student_performance_code

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
#reading csv
stud_perf <- read.csv("C:/Users/cmanz/OneDrive/Documents/Ryerson
stuff/cind820/student dataset/student-perf.csv", header = T, stringsAsFactors
= F,                          na.strings = c("", "NA"), sep = ";")
head(stud_perf)

##   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
```

```
#checking the datatypes of the attributes
str(stud_perf)
```

```
## 'data.frame':    1044 obs. of  33 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : chr  "at_home" "at_home" "at_home" "health" ...
##  $ Fjob      : chr  "teacher" "other" "other" "services" ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "yes" "yes" ...
##  $ activities: chr  "no" "no" "no" "yes" ...
##  $ nursery   : chr  "yes" "no" "yes" "yes" ...
##  $ higher    : chr  "yes" "yes" "yes" "yes" ...
##  $ internet  : chr  "no" "yes" "yes" "yes" ...
##  $ romantic  : chr  "no" "no" "no" "yes" ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
##  $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

```
#checking for missing values
sum(is.na(stud_perf))
```

```
## [1] 0
```

```r
#looking for correlation between numeric attributes except final grade(G3)
cor(stud_perf[, c('age', 'Medu', 'Fedu', 'traveltime', 'studytime',
'failures', 'famrel', 'freetime', 'goout', 'Dalc', 'Walc', 'health',
'absences',                      'G1', 'G2')])
```

```
##                    age         Medu          Fedu   traveltime
studytime
## age         1.000000000 -0.130196115 -0.1385207614  0.049215707 -
0.007870098
## Medu       -0.130196115  1.000000000  0.6420631457 -0.238180728
0.090616377
## Fedu       -0.138520761  0.642063146  1.0000000000 -0.196328161
0.033457874
## traveltime  0.049215707 -0.238180728 -0.1963281605  1.000000000 -
0.081328016
## studytime  -0.007870098  0.090616377  0.0334578745 -0.081328016
1.000000000
## failures    0.282363566 -0.187769404 -0.1913904210  0.087177495 -
0.152023523
## famrel      0.007161921  0.015003618  0.0130659150 -0.012577522
0.012324093
## freetime    0.002645147  0.001054219  0.0021417298 -0.007402578 -
0.094429345
## goout       0.118510124  0.025614278  0.0300748764  0.049739783 -
0.072940739
## Dalc        0.133452990  0.001515097 -0.0001648393  0.109423016 -
0.159664641
## Walc        0.098291406 -0.029330541  0.0195239342  0.084292404 -
0.229073148
## health     -0.029129265 -0.013254090  0.0342882377 -0.029001978 -
0.063044459
## absences    0.153195647  0.059707676  0.0408288855 -0.022668699 -
0.075593669
## G1         -0.124121249  0.226100602  0.1958980209 -0.121053301
0.211313915
## G2         -0.119474744  0.224661748  0.1826339619 -0.140162973
0.183166702
##               failures       famrel     freetime       goout          Dalc
## age         0.28236357  0.007161921  0.002645147  0.11851012  0.1334529897
## Medu       -0.18776940  0.015003618  0.001054219  0.02561428  0.0015150967
## Fedu       -0.19139042  0.013065915  0.002141730  0.03007488 -0.0001648393
## traveltime  0.08717749 -0.012577522 -0.007402578  0.04973978  0.1094230162
## studytime  -0.15202352  0.012324093 -0.094429345 -0.07294074 -0.1596646413
## failures    1.00000000 -0.053676457  0.102678757  0.07468331  0.1163357901
## famrel     -0.05367646  1.000000000  0.136900650  0.08061921 -0.0764826572
## freetime    0.10267876  0.136900650  1.000000000  0.32355575  0.1449791279
## goout       0.07468331  0.080619212  0.323555753  1.00000000  0.2531348291
## Dalc        0.11633579 -0.076482657  0.144979128  0.25313483  1.0000000000
## Walc        0.10743159 -0.100663375  0.130377028  0.39979373  0.6278138380
## health      0.04831102  0.104100776  0.081517225 -0.01373623  0.0655153422
```

