Practice-6

Harsh

29/06/2020

```r
#Importing all libraries required

#install.packages("rpart")
#install.packages("rpart.plot")

library(psych)
```

```
## Warning: package 'psych' was built under R version 3.6.3
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.6.3

## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.6.3

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```r
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 3.6.3
```

```r
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 3.6.3
```

```r
library(RWeka)
```

```
## Warning: package 'RWeka' was built under R version 3.6.3
```

Problem 1: Download the data set on student achievement in secondary education math education of two Portuguese schools (use the data set Students Math).

1. Create scatter plots and pairwise correlations between age, absences, G1, and G2 and final grade (G3) using the pairs.panels() function in R.
2. Build a multiple regression model predicting final math grade (G3) using as many features as you like but you must use at least four. Include at least one categorical variables and be sure to properly convert it to dummy codes. Select the features that you believe are useful – you do not have to include all features.
3. Using the model from (2), use stepwise backward elimination to remove all non-significant variables and then state the final model as an equation. State the backward elimination measure you applied (p-value, AIC, Adjusted R2). This tutorial shows how to use various feature elimination techniques.
4. Calculate the 95% confidence interval for a prediction – you may choose any data you wish for some new student.
5. What is the RMSE for this model – use the entire data set for both training and validation. You may find the residuals() function useful. Alternatively, you can inspect the model object, e.g., if your model is in the variable m, then the residuals (errors) are in m$residualsandyourpredictedvalues(fittedvalues)areinm$fitted.values.

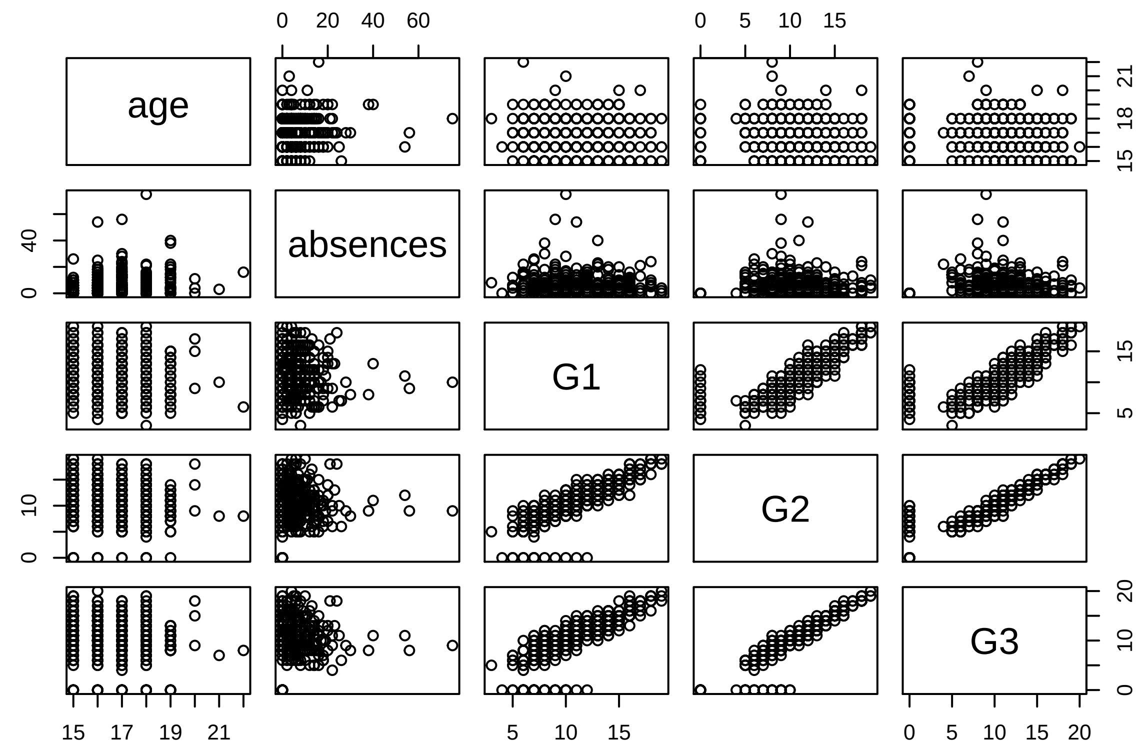
```
#Importing student Data
student_math <- read.csv("C:\\Users\\harsh\\Desktop\\Introduction to Machine learning and Data Mining\\

#Exploring Data
head(student_math)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        3       yes     no  yes         no
## 4   mother          1         3        0        no    yes  yes        yes
## 5   father          1         2        0        no    yes  yes         no
## 6   mother          1         2        0        no    yes  yes        yes
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
## 5     yes    yes       no       no      4        3     2    1    2      5
## 6     yes    yes      yes       no      5        4     2    1    2      5
##   absences G1 G2 G3
## 1        6  5  6  6
## 2        4  5  5  6
## 3       10  7  8 10
## 4        2 15 14 15
## 5        4  6 10 10
## 6       10 15 15 15
```

