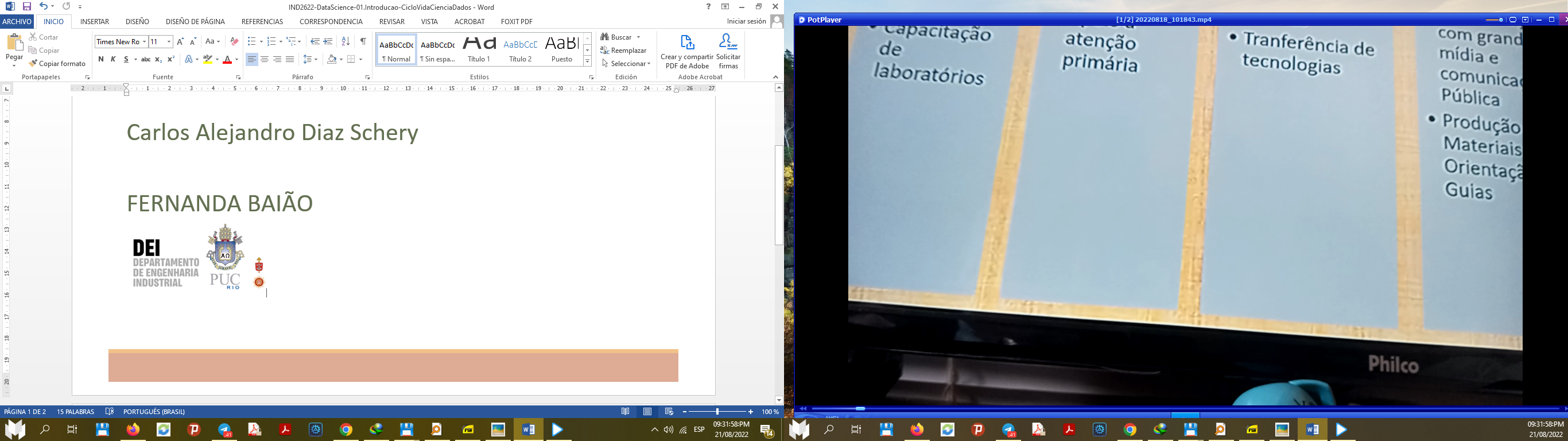
Ciência de Dados para Processos de Negócio

Yiselis Rodriguez Vignon

e

Carlos Alejandro Diaz Schery

Tarefa 2: Practical task 2- (Sala de aula invertida)

FERNANDA BAIÃO

**Title and Authors**

Yiselis Rodriguez Vignon e Carlos Alejandro Diaz Schery

**Pre-processing**

1. Transform the categorical features into binary variables (dummy variables). Dummy variables are accepted by almost any classifier. (1 point)

**Parcel used**

library(ISLR)

require(tree)

library(rpart)

library(rpart.plot)

install.packages("mccr")

library(mccr)

require(mccr)

ds <- read.csv("H:/CLASE CARLOS MAESTRIA 2022/IND 2622 FERNANDA CIÊNCIA DE DADOS PARA 5ªf., 13-16 h/Preprocesamiento y posprocesamiento de datos/para hoy/TAREA 2/tarea renato/ds\_salaries.csv")

**Transform-binary variables**

* ds$company\_size <- ifelse(ds$company\_size=="S", 0,

ifelse(ds$company\_size=="M", 1, 2))

* ds$experience\_level <- ifelse(ds$experience\_level=="EN", 1,

ifelse(ds$experience\_level=="MI", 2,

ifelse(ds$experience\_level=="SE", 3, 4)))

* ds$employment\_type\_PT <- ifelse(ds$employment\_type == "PT", 1, 0)

ds$employment\_type\_FT <- ifelse(ds$employment\_type == "FT", 1, 0)

ds$employment\_type\_CT <- ifelse(ds$employment\_type == "CT", 1, 0)

ds$employment\_type\_FL <- ifelse(ds$employment\_type == "FL", 1, 0)

* ds$job\_title <- ifelse(ds$job\_title== "3D Computer Vision Researcher", 1, 0)

ds$job\_title <- ifelse(ds$job\_title =="AI Scientist", 1, 0)

**Two examples of the variable examined are this (**job\_title**)**

* ds$salary\_currency <- ifelse(ds$salary\_currency == "INR", 1, 0)

ds$salary\_currency <- ifelse(ds$salary\_currency == "DKK", 1, 0)

**Two examples of the variable examined are this (**salary\_currency**)**

* ds$employee\_residence <- ifelse(ds$employee\_residence == "AE", 1, 0)

ds$employee\_residence <- ifelse(ds$employee\_residence == "AR", 1, 0)

**Two examples of the variable examined are this (**employee\_residence**)**

* ds\_small <- ds[c("work\_year", "work\_month",

"salary","salary\_in\_usd","remote\_ratio", "experience\_level","employment\_type\_PT","employment\_type\_FT","employment\_type\_CT","employment\_type\_FL","job\_title","salary\_currency", "employee\_residence", "company\_size")]

#make this example reproducible

set.seed(1)

#use 70% of dataset as training set and 30% as test set

sample <- sample(c(TRUE, FALSE), nrow(ds\_small), replace=TRUE, prob=c(0.7,0.3))

data\_train <- ds\_small[sample, ]

data\_test <- ds\_small[!sample, ]

**Processing**

2. Fit a Decision Tree to classify the size of company in which each data scientist work. This is represented by the feature ‘company\_size’, which is the target variable. (3 points)

2.1. Set the maximum depth of the tree to 3. (1 point)