```
## absences    0.09999785 -0.062170662 -0.032078736  0.05614214  0.1328671345
## G1         -0.37417487  0.036947274 -0.051984712 -0.10116347 -0.1509425374
## G2         -0.37717218  0.042053621 -0.068951886 -0.10841089 -0.1315764840
##                   Walc      health    absences          G1          G2
## age         0.09829141 -0.02912927  0.15319565 -0.12412125 -0.11947474
## Medu       -0.02933054 -0.01325409  0.05970768  0.22610060  0.22466175
## Fedu        0.01952393  0.03428824  0.04082889  0.19589802  0.18263396
## traveltime  0.08429240 -0.02900198 -0.02266870 -0.12105330 -0.14016297
## studytime  -0.22907315 -0.06304446 -0.07559367  0.21131391  0.18316670
## failures    0.10743159  0.04831102  0.09999785 -0.37417487 -0.37717218
## famrel     -0.10066338  0.10410078 -0.06217066  0.03694727  0.04205362
## freetime    0.13037703  0.08151722 -0.03207874 -0.05198471 -0.06895189
## goout       0.39979373 -0.01373623  0.05614214 -0.10116347 -0.10841089
## Dalc        0.62781384  0.06551534  0.13286713 -0.15094254 -0.13157648
## Walc        1.00000000  0.10666944  0.13970313 -0.14240140 -0.12811435
## health      0.10666944  1.00000000 -0.02747860 -0.06047794 -0.08800109
## absences    0.13970313 -0.02747860  1.00000000 -0.09242463 -0.08933169
## G1         -0.14240140 -0.06047794 -0.09242463  1.00000000  0.85873875
## G2         -0.12811435 -0.08800109 -0.08933169  0.85873875  1.00000000
```

```
#graphing frequency distribution of final grade(G3)
library(epiDisplay)
```

```
## Warning: package 'epiDisplay' was built under R version 4.1.3
```

```
## Loading required package: foreign
```

```
## Loading required package: survival
```

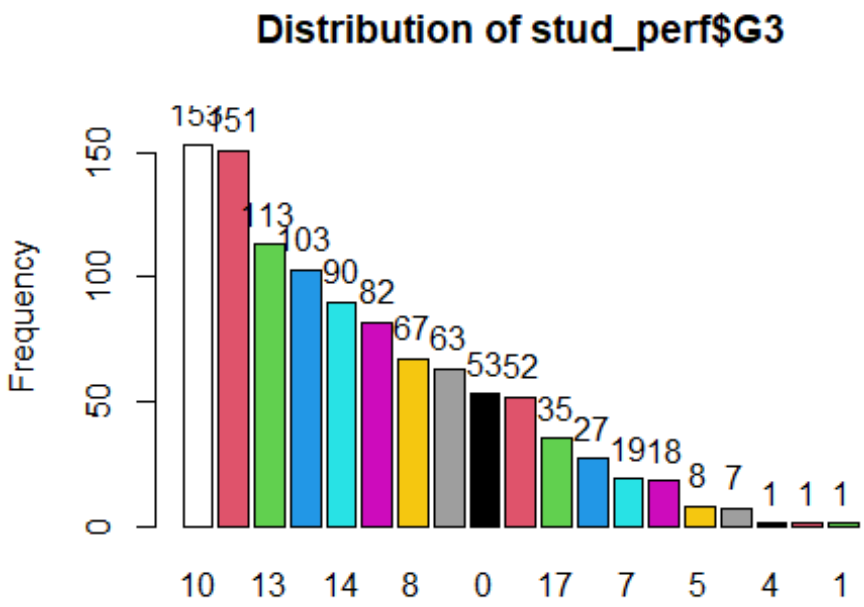
```
## Loading required package: MASS
```

```
## Warning: package 'MASS' was built under R version 4.1.2
```

```
## Loading required package: nnet
```

```
## Warning: package 'nnet' was built under R version 4.1.3
```

```
tab1(stud_perf$G3, sort.group = "decreasing", cum.percent = T)
```




```
## stud_perf$G3 :
##          Frequency Percent Cum. percent
## 10             153    14.7         14.7
## 11             151    14.5         29.1
## 13             113    10.8         39.9
## 12             103     9.9         49.8
## 14              90     8.6         58.4
## 15              82     7.9         66.3
## 8               67     6.4         72.7
## 9               63     6.0         78.7
## 0               53     5.1         83.8
## 16              52     5.0         88.8
## 17              35     3.4         92.1
## 18              27     2.6         94.7
## 7               19     1.8         96.6
## 6               18     1.7         98.3
## 5                8     0.8         99.0
## 19               7     0.7         99.7
## 4                1     0.1         99.8
## 20               1     0.1         99.9
## 1                1     0.1        100.0
##   Total       1044   100.0        100.0

#assigning numeric values to Mjob(mother's job) and Fjob(father's job).
