Red Wine Quality

R project

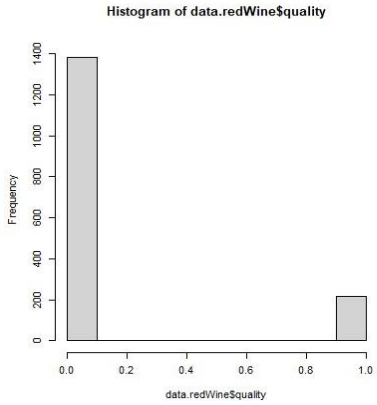
By Tomer Teprovich

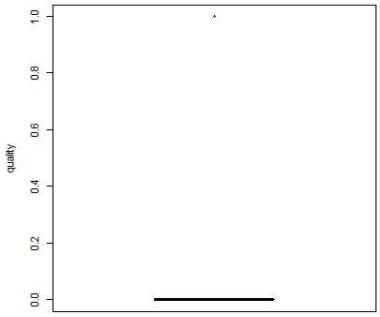
The Data

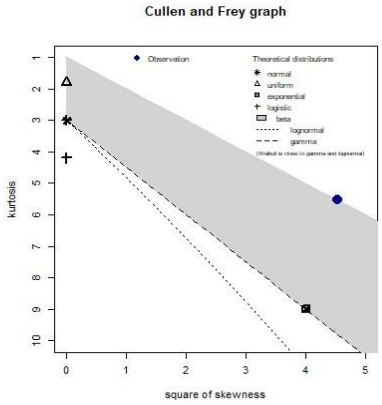
- https://archive.ics.uci.edu/ml/datasets/wine+quality
- 1. Alcohol: the amount of alcohol in wine
- 2. Volatile acidity: are high acetic acid in wine which leads to an unpleasant vinegar taste
- 3. Sulphates: a wine additive that contributes to SO2 levels and acts as an antimicrobial and antioxidant
- 4. Citric Acid: acts as a preservative to increase acidity (small quantities add freshness and flavor to wines)
- 5. Total Sulfur Dioxide: is the amount of free + bound forms of SO2
- 6. Density: sweeter wines have a higher density
- 7. Chlorides: the amount of salt in the wine
- 8. Fixed acidity: are non-volatile acids that do not evaporate readily
- 9. pH: the level of acidity
- 10. Free Sulfur Dioxide: it prevents microbial growth and the oxidation of wine
- 11. Residual sugar: is the amount of sugar remaining after fermentation stops. The key is to have a perfect balance between — sweetness and sourness (wines > 45g/ltrs are sweet)

