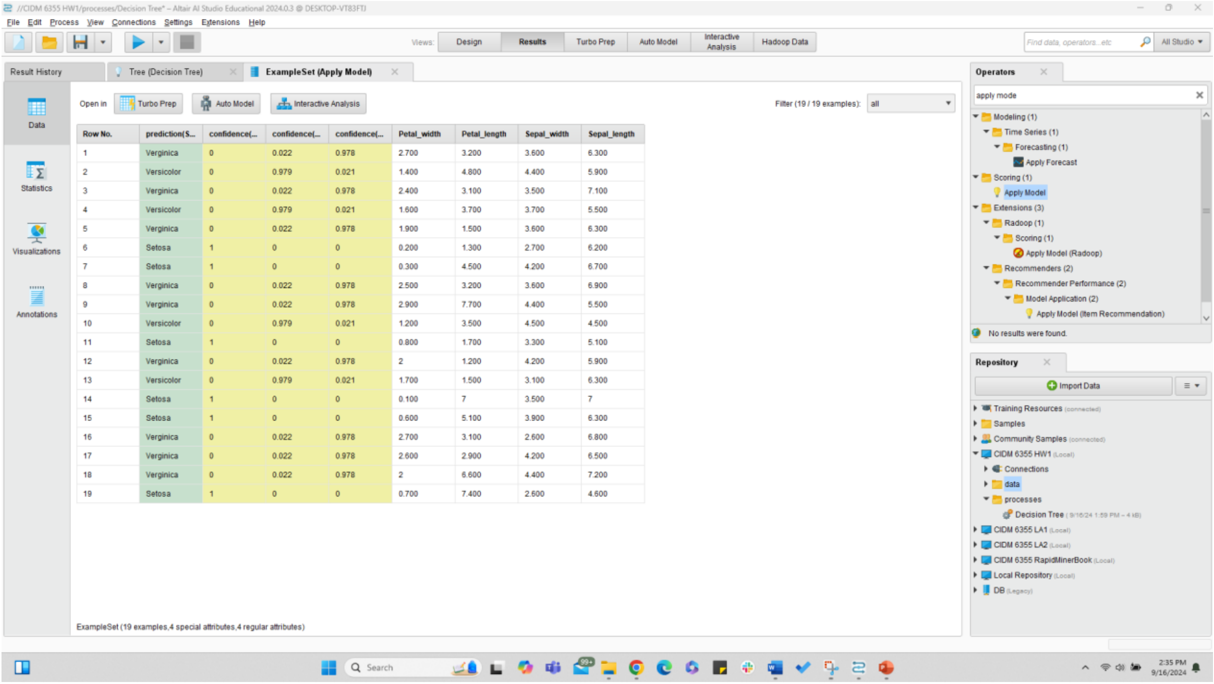
1) Screenshots in RapidMiner Lab (10 points)

- Screenshot 1: A screenshot of the decision tree graph with date and time at Step 1.8 (5 points)


//CIDM 6355 HW1/processes/Decision Tree* – Altair AI Studio Educational 2024.0.3 @ DESKTOP-VT83FTJ
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Interactive Analysis Hadoop Data
Result History
Tree (Decision Tree)
ExampleSet (Apply Model)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model Interactive Analysis
Operators
Repository
Import Data
2:35 PM
9/16/2024

