Problem 1: Cartfit Decision Tree:

> # Run to identify the file path of the file

> file.choose()

[1] "C:\\Users\\navya\\Downloads\\problem\_set-6-Clem3Training.csv"

> # Read in the data, install and load required package -

> dat <- read.csv(file.choose(),header=T)

> install.packages("rpart")

WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/

Installing package into ‘C:/Users/navya/Documents/R/win-library/3.6’

(as ‘lib’ is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.6/rpart\_4.1-15.zip'

Content type 'application/zip' length 769864 bytes (751 KB)

downloaded 751 KB

package ‘rpart’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\navya\AppData\Local\Temp\RtmpimjKBF\downloaded\_packages

> library(rpart)

> # Collapse some of the categories by giving them the same factor label

> # Display all levels for marital status and work class

> levels(dat$marital.status)

[1] "Divorced"

[2] "Married-AF-spouse"

[3] "Married-civ-spouse"

[4] "Married-spouse-absent"

[5] "Never-married"

[6] "Separated"

[7] "Widowed"

> levels(dat$workclass)

[1] "?" "Federal-gov"

[3] "Local-gov" "Never-worked"

[5] "Private" "Self-emp-inc"

[7] "Self-emp-not-inc" "State-gov"

[9] "Without-pay"

> # Reassign level names

> levels(dat$marital.status)[2:4] <- "Married"

> levels(dat$workclass)[c(2,3,8)] <- "Gov"

> levels(dat$workclass)[c(5, 6)] <- "Self"

> # Standardize the numeric variables

> dat$age.z <- (dat$age - mean(dat$age))/sd(dat$age)

> dat$education.num.z <- (dat$education.num -

+ mean(dat$education.num))/sd(dat$education.num)

> dat$capital.gain.z <- (dat$capital.gain - mean(dat$capital.gain))/sd(dat$capital.gain)

> dat$capital.loss.z <- (dat$capital.loss - mean(dat$capital.loss))/sd(dat$capital.loss)

> dat$hours.per.week.z <- (dat$hours.per.week -

+ mean(dat$hours.per.week))/sd(dat$hours.per.week)

> # Use predictors to classify whether or not a personâ€™s income is less than $50K

> cartfit <- rpart(income ~ age.z + education.num.z + capital.gain.z + capital.loss.z + hours.per.week.z + race + sex + workclass + marital.status, data = dat, method = "class")

> # Print a summary of the results

> print(cartfit)

n= 25000

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 25000 5984 <=50K. (0.76064000 0.23936000)

2) marital.status=Divorced,Never-married,Separated,Widowed 13215 845 <=50K. (0.93605751 0.06394249)

4) capital.gain.z< 0.8082312 12986 625 <=50K. (0.95187125 0.04812875) \*

5) capital.gain.z>=0.8082312 229 9 >50K. (0.03930131 0.96069869) \*

3) marital.status=Married 11785 5139 <=50K. (0.56393721 0.43606279)

6) education.num.z< 0.9458454 8296 2672 <=50K. (0.67791707 0.32208293)

12) capital.gain.z< 0.5352109 7894 2280 <=50K. (0.71117304 0.28882696)