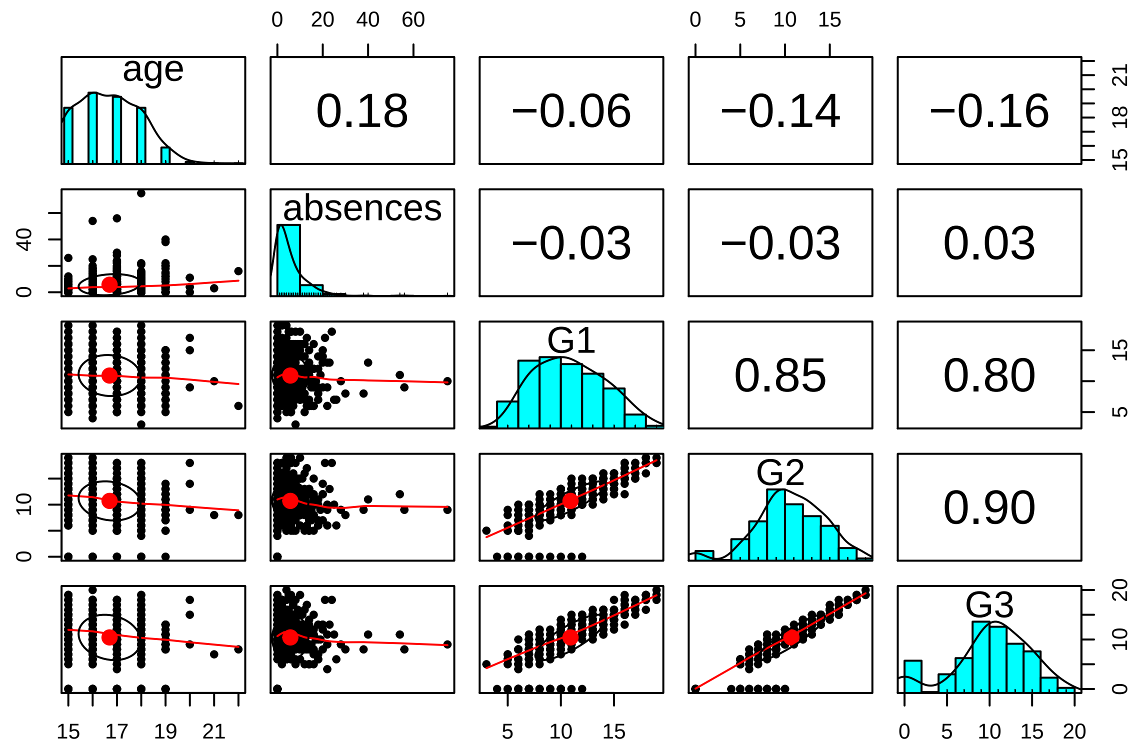
```r
#Checking the correlation between different features of the data
cor(student_math[c("age", "absences", "G1", "G2", "G3")])
```

```
##                 age    absences         G1         G2          G3
## age       1.0000000  0.17523008 -0.0640815 -0.1434740 -0.16157944
## absences  0.1752301  1.00000000 -0.0310029 -0.0317767  0.03424732
## G1       -0.0640815 -0.03100290  1.0000000  0.8521181  0.80146793
## G2       -0.1434740 -0.03177670  0.8521181  1.0000000  0.90486799
## G3       -0.1615794  0.03424732  0.8014679  0.9048680  1.00000000
```

```r
#Scatter plot between different features
pairs(student_math[c("age", "absences", "G1", "G2", "G3")])
```




```r
#Pair.panel function is used to plot histogram and it provides correlation between different features
pairs.panels(student_math[c("age", "absences", "G1", "G2", "G3")])
```




```r
#Exploratory analysis of student_math data
summary(student_math)
```

```
##  school   sex          age        address famsize   Pstatus      Medu      
##  GP:349   F:208   Min.   :15.0   R: 88   GT3:281   A: 41    Min.   :0.000  
##  MS: 46   M:187   1st Qu.:16.0   U:307   LE3:114   T:354    1st Qu.:2.000  
##                   Median :17.0                              Median :3.000  
##                   Mean   :16.7                              Mean   :2.749  
##                   3rd Qu.:18.0                              3rd Qu.:4.000  
##                   Max.   :22.0                              Max.   :4.000  
##       Fedu             Mjob           Fjob           reason      guardian  
##  Min.   :0.000   at_home : 59   at_home : 20   course    :145   father: 90  
##  1st Qu.:2.000   health  : 34   health  : 18   home      :109   mother:273  
##  Median :2.000   other   :141   other   :217   other     : 36   other : 32  
##  Mean   :2.522   services:103   services:111   reputation:105               
##  3rd Qu.:3.000   teacher : 58   teacher : 29                                
##  Max.   :4.000                                                              
##    traveltime      studytime        failures      schoolsup famsup     paid    
##  Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :344   no :153   no :214  
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 51   yes:242   yes:181  
##  Median :1.000   Median :2.000   Median :0.0000                                
##  Mean   :1.448   Mean   :2.035   Mean   :0.3342                                
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000                                
##  Max.   :4.000   Max.   :4.000   Max.   :3.0000                                
##  activities nursery    higher    internet  romantic      famrel     
##  no :194     no : 81   no : 20   no : 66   no :263   Min.   :1.000  
```

```
##  yes:201    yes:314   yes:375   yes:329   yes:132   1st Qu.:4.000
##                                                     Median :4.000
##                                                     Mean   :3.944
##                                                     3rd Qu.:5.000
##                                                     Max.   :5.000
##     freetime         goout            Dalc            Walc
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :3.000   Median :3.000   Median :1.000   Median :2.000
##  Mean   :3.235   Mean   :3.109   Mean   :1.481   Mean   :2.291
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##      health         absences            G1              G2
##  Min.   :1.000   Min.   : 0.000   Min.   : 3.00   Min.   : 0.00
##  1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00
##  Median :4.000   Median : 4.000   Median :11.00   Median :11.00
##  Mean   :3.554   Mean   : 5.709   Mean   :10.91   Mean   :10.71
##  3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00
##  Max.   :5.000   Max.   :75.000   Max.   :19.00   Max.   :19.00
##        G3
##  Min.   : 0.00
##  1st Qu.: 8.00
##  Median :11.00
##  Mean   :10.42
##  3rd Qu.:14.00
##  Max.   :20.00
```

```r
#Selecting relevant features
math_features <- student_math[,c(2,14,16,17,18,23,24,25,29,30,31,32,33)]

#Converting categorical variables to dummy codes using factor
math_features$sex <- as.factor(math_features$sex)
math_features$schoolsup <- as.factor(math_features$schoolsup)
math_features$famsup <- as.factor(math_features$famsup)
math_features$paid <- as.factor(math_features$paid)
math_features$romantic <- as.factor(math_features$romantic)

#Multiple regression model using lm() function. Observed R-squared values is 0.8356
#and we see that the p-values is quite low
math_g3_pred <- lm(G3~sex+studytime+schoolsup+famsup+paid+romantic+famrel+freetime+health+absences+G1+G2
summary(math_g3_pred)
```

```
##
## Call:
## lm(formula = G3 ~ sex + studytime + schoolsup + famsup + paid +
##     romantic + famrel + freetime + health + absences + G1 + G2,
##     data = math_features)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3194 -0.4608  0.2894  0.9535  3.8932
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -3.71833    0.69456  -5.353 1.49e-07 ***
## sexM          0.10543    0.21213   0.497  0.61948
## studytime    -0.15404    0.12436  -1.239  0.21624
## schoolsupyes  0.51095    0.29741   1.718  0.08661 .
## famsupyes     0.14637    0.20887   0.701  0.48388
## paidyes       0.15748    0.20482   0.769  0.44244
## romanticyes  -0.36712    0.20874  -1.759  0.07942 .
## famrel        0.31208    0.10856   2.875  0.00427 **
## freetime      0.04278    0.09988   0.428  0.66864
## health        0.07095    0.07015   1.011  0.31242
## absences      0.04072    0.01212   3.359  0.00086 ***
## G1            0.18574    0.05720   3.247  0.00127 **
## G2            0.97181    0.05002  19.430  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.887 on 382 degrees of freedom
## Multiple R-squared:  0.8356, Adjusted R-squared:  0.8304
## F-statistic: 161.8 on 12 and 382 DF,  p-value: < 2.2e-16
```

```r
#Using backward step elimination method to remove non-significant features.
#We see that out of 12 features 6 have been removed because of low AIC
step(math_g3_pred, direction = "backward")
```

```
