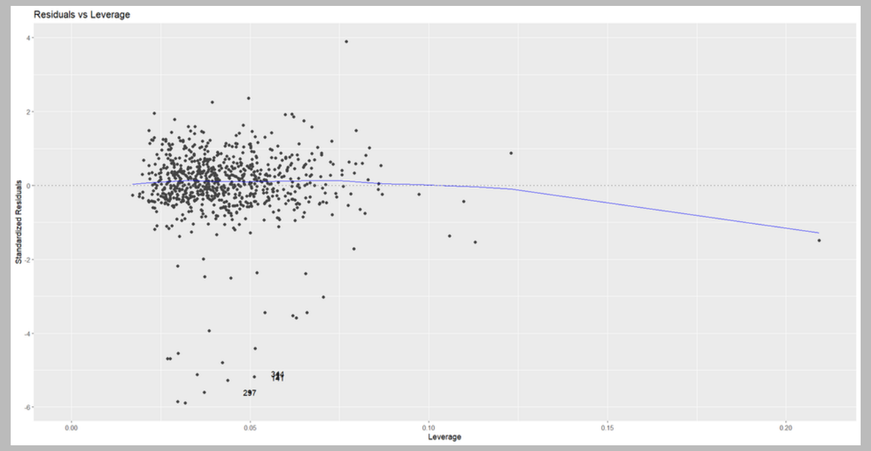
When creating the same model, but omitting the grades of the first and second trimester, we have some interesting changes:

Things that become more significant	Things that become less significant
Mother job	School
	Weekend alcohol consumption
	Absences

Supervised Learning – Alcohol consumption data

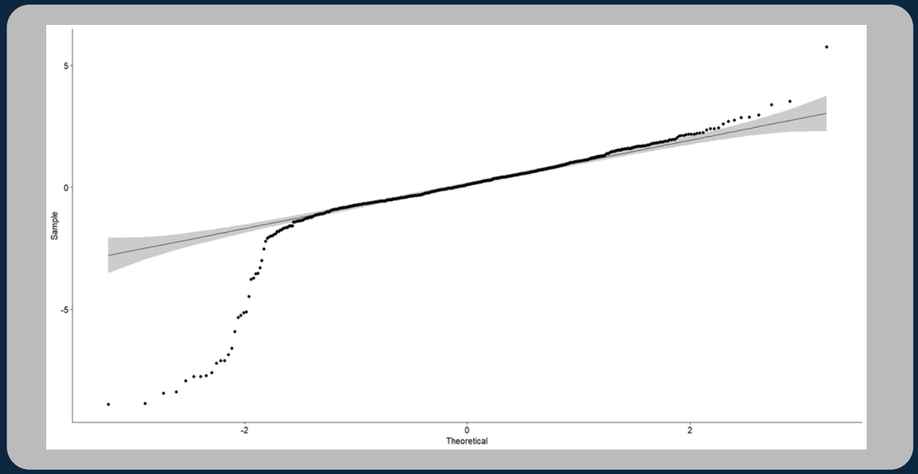
REGRESSION





In the Residuals vs Fitted plot there are some outliers, but the general relationship is linear, as the line is orizontal.
A similar thing can be seen in the Residuals vs Leverage plot.