# 1 - at_home
# 2 - services
# 3 - other
```

```
# 4 - teacher
# 5 - health
stud_perf$Mjob[stud_perf$Mjob == 'at_home'] = 1
stud_perf$Mjob[stud_perf$Mjob == 'services'] = 2
stud_perf$Mjob[stud_perf$Mjob == 'other'] = 3
stud_perf$Mjob[stud_perf$Mjob == 'teacher'] = 4
stud_perf$Mjob[stud_perf$Mjob == 'health'] = 5

stud_perf$Fjob[stud_perf$Fjob == 'at_home'] = 1
stud_perf$Fjob[stud_perf$Fjob == 'services'] = 2
stud_perf$Fjob[stud_perf$Fjob == 'other'] = 3
stud_perf$Fjob[stud_perf$Fjob == 'teacher'] = 4
stud_perf$Fjob[stud_perf$Fjob == 'health'] = 5

head(stud_perf)
```

```
##   school sex age address famsize Pstatus Medu Fedu Mjob Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4    1    4     course
## 2     GP   F  17       U     GT3       T    1    1    1    3     course
## 3     GP   F  15       U     LE3       T    1    1    1    3      other
## 4     GP   F  15       U     GT3       T    4    2    5    2       home
## 5     GP   F  16       U     GT3       T    3    3    3    3       home
## 6     GP   M  16       U     LE3       T    4    3    2    3 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
```

```
#assigning binary values to yes or no attributes (schoolsup, famsup, paid,
activities, nursery, higher, internet, romantic)
stud_perf$schoolsup[stud_perf$schoolsup == 'yes'] = 1
stud_perf$schoolsup[stud_perf$schoolsup == 'no'] = 0
```

```
stud_perf$famsup[stud_perf$famsup == 'yes'] = 1
stud_perf$famsup[stud_perf$famsup == 'no'] = 0

stud_perf$paid[stud_perf$paid == 'yes'] = 1
stud_perf$paid[stud_perf$paid == 'no'] = 0

stud_perf$activities[stud_perf$activities == 'yes'] = 1
stud_perf$activities[stud_perf$activities == 'no'] = 0

stud_perf$nursery[stud_perf$nursery == 'yes'] = 1
stud_perf$nursery[stud_perf$nursery == 'no'] = 0

stud_perf$higher[stud_perf$higher == 'yes'] = 1
stud_perf$higher[stud_perf$higher == 'no'] = 0

stud_perf$internet[stud_perf$internet == 'yes'] = 1
stud_perf$internet[stud_perf$internet == 'no'] = 0

stud_perf$romantic[stud_perf$romantic == 'yes'] = 1
stud_perf$romantic[stud_perf$romantic == 'no'] = 0

head(stud_perf)
```

```
##   school sex age address famsize Pstatus Medu Fedu Mjob Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4    1    4     course
## 2     GP   F  17       U     GT3       T    1    1    1    3     course
## 3     GP   F  15       U     LE3       T    1    1    1    3      other
## 4     GP   F  15       U     GT3       T    4    2    5    2       home
## 5     GP   F  16       U     GT3       T    3    3    3    3       home
## 6     GP   M  16       U     LE3       T    4    3    2    3 reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0         1      0    0          0
## 2   father          1         2        0         0      1    0          0
## 3   mother          1         2        3         1      0    1          0
## 4   mother          1         3        0         0      1    1          1
## 5   father          1         2        0         0      1    1          0
## 6   mother          1         2        0         0      1    1          1
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1       1      1        0        0      4        3     4    1    1      3
## 2       0      1        1        0      5        3     3    1    1      3
## 3       1      1        1        0      4        3     2    2    3      3
## 4       1      1        1        1      3        2     2    1    1      5
## 5       1      1        0        0      4        3     2    1    2      5
## 6       1      1        1        0      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
```

```
#assigning pass or fail to the grades columns
#for G1
stud_perf$G1[stud_perf$G1 == 0] = 'Fail'
stud_perf$G1[stud_perf$G1 == 1] = 'Fail'
stud_perf$G1[stud_perf$G1 == 2] = 'Fail'
stud_perf$G1[stud_perf$G1 == 3] = 'Fail'
stud_perf$G1[stud_perf$G1 == 4] = 'Fail'
stud_perf$G1[stud_perf$G1 == 5] = 'Fail'
stud_perf$G1[stud_perf$G1 == 6] = 'Fail'
stud_perf$G1[stud_perf$G1 == 7] = 'Fail'
stud_perf$G1[stud_perf$G1 == 8] = 'Fail'
stud_perf$G1[stud_perf$G1 == 9] = 'Fail'
stud_perf$G1[stud_perf$G1 == 10] = 'Pass'
stud_perf$G1[stud_perf$G1 == 11] = 'Pass'
stud_perf$G1[stud_perf$G1 == 12] = 'Pass'
stud_perf$G1[stud_perf$G1 == 13] = 'Pass'
stud_perf$G1[stud_perf$G1 == 14] = 'Pass'
stud_perf$G1[stud_perf$G1 == 15] = 'Pass'
stud_perf$G1[stud_perf$G1 == 16] = 'Pass'