24) capital.loss.z< 4.254168 7615 2076 <=50K. (0.72738017 0.27261983) \*

25) capital.loss.z>=4.254168 279 75 >50K. (0.26881720 0.73118280) \*

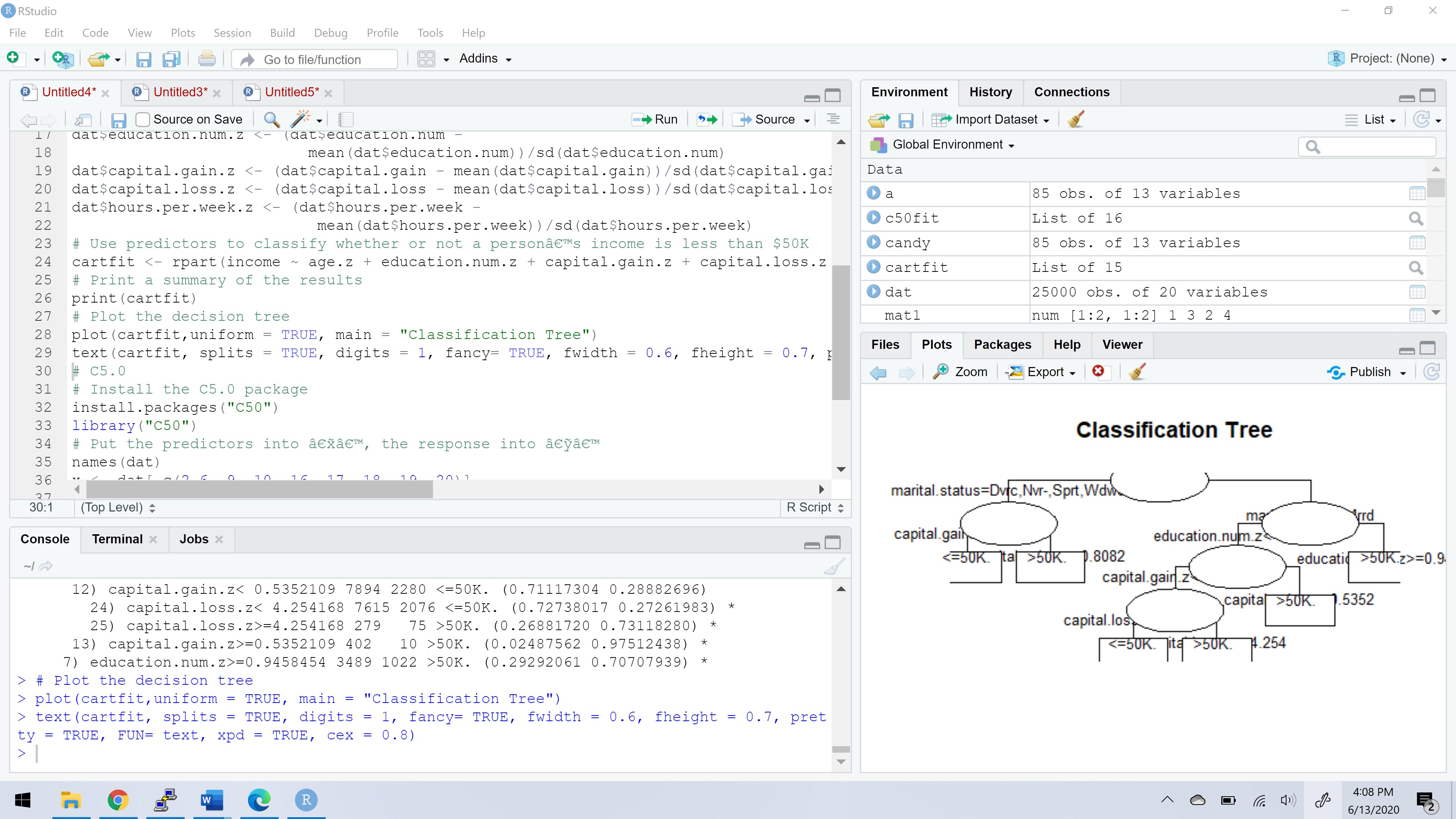
13) capital.gain.z>=0.5352109 402 10 >50K. (0.02487562 0.97512438) \*

7) education.num.z>=0.9458454 3489 1022 >50K. (0.29292061 0.70707939) \*

> # Plot the decision tree

> plot(cartfit,uniform = TRUE, main = "Classification Tree")

> text(cartfit, splits = TRUE, digits = 1, fancy= TRUE, fwidth = 0.6, fheight = 0.7, pretty = TRUE, FUN= text, xpd = TRUE, cex = 0.8)



> # C5.0

> # Install the C5.0 package

> install.packages("C50")

WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/

Installing package into ‘C:/Users/navya/Documents/R/win-library/3.6’

(as ‘lib’ is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.6/C50\_0.1.3.1.zip'

Content type 'application/zip' length 557666 bytes (544 KB)

downloaded 544 KB

package ‘C50’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\navya\AppData\Local\Temp\RtmpimjKBF\downloaded\_packages

> library("C50")

> # Put the predictors into â€˜xâ€™, the response into â€˜yâ€™

> names(dat)

[1] "age" "workclass"

[3] "demogweight" "education"

[5] "education.num" "marital.status"

[7] "occupation" "relationship"

[9] "race" "sex"

[11] "capital.gain" "capital.loss"

[13] "hours.per.week" "native.country"

[15] "income" "age.z"

[17] "education.num.z" "capital.gain.z"

[19] "capital.loss.z" "hours.per.week.z"

> x <- dat[,c(2,6, 9, 10, 16, 17, 18, 19, 20)]

> names(x)

[1] "workclass" "marital.status"

[3] "race" "sex"

[5] "age.z" "education.num.z"

[7] "capital.gain.z" "capital.loss.z"

[9] "hours.per.week.z"

> y <- dat$income

> # Run C5.0

> c50fit <- C5.0(x, y)

> summary(c50fit)

Call:

C5.0.default(x = x, y = y)

C5.0 [Release 2.07 GPL Edition] Sat Jun 13 16:01:26 2020

-------------------------------

Class specified by attribute `outcome'

Read 25000 cases (10 attributes) from undefined.data

Decision tree:

capital.gain.z > 0.7694287:

:...hours.per.week.z > -0.4396555: >50K. (973/10)

: hours.per.week.z <= -0.4396555:

: :...age.z <= -0.8479775: <=50K. (6/1)

: age.z > -0.8479775: >50K. (86/3)

capital.gain.z <= 0.7694287:

:...marital.status in {Divorced,Never-married,Separated,Widowed}:

:...capital.loss.z > 5.282196:

: :...capital.loss.z <= 5.646056: <=50K. (38/12)

: : capital.loss.z > 5.646056:

: : :...capital.loss.z <= 7.270963: >50K. (28)

: : capital.loss.z > 7.270963: <=50K. (8/1)

: capital.loss.z <= 5.282196:

: :...capital.gain.z > 0.4757047:

: :...capital.gain.z <= 0.494004: >50K. (19)

: : capital.gain.z > 0.494004: <=50K. (67/6)

: capital.gain.z <= 0.4757047:

: :...education.num.z <= 0.7503064: <=50K. (10336/211)

: education.num.z > 0.7503064:

: :...education.num.z <= 1.532462:

: :...hours.per.week.z <= 0.2107898: <=50K. (1598/109)

: : hours.per.week.z > 0.2107898:

: : :...sex = Female: <=50K. (314/49)

: : sex = Male:

: : :...education.num.z <= 1.141384: <=50K. (328/83)

: : education.num.z > 1.141384:

: : :...age.z > 1.197644: <=50K. (14/2)

: : age.z <= 1.197644:

: : :...age.z <= -0.4096299: <=50K. (15/1)

: : age.z > -0.4096299: >50K. (50/14)

: education.num.z > 1.532462:

: :...age.z <= -0.4826879: <=50K. (57/7)

: age.z > -0.4826879:

: :...age.z > 1.124586: <=50K. (22/3)

: age.z <= 1.124586:

: :...marital.status in {Never-married,

: : Widowed}: >50K. (50/13)

: marital.status = Separated:

: :...sex = Female: <=50K. (6/2)

: : sex = Male: >50K. (7/1)

: marital.status = Divorced:

: :...education.num.z <= 1.92354:

: :...sex = Female: <=50K. (8)

: : sex = Male: >50K. (10/4)

: education.num.z > 1.92354:

: :...hours.per.week.z <= 1.755597: >50K. (9/2)