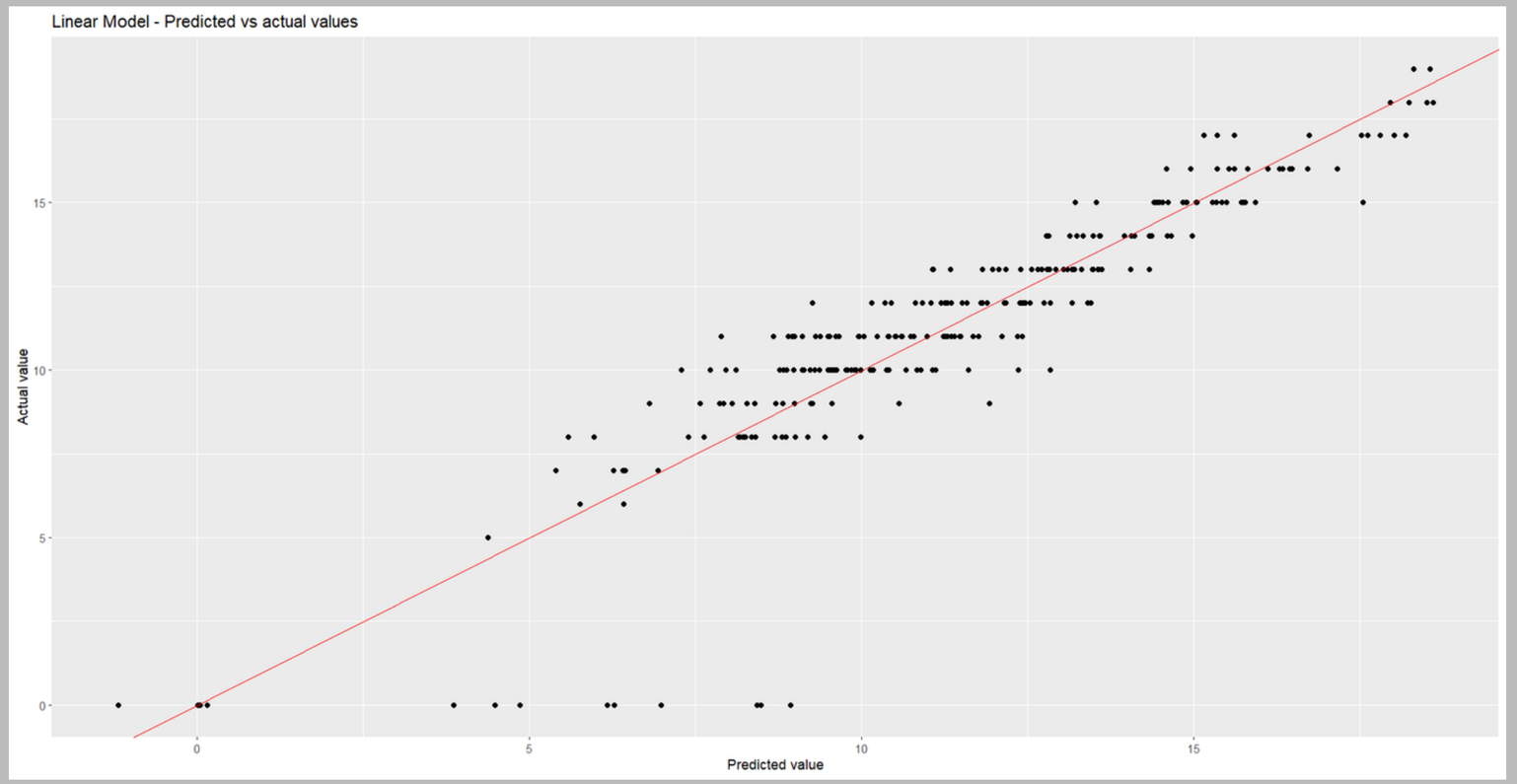
The general idea is that the blue line should stay as close as possible to the dashed line, which represents the Cook's distance



0.766 The value of the Shapiro – Wilk test. We can assume normality.

Supervised Learning – Alcohol consumption data

REGRESSION



Improvement of the R – Squared after the eXtreme Gradient Boosting and cross – validation (5 folds):