DATA PLOTS

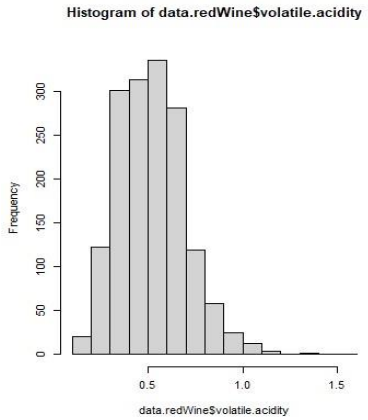
Quality- Y index

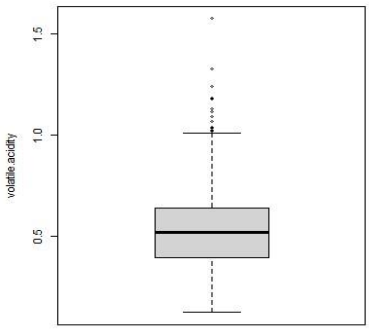


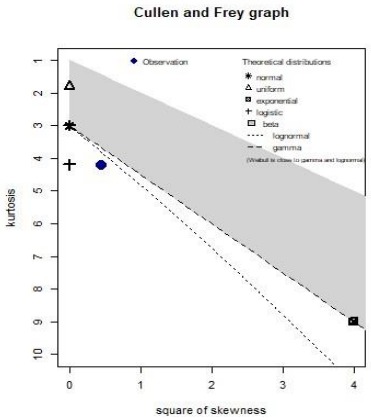




Volatile acidity

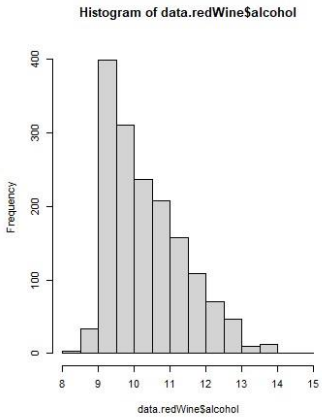


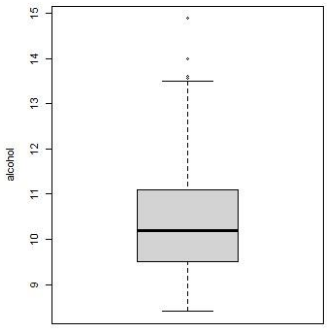


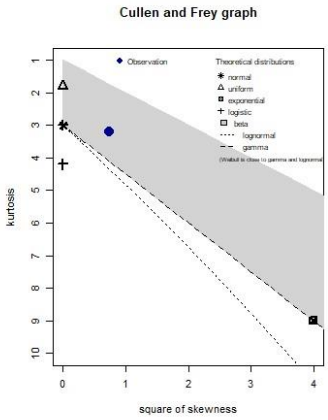


DATA PLOTS

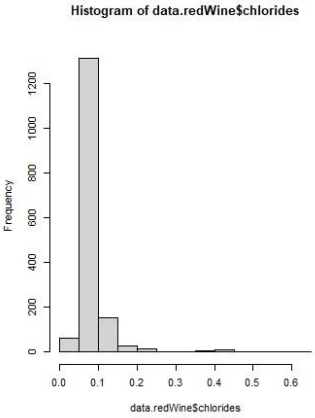
Alcohol

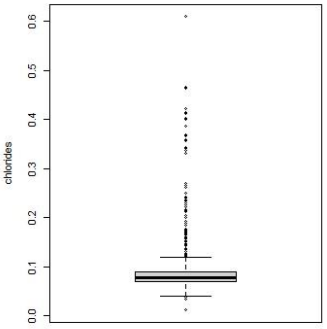


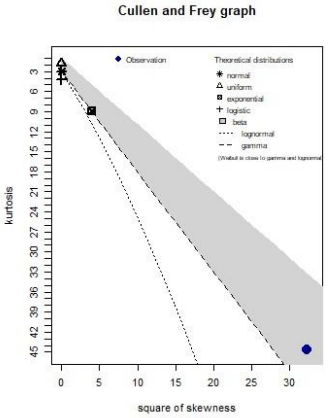




Chlorides

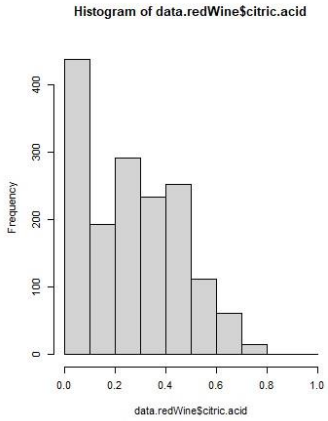


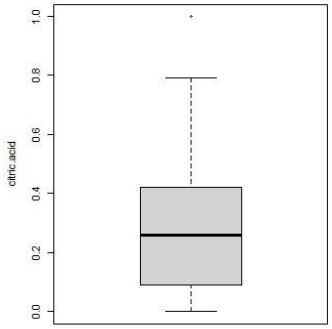


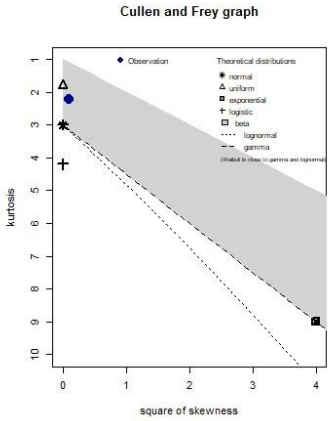


DATA PLOTS

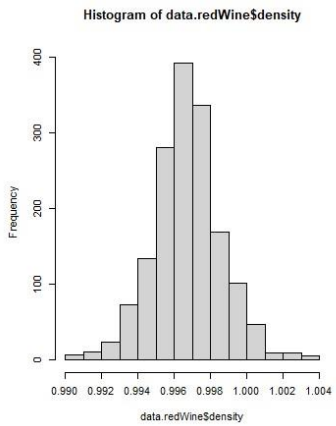
Citric Acid

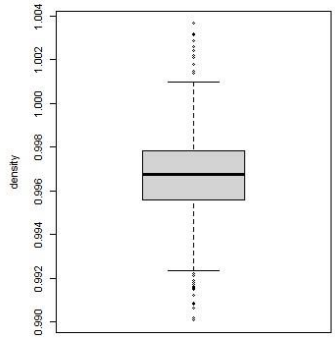


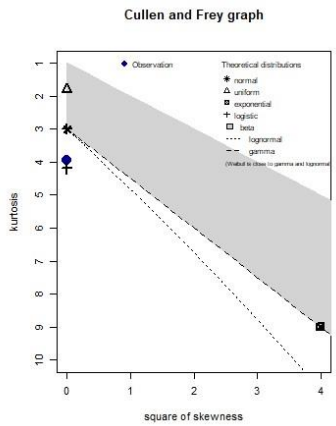




Density

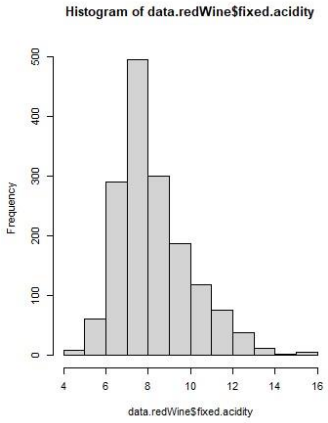


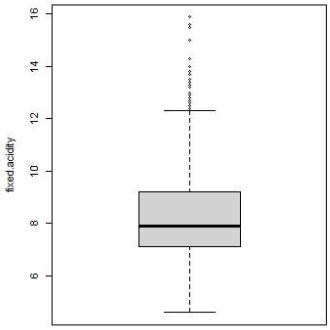


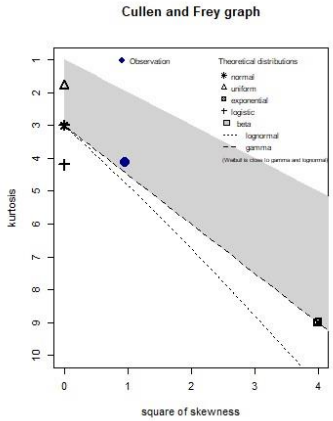


DATA PLOTS