stud_perf$G1[stud_perf$G1 == 17] = 'Pass'
stud_perf$G1[stud_perf$G1 == 18] = 'Pass'
stud_perf$G1[stud_perf$G1 == 19] = 'Pass'
stud_perf$G1[stud_perf$G1 == 20] = 'Pass'

#for G2
stud_perf$G2[stud_perf$G2 == 0] = 'Fail'
stud_perf$G2[stud_perf$G2 == 1] = 'Fail'
stud_perf$G2[stud_perf$G2 == 2] = 'Fail'
stud_perf$G2[stud_perf$G2 == 3] = 'Fail'
stud_perf$G2[stud_perf$G2 == 4] = 'Fail'
stud_perf$G2[stud_perf$G2 == 5] = 'Fail'
stud_perf$G2[stud_perf$G2 == 6] = 'Fail'
stud_perf$G2[stud_perf$G2 == 7] = 'Fail'
stud_perf$G2[stud_perf$G2 == 8] = 'Fail'
stud_perf$G2[stud_perf$G2 == 9] = 'Fail'
stud_perf$G2[stud_perf$G2 == 10] = 'Pass'
stud_perf$G2[stud_perf$G2 == 11] = 'Pass'
stud_perf$G2[stud_perf$G2 == 12] = 'Pass'
stud_perf$G2[stud_perf$G2 == 13] = 'Pass'
stud_perf$G2[stud_perf$G2 == 14] = 'Pass'
stud_perf$G2[stud_perf$G2 == 15] = 'Pass'
stud_perf$G2[stud_perf$G2 == 16] = 'Pass'
stud_perf$G2[stud_perf$G2 == 17] = 'Pass'
stud_perf$G2[stud_perf$G2 == 18] = 'Pass'
stud_perf$G2[stud_perf$G2 == 19] = 'Pass'
stud_perf$G2[stud_perf$G2 == 20] = 'Pass'
```

```r
#for G3
stud_perf$G3[stud_perf$G3 == 0] = 'Fail'
stud_perf$G3[stud_perf$G3 == 1] = 'Fail'
stud_perf$G3[stud_perf$G3 == 2] = 'Fail'
stud_perf$G3[stud_perf$G3 == 3] = 'Fail'
stud_perf$G3[stud_perf$G3 == 4] = 'Fail'
stud_perf$G3[stud_perf$G3 == 5] = 'Fail'
stud_perf$G3[stud_perf$G3 == 6] = 'Fail'
stud_perf$G3[stud_perf$G3 == 7] = 'Fail'
stud_perf$G3[stud_perf$G3 == 8] = 'Fail'
stud_perf$G3[stud_perf$G3 == 9] = 'Fail'
stud_perf$G3[stud_perf$G3 == 10] = 'Pass'
stud_perf$G3[stud_perf$G3 == 11] = 'Pass'
stud_perf$G3[stud_perf$G3 == 12] = 'Pass'
stud_perf$G3[stud_perf$G3 == 13] = 'Pass'
stud_perf$G3[stud_perf$G3 == 14] = 'Pass'
stud_perf$G3[stud_perf$G3 == 15] = 'Pass'
stud_perf$G3[stud_perf$G3 == 16] = 'Pass'
stud_perf$G3[stud_perf$G3 == 17] = 'Pass'
stud_perf$G3[stud_perf$G3 == 18] = 'Pass'
stud_perf$G3[stud_perf$G3 == 19] = 'Pass'
stud_perf$G3[stud_perf$G3 == 20] = 'Pass'

head(stud_perf)
```

```
##   school sex age address famsize Pstatus Medu Fedu Mjob Fjob     reason
## 1     GP   F  18       U     GT3       A    4    4    1    4     course
## 2     GP   F  17       U     GT3       T    1    1    1    3     course
## 3     GP   F  15       U     LE3       T    1    1    1    3      other
## 4     GP   F  15       U     GT3       T    4    2    5    2       home
## 5     GP   F  16       U     GT3       T    3    3    3    3       home
## 6     GP   M  16       U     LE3       T    4    3    2    3 reputation
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0         1      0    0          0
## 2   father          1         2        0         0      1    0          0
## 3   mother          1         2        3         1      0    1          0
## 4   mother          1         3        0         0      1    1          1
## 5   father          1         2        0         0      1    1          0
## 6   mother          1         2        0         0      1    1          1
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1       1      1        0        0      4        3     4    1    1      3
## 2       0      1        1        0      5        3     3    1    1      3
## 3       1      1        1        0      4        3     2    2    3      3
## 4       1      1        1        1      3        2     2    1    1      5
## 5       1      1        0        0      4        3     2    1    2      5
## 6       1      1        1        0      5        4     2    1    2      5
##   absences   G1   G2   G3
## 1        6 Fail Fail Fail
## 2        4 Fail Fail Fail
## 3       10 Fail Fail Pass
```

```
## 4        2 Pass Pass Pass
## 5        4 Fail Pass Pass
## 6       10 Pass Pass Pass
```

```
#changing specific columns to numeric
stud_perf$Mjob <- as.numeric(as.character(stud_perf$Mjob))
stud_perf$Fjob <- as.numeric(as.character(stud_perf$Fjob))
stud_perf$schoolsup <- as.numeric(as.character(stud_perf$schoolsup))
stud_perf$famsup <- as.numeric(as.character(stud_perf$famsup))
stud_perf$paid <- as.numeric(as.character(stud_perf$paid))
stud_perf$activities <- as.numeric(as.character(stud_perf$activities))
stud_perf$nursery <- as.numeric(as.character(stud_perf$nursery))
stud_perf$higher <- as.numeric(as.character(stud_perf$higher))
stud_perf$internet <- as.numeric(as.character(stud_perf$internet))
stud_perf$romantic <- as.numeric(as.character(stud_perf$romantic))
```

```
str(stud_perf)
```