0.8522 ⟶ 0.9548

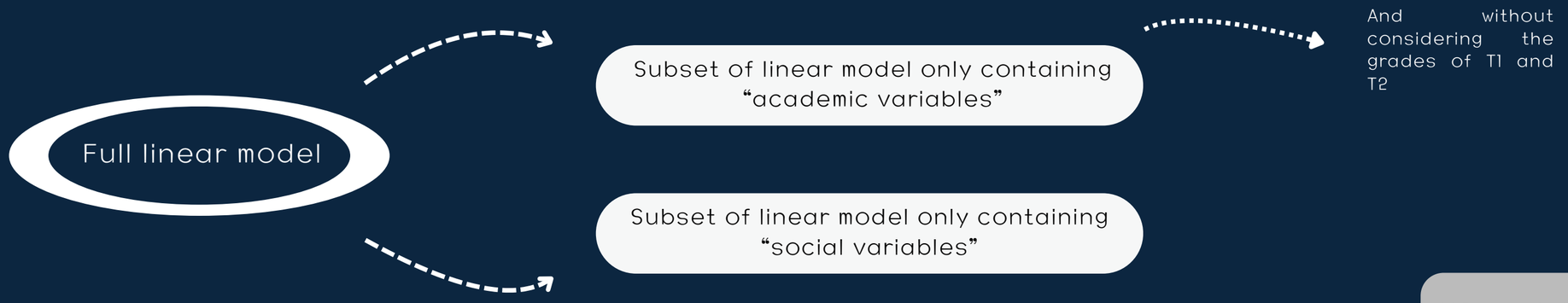
By looking at the boosted model we have a proof that the academic features have more impact on the model compared to the social variables.

	var	rel.inf
1	Grades_t2	83.18836051
2	Absences	7.36495173
3	Grades_t1	2.45388263
4	Subject	1.81593051

Supervised Learning – Alcohol consumption data

REGRESSION

I wanted a proof of the fact that academic variables influence the final grade more than social variables, so I created two new linear models: one consisting of a subset of the full model, but only containing academic variables, and one containing only social variables. I then compared the goodness of each model, by using the Mallows CP and the AIC.



A small Mallows' Cp value indicates that the model is relatively precise (has small variance) in estimating the true regression coefficients and predicting future responses.

Mallows CP

21.2	117.3
Academic variables LM	Social variables LM

Akaike Information Criterion (AIC)

4173	4264
Academic variables LM	Social variables LM

AIC determines the relative information value of the model using the maximum likelihood estimate and the number of parameters in the model.

Supervised Learning – Alcohol consumption data

REGRESSION

It is interesting to see how a Random Forest regression model performs in comparison to a linear model.
I compared it with both the full linear model and with a linear model that doesn't contain the T1 and T2 grades.

The accuracy has slightly improved in one model:

0.2718 → 0.3126 (+ 4%)