## Start:  AIC=514.3
## G3 ~ sex + studytime + schoolsup + famsup + paid + romantic +
##     famrel + freetime + health + absences + G1 + G2
##
##             Df Sum of Sq    RSS    AIC
## - freetime   1      0.65 1360.5 512.49
## - sex        1      0.88 1360.7 512.56
## - famsup     1      1.75 1361.5 512.81
## - paid       1      2.10 1361.9 512.91
## - health     1      3.64 1363.4 513.36
## - studytime  1      5.46 1365.3 513.89
## <none>                   1359.8 514.30
## - schoolsup  1     10.51 1370.3 515.34
## - romantic   1     11.01 1370.8 515.49
## - famrel     1     29.42 1389.2 520.76
## - G1         1     37.54 1397.3 523.06
## - absences   1     40.17 1400.0 523.80
## - G2         1   1343.89 2703.7 783.78
##
## Step:  AIC=512.49
## G3 ~ sex + studytime + schoolsup + famsup + paid + romantic +
##     famrel + health + absences + G1 + G2
##
##             Df Sum of Sq    RSS    AIC
## - sex        1      1.24 1361.7 510.85
## - famsup     1      1.91 1362.4 511.05
## - paid       1      2.02 1362.5 511.08
## - health     1      3.70 1364.2 511.56
## - studytime  1      5.86 1366.3 512.19
## <none>                   1360.5 512.49
```

```
## - schoolsup  1     10.45 1370.9 513.51
## - romantic   1     10.89 1371.3 513.64
## - famrel     1     31.32 1391.8 519.48
## - G1         1     37.92 1398.4 521.35
## - absences   1     39.76 1400.2 521.87
## - G2         1   1343.41 2703.9 781.80
##
## Step:  AIC=510.85
## G3 ~ studytime + schoolsup + famsup + paid + romantic + famrel +
##     health + absences + G1 + G2
##
##             Df Sum of Sq    RSS    AIC
## - famsup     1      1.70 1363.4 509.34
## - paid       1      1.83 1363.5 509.38
## - health     1      4.35 1366.0 510.11
## <none>                   1361.7 510.85
## - studytime  1      8.26 1370.0 511.24
## - schoolsup  1      9.81 1371.5 511.68
## - romantic   1     11.53 1373.2 512.18
## - famrel     1     32.11 1393.8 518.06
## - G1         1     38.25 1399.9 519.79
## - absences   1     39.03 1400.7 520.01
## - G2         1   1350.77 2712.5 781.06
##
## Step:  AIC=509.34
## G3 ~ studytime + schoolsup + paid + romantic + famrel + health +
##     absences + G1 + G2
##
##             Df Sum of Sq    RSS    AIC
## - paid       1      3.25 1366.6 508.28
## - health     1      4.71 1368.1 508.71
## <none>                   1363.4 509.34
## - studytime  1      7.50 1370.9 509.51
## - schoolsup  1     10.68 1374.1 510.43
## - romantic   1     11.55 1374.9 510.67
## - famrel     1     31.68 1395.1 516.42
## - G1         1     37.84 1401.2 518.16
## - absences   1     39.49 1402.9 518.62
## - G2         1   1349.22 2712.6 779.08
##
## Step:  AIC=508.28
## G3 ~ studytime + schoolsup + romantic + famrel + health + absences +
##     G1 + G2
##
##             Df Sum of Sq    RSS    AIC
## - health     1      4.28 1370.9 507.52
## - studytime  1      6.14 1372.8 508.05
## <none>                   1366.6 508.28
## - schoolsup  1     10.17 1376.8 509.21
## - romantic   1     11.37 1378.0 509.56
## - famrel     1     31.91 1398.5 515.40
## - G1         1     35.75 1402.4 516.48
## - absences   1     39.88 1406.5 517.65
## - G2         1   1396.04 2762.7 784.30
```

```
##
## Step:  AIC=507.52
## G3 ~ studytime + schoolsup + romantic + famrel + absences + G1 +
##     G2
##
##             Df Sum of Sq    RSS    AIC
## - studytime  1      6.91 1377.8 507.50
## <none>                   1370.9 507.52
## - schoolsup  1      9.71 1380.6 508.31
## - romantic   1     11.00 1381.9 508.67
## - famrel     1     34.48 1405.4 515.33
## - G1         1     35.92 1406.8 515.73
## - absences   1     38.99 1409.9 516.60
## - G2         1   1391.78 2762.7 782.30
##
## Step:  AIC=507.5
## G3 ~ schoolsup + romantic + famrel + absences + G1 + G2
##
##             Df Sum of Sq    RSS    AIC
## <none>                   1377.8 507.50
## - schoolsup  1      8.47 1386.3 507.92
## - romantic   1     12.50 1390.3 509.07
## - famrel     1     33.35 1411.2 514.95
## - G1         1     33.41 1411.2 514.97
## - absences   1     41.53 1419.3 517.23
## - G2         1   1391.35 2769.2 781.23
```

```
##
## Call:
## lm(formula = G3 ~ schoolsup + romantic + famrel + absences +
##     G1 + G2, data = math_features)
##
## Coefficients:
##  (Intercept)  schoolsupyes   romanticyes        famrel      absences
##     -3.37474       0.45186      -0.38852       0.32649       0.04111
##           G1            G2
##      0.17304       0.97548
```

```
#Testing multiple regression model for new selected features.
#We observe that the R-squared values has reduced from 0.8356 to 0.8334 but the p-value remains the sam
new_math_g3_pred <- lm(G3~schoolsup+romantic+famrel+absences+G1+G2, data = math_features)
summary(new_math_g3_pred)
```

```
##
## Call:
## lm(formula = G3 ~ schoolsup + romantic + famrel + absences +
##     G1 + G2, data = math_features)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4165 -0.3955  0.2811  0.9153  3.5797
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -3.37474    0.55390  -6.093 2.68e-09 ***
## schoolsupyes  0.45186    0.29257   1.544 0.123293    
## romanticyes  -0.38852    0.20705  -1.876 0.061349 .  
## famrel        0.32649    0.10654   3.065 0.002332 ** 
## absences      0.04111    0.01202   3.420 0.000693 ***
## G1            0.17304    0.05642   3.067 0.002312 ** 
## G2            0.97548    0.04928  19.794  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.884 on 388 degrees of freedom
## Multiple R-squared:  0.8334, Adjusted R-squared:  0.8308 
## F-statistic: 323.5 on 6 and 388 DF,  p-value: < 2.2e-16
```

```r
#As we observed in the above model that the residual standard error is 1.884,
#we assign this to a new variable
SE <- 1.884

#We select a test data for prediction
sample_data_CI <- student_math[395,]
sample_data_CI
```

```
##     school sex age address famsize Pstatus Medu Fedu  Mjob    Fjob reason
## 395     MS   M  19       U     LE3       T    1    1 other at_home course
##     guardian traveltime studytime failures schoolsup famsup paid activities
## 395   father          1         1        0        no     no   no         no
##     nursery higher internet romantic famrel freetime goout Dalc Walc health
## 395     yes    yes      yes       no      3        2     3    3    3      5
##     absences G1 G2 G3
## 395        5  8  9  9
```

```r
#Using test sample to predict the G3 grade for row 395
CI_pred <- predict(new_math_g3_pred, sample_data_CI)

#Calculating upper and lower boundary for 95% confidence interval
lower_CI <- unname(CI_pred - (1.96 * SE))
upper_CI <- unname(CI_pred + (1.96 * SE))

#As we can see that predicted value lies in between the upper and
#lower boundaries of 95% confidence interval
CI_pred
```

```
##      395 
## 7.973976
```

```r
lower_CI
```

```
## [1] 4.281336
```

```r
upper_CI
```

```
## [1] 11.66662
```

```r
#Calculating RMSE value for whole data.
#We use residual function to get the error values.
model <- lm(G3~., data = student_math)

RMSE <- sqrt(mean(model$residuals^2))
RMSE
```

```
## [1] 1.796979
```