```
## 'data.frame':    1044 obs. of  33 variables:
##  $ school    : chr  "GP" "GP" "GP" "GP" ...
##  $ sex       : chr  "F" "F" "F" "F" ...
##  $ age       : int  18 17 15 15 16 16 16 17 15 15 ...
##  $ address   : chr  "U" "U" "U" "U" ...
##  $ famsize   : chr  "GT3" "GT3" "LE3" "GT3" ...
##  $ Pstatus   : chr  "A" "T" "T" "T" ...
##  $ Medu      : int  4 1 1 4 3 4 2 4 3 3 ...
##  $ Fedu      : int  4 1 1 2 3 3 2 4 2 4 ...
##  $ Mjob      : num  1 1 1 5 3 2 3 3 2 3 ...
##  $ Fjob      : num  4 3 3 2 3 3 3 4 3 3 ...
##  $ reason    : chr  "course" "course" "other" "home" ...
##  $ guardian  : chr  "mother" "father" "mother" "mother" ...
##  $ traveltime: int  2 1 1 1 1 1 1 2 1 1 ...
##  $ studytime : int  2 2 2 3 2 2 2 2 2 2 ...
##  $ failures  : int  0 0 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : num  1 0 1 0 0 0 0 1 0 0 ...
##  $ famsup    : num  0 1 0 1 1 1 0 1 1 1 ...
##  $ paid      : num  0 0 1 1 1 1 0 0 1 1 ...
##  $ activities: num  0 0 0 1 0 1 0 0 0 1 ...
##  $ nursery   : num  1 0 1 1 1 1 1 1 1 1 ...
##  $ higher    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ internet  : num  0 1 1 1 0 1 1 0 1 1 ...
##  $ romantic  : num  0 0 0 1 0 0 0 0 0 0 ...
##  $ famrel    : int  4 5 4 3 4 5 4 4 4 5 ...
##  $ freetime  : int  3 3 3 2 3 4 4 1 2 5 ...
##  $ goout     : int  4 3 2 2 2 2 4 4 2 1 ...
##  $ Dalc      : int  1 1 2 1 1 1 1 1 1 1 ...
##  $ Walc      : int  1 1 3 1 2 2 1 1 1 1 ...
##  $ health    : int  3 3 3 5 5 5 3 1 1 5 ...
##  $ absences  : int  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : chr  "Fail" "Fail" "Fail" "Pass" ...
```