LM without T1 and T2 grades

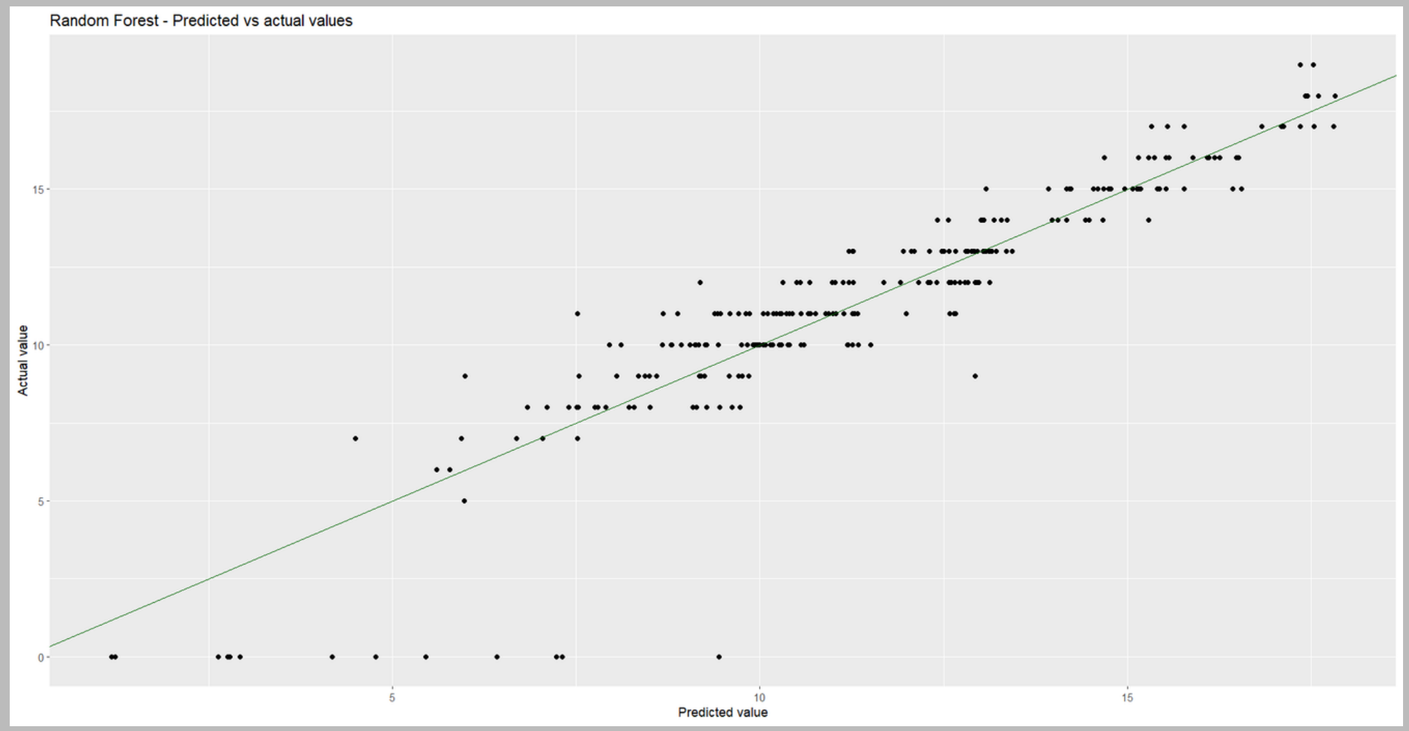
Random Forest without T1 and T2 grades

And surprisingly decreased in the other model

0.8522 → 0.8469 (–1%)

LM with T1 and T2 grades

Random Forest with T1 and T2 grades



Supervised Learning – Alcohol consumption data

CLASSIFICATION

I used 4 different classification methods: Random Forest, Linear Discriminant Analysis, K – Nearest Neighbors and Recursive Partitioning. I then compared the accuracy of each of these methods.