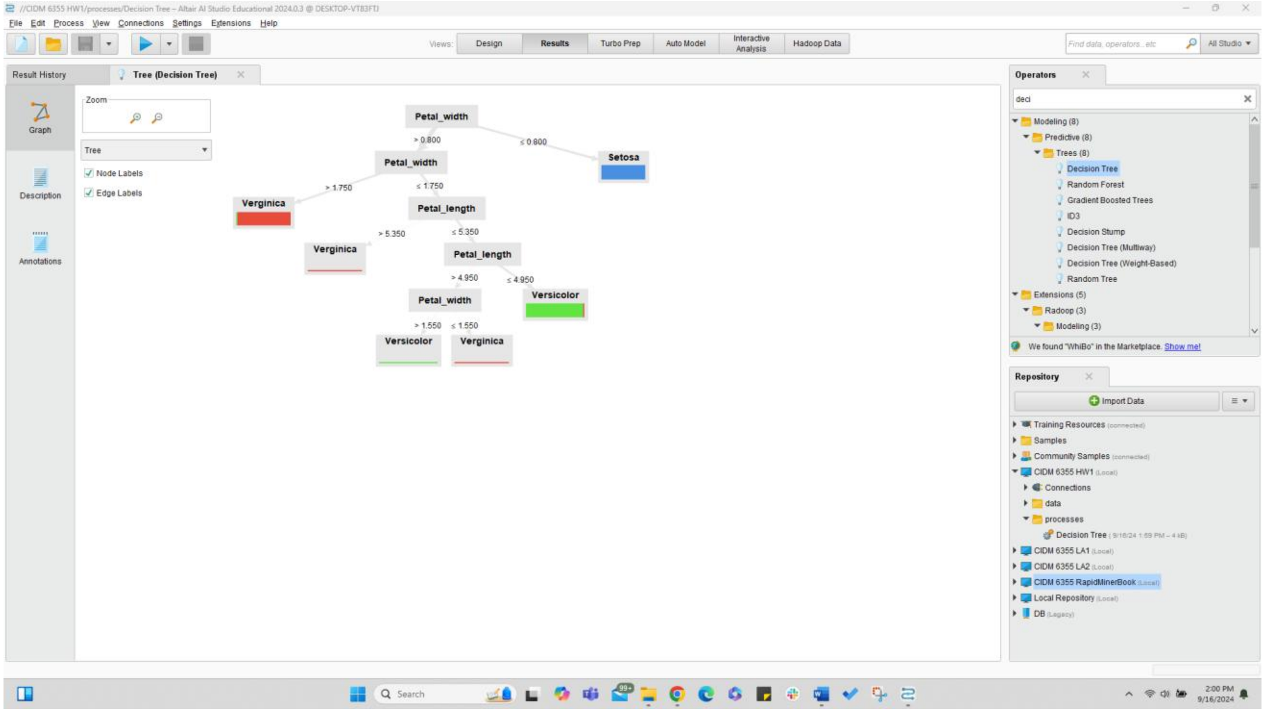
Filter (19 / 19 examples): all

Row No.	prediction(S...	confidence(...	confidence(...	confidence(...	Petal_width	Petal_length	Sepal_width	Sepal_length
1	Verginica	0	0.022	0.978	2.700	3.200	3.600	6.300
2	Versicolor	0	0.979	0.021	1.400	4.800	4.400	5.900
3	Verginica	0	0.022	0.978	2.400	3.100	3.500	7.100
4	Versicolor	0	0.979	0.021	1.600	3.700	3.700	5.500
5	Verginica	0	0.022	0.978	1.900	1.500	3.600	6.300
6	Setosa	1	0	0	0.200	1.300	2.700	6.200
7	Setosa	1	0	0	0.300	4.500	4.200	6.700
8	Verginica	0	0.022	0.978	2.500	3.200	3.600	6.900
9	Verginica	0	0.022	0.978	2.900	7.700	4.400	5.500
10	Versicolor	0	0.979	0.021	1.200	3.500	4.500	4.500
11	Setosa	1	0	0	0.800	1.700	3.300	5.100
12	Verginica	0	0.022	0.978	2	1.200	4.200	5.900
13	Versicolor	0	0.979	0.021	1.700	1.500	3.100	6.300
14	Setosa	1	0	0	0.100	7	3.500	7
15	Setosa	1	0	0	0.600	5.100	3.900	6.300
16	Verginica	0	0.022	0.978	2.700	3.100	2.600	6.800
17	Verginica	0	0.022	0.978	2.600	2.900	4.200	6.500
18	Verginica	0	0.022	0.978	2	6.600	4.400	7.200
19	Setosa	1	0	0	0.700	7.400	2.600	4.600

ExampleSet (19 examples,4 special attributes,4 regular attributes)

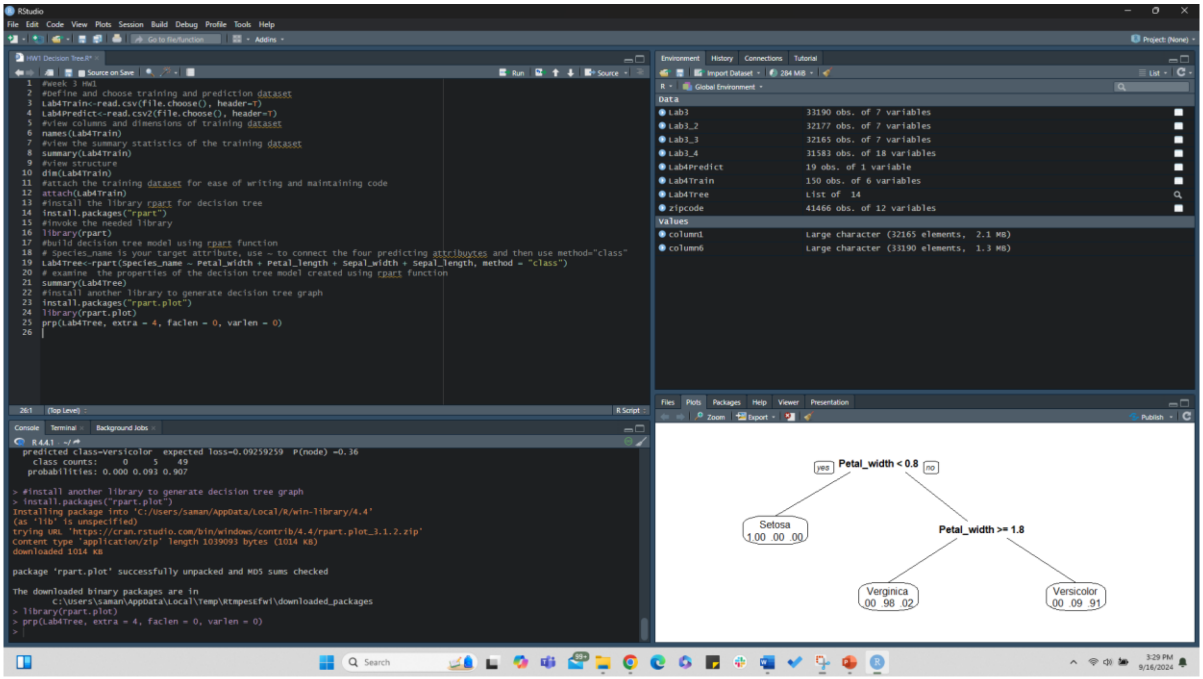
- Screenshot 2: A screenshot of prediction results for the 19 observations with date and time in Step 2.2 (5 points)

2) Deliverables in R Lab (50 points)



- **Deliverable R1**: take a screenshot of your decision tree model with date and time (5 points).

- **Deliverable R2**: take a screenshot of your decision tree graph with date and time and briefly describe it.