Problem 2 : For this problem, the following short tutorial might be helpful in interpreting the logistic regression output.

1. Using the same data set as in Problem (1), add another column, PF – pass-fail. Mark any student whose final grade is less than 10 as F, otherwise as P and then build a dummy code variable for that new column. Use the new dummy variable column as the response variable.
2. Build a binomial logistic regression model classifying a student as passing or failing. Eliminate any non-significant variable using an elimination approach of your choice. Use as many features as you like but you must use at least four – choose the ones you believe are most useful.
3. State the regression equation.
4. What is the accuracy of your model? Use the entire data set for both training and validation.

```r
#Creating a duplicate of student_math dataset
student_math_PF <- student_math

#Creating a new column of pass and fail
student_math_PF$PF <- ifelse(student_math_PF$G3 < 10, "F", "P")

#Converting the categorical variable to dummy code using as.factor() function
student_math_PF$PF <- as.factor(student_math_PF$PF)

#Exploring new data
head(student_math_PF)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher     course
## 2     GP   F  17       U     GT3       T    1    1  at_home    other     course
## 3     GP   F  15       U     LE3       T    1    1  at_home    other      other
## 4     GP   F  15       U     GT3       T    4    2   health services       home
## 5     GP   F  16       U     GT3       T    3    3    other    other       home
## 6     GP   M  16       U     LE3       T    4    3 services    other reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
## 2   father          1         2        0        no    yes   no         no
## 3   mother          1         2        3       yes     no  yes         no
## 4   mother          1         3        0        no    yes  yes        yes
## 5   father          1         2        0        no    yes  yes         no
## 6   mother          1         2        0        no    yes  yes        yes
```

```
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
## 2      no    yes      yes       no      5        3     3    1    1      3
## 3     yes    yes      yes       no      4        3     2    2    3      3
## 4     yes    yes      yes      yes      3        2     2    1    1      5
## 5     yes    yes       no       no      4        3     2    1    2      5
## 6     yes    yes      yes       no      5        4     2    1    2      5
##   absences G1 G2 G3 PF
## 1        6  5  6  6  F
## 2        4  5  5  6  F
## 3       10  7  8 10  P
## 4        2 15 14 15  P
## 5        4  6 10 10  P
## 6       10 15 15 15  P
```

```r
#Counting total elements in PF column
table(student_math_PF$PF)
```

```
##
##   F   P
## 130 265
```

```r
#Binomial logistic regression of pass or fail using selected relevant features
glm_pred <- glm(PF~sex+studytime+schoolsup+famsup+paid+romantic+famrel+freetime+health+absences+G1+G2, d

#We observe that the AIC of the model is 153.92
summary(glm_pred)
```

```
##
## Call:
## glm(formula = PF ~ sex + studytime + schoolsup + famsup + paid +
##     romantic + famrel + freetime + health + absences + G1 + G2,
##     family = "binomial", data = student_math_PF)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.88309  -0.02830   0.00328   0.13082   2.36556
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -20.16349    3.27547  -6.156 7.47e-10 ***
## sexM          -0.39261    0.52814  -0.743  0.45725
## studytime     -0.89826    0.34539  -2.601  0.00930 **
## schoolsupyes   0.02781    0.59438   0.047  0.96268
## famsupyes     -0.39508    0.50775  -0.778  0.43651
## paidyes        0.08593    0.49225   0.175  0.86142
## romanticyes   -0.74386    0.52499  -1.417  0.15651
## famrel         0.85086    0.31590   2.693  0.00707 **
## freetime      -0.15098    0.25705  -0.587  0.55696
## health        -0.20115    0.17196  -1.170  0.24211
## absences      -0.03572    0.02704  -1.321  0.18654
## G1             0.25958    0.17212   1.508  0.13151
## G2             1.99651    0.32051   6.229 4.69e-10 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 500.50  on 394  degrees of freedom
## Residual deviance: 127.92  on 382  degrees of freedom
## AIC: 153.92
##
## Number of Fisher Scoring iterations: 8
```

```r
#Using backward elimination method to remove irrelevant features.
#We observe that out of 12 features we are left with 4 significant features
step(glm_pred, direction="backward")
```

```