These are the results, in terms of accuracy:

RANDOM FOREST	LINEAR DISCRIMINANT ANALYSIS
76.8%	68.5%
K – NEAREST NEIGHBORS	**RECURSIVE PARTITIONING**
69.1%	72.4%

QUESTION

CAN WE DETERMINE, FOR EACH PERSON AND BASED ON HIS ACADEMIC AND SOCIAL CHARACTERISTICS, THE TYPE OF ALCOHOL CONSUMER HE IS?

Three levels of alcoholism:

HIGH

MODERATE

LOW

Supervised Learning – Alcohol consumption data

CLASSIFICATION

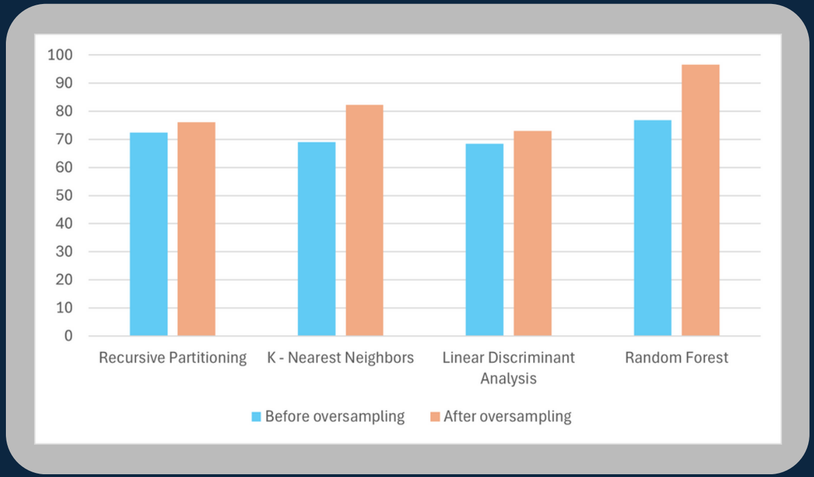
The total number of students interviewed is quite low