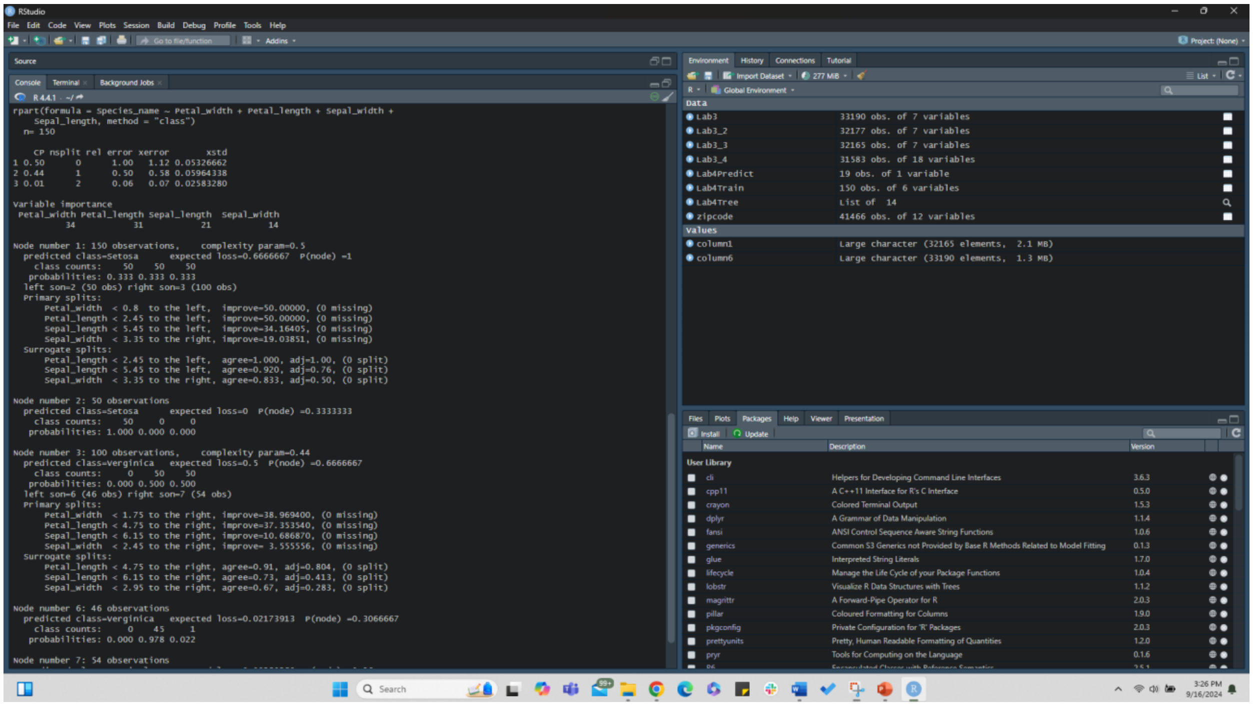


```r
#Week 3 HW1
#Define and choose training and prediction dataset
Lab4Train<-read.csv(file.choose(), header=T)
Lab4Predict<-read.csv2(file.choose(), header=T)
#view columns and dimensions of training dataset
names(Lab4Train)
#view the summary statistics of the training dataset
summary(Lab4Train)
#view structure
dim(Lab4Train)
#attach the training dataset for ease of writing and maintaining code
attach(Lab4Train)
#install the library rpart for decision tree
install.packages("rpart")
#invoke the needed library
library(rpart)
#build decision tree model using rpart function
# Species_name is your target attribute, use ~ to connect the four predicting attribuytes and then use method="class"
Lab4Tree<-rpart(Species_name ~ Petal_width + Petal_length + Sepal_width + Sepal_length, method = "class")
# examine  the properties of the decision tree model created using rpart function
summary(Lab4Tree)
#install another library to generate decision tree graph
install.packages("rpart.plot")
library(rpart.plot)
prp(Lab4Tree, extra = 4, faclen = 0, varlen = 0)
```

```
  predicted class=Versicolor  expected loss=0.09259259  P(node) =0.36
    class counts:     0     5    49
   probabilities: 0.000 0.093 0.907

> #install another library to generate decision tree graph
> install.packages("rpart.plot")
Installing package into 'C:/Users/saman/AppData/Local/R/win-library/4.4'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/4.4/rpart.plot_3.1.2.zip'
Content type 'application/zip' length 1039093 bytes (1014 KB)
downloaded 1014 KB

package 'rpart.plot' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\saman\AppData\Local\Temp\RtmpesEfwi\downloaded_packages
> library(rpart.plot)
> prp(Lab4Tree, extra = 4, faclen = 0, varlen = 0)
> 
```

```
Data
Lab3           33190 obs. of 7 variables
Lab3_2         32177 obs. of 7 variables
Lab3_3         32165 obs. of 7 variables
Lab3_4         31583 obs. of 18 variables
Lab4Predict    19 obs. of 1 variable
Lab4Train      150 obs. of 6 variables
Lab4Tree       List of  14
zipcode        41466 obs. of 12 variables
Values
column1        Large character (32165 elements,  2.1 MB)
column6        Large character (33190 elements,  1.3 MB)
```

Your description must include the root node, split nodes, and leaf nodes. (10 points: 5 points for your screenshot and 5 points for your description).