## Start:  AIC=153.92
## PF ~ sex + studytime + schoolsup + famsup + paid + romantic +
##     famrel + freetime + health + absences + G1 + G2
##
##             Df Deviance    AIC
## - schoolsup  1   127.92 151.92
## - paid       1   127.95 151.95
## - freetime   1   128.27 152.27
## - sex        1   128.48 152.48
## - famsup     1   128.53 152.53
## - health     1   129.31 153.31
## <none>           127.92 153.92
## - romantic   1   129.96 153.96
## - absences   1   130.04 154.04
## - G1         1   130.30 154.30
## - studytime  1   135.44 159.44
## - famrel     1   136.06 160.06
## - G2         1   231.87 255.87
##
## Step:  AIC=151.92
## PF ~ sex + studytime + famsup + paid + romantic + famrel + freetime +
##     health + absences + G1 + G2
##
##             Df Deviance    AIC
## - paid       1   127.95 149.95
## - freetime   1   128.27 150.27
## - sex        1   128.48 150.48
## - famsup     1   128.54 150.54
## - health     1   129.35 151.35
## <none>           127.92 151.92
## - romantic   1   130.02 152.02
## - absences   1   130.06 152.06
## - G1         1   130.34 152.34
## - studytime  1   135.53 157.53
## - famrel     1   136.19 158.19
## - G2         1   233.31 255.31
##
## Step:  AIC=149.95
## PF ~ sex + studytime + famsup + romantic + famrel + freetime +
```

```
##     health + absences + G1 + G2
##
##             Df Deviance    AIC
## - freetime   1   128.29 148.29
## - sex        1   128.54 148.54
## - famsup     1   128.54 148.54
## - health     1   129.36 149.36
## <none>           127.95 149.95
## - absences   1   130.11 150.11
## - romantic   1   130.11 150.11
## - G1         1   130.34 150.34
## - studytime  1   135.61 155.61
## - famrel     1   136.21 156.21
## - G2         1   235.49 255.49
##
## Step:  AIC=148.29
## PF ~ sex + studytime + famsup + romantic + famrel + health +
##     absences + G1 + G2
##
##             Df Deviance    AIC
## - famsup     1   128.95 146.95
## - sex        1   129.13 147.13
## - health     1   129.71 147.71
## - absences   1   130.22 148.22
## <none>           128.29 148.29
## - romantic   1   130.40 148.40
## - G1         1   130.43 148.43
## - studytime  1   135.74 153.74
## - famrel     1   136.29 154.29
## - G2         1   238.02 256.02
##
## Step:  AIC=146.95
## PF ~ sex + studytime + romantic + famrel + health + absences +
##     G1 + G2
##
##             Df Deviance    AIC
## - sex        1   129.49 145.49
## - health     1   130.40 146.40
## - romantic   1   130.95 146.95
## <none>           128.95 146.95
## - absences   1   130.96 146.96
## - G1         1   131.30 147.30
## - studytime  1   136.15 152.15
## - famrel     1   136.94 152.94
## - G2         1   238.18 254.18
##
## Step:  AIC=145.49
## PF ~ studytime + romantic + famrel + health + absences + G1 +
##     G2
##
##             Df Deviance    AIC
## - health     1   131.02 145.02
## - romantic   1   131.38 145.38
## - absences   1   131.39 145.39
```

```
## <none>           129.49 145.49
## - G1         1   131.70 145.70
## - studytime  1   136.22 150.22
## - famrel     1   137.60 151.60
## - G2         1   238.63 252.63
##
## Step:  AIC=145.02
## PF ~ studytime + romantic + famrel + absences + G1 + G2
##
##             Df Deviance    AIC
## - absences   1   132.86 144.86
## - romantic   1   132.88 144.88
## <none>           131.02 145.02
## - G1         1   133.36 145.36
## - studytime  1   137.10 149.10
## - famrel     1   138.60 150.60
## - G2         1   240.10 252.10
##
## Step:  AIC=144.86
## PF ~ studytime + romantic + famrel + G1 + G2
##
##             Df Deviance    AIC
## - G1         1   134.68 144.68
## <none>           132.86 144.86
## - romantic   1   136.64 146.64
## - studytime  1   138.11 148.11
## - famrel     1   141.06 151.06
## - G2         1   243.62 253.62
##
## Step:  AIC=144.68
## PF ~ studytime + romantic + famrel + G2
##
##             Df Deviance    AIC
## <none>           134.68 144.68
## - romantic   1   137.85 145.85
## - studytime  1   139.37 147.37
## - famrel     1   143.53 151.53
## - G2         1   493.61 501.61
```

```
##
## Call:  glm(formula = PF ~ studytime + romantic + famrel + G2, family = "binomial",
##     data = student_math_PF)
##
## Coefficients:
## (Intercept)    studytime  romanticyes       famrel           G2
##    -21.2057      -0.6036      -0.8220       0.8104       2.1259
##
## Degrees of Freedom: 394 Total (i.e. Null);  390 Residual
## Null Deviance:       500.5
## Residual Deviance: 134.7     AIC: 144.7
```

```r
#Testing model with new features
glm_pred_new <- glm(PF~studytime+romantic+famrel+G2, data=student_math_PF, family="binomial")
```

```r
#We see that AIC has reduced to 144.68 from 153.92
summary(glm_pred_new)
```

```
##
## Call:
## glm(formula = PF ~ studytime + romantic + famrel + G2, family = "binomial",
##     data = student_math_PF)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.90805  -0.03616   0.00473   0.12694   2.37845
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -21.2057     3.1221  -6.792 1.11e-11 ***
## studytime    -0.6036     0.2883  -2.093  0.03632 *
## romanticyes  -0.8220     0.4665  -1.762  0.07809 .
## famrel        0.8104     0.2892   2.803  0.00507 **
## G2            2.1259     0.2903   7.322 2.43e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 500.50  on 394  degrees of freedom
## Residual deviance: 134.68  on 390  degrees of freedom
## AIC: 144.68
##
## Number of Fisher Scoring iterations: 8
```

```r
#a <- anova(glm_pred, glm_pred_new)
```