: hours.per.week.z > 1.755597: <=50K. (2)

marital.status = Married:

:...capital.loss.z > 4.175664:

:...capital.loss.z <= 4.711484: >50K. (439/10)

: capital.loss.z > 4.711484:

: :...capital.loss.z <= 5.175032: <=50K. (48)

: capital.loss.z > 5.175032:

: :...education.num.z > 0.7503064: >50K. (47/1)

: education.num.z <= 0.7503064:

: :...education.num.z <= -0.8140052: <=50K. (7)

: education.num.z > -0.8140052:

: :...capital.loss.z > 5.803064: <=50K. (7)

: capital.loss.z <= 5.803064:

: :...capital.loss.z > 5.745743: >50K. (9)

: capital.loss.z <= 5.745743:

: :...age.z <= 0.613181: <=50K. (10)

: age.z > 0.613181: >50K. (13/4)

capital.loss.z <= 4.175664:

:...capital.gain.z > 0.5241912:

:...capital.gain.z <= 0.7118593: >50K. (77)

: capital.gain.z > 0.7118593:

: :...capital.gain.z <= 0.7246822: >50K. (4)

: capital.gain.z > 0.7246822: <=50K. (4)

capital.gain.z <= 0.5241912:

:...education.num.z > 0.7503064:

:...capital.loss.z > 1.342044: <=50K. (25/5)

: capital.loss.z <= 1.342044:

: :...capital.gain.z > 0.2690694:

: :...capital.gain.z <= 0.4023739: <=50K. (25)

: : capital.gain.z > 0.4023739:

: : :...capital.gain.z <= 0.4444489: >50K. (12/1)

: : capital.gain.z > 0.4444489: <=50K. (17)

: capital.gain.z <= 0.2690694:

: :...hours.per.week.z <= -0.6835725:

: :...sex = Female: >50K. (69/28)

: : sex = Male:

: : :...education.num.z <= 1.532462: <=50K. (140/31)

: : education.num.z > 1.532462: >50K. (31/12)

: hours.per.week.z > -0.6835725:

: :...age.z <= -0.4096299:

: :...age.z <= -0.9940934: <=50K. (50/14)

: : age.z > -0.9940934: [S1]

: age.z > -0.4096299:

: :...race in {Black,White}: >50K. (1747/507)

: race = Other: <=50K. (6/1)

: race = Amer-Indian-Eskimo:

: :...hours.per.week.z <= 0.2920955: <=50K. (4)

: : hours.per.week.z > 0.2920955: >50K. (3)

: race = Asian-Pac-Islander:

: :...sex = Male: >50K. (95/41)

: sex = Female:

: :...age.z <= 0.1748335: >50K. (6/1)

: age.z > 0.1748335: <=50K. (15/2)

education.num.z <= 0.7503064:

:...capital.loss.z > 3.68221: <=50K. (92)

capital.loss.z <= 3.68221:

:...capital.gain.z > 0.4444489: <=50K. (47)

capital.gain.z <= 0.4444489:

:...capital.gain.z > 0.4023739:

:...hours.per.week.z <= 0.9425408: >50K. (41/5)

: hours.per.week.z > 0.9425408: <=50K. (4/1)

capital.gain.z <= 0.4023739:

:...capital.gain.z > 0.2690694: <=50K. (131)

capital.gain.z <= 0.2690694:

:...capital.gain.z > 0.2543766: >50K. (50/2)

capital.gain.z <= 0.2543766:

:...capital.gain.z > -0.09184103: <=50K. (118)

capital.gain.z <= -0.09184103:

:...hours.per.week.z <= -0.5209612: <=50K. (736/92)

hours.per.week.z > -0.5209612:

:...age.z <= -0.2635141: <=50K. (2178/408)

age.z > -0.2635141: [S2]

SubTree [S1]

race in {Asian-Pac-Islander,Other}: <=50K. (26/10)

race in {Amer-Indian-Eskimo,White}: >50K. (377/162)

race = Black:

:...education.num.z <= 1.141384: >50K. (19/8)

education.num.z > 1.141384: <=50K. (2)

SubTree [S2]

race in {Amer-Indian-Eskimo,Other}: <=50K. (55/7)

race in {Asian-Pac-Islander,Black,White}:

:...education.num.z <= -0.4229273: <=50K. (2659/777)

education.num.z > -0.4229273:

:...hours.per.week.z > 0.2107898:

:...workclass in {?,Gov,Never-worked,Private}: >50K. (410.8/181.4)

: workclass = Without-pay: <=50K. (1)

: workclass = Self:

: :...education.num.z <= 0.3592285: <=50K. (175.2/74.5)

: education.num.z > 0.3592285:

: :...hours.per.week.z <= 1.349069: >50K. (11/2)

: hours.per.week.z > 1.349069: <=50K. (4)

hours.per.week.z <= 0.2107898:

:...workclass in {?,Gov,Never-worked,Without-pay}: <=50K. (190/93.5)

workclass = Self:

:...sex = Female: >50K. (8.4/2.3)

: sex = Male: <=50K. (85.6/33.6)

workclass = Private:

:...age.z > 0.613181:

:...age.z <= 1.70905: >50K. (206.5/94.8)

: age.z > 1.70905: <=50K. (20.7/3.7)

age.z <= 0.613181:

:...sex = Male: <=50K. (337.4/118.7)

sex = Female:

:...education.num.z <= -0.03184938: <=50K. (41.7/17)

education.num.z > -0.03184938: >50K. (11.7/2)

Evaluation on training data (25000 cases):

Decision Tree

----------------

Size Errors

78 3286(13.1%) <<

(a) (b) <-classified as

---- ----

17902 1114 (a): class <=50K.

2172 3812 (b): class >50K.

Attribute usage:

100.00% capital.gain.z

95.74% marital.status

95.74% capital.loss.z

92.81% education.num.z

52.65% hours.per.week.z

36.44% age.z

26.07% race

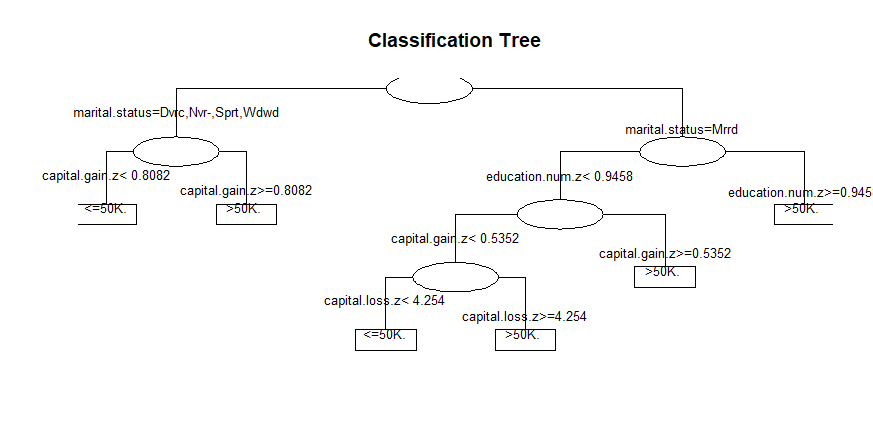
6.43% sex

5.90% workclass

Time: 0.3 secs

> #write file for problem 2

> write.csv(dat, file="C:\\Users\\navya\\Documents\\CLEM3TrainingP2.csv")



The decision tree is most useful predictive machine learning technique, that is used for both classification and regression. That is the reason it is called CARTfit decision tree. This model representation is a binary tree representation. Cartfit decision tree has a parent named “Classification Tree” and marital status are the child nodes, which is right and left of the parent tree. Each node has their own child nodes which are education, capital gain etc., creating a Cart model involves selecting input variables and split points which follows greedy approach. Tree construction ends with a leaf node.