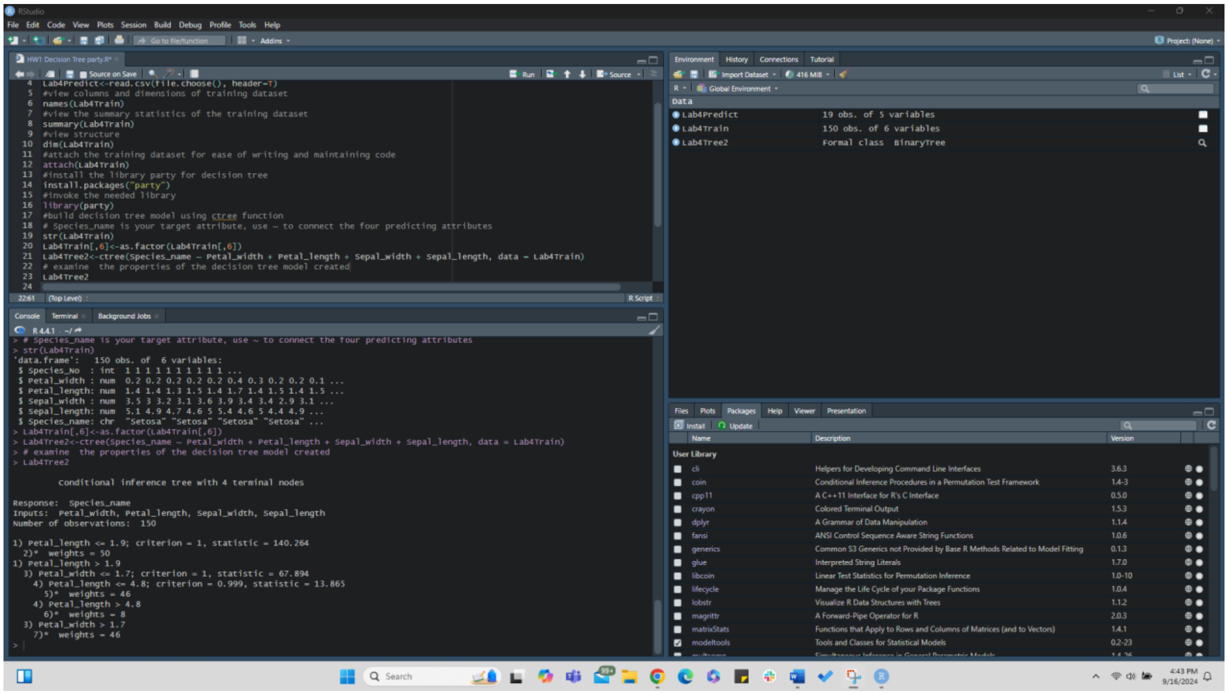
```
rpart(formula = Species_name ~ Petal_width + Petal_length + Sepal_width +
    Sepal_length, method = "class")
  n= 150

    CP nsplit rel error xerror       xstd
1 0.50      0      1.00   1.12 0.05326662
2 0.44      1      0.50   0.58 0.05964338
3 0.01      2      0.06   0.07 0.02583280

Variable importance
 Petal_width Petal_length Sepal_length  Sepal_width
          34           31           21           14

Node number 1: 150 observations,    complexity param=0.5
  predicted class=Setosa      expected loss=0.6666667  P(node) =1
    class counts:    50    50    50
   probabilities: 0.333 0.333 0.333
  left son=2 (50 obs) right son=3 (100 obs)
  Primary splits:
      Petal_width  < 0.8  to the left,  improve=50.00000, (0 missing)
      Petal_length < 2.45 to the left,  improve=50.00000, (0 missing)
      Sepal_length < 5.45 to the left,  improve=34.16405, (0 missing)
      Sepal_width  < 3.35 to the right, improve=19.03851, (0 missing)
  Surrogate splits:
      Petal_length < 2.45 to the left,  agree=1.000, adj=1.00, (0 split)
      Sepal_length < 5.45 to the left,  agree=0.920, adj=0.76, (0 split)
      Sepal_width  < 3.35 to the right, agree=0.833, adj=0.50, (0 split)

Node number 2: 50 observations
  predicted class=Setosa      expected loss=0  P(node) =0.3333333
    class counts:    50     0     0
   probabilities: 1.000 0.000 0.000

Node number 3: 100 observations,    complexity param=0.44
  predicted class=Verginica   expected loss=0.5  P(node) =0.6666667
    class counts:     0    50    50
   probabilities: 0.000 0.500 0.500
  left son=6 (46 obs) right son=7 (54 obs)
  Primary splits:
      Petal_width  < 1.75 to the right, improve=38.969400, (0 missing)
      Petal_length < 4.75 to the right, improve=37.353540, (0 missing)
      Sepal_length < 6.15 to the right, improve=10.686870, (0 missing)
      Sepal_width  < 2.45 to the right, improve= 3.555556, (0 missing)
  Surrogate splits:
      Petal_length < 4.75 to the right, agree=0.91, adj=0.804, (0 split)
      Sepal_length < 6.15 to the right, agree=0.73, adj=0.413, (0 split)
      Sepal_width  < 2.95 to the right, agree=0.67, adj=0.283, (0 split)

Node number 6: 46 observations
  predicted class=Verginica   expected loss=0.02173913  P(node) =0.3066667
    class counts:     0    45     1
   probabilities: 0.000 0.978 0.022

Node number 7: 54 observations
```

```
Data
Lab3           33190 obs. of 7 variables
Lab3_2         32177 obs. of 7 variables
Lab3_3         32165 obs. of 7 variables
Lab3_4         31583 obs. of 18 variables
Lab4Predict    19 obs. of 1 variable
Lab4Train      150 obs. of 6 variables
Lab4Tree       List of  14
zipcode        41466 obs. of 12 variables
Values
column1        Large character (32165 elements,  2.1 MB)
column6        Large character (33190 elements,  1.3 MB)
```

	Name	Description	Version
	User Library		
□	cli	Helpers for Developing Command Line Interfaces	3.6.3
□	cpp11	A C++11 Interface for R's C Interface	0.5.0
□	crayon	Colored Terminal Output	1.5.3
□	dplyr	A Grammar of Data Manipulation	1.1.4
□	fansi	ANSI Control Sequence Aware String Functions	1.0.6
□	generics	Common S3 Generics not Provided by Base R Methods Related to Model Fitting	0.1.3
□	glue	Interpreted String Literals	1.7.0
□	lifecycle	Manage the Life Cycle of your Package Functions	1.0.4
□	lobstr	Visualize R Data Structures with Trees	1.1.2
□	magrittr	A Forward-Pipe Operator for R	2.0.3
□	pillar	Coloured Formatting for Columns	1.9.0
□	pkgconfig	Private Configuration for 'R' Packages	2.0.3
□	prettyunits	Pretty, Human Readable Formatting of Quantities	1.2.0
□	pryr	Tools for Computing on the Language	0.1.6

The root note of this decision tree is Petal_width < 0.8 at the top of the tree. There is only one other split node (not including the root node), which is Petal_width >= 1.8. The leaf nodes shown circled and are Setosa, Verginica, and Versicolor.