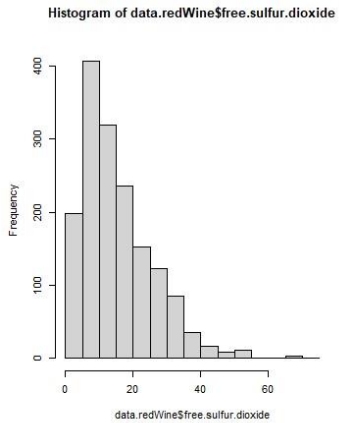
Fixed acidity

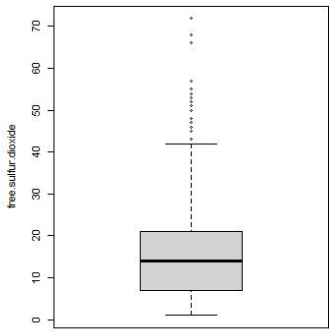


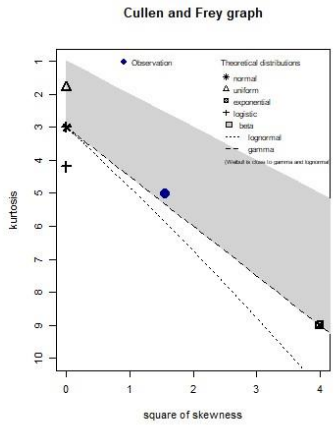




Free Sulfur Dioxide

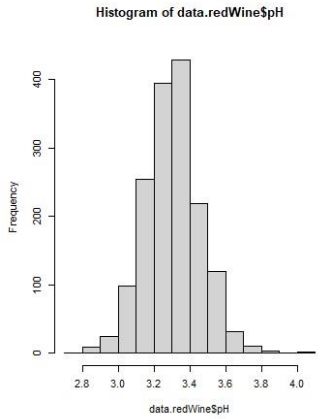


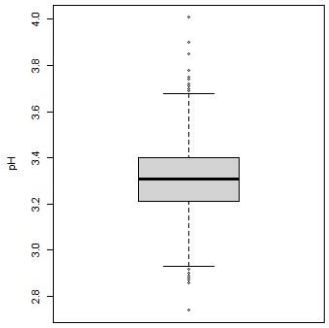


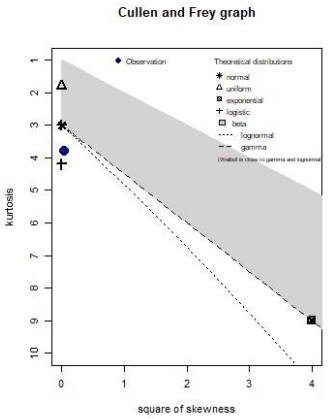


DATA PLOTS

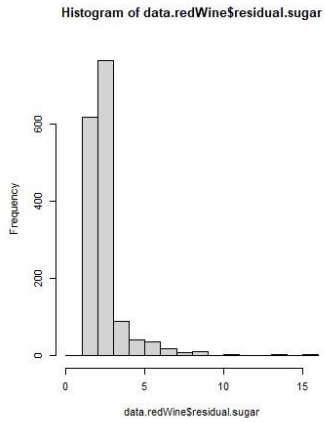
pH

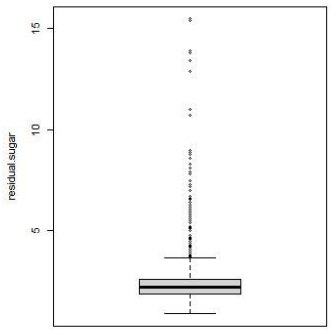


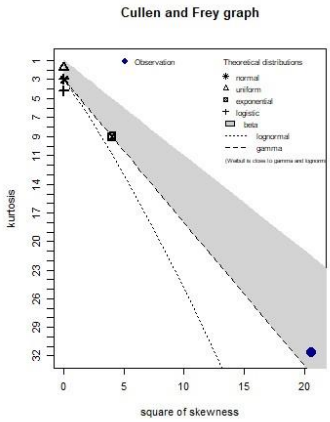




Residual sugar

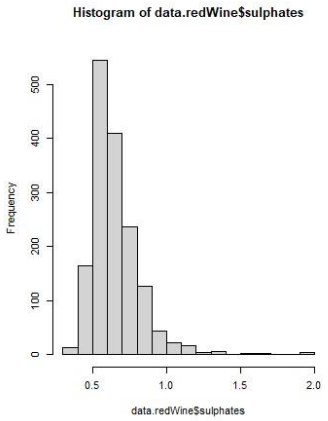


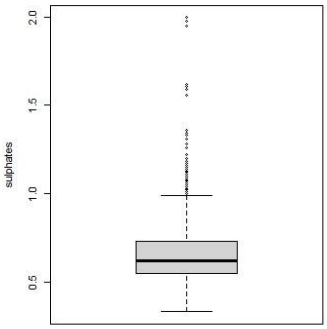


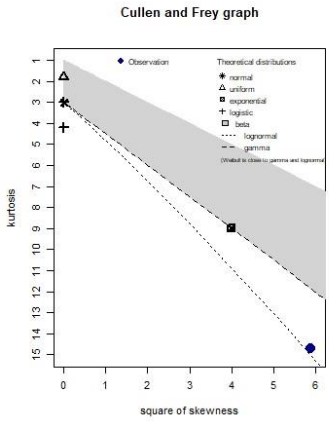


DATA PLOTS

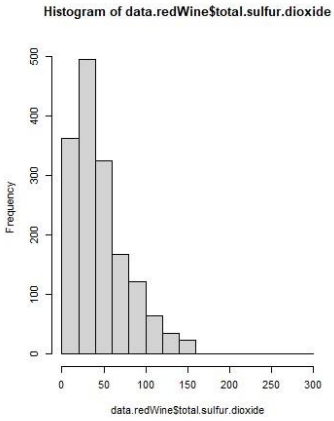
Sulphates

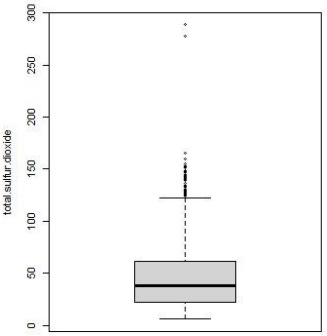


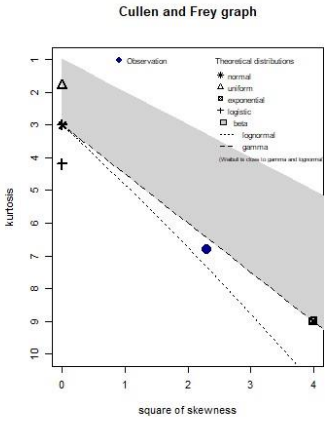




Total Sulfur Dioxide







DATA PLOTS

summary

```
> summary(data.redwine)
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar
 Min.    : 4.60   Min.    :0.1200   Min.    :0.000   Min.    : 0.900
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
   chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
 Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
 1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
 Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
 Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
 3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
 Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
       pH          sulphates         alcohol         quality
 Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :0.0000
 1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:0.0000
 Median :3.310   Median :0.6200   Median :10.20   Median :0.0000
 Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :0.1357
 3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:0.0000
 Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :1.0000
```