**Problem 2: Rattle Decision tree:**

> install.packages("rattle")

WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/

Installing package into ‘C:/Users/navya/Documents/R/win-library/3.6’

(as ‘lib’ is unspecified)

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.6/rattle\_5.4.0.zip'

Content type 'application/zip' length 5464776 bytes (5.2 MB)

downloaded 5.2 MB

package ‘rattle’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\navya\AppData\Local\Temp\RtmpEbftgz\downloaded\_packages

> library(rattle)

Loading required package: tibble

Loading required package: bitops

Rattle: A free graphical interface for data science with R.

Version 5.4.0 Copyright (c) 2006-2020 Togaware Pty Ltd.

Type 'rattle()' to shake, rattle, and roll your data.

> rattle()

Loading required package: RGtk2

> install.packages("RGtk2")

WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/

Installing package into ‘C:/Users/navya/Documents/R/win-library/3.6’

(as ‘lib’ is unspecified)

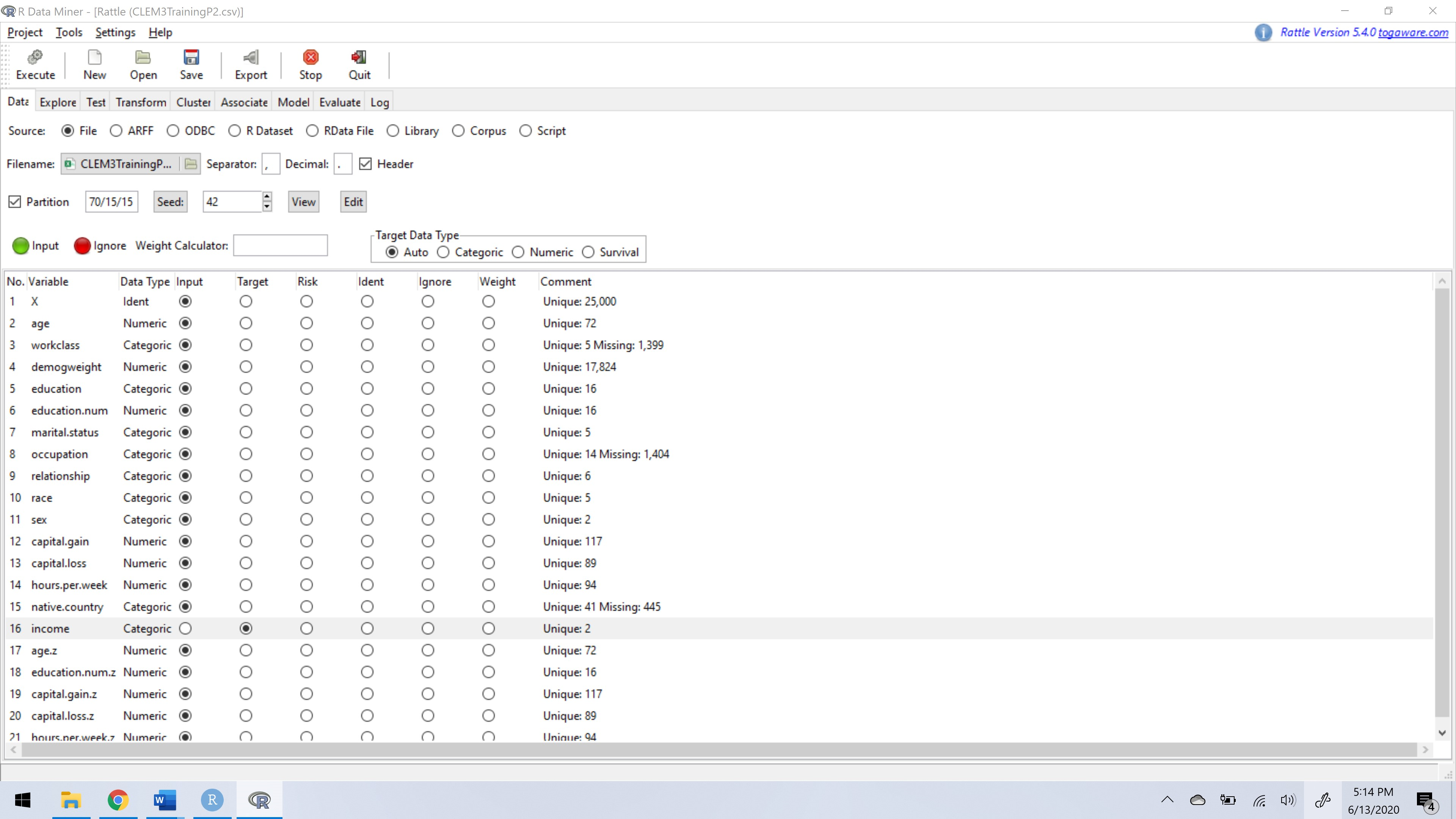
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.6/RGtk2\_2.20.36.zip'