- **Deliverable R3**: after you apply the decision tree model to your prediction dataset, take a screenshot of the prediction result with date and time and briefly describe how the result help you determine the predicted class of each case. (10 points: 5 points for your screenshot and 5 points for your description).
Deliverable R3 description: For each row, the results show a confidence percentage of each species name. Where the highest confidence percentage is, is where I determine the prediction. For example, if we look at row 4, the confidence percentages are as follows: Setosa:0, Verginica: 0.09259259, and Versicolor: 0.90740741. Since Versicolor has the highest confidence percentage, we can conclude that Versicolor is the species name.



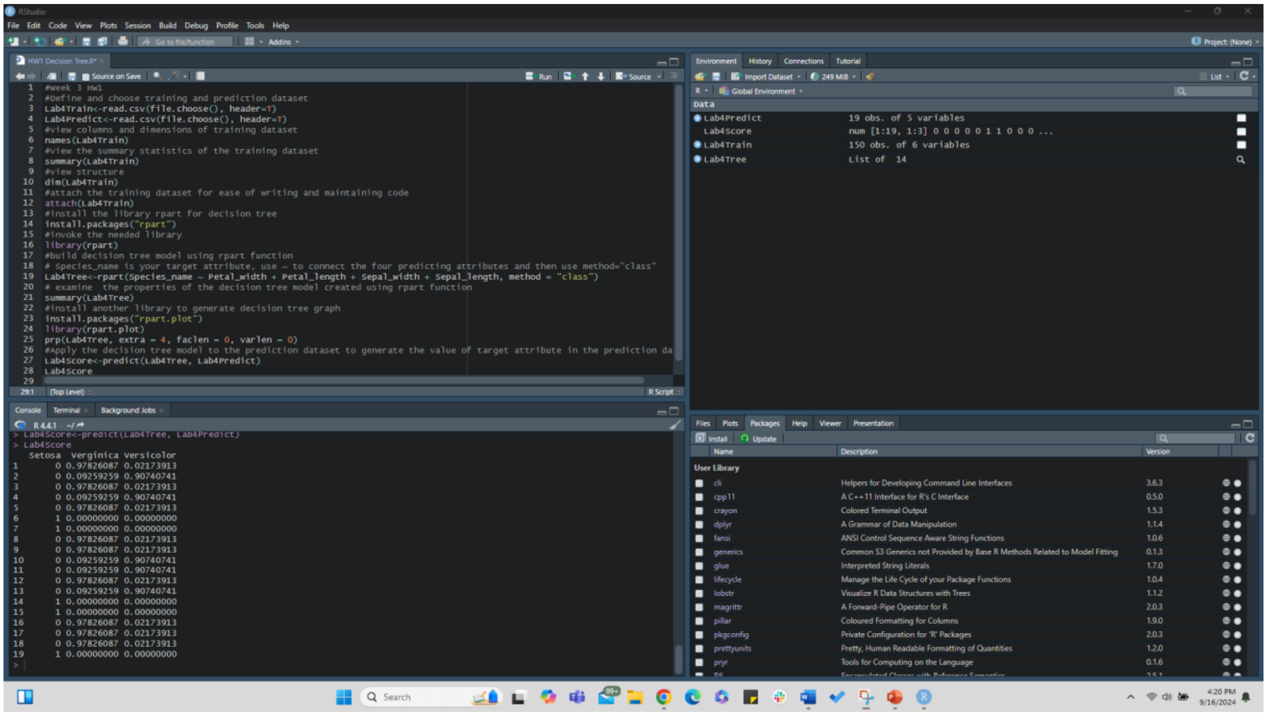
```
Lab4Predict<-read.csv(file.choose(), header=T)
#view columns and dimensions of training dataset
names(Lab4Train)
#view the summary statistics of the training dataset
summary(Lab4Train)
#view structure
dim(Lab4Train)
#attach the training dataset for ease of writing and maintaining code
attach(Lab4Train)
#install the library party for decision tree
install.packages("party")
#invoke the needed library
library(party)
#build decision tree model using ctree function
# Species_name is your target attribute, use ~ to connect the four predicting attributes
str(Lab4Train)
Lab4Train[,6]<-as.factor(Lab4Train[,6])
Lab4Tree2<-ctree(Species_name ~ Petal_width + Petal_length + Sepal_width + Sepal_length, data = Lab4Train)
# examine  the properties of the decision tree model created
Lab4Tree2
```

```
> # Species_name is your target attribute, use ~ to connect the four predicting attributes
> str(Lab4Train)
'data.frame':   150 obs. of  6 variables:
 $ Species_No  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Petal_width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Petal_length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Sepal_width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Sepal_length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Species_name: chr  "Setosa" "Setosa" "Setosa" "Setosa" ...
> Lab4Train[,6]<-as.factor(Lab4Train[,6])
> Lab4Tree2<-ctree(Species_name ~ Petal_width + Petal_length + Sepal_width + Sepal_length, data = Lab4Train)
> # examine  the properties of the decision tree model created
> Lab4Tree2

	 Conditional inference tree with 4 terminal nodes

Response:  Species_name 
Inputs:  Petal_width, Petal_length, Sepal_width, Sepal_length 
Number of observations:  150 

1) Petal_length <= 1.9; criterion = 1, statistic = 140.264
  2)*  weights = 50 
1) Petal_length > 1.9
  3) Petal_width <= 1.7; criterion = 1, statistic = 67.894
    4) Petal_length <= 4.8; criterion = 0.999, statistic = 13.865
      5)*  weights = 46 
    4) Petal_length > 4.8
      6)*  weights = 8 
  3) Petal_width > 1.7
    7)*  weights = 46 
> |
```