Regression equations:

1. glm_pred <- glm(PF~sex+studytime+schoolsup+famsup+paid+romantic+famrel+freetime+health+absences+G1+G data=student_math_PF, family="binomial")
2. glm_pred_new <- glm(PF~studytime+romantic+famrel+G2, data=student_math_PF, family="binomial")

```r
#Testing the accuracy of the model by using the PF column as response
glm_predict <- round(predict(glm_pred_new, newdata= student_math_PF, type="response"),0)
student_math_PF$glm_predict <- unname(glm_predict)

student_math_PF$PF <- as.numeric(ifelse(student_math_PF$PF == "F", 0, 1))

#The observed accuracy of the model is 92.66%
confusionMatrix(table(student_math_PF$glm_predict, student_math_PF$PF))
```

```
## Confusion Matrix and Statistics
##
##
```

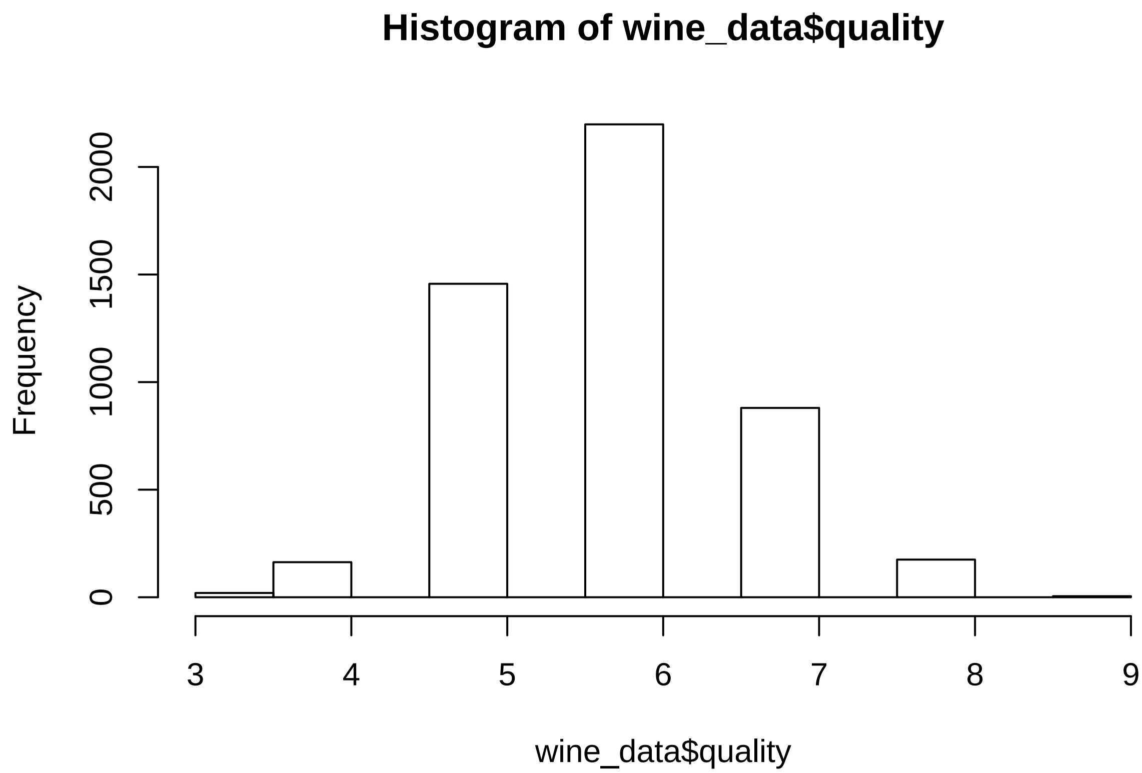
```
##       0   1
##   0 118  17
##   1  12 248
##
##                Accuracy : 0.9266
##                  95% CI : (0.8963, 0.9503)
##     No Information Rate : 0.6709
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.8354
##
##  Mcnemar's Test P-Value : 0.4576
##
##             Sensitivity : 0.9077
##             Specificity : 0.9358
##          Pos Pred Value : 0.8741
##          Neg Pred Value : 0.9538
##              Prevalence : 0.3291
##          Detection Rate : 0.2987
##    Detection Prevalence : 0.3418
##       Balanced Accuracy : 0.9218
##
##        'Positive' Class : 0
##
```

Problem 3 : 1. Implement the example from the textbook on pages 205 to 217 for the data set on white wines. 2. Calculate the RMSE for the model.

```r
#Importing wine dataset
wine_data <- read.csv("C:\\Users\\harsh\\Desktop\\Introduction to Machine learning and Data Mining\\Prac

#Exploring wine data
str(wine_data)
```

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

```r
#We can see the data is normally distributed where 5-6 are the most common values
hist(wine_data$quality)
```




```r
#Splitting the data into training and testing dataset
wine_train <- wine_data[1:3750,]
wine_test <- wine_data[3751:4898,]

#Using rpart function for generating classification tree of wine dataset
#where quality is selected as independent variable
model <- rpart(quality ~ ., data = wine_train)
model
```

```
## n= 3750
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 3750 3140.06000 5.886933
##    2) alcohol< 10.85 2473 1510.66200 5.609381
##      4) volatile.acidity>=0.2425 1406  740.15080 5.402560
##        8) volatile.acidity>=0.4225 182   92.99451 4.994505 *
##        9) volatile.acidity< 0.4225 1224  612.34560 5.463235 *
##      5) volatile.acidity< 0.2425 1067  631.12090 5.881912 *
##    3) alcohol>=10.85 1277 1069.95800 6.424432
##      6) free.sulfur.dioxide< 11.5 93   99.18280 5.473118 *
##      7) free.sulfur.dioxide>=11.5 1184  879.99920 6.499155
##       14) alcohol< 11.85 611  447.38130 6.296236 *
##       15) alcohol>=11.85 573  380.63180 6.715532 *
```