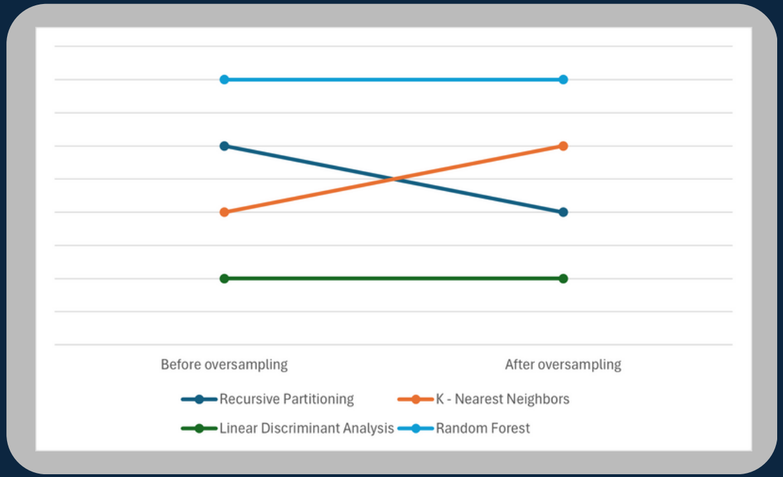
Most of the students are low alcohol consumers

Small sample size

Class imbalance

OVERSAMPLING





Some models have been more affected by oversampling than others, but overall it brought to an increase in accuracy. The most affected model was Random Forest, with an increase in accuracyof almost 0.2.

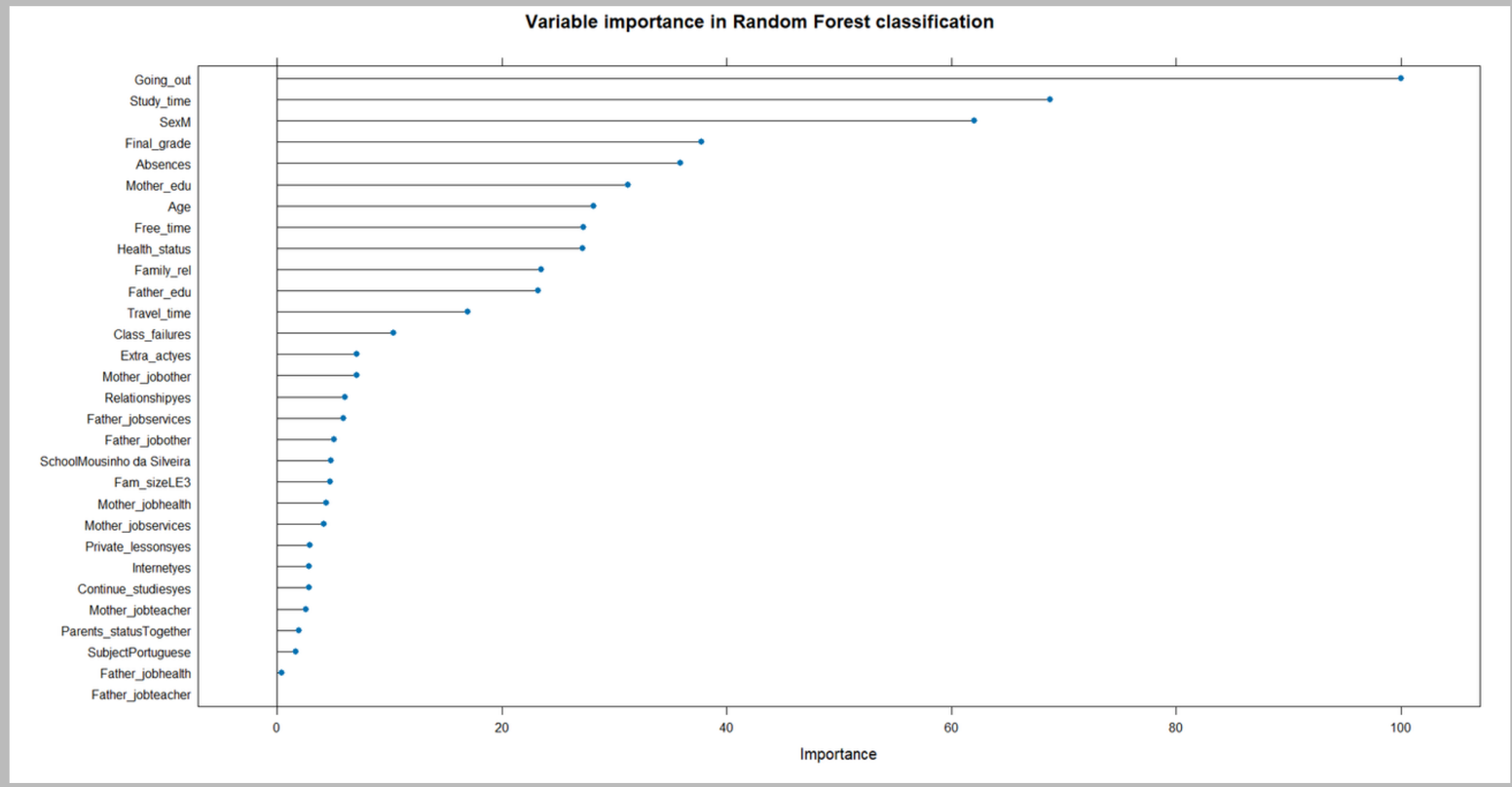
Comparing the ranking of classification methods before and after oversampling, we can see that KNN became more efficient than LDA.