```
##  $ G2        : chr  "Fail" "Fail" "Fail" "Pass" ...
##  $ G3        : chr  "Fail" "Fail" "Pass" "Pass" ...

#normalizing the numeric attributes
minmaxNorm <- function(x) {
  (x - min(x)) / (max(x) - min(x))
}

studperf_norm1 <- as.data.frame(lapply(stud_perf[7:10], minmaxNorm))
head(studperf_norm1)

##   Medu Fedu Mjob Fjob
## 1 1.00 1.00 0.00 0.75
## 2 0.25 0.25 0.00 0.50
## 3 0.25 0.25 0.00 0.50
## 4 1.00 0.50 1.00 0.25
## 5 0.75 0.75 0.50 0.50
## 6 1.00 0.75 0.25 0.50

studperf_norm2 <- as.data.frame(lapply(stud_perf[13:30], minmaxNorm))
head(studperf_norm2)

##   traveltime studytime failures schoolsup famsup paid activities nursery
higher
## 1  0.3333333 0.3333333        0         1      0    0          0       1
1
## 2  0.0000000 0.3333333        0         0      1    0          0       0
1
## 3  0.0000000 0.3333333        1         1      0    1          0       1
1
## 4  0.0000000 0.6666667        0         0      1    1          1       1
1
## 5  0.0000000 0.3333333        0         0      1    1          0       1
1
## 6  0.0000000 0.3333333        0         0      1    1          1       1
1
##   internet romantic famrel freetime goout Dalc Walc health   absences
## 1        0        0   0.75     0.50  0.75 0.00 0.00    0.5 0.08000000
## 2        1        0   1.00     0.50  0.50 0.00 0.00    0.5 0.05333333
## 3        1        0   0.75     0.50  0.25 0.25 0.50    0.5 0.13333333
## 4        1        1   0.50     0.25  0.25 0.00 0.00    1.0 0.02666667
## 5        0        0   0.75     0.50  0.25 0.00 0.25    1.0 0.05333333
## 6        1        0   1.00     0.75  0.25 0.00 0.25    1.0 0.13333333

#merging the normalized data frames side by side
studperf_norm <- cbind(studperf_norm1, studperf_norm2)
head(studperf_norm)

##   Medu Fedu Mjob Fjob traveltime studytime failures schoolsup famsup paid
## 1 1.00 1.00 0.00 0.75  0.3333333 0.3333333        0         1      0    0
## 2 0.25 0.25 0.00 0.50  0.0000000 0.3333333        0         0      1    0
```

```
## 3 0.25 0.25 0.00 0.50  0.0000000 0.3333333        1         1      0    1
## 4 1.00 0.50 1.00 0.25  0.0000000 0.6666667        0         0      1    1
## 5 0.75 0.75 0.50 0.50  0.0000000 0.3333333        0         0      1    1
## 6 1.00 0.75 0.25 0.50  0.0000000 0.3333333        0         0      1    1
##   activities nursery higher internet romantic famrel freetime goout Dalc
Walc
## 1          0       1      1        0        0   0.75     0.50  0.75 0.00
0.00
## 2          0       0      1        1        0   1.00     0.50  0.50 0.00
0.00
## 3          0       1      1        1        0   0.75     0.50  0.25 0.25
0.50
## 4          1       1      1        1        1   0.50     0.25  0.25 0.00
0.00
## 5          0       1      1        0        0   0.75     0.50  0.25 0.00
0.25
## 6          1       1      1        1        0   1.00     0.75  0.25 0.00
0.25
##   health   absences
## 1    0.5 0.08000000
## 2    0.5 0.05333333
## 3    0.5 0.13333333
## 4    1.0 0.02666667
## 5    1.0 0.05333333
## 6    1.0 0.13333333
```

```
#creating the train and test sets
train_ind <- sample(1:nrow(stud_perf), 0.7 * nrow(stud_perf))
train.perf <- stud_perf[train_ind, ]
test.perf <- stud_perf[-train_ind, ]

head(train.perf)
```

```
##     school sex age address famsize Pstatus Medu Fedu Mjob Fjob reason
guardian
## 415     GP   M  16       U     LE3       T    4    3    5    3   home
father
## 359     MS   M  18       U     LE3       T    1    1    3    2   home
father
## 326     GP   M  18       U     GT3       T    4    4    3    3 course
mother
## 692     GP   M  18       U     GT3       T    2    1    3    3   home
mother
## 794     GP   F  18       U     GT3       T    2    3    1    3 course
mother
## 110     GP   F  16       U     LE3       T    4    4    5    5  other
mother
##     traveltime studytime failures schoolsup famsup paid activities nursery
## 415          1         1        0         0      0    0          1       1
## 359          2         1        0         0      0    0          0       0
```

```