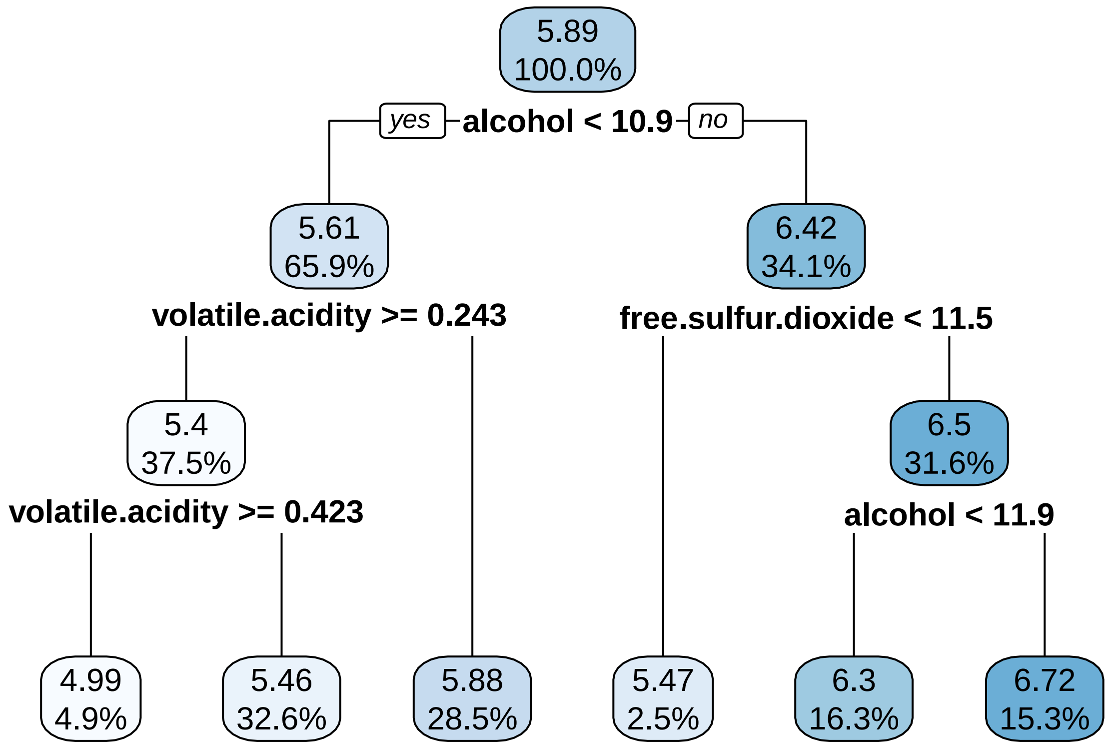
```r
#Summary provides details of each and every node and
#number of observations present in the specific node
summary(model)
```

```
## Call:
## rpart(formula = quality ~ ., data = wine_train)
##   n= 3750
##
##           CP nsplit rel error    xerror       xstd
## 1 0.17816211      0 1.0000000 1.0006941 0.02389332
## 2 0.04439109      1 0.8218379 0.8263982 0.02240013
## 3 0.02890893      2 0.7774468 0.7908293 0.02215446
## 4 0.01655575      3 0.7485379 0.7623477 0.02098749
## 5 0.01108600      4 0.7319821 0.7471944 0.02051056
## 6 0.01000000      5 0.7208961 0.7412701 0.02031185
##
## Variable importance
##              alcohol              density            chlorides
##                   38                   23                   12
##     volatile.acidity total.sulfur.dioxide  free.sulfur.dioxide
##                   12                    7                    6
##            sulphates                   pH       residual.sugar
##                    1                    1                    1
##
## Node number 1: 3750 observations,    complexity param=0.1781621
##   mean=5.886933, MSE=0.8373493
##   left son=2 (2473 obs) right son=3 (1277 obs)
##   Primary splits:
##       alcohol              < 10.85    to the left,  improve=0.17816210, (0 missing)
##       density              < 0.992385 to the right, improve=0.11980970, (0 missing)
##       chlorides            < 0.0395   to the right, improve=0.08199995, (0 missing)
##       total.sulfur.dioxide < 153.5    to the right, improve=0.03875440, (0 missing)
##       free.sulfur.dioxide  < 11.75    to the left,  improve=0.03632119, (0 missing)
##   Surrogate splits:
##       density              < 0.99201  to the right, agree=0.869, adj=0.614, (0 split)
##       chlorides            < 0.0375   to the right, agree=0.773, adj=0.334, (0 split)
##       total.sulfur.dioxide < 102.5    to the right, agree=0.705, adj=0.132, (0 split)
##       sulphates            < 0.345    to the right, agree=0.670, adj=0.031, (0 split)
##       fixed.acidity        < 5.25     to the right, agree=0.662, adj=0.009, (0 split)
##
## Node number 2: 2473 observations,    complexity param=0.04439109
##   mean=5.609381, MSE=0.6108623
##   left son=4 (1406 obs) right son=5 (1067 obs)
##   Primary splits:
##       volatile.acidity    < 0.2425   to the right, improve=0.09227123, (0 missing)
##       free.sulfur.dioxide < 13.5     to the left,  improve=0.04177240, (0 missing)
##       alcohol             < 10.15    to the left,  improve=0.03313802, (0 missing)
##       citric.acid         < 0.205    to the left,  improve=0.02721200, (0 missing)
##       pH                  < 3.325    to the left,  improve=0.01860335, (0 missing)
##   Surrogate splits:
##       total.sulfur.dioxide < 111.5    to the right, agree=0.610, adj=0.097, (0 split)
##       pH                   < 3.295    to the left,  agree=0.598, adj=0.067, (0 split)
##       alcohol              < 10.05    to the left,  agree=0.590, adj=0.049, (0 split)
```

