Data mining - second lab assignment report

Exercise 1. Feature selection.

In this exercise we had to make our own implementation of FIsher's score and Gini index. The Fisher's score was tested on the *iris* dataset, where we can see that the highest value of the score was for features 3 and 4, which have the highest discriminatory power.

```
> FisherScore(iris.dat[1:2], iris.lab)
[1] 0.0233007
> FisherScore(iris.dat[2:3], iris.lab)
[1] 0.4637003
> FisherScore(iris.dat[c(2,4)], iris.lab)
[1] 0.06058037
> FisherScore(iris.dat[3:4], iris.lab)
[1] 0.7152423
```

While the gini index was tested on the *golf* dataset which i used for the next task also.

```
> gini.index(X, y, feature_name = "Outlook")
[1] 0.3428571
> gini.index(X, y, feature_name = "Temp.")
[1] 0.4404762
> gini.index(X, y, feature_name = "Wind")
[1] 0.4285714
```

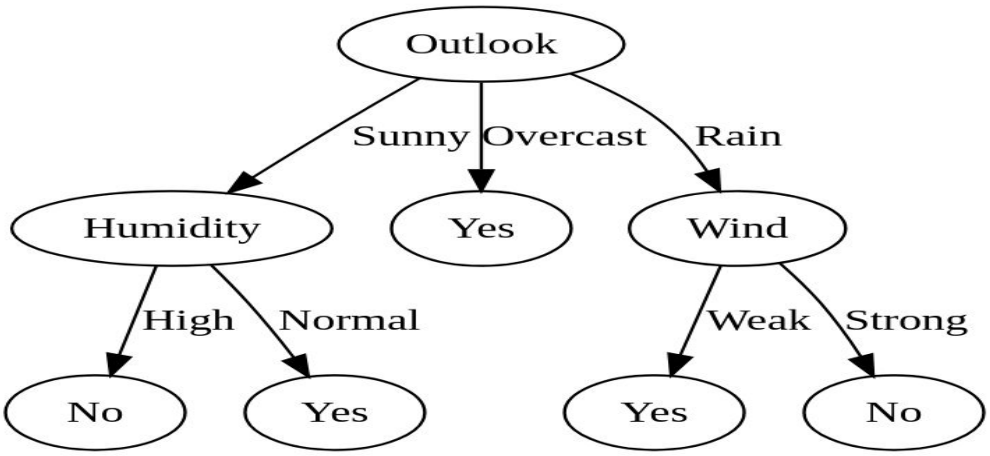
The "Outlook" feature has the smallest gini index which implies greater discrimination.

Exercise 2. Classification.

In this task we had to implement the decision tree classifier and I also used my own implementation of the gini index as the decision rule maker. The dataset used for this task is the *golf* dataset which the first few rows can be seen below.

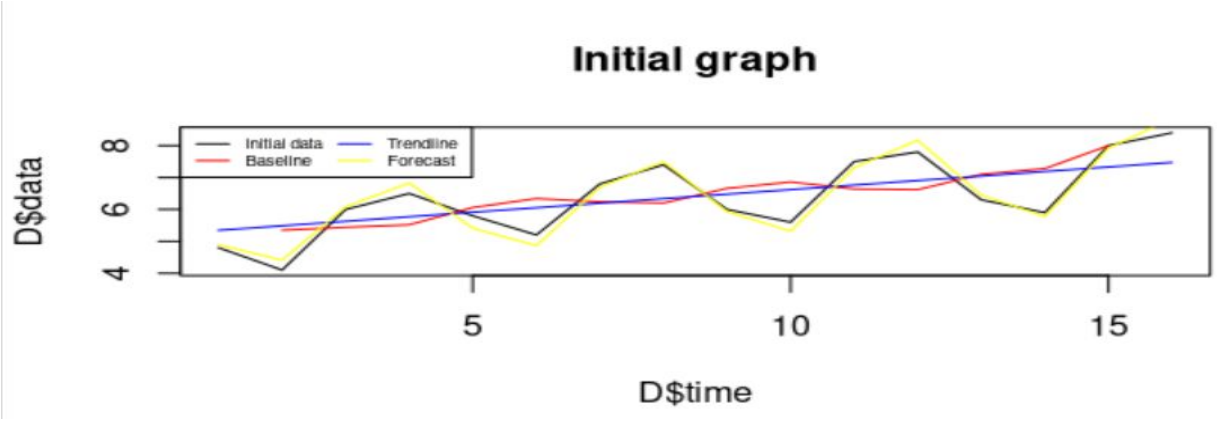
Outlook	Temp.	Humidity	Wind	Decision
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes

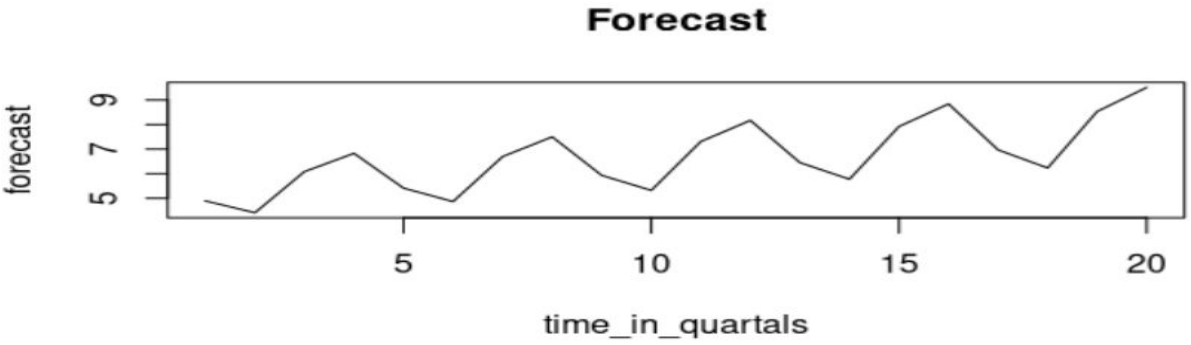
The final output is a decision tree seen below.



Exercise 3. Time series.

The third task was to implement the example discussed during the practice. The idea was to make a model for predicting a time series and doing a forecast.





Sources used:

https://sefiks.com/2018/08/27/a-step-by-step-cart-decision-tree-example/?fbclid=IwAR1r0wBv5qN1jOlNJRvhyp80Fx5NoZWJ0VoZIMyNBCa1exLno0QFm1IJmYg
https://cran.r-project.org/web/packages/data.tree/vignettes/data.tree.html#climbing-a-tree-tree-navigation
Data mining lectures at https://moodle.taltech.ee/course/view.php?id=31036
https://www.researchgate.net/publication/283867183_Data_Mining_Algorithms_Explained_Using_R