Data	
Lab4Predict	19 obs. of 5 variables
Lab4Train	150 obs. of 6 variables
Lab4Tree2	Formal class BinaryTree

	Name	Description	Version
	User Library		
□	cli	Helpers for Developing Command Line Interfaces	3.6.3
□	coin	Conditional Inference Procedures in a Permutation Test Framework	1.4-3
□	cpp11	A C++11 Interface for R's C Interface	0.5.0
□	crayon	Colored Terminal Output	1.5.3
□	dplyr	A Grammar of Data Manipulation	1.1.4
□	fansi	ANSI Control Sequence Aware String Functions	1.0.6
□	generics	Common S3 Generics not Provided by Base R Methods Related to Model Fitting	0.1.3
□	glue	Interpreted String Literals	1.7.0
□	libcoin	Linear Test Statistics for Permutation Inference	1.0-10
□	lifecycle	Manage the Life Cycle of your Package Functions	1.0.4
□	lobstr	Visualize R Data Structures with Trees	1.1.2
□	magrittr	A Forward-Pipe Operator for R	2.0.3
□	matrixStats	Functions that Apply to Rows and Columns of Matrices (and to Vectors)	1.4.1
☑	modeltools	Tools and Classes for Statistical Models	0.2-23

- **Deliverable R4**: take a screenshot of your decision tree model with date and time. Try to use the resources provided to understand its output (5 points).



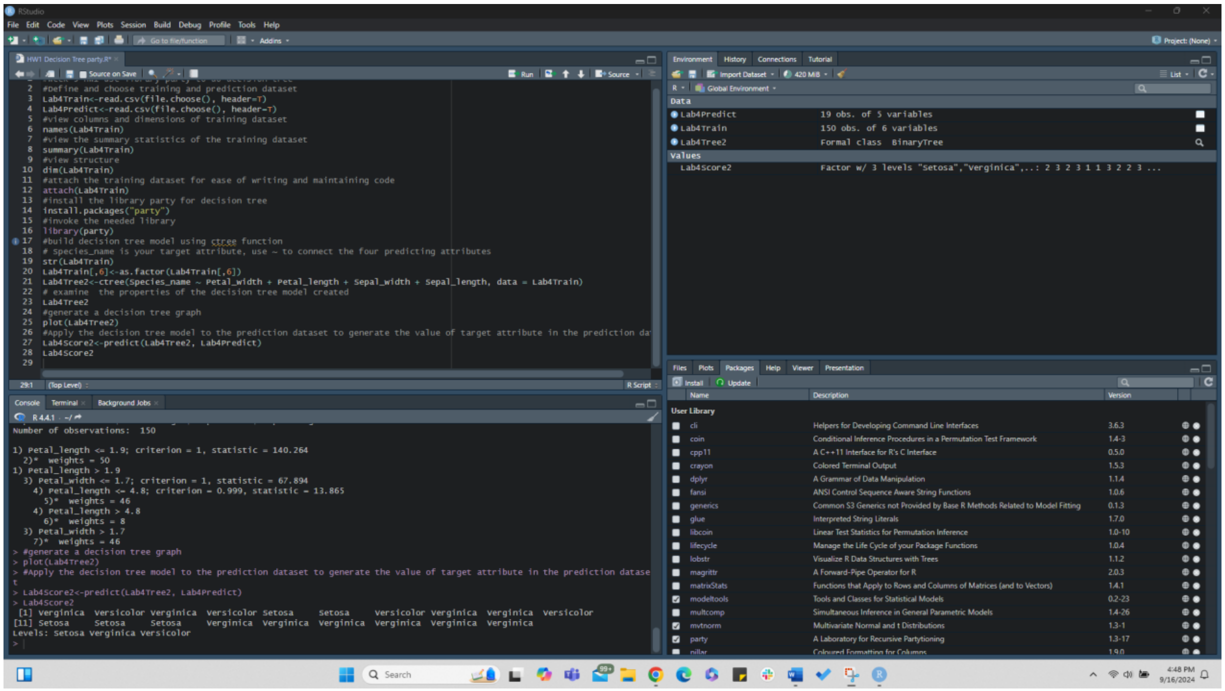
```
#Week 3 HW1
#Define and choose training and prediction dataset
Lab4Train<-read.csv(file.choose(), header=T)
Lab4Predict<-read.csv(file.choose(), header=T)
#view columns and dimensions of training dataset
names(Lab4Train)
#view the summary statistics of the training dataset
summary(Lab4Train)
#view structure
dim(Lab4Train)
#attach the training dataset for ease of writing and maintaining code
attach(Lab4Train)
#install the library rpart for decision tree
install.packages("rpart")
#invoke the needed library
library(rpart)
#build decision tree model using rpart function
# Species_name is your target attribute, use ~ to connect the four predicting attributes and then use method="class"
Lab4Tree<-rpart(Species_name ~ Petal_width + Petal_length + Sepal_width + Sepal_length, method = "class")
# examine  the properties of the decision tree model created using rpart function
summary(Lab4Tree)
#install another library to generate decision tree graph
install.packages("rpart.plot")
library(rpart.plot)
prp(Lab4Tree, extra = 4, faclen = 0, varlen = 0)
#Apply the decision tree model to the prediction dataset to generate the value of target attribute in the prediction da
Lab4Score<-predict(Lab4Tree, Lab4Predict)
Lab4Score
```

```
> Lab4Score<-predict(Lab4Tree, Lab4Predict)
> Lab4Score
   Setosa  Verginica Versicolor
1       0 0.97826087 0.02173913
2       0 0.09259259 0.90740741
3       0 0.97826087 0.02173913
4       0 0.09259259 0.90740741
5       0 0.97826087 0.02173913
6       1 0.00000000 0.00000000
7       1 0.00000000 0.00000000
8       0 0.97826087 0.02173913
9       0 0.97826087 0.02173913
10      0 0.09259259 0.90740741
11      0 0.09259259 0.90740741
12      0 0.97826087 0.02173913
13      0 0.09259259 0.90740741
14      1 0.00000000 0.00000000
15      1 0.00000000 0.00000000
16      0 0.97826087 0.02173913
17      0 0.97826087 0.02173913
18      0 0.97826087 0.02173913
19      1 0.00000000 0.00000000
> |
```