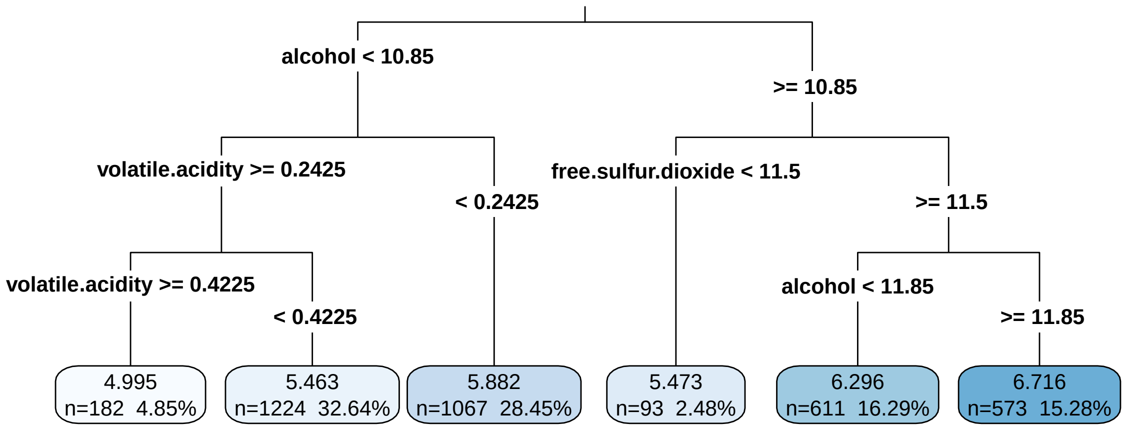
```
##       sulphates            < 0.715    to the left,  agree=0.584, adj=0.037, (0 split)
##       residual.sugar       < 1.85     to the right, agree=0.581, adj=0.029, (0 split)
##
## Node number 3: 1277 observations,    complexity param=0.02890893
##   mean=6.424432, MSE=0.8378682
##   left son=6 (93 obs) right son=7 (1184 obs)
##   Primary splits:
##       free.sulfur.dioxide  < 11.5     to the left,  improve=0.08484051, (0 missing)
##       alcohol              < 11.85    to the left,  improve=0.06149941, (0 missing)
##       fixed.acidity        < 7.35     to the right, improve=0.04259695, (0 missing)
##       residual.sugar       < 1.275    to the left,  improve=0.02795662, (0 missing)
##       total.sulfur.dioxide < 67.5     to the left,  improve=0.02541719, (0 missing)
##   Surrogate splits:
##       total.sulfur.dioxide < 48.5     to the left,  agree=0.937, adj=0.14, (0 split)
##
## Node number 4: 1406 observations,    complexity param=0.011086
##   mean=5.40256, MSE=0.526423
##   left son=8 (182 obs) right son=9 (1224 obs)
##   Primary splits:
##       volatile.acidity     < 0.4225   to the right, improve=0.04703189, (0 missing)
##       free.sulfur.dioxide  < 17.5     to the left,  improve=0.04607770, (0 missing)
##       total.sulfur.dioxide < 86.5     to the left,  improve=0.02894310, (0 missing)
##       alcohol              < 10.25    to the left,  improve=0.02890077, (0 missing)
##       chlorides            < 0.0455   to the right, improve=0.02096635, (0 missing)
##   Surrogate splits:
##       density       < 0.99107  to the left,  agree=0.874, adj=0.027, (0 split)
##       citric.acid   < 0.11     to the left,  agree=0.873, adj=0.022, (0 split)
##       fixed.acidity < 9.85     to the right, agree=0.873, adj=0.016, (0 split)
##       chlorides     < 0.206    to the right, agree=0.871, adj=0.005, (0 split)
##
## Node number 5: 1067 observations
##   mean=5.881912, MSE=0.591491
##
## Node number 6: 93 observations
##   mean=5.473118, MSE=1.066482
##
## Node number 7: 1184 observations,    complexity param=0.01655575
##   mean=6.499155, MSE=0.7432425
##   left son=14 (611 obs) right son=15 (573 obs)
##   Primary splits:
##       alcohol        < 11.85    to the left,  improve=0.05907511, (0 missing)
##       fixed.acidity  < 7.35     to the right, improve=0.04400660, (0 missing)
##       density        < 0.991395 to the right, improve=0.02522410, (0 missing)
##       residual.sugar < 1.225    to the left,  improve=0.02503936, (0 missing)
##       pH             < 3.245    to the left,  improve=0.02417936, (0 missing)
##   Surrogate splits:
##       density              < 0.991115 to the right, agree=0.710, adj=0.401, (0 split)
##       volatile.acidity     < 0.2675   to the left,  agree=0.665, adj=0.307, (0 split)
##       chlorides            < 0.0365   to the right, agree=0.631, adj=0.237, (0 split)
##       total.sulfur.dioxide < 126.5    to the right, agree=0.566, adj=0.103, (0 split)
##       residual.sugar       < 1.525    to the left,  agree=0.560, adj=0.091, (0 split)
##
## Node number 8: 182 observations
##   mean=4.994505, MSE=0.5109588
```

```
##
## Node number 9: 1224 observations
##   mean=5.463235, MSE=0.5002823
##
## Node number 14: 611 observations
##   mean=6.296236, MSE=0.7322117
##
## Node number 15: 573 observations
##   mean=6.715532, MSE=0.6642788
```

```r
#rpart.plot function is used to plot the classification tree
rpart.plot(model, digits = 3)
```




```r
#fallen.leaves parameter forces the leaf nodes to be aligned at the
#bottom of the plot, while the type and extra parameters affect the
#way the decisions and nodes are labeled
rpart.plot(model, digits = 4, fallen.leaves = TRUE, type = 3, extra = 101)
```




```r
#Testing the rpart model using testing data
predict <- predict(model, wine_test)

#Based on the observation we see that the extreme cases are not handled properly
#as the max value for both variables vary
summary(predict)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.995   5.463   5.882   5.999   6.296   6.716
```

```r
summary(wine_test$quality)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.848   6.000   8.000
```

```r
#Correlation between predicted and acutal quality compares how well
#the prediction has taken place
cor(predict, wine_test$quality)
```

```
## [1] 0.4931608
```

```r
#Mean absolute error function
MAE <- function(actual, predicted)
```

```r
{
  mean(abs(actual - predicted))
}

#Calculating MAE for the predicted model
MAE(predict, wine_test$quality)
```

```
## [1] 0.5732104
```

```r
#Mean of quality rating
mean(wine_train$quality)
```

```
## [1] 5.886933
```

```r
#Checking error for 5.88 i.e mean value
MAE(5.88, wine_test$quality)
```

```
## [1] 0.5778397
```

```r
#For some reason the model performance did not improve. The values observed were a bit insignificant.
m5p <- M5P(quality ~ ., data = wine_train)
summary(m5p)
```

```
##
## === Summary ===
##
## Correlation coefficient                  -0.2414
## Mean absolute error                     102.3629
## Root mean squared error                 129.5719
## Relative absolute error               14704.2234 %
## Root relative squared error           14159.8116 %
## Total Number of Instances              3750
```

```r
#Evaluation of model using testing data
p.m5p <- predict(m5p, wine_test)
summary(p.m5p)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -539.90 -165.65 -107.07 -112.27  -33.70   32.49
```

```r
cor(p.m5p, wine_test$quality)
```

```
## [1] -0.2036594
```

```r
#The MAE of the model was a bit off. The observed value is 118.68
#which is huge compared to the above model
MAE(wine_test$quality, p.m5p)
```

```
## [1] 118.6835
```

```r
#RMSE function
RMSE <- function(actual, pred)
{
  return(sqrt(sum(actual-pred)^2/length(actual)))
}

#The RMSE error of the model was observed as 4002.081 which means the model is broken
RMSE(wine_test$quality, p.m5p)
```

```
## [1] 4002.081
```