Supervised Learning – Alcohol consumption data

CLASSIFICATION

We can notice the difference between the importance of variables in the classification model in comparison to the regression model.

	Overall
Going_out	100.000
Study_time	68.750
SexM	62.003
Final_grade	37.775
Absences	35.880
Mother_edu	31.231
Age	28.185
Free_time	27.281
Health_status	27.205
Family_rel	23.486

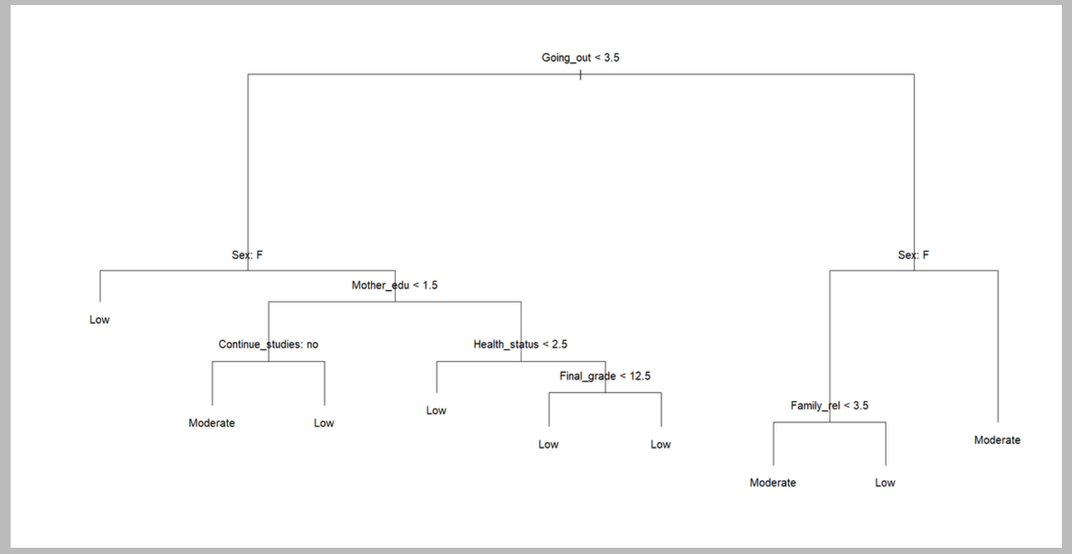


Supervised Learning – Alcohol consumption data

DECISION TREES

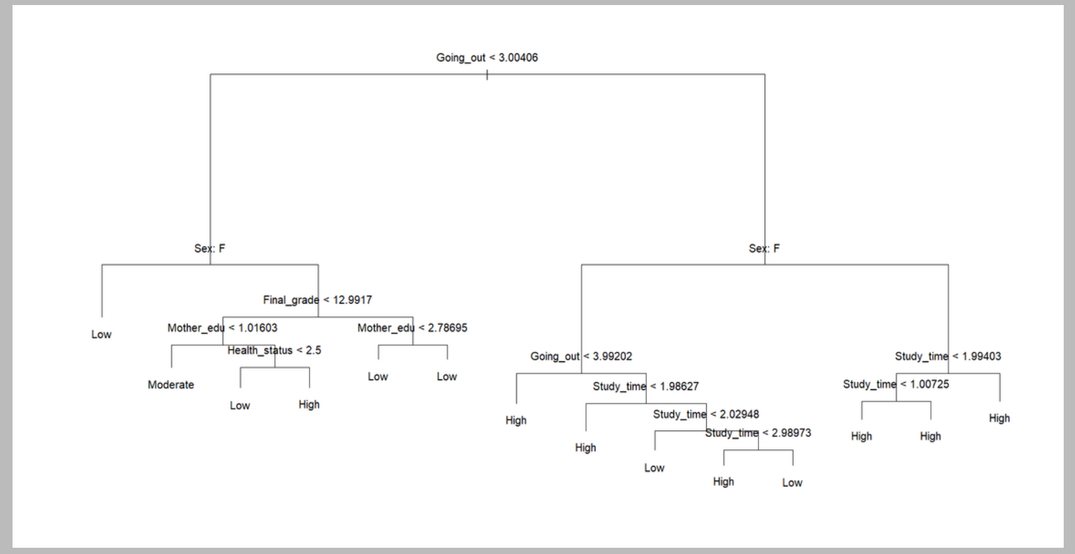
Decision tree with the original data

Less high consumption people → Difficult to predict by the decision tree



Decision tree with the oversampled data

Low and high consumption have been better balanced → Moderate class is difficult to predict