## 326          1         3        0         0      0    0          1       1
## 692          1         2        0         0      0    0          1       1
## 794          1         3        0         0      1    0          0       1
## 110          1         3        0         0      1    1          1       1
##     higher internet romantic famrel freetime goout Dalc Walc health
absences
## 415      1        1        0      3        1     3    1    3      5
6
## 359      1        1        1      3        3     2    1    2      3
4
## 326      1        1        0      4        3     3    2    2      3
3
## 692      1        1        0      5        2     4    1    2      4
2
## 794      1        1        0      4        3     3    1    2      3
0
## 110      1        1        1      5        4     5    1    1      4
4
##       G1   G2   G3
## 415 Pass Pass Pass
## 359 Pass Pass Pass
## 326 Fail Pass Pass
## 692 Pass Pass Pass
## 794 Pass Pass Pass
## 110 Pass Pass Pass

head(test.perf)

##    school sex age address famsize Pstatus Medu Fedu Mjob Fjob reason
guardian
## 1      GP   F  18       U     GT3       A    4    4    1    4 course
mother
## 4      GP   F  15       U     GT3       T    4    2    5    2   home
mother
## 7      GP   M  16       U     LE3       T    2    2    3    3   home
mother
## 8      GP   F  17       U     GT3       A    4    4    3    4   home
mother
## 9      GP   M  15       U     LE3       A    3    2    2    3   home
mother
## 16     GP   F  16       U     GT3       T    4    4    5    3   home
mother
##    traveltime studytime failures schoolsup famsup paid activities nursery
## 1           2         2        0         1      0    0          0       1
## 4           1         3        0         0      1    1          1       1
## 7           1         2        0         0      0    0          0       1
## 8           2         2        0         1      1    0          0       1
## 9           1         2        0         0      1    1          0       1
## 16          1         1        0         0      1    0          0       1
##    higher internet romantic famrel freetime goout Dalc Walc health
```

```
absences
## 1       1        0        0      4        3     4    1    1      3
6
## 4       1        1        1      3        2     2    1    1      5
2
## 7       1        1        0      4        4     4    1    1      3
0
## 8       1        0        0      4        1     4    1    1      1
6
## 9       1        1        0      4        2     2    1    1      1
0
## 16      1        1        0      4        4     4    1    2      2
4
##      G1   G2   G3
## 1  Fail Fail Fail
## 4  Pass Pass Pass
## 7  Pass Pass Pass
## 8  Fail Fail Fail
## 9  Pass Pass Pass
## 16 Pass Pass Pass

#creating the regression model
glm_model <-
glm(as.factor(G3)~Medu+Fedu+Mjob+Fjob+traveltime+studytime+failures+schoolsup
+famsup+paid+activities+nursery+higher+internet+romantic+famrel+freetime+goou
t+Dalc+Walc+health+absences, family = "binomial", data = train.perf)

summary(glm_model)

##
## Call:
## glm(formula = as.factor(G3) ~ Medu + Fedu + Mjob + Fjob + traveltime +
##     studytime + failures + schoolsup + famsup + paid + activities +
##     nursery + higher + internet + romantic + famrel + freetime +
##     goout + Dalc + Walc + health + absences, family = "binomial",
##     data = train.perf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4023   0.3259   0.4787   0.6434   2.0904
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.10337    0.91374   2.302  0.02134 *
## Medu         0.09363    0.13418   0.698  0.48533
## Fedu         0.22562    0.12413   1.818  0.06912 .
## Mjob        -0.09574    0.10310  -0.929  0.35307
## Fjob        -0.07292    0.12892  -0.566  0.57166
## traveltime   0.08422    0.15038   0.560  0.57546
## studytime    0.11781    0.12841   0.917  0.35892
```

```