correlation matrix

KNN

1. חילקתי את הדאטה לטריין וטסט 70% ו 30% בהתאמה.
2. בניתי שלושה מודלים של KNN עם כל המשתנים המסבירים וכמות שכנים שונה 1 , 3 , 5
3. לאחר מכן בניתי Confusion matrix לכל אחד מהמודלים

```
> table(knn.1 ,test.y)
     test.y
knn.1   0   1
    0 384  29
    1  26  41
> table(knn.3 ,test.y)
     test.y
knn.3   0   1
    0 394  46
    1  16  24
> table(knn.5 ,test.y)
     test.y
knn.5   0   1
    0 396  50
    1  14  20
```

4. לבסוף בדקתי proportion of correct classification לכל אחד מהמודלים

```
> # proportion of correct classification for k = 1, 3, 5
> 100 * sum(test.y == knn.1)/length(test.y)  # For knn = 1
[1] 88.54167
> 100 * sum(test.y == knn.3)/length(test.y)  # For knn = 3
[1] 87.08333
> 100 * sum(test.y == knn.5)/length(test.y)  # For knn = 5
[1] 86.66667
>
```

linear regression

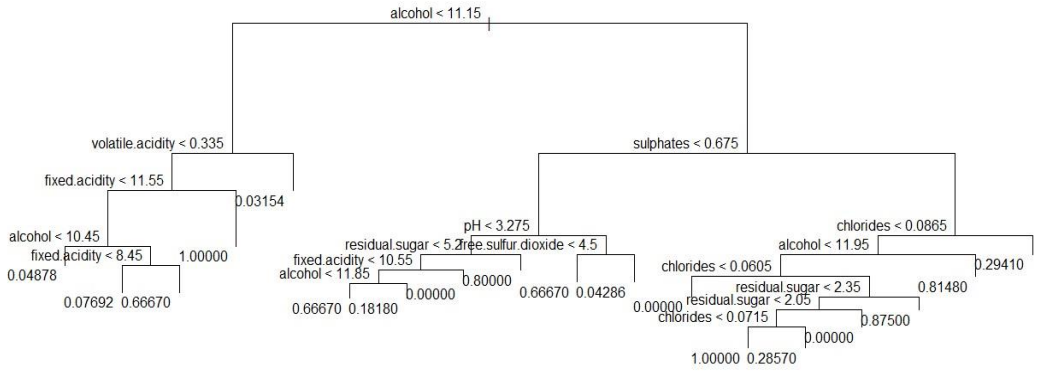
1. חילקתי את הדאטה לטריין וטסט 70% ו 30% בהתאמה.
2. לאחר מכן הוצאתי Summary

```
Call:
lm(formula = quality ~ ., data = train.set)

Residuals:
     Min       1Q   Median       3Q      Max
-0.59331 -0.17566 -0.04142  0.03934  1.00689

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.513e+01  1.161e+01   3.025 0.002540 **
fixed.acidity         4.372e-02  1.454e-02   3.007 0.002696 **
volatile.acidity     -2.310e-01  6.779e-02  -3.407 0.000680 ***
citric.acid          -1.795e-02  8.341e-02  -0.215 0.829650
residual.sugar        3.005e-02  8.167e-03   3.680 0.000244 ***
chlorides            -4.199e-01  2.423e-01  -1.733 0.083411 .
free.sulfur.dioxide  -7.322e-04  1.214e-03  -0.603 0.546430
total.sulfur.dioxide -3.344e-04  4.060e-04  -0.824 0.410400
density              -3.630e+01  1.185e+01  -3.063 0.002244 **
pH                   -3.914e-03  1.059e-01  -0.037 0.970528
sulphates             2.810e-01  6.154e-02   4.567 5.51e-06 ***
alcohol               7.357e-02  1.458e-02   5.048 5.23e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3009 on 1107 degrees of freedom
Multiple R-squared:  0.2152,    Adjusted R-squared:  0.2074
F-statistic:  27.6 on 11 and 1107 DF,  p-value: < 2.2e-16
```