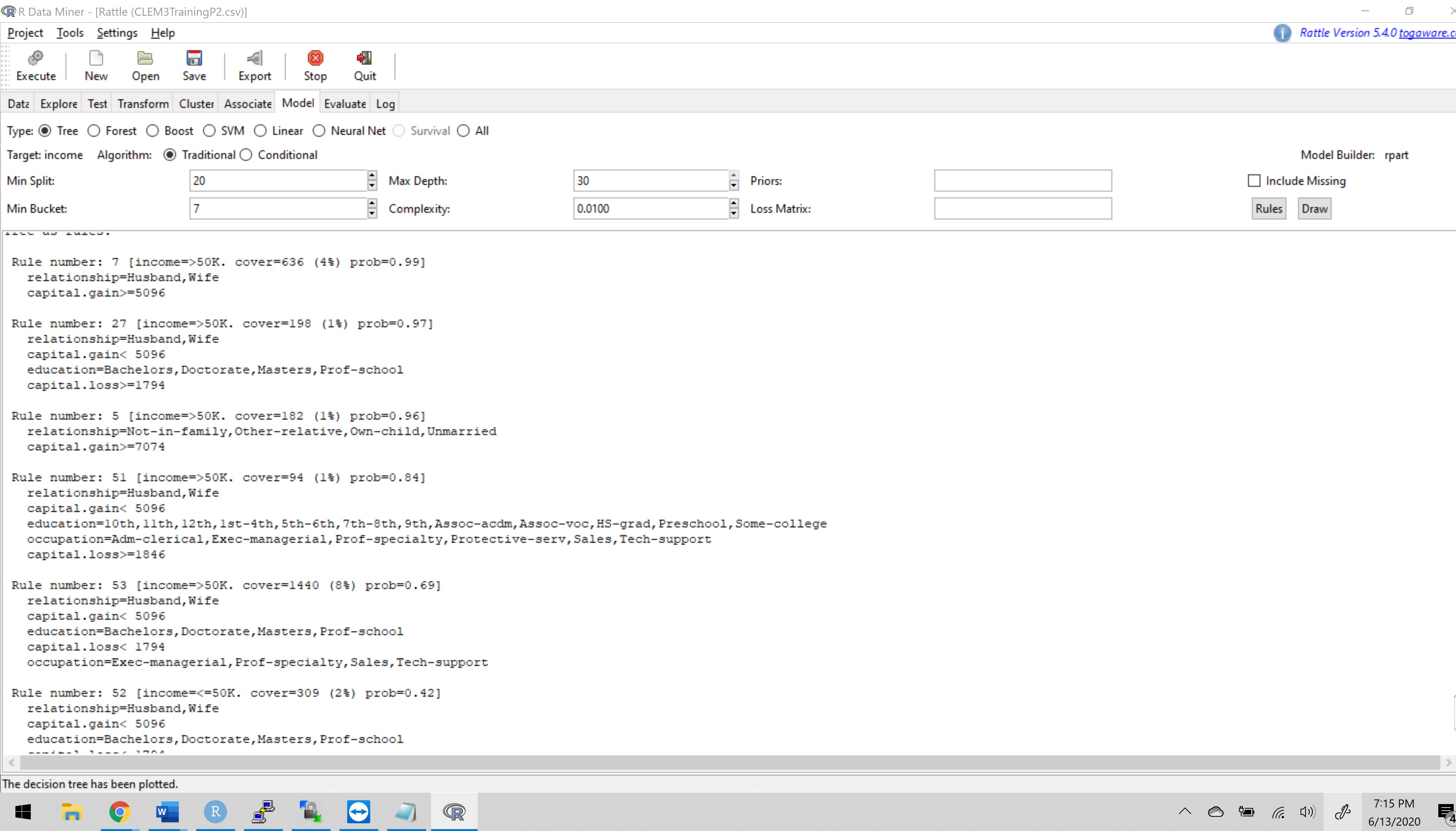
Content type 'application/zip' length 16558916 bytes (15.8 MB)