Data	
Lab4Predict	19 obs. of 5 variables
Lab4Score	num [1:19, 1:3] 0 0 0 0 0 1 1 0 0 0 ...
Lab4Train	150 obs. of 6 variables
Lab4Tree	List of 14

	Name	Description	Version
	User Library		
□	cli	Helpers for Developing Command Line Interfaces	3.6.3
□	cpp11	A C++11 Interface for R's C Interface	0.5.0
□	crayon	Colored Terminal Output	1.5.3
□	dplyr	A Grammar of Data Manipulation	1.1.4
□	fansi	ANSI Control Sequence Aware String Functions	1.0.6
□	generics	Common S3 Generics not Provided by Base R Methods Related to Model Fitting	0.1.3
□	glue	Interpreted String Literals	1.7.0
□	lifecycle	Manage the Life Cycle of your Package Functions	1.0.4
□	lobstr	Visualize R Data Structures with Trees	1.1.2
□	magrittr	A Forward-Pipe Operator for R	2.0.3
□	pillar	Coloured Formatting for Columns	1.9.0
□	pkgconfig	Private Configuration for 'R' Packages	2.0.3
□	prettyunits	Pretty, Human Readable Formatting of Quantities	1.2.0
□	pryr	Tools for Computing on the Language	0.1.6

- **Deliverable R5**: take a screenshot of your decision tree graph with date and time (5 points).



```
#Define and choose training and prediction dataset
Lab4Train<-read.csv(file.choose(), header=T)
Lab4Predict<-read.csv(file.choose(), header=T)
#view columns and dimensions of training dataset
names(Lab4Train)
#view the summary statistics of the training dataset
summary(Lab4Train)
#view structure
dim(Lab4Train)
#attach the training dataset for ease of writing and maintaining code
attach(Lab4Train)
#install the library party for decision tree
install.packages("party")
#invoke the needed library
library(party)
#build decision tree model using ctree function
# Species_name is your target attribute, use ~ to connect the four predicting attributes
str(Lab4Train)
Lab4Train[,6]<-as.factor(Lab4Train[,6])
Lab4Tree2<-ctree(Species_name ~ Petal_width + Petal_length + Sepal_width + Sepal_length, data = Lab4Train)
# examine  the properties of the decision tree model created
Lab4Tree2
#generate a decision tree graph
plot(Lab4Tree2)
#Apply the decision tree model to the prediction dataset to generate the value of target attribute in the prediction da
Lab4Score2<-predict(Lab4Tree2, Lab4Predict)
Lab4Score2
```

```
Number of observations:  150

1) Petal_length <= 1.9; criterion = 1, statistic = 140.264
  2)*  weights = 50
1) Petal_length > 1.9
  3) Petal_width <= 1.7; criterion = 1, statistic = 67.894
    4) Petal_length <= 4.8; criterion = 0.999, statistic = 13.865
      5)*  weights = 46
    4) Petal_length > 4.8
      6)*  weights = 8
  3) Petal_width > 1.7
    7)*  weights = 46
> #generate a decision tree graph
> plot(Lab4Tree2)
> #Apply the decision tree model to the prediction dataset to generate the value of target attribute in the prediction datase
t
> Lab4Score2<-predict(Lab4Tree2, Lab4Predict)
> Lab4Score2
 [1] Verginica  Versicolor Verginica  Versicolor Setosa     Setosa     Versicolor Verginica  Verginica  Versicolor
[11] Setosa     Setosa     Setosa     Verginica  Verginica  Verginica  Verginica  Verginica  Verginica 
Levels: Setosa Verginica Versicolor
> 
```

```
Data
Lab4Predict          19 obs. of 5 variables
Lab4Train            150 obs. of 6 variables
Lab4Tree2            Formal class  BinaryTree
Values
Lab4Score2           Factor w/ 3 levels "Setosa","Verginica",..: 2 3 2 3 1 1 3 2 2 3 ...
```