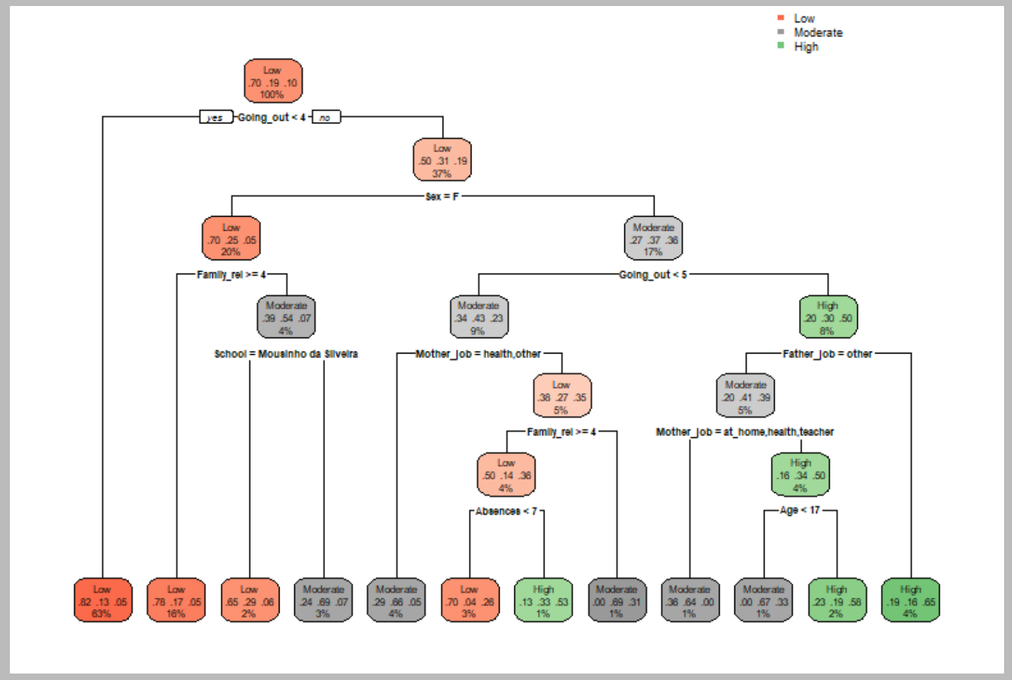


Supervised Learning – Alcohol consumption data

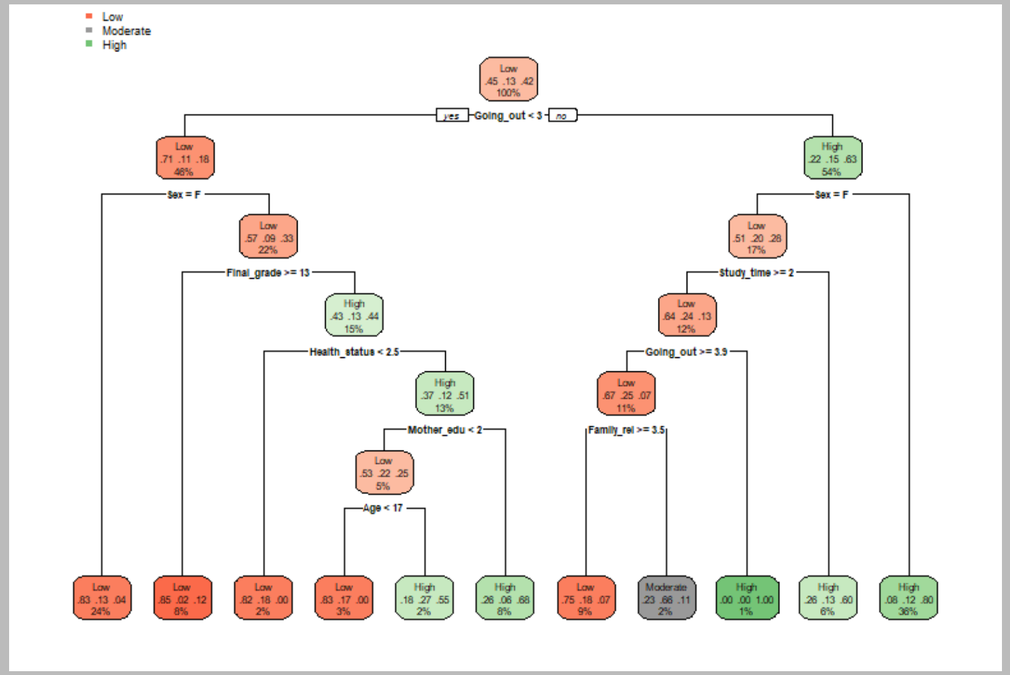
DECISION TREES

When plotting Recursive Partititoning trees the result is slightly improved, in particular with the original data:

Decision tree with the original data



Decision tree with the oversampled data



Supervised Learning – Alcohol consumption data

INTERPRETING THE RESULTS OF THE SUPERVISED LEARNING

Overall, the regression model was good at predicting the final grade of the students. The usage of XGboost and cross-validation helped improving the accuracy even more. Particularly interesting was analyzing the two types of variables contributing to the prediction: the academic variables and the social variables; by creating two independent models with these variables, I was able to prove the results obtained in the full model with the p-value and the Anova test: the school variables have more influence on the social variables on the final grade of a student.
It was interesting also using the Random Forest regression, in order to compare the results with the Linear Model.

All the four classification models were accurate in predicting the three different classes, with the Random Forest being more accurate than the others. The usage of oversampling was very useful in balancing the classes and increasing the size of the data; in addition, it increased the accuracy of all the models, in particular the one of the Random Forest.
Decision trees have been helpful in representing visually the characteristics of the classes, and the steps needed to classify a student into one of the three classes.

THANK YOU