## failures    -0.88154    0.14151  -6.229 4.68e-10 ***
## schoolsup   -0.11379    0.31885  -0.357  0.72119
## famsup      -0.37594    0.21951  -1.713  0.08678 .
## paid        -0.66282    0.24187  -2.740  0.00614 **
## activities   0.01786    0.20651   0.086  0.93108
## nursery     -0.51671    0.27910  -1.851  0.06412 .
## higher       0.90143    0.33420   2.697  0.00699 **
## internet     0.02763    0.25784   0.107  0.91468
## romantic    -0.12715    0.21149  -0.601  0.54769
## famrel       0.05723    0.11253   0.509  0.61103
## freetime    -0.06149    0.10478  -0.587  0.55733
## goout       -0.22100    0.10033  -2.203  0.02761 *
## Dalc        -0.04020    0.14347  -0.280  0.77932
## Walc         0.06404    0.11013   0.581  0.56092
## health      -0.08149    0.07541  -1.081  0.27985
## absences    -0.04191    0.01394  -3.007  0.00263 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 762.65  on 729  degrees of freedom
## Residual deviance: 644.30  on 707  degrees of freedom
## AIC: 690.3
##
## Number of Fisher Scoring iterations: 4
```

```
#confusion matrix for the regression model
predicted <- predict(glm_model, test.perf, type = "response")

predicted_class <- ifelse(predicted >= 0.5, 1, 0)
ConfusionMatrix <- table(actual = test.perf$G3, predicted = predicted_class)
ConfusionMatrix
```

```
##       predicted
## actual   0   1
##   Fail  19  53
##   Pass   7 235
```

```
#finding accuracy, precision, recall, sensitivity and specificity using the
confusion matrix
#accuracy
acc <- sum(diag(ConfusionMatrix))/nrow(test.perf)

#precision
prec <- ConfusionMatrix[2,2]/sum(ConfusionMatrix[2,2]+ConfusionMatrix[2,1])

#recall
recall <- ConfusionMatrix[2,2]/sum(ConfusionMatrix[2,2]+ConfusionMatrix[1,2])
```

```r
#sensitivity
sens <- ConfusionMatrix[1,1]/sum(ConfusionMatrix[1,1]+ConfusionMatrix[2,1])

#specificity
spec <- ConfusionMatrix[2,2]/sum(ConfusionMatrix[1,2]+ConfusionMatrix[2,2])

acc
```

```
## [1] 0.8089172
```

```r
prec
```

```
## [1] 0.9710744
```

```r
recall
```

```
## [1] 0.8159722
```

```r
sens
```

```
## [1] 0.7307692
```

```r
spec
```

```
## [1] 0.8159722
```

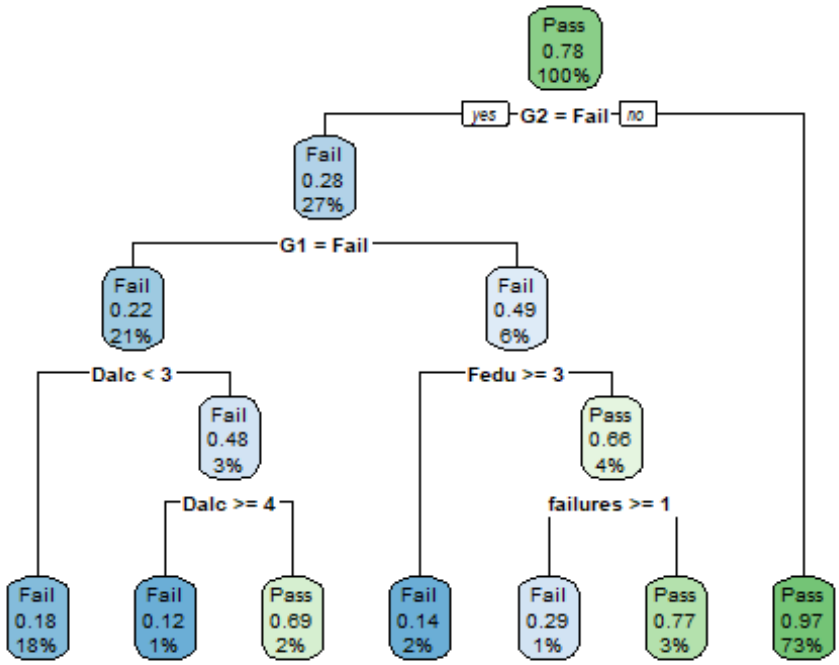
```r
#finding f score
fscore <- (2*prec*recall)/(prec+recall)
fscore
```

```
## [1] 0.8867925
```

```r
#creating a decision tree to predict G3
#install.packages("rpart.plot")
library(rpart)
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.1.3
```

```r
tree <- rpart(G3~., data = train.perf, method = 'class')
rpart.plot(tree)
```




```
#matrix for the decision tree
matrix_tree <-predict(tree, test.perf, type = 'class')
table_mat <- table(test.perf$G3, matrix_tree)
table_mat
```

```
##       matrix_tree
##        Fail Pass
##   Fail   58   14
##   Pass   14  228
```

```
#accuracy of the decision tree matrix
acc_tree <- sum(diag(table_mat))/sum(table_mat)

#precision
prec_tree <- table_mat[2,2]/sum(table_mat[2,2]+table_mat[2,1])

#recall
recall_tree <- table_mat[2,2]/sum(table_mat[2,2]+table_mat[1,2])

#sensitivity
sens_tree <- table_mat[1,1]/sum(table_mat[1,1]+table_mat[2,1])

#specificity
spec_tree <- table_mat[2,2]/sum(table_mat[1,2]+table_mat[2,2])

acc_tree
```

```
## [1] 0.910828

prec_tree

## [1] 0.9421488

recall_tree

## [1] 0.9421488

sens_tree

## [1] 0.8055556

spec_tree

## [1] 0.9421488

#random forest
#install.packages("caret")
#install.packages("e1071")
#install.packages("randomForest")
library(caret)

## Warning: package 'caret' was built under R version 4.1.3

## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:epiDisplay':
##
##     alpha

## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:epiDisplay':
##
##     dotplot

##
## Attaching package: 'caret'

## The following object is masked from 'package:survival':
##
##     cluster

library(e1071)

## Warning: package 'e1071' was built under R version 4.1.3
```

```
library(randomForest)

## Warning: package 'randomForest' was built under R version 4.1.3

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin

trControl <- trainControl(method = "cv", number = 10, search = "grid")

rf_default <- train(G3~., data = train.perf, method = "rf", metric =
"Accuracy", trControl = trControl)

print(rf_default)

## Random Forest
##
## 730 samples
##  32 predictor
##   2 classes: 'Fail', 'Pass'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 657, 657, 658, 657, 656, 657, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.8699844  0.5641578
##   18    0.9069182  0.7351383
##   35    0.9012886  0.7184730
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 18.
```