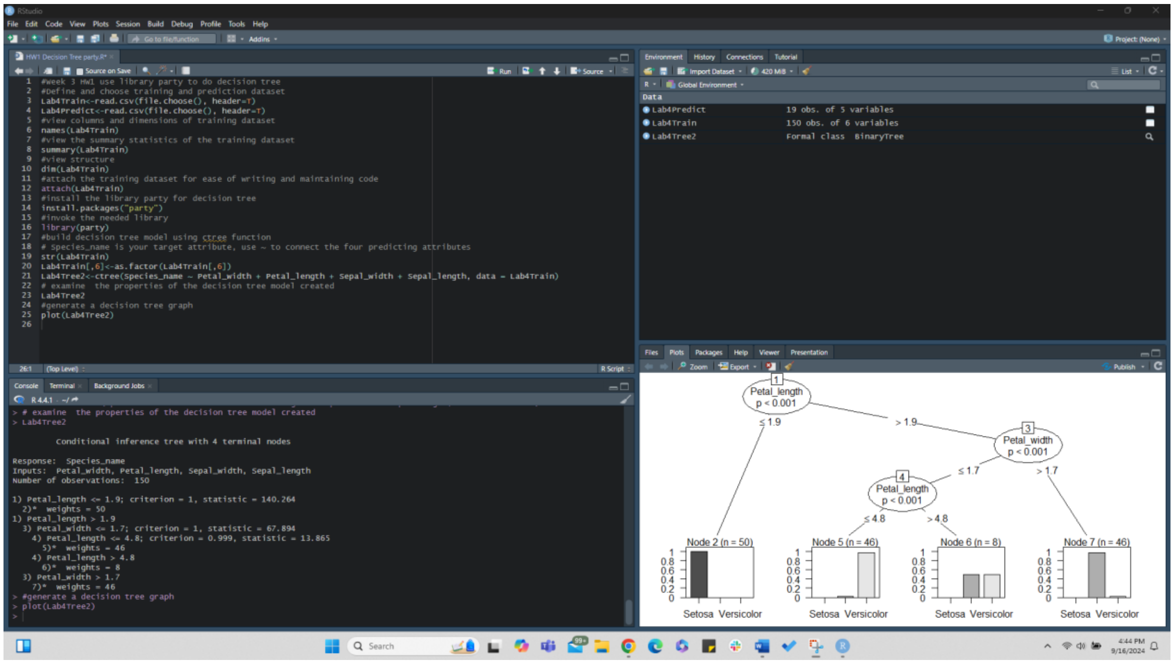
User Library

	Name	Description	Version
□	cli	Helpers for Developing Command Line Interfaces	3.6.3
□	coin	Conditional Inference Procedures in a Permutation Test Framework	1.4-3
□	cpp11	A C++11 Interface for R's C Interface	0.5.0
□	crayon	Colored Terminal Output	1.5.3
□	dplyr	A Grammar of Data Manipulation	1.1.4
□	fansi	ANSI Control Sequence Aware String Functions	1.0.6
□	generics	Common S3 Generics not Provided by Base R Methods Related to Model Fitting	0.1.3
□	glue	Interpreted String Literals	1.7.0
□	libcoin	Linear Test Statistics for Permutation Inference	1.0-10
□	lifecycle	Manage the Life Cycle of your Package Functions	1.0.4
□	lobstr	Visualize R Data Structures with Trees	1.1.2
□	magrittr	A Forward-Pipe Operator for R	2.0.3
□	matrixStats	Functions that Apply to Rows and Columns of Matrices (and to Vectors)	1.4.1
☑	modeltools	Tools and Classes for Statistical Models	0.2-23
□	multcomp	Simultaneous Inference in General Parametric Models	1.4-26
☑	mvtnorm	Multivariate Normal and t Distributions	1.3-1
☑	party	A Laboratory for Recursive Partytioning	1.3-17
□	pillar	Coloured Formatting for Columns	1.9.0

4:48 PM
9/16/2024

- **Deliverable R6**: after you apply the decision tree model to your prediction dataset, and take a screenshot of the prediction result with date and time (5 points).

- **Deliverable R7:** Choose one of the two decision tree models generated in R and compare it with the decision tree model generated in RapidMiner. Identify and discuss at least three differences between the



```
#Week 3 HW1 use library party to do decision tree
#Define and choose training and prediction dataset
Lab4Train<-read.csv(file.choose(), header=T)
Lab4Predict<-read.csv(file.choose(), header=T)
#view columns and dimensions of training dataset
names(Lab4Train)
#view the summary statistics of the training dataset
summary(Lab4Train)
#view structure
dim(Lab4Train)
#attach the training dataset for ease of writing and maintaining code
attach(Lab4Train)
#install the library party for decision tree
install.packages("party")
#invoke the needed library
library(party)
#build decision tree model using ctree function
# Species_name is your target attribute, use ~ to connect the four predicting attributes
str(Lab4Train)
Lab4Train[,6]<-as.factor(Lab4Train[,6])
Lab4Tree2<-ctree(Species_name ~ Petal_width + Petal_length + Sepal_width + Sepal_length, data = Lab4Train)
# examine  the properties of the decision tree model created
Lab4Tree2
#generate a decision tree graph
plot(Lab4Tree2)
```

```
> # examine  the properties of the decision tree model created
> Lab4Tree2

         Conditional inference tree with 4 terminal nodes

Response:  Species_name 
Inputs:  Petal_width, Petal_length, Sepal_width, Sepal_length 
Number of observations:  150 

1) Petal_length <= 1.9; criterion = 1, statistic = 140.264
  2)*  weights = 50 
1) Petal_length > 1.9
  3) Petal_width <= 1.7; criterion = 1, statistic = 67.894
    4) Petal_length <= 4.8; criterion = 0.999, statistic = 13.865
      5)*  weights = 46 
    4) Petal_length > 4.8
      6)*  weights = 8 
  3) Petal_width > 1.7
    7)*  weights = 46 
> #generate a decision tree graph
> plot(Lab4Tree2)
> 
```

```
Data
Lab4Predict          19 obs. of 5 variables
Lab4Train            150 obs. of 6 variables
Lab4Tree2            Formal class  BinaryTree
```

4:44 PM
9/16/2024

two models. When discussing each difference, please include both R and RM. For example, "R does …, but RM does not …." (10 points).

I am going to be comparing the RM decision tree and R library party decision tree. One of the first major differences I noticed was the structure of the decision tree itself. The decision tree in RM had more split nodes and leaf nodes than the one in R. Another difference between the two was the predictions. RM found the following predictions (in count): Setosa-6, Versicolor-4, and Virginica-9. While R found (in count): Setosa-5, Versicolor-4, and Virginica-10. We could say that there is not a major difference between these numbers, so either model still works. Lastly, back to the decision tree, the two models used different root notes. RM used Petal_width while R used Petal_length. This could have led to the minor differences in the prediction results.