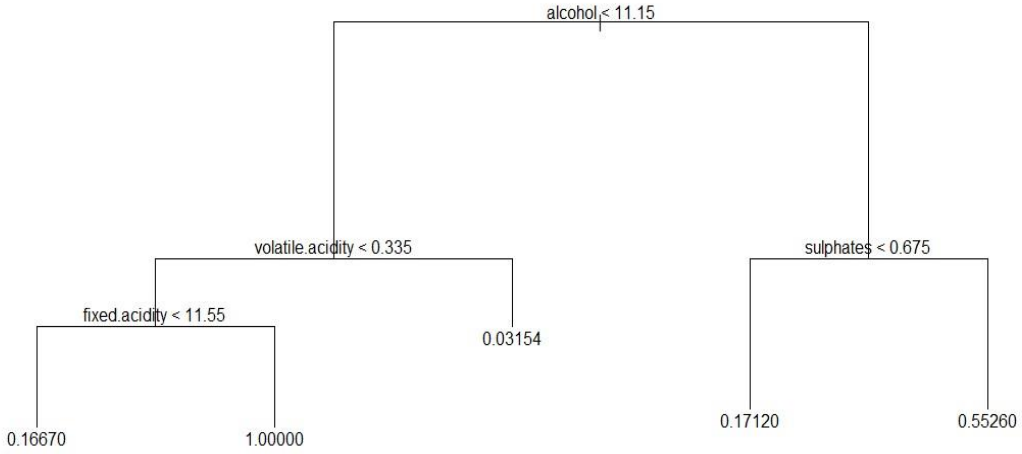
3. לא הורדתי משתנים בעלי P-value גבוהה או שה-Signif נמוך (הייתי צריך להוריד).
4. קיבלתי MSE 0.08932080

Tree

1. חילקתי את הדאטה לטריין וטסט 50% כל אחד.
2. לאחר מכן בניתי מודל עץ עם כל המשתנים המסברים



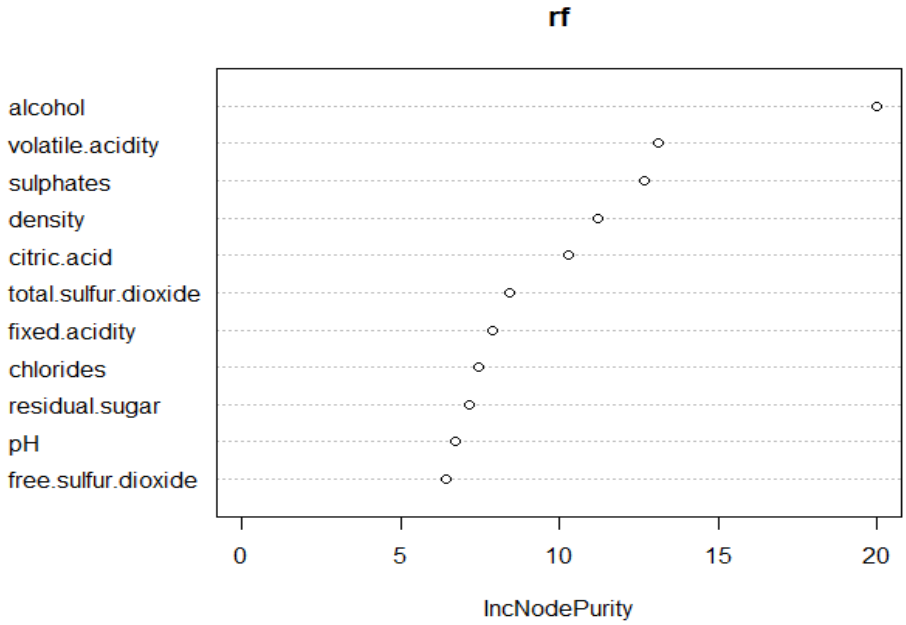
3. ולבסוף גזמתי את העץ ל5



4. קיבלתי MSE = 0.1133211

random forest

1. חילקתי את הדאטה לטריין וטסט 70% ו 30% בהתאמה.
2. לאחר מכן בניתי מודל Random Foreest עם כל המשתנים המסבירים ו1500 עצים.
3. ואז הוצאתי את כמות העצים האופטימלית
4. בדקתי איך המשתנים משפעים



5. עשיתי Cross Validation ולקחתי את מספר המשתנים לכל החלטה (mtry)
6. ולבסוף בניתי מודל Random Foreest חדש עם כמות העצים האופטימלי ומינימום mtry
7. קיבלתי MSE = 0.06032756 שהוא המודל הכי מוצלח!!!