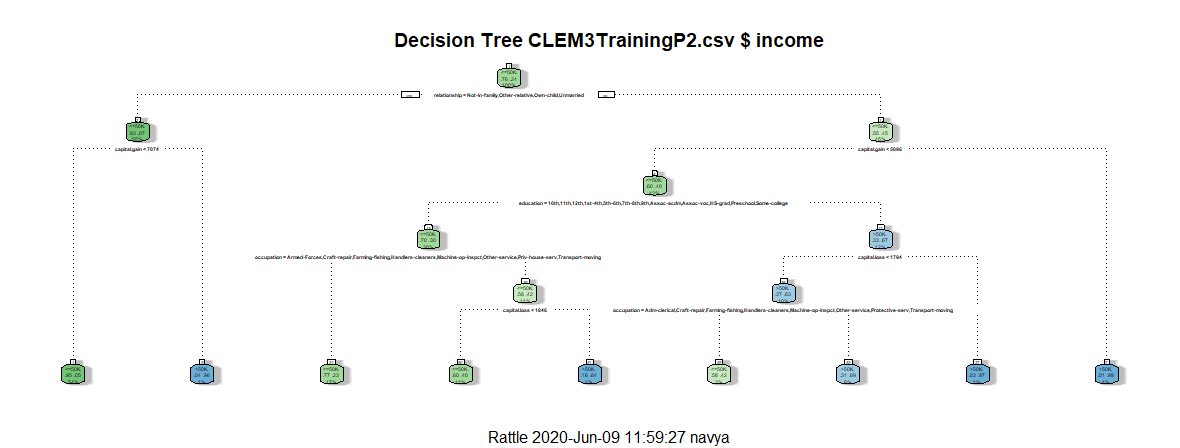
downloaded 15.8 MB

package ‘RGtk2’ successfully unpacked and MD5 sums checked

> library(RGtk2)







Rattle is a package written in R which Provides a graphical user interface to very many other R packages that provide functionality for data mining.

At the start, the root node has a depth over 3. The classification is done based on relationships which is <=50. The node proportion are <=50, then we move towards left child node whose depth is 2. we move until the stopping criteria is reached.

Difference between Cartfit and Rattle Decision Trees are:

1. Both trees follow a top-down greedy approach known as recursive binary splitting.
2. Rattle trees also provide counts of how many items are classified under that division.
3. Regression trees are used when dependent variable is continuous.
4. Classification tress are used when dependent variable is categorical.
5. Rattle trees provide the percentage of data at the node details.
6. This splitting process is continued until a user defined stopping criteria is reached.

**Problem 3: TreePlan Decision Tree:**



This tree consists of decision node, event nodes, and terminal nodes which are connected by branches. Each branch is surrounded by cells which consists of formulae. Both cartfit and rattle decision trees divide the independent variables into distinct and non-overlapping region until the stopping criteria is reached. All the three decision trees are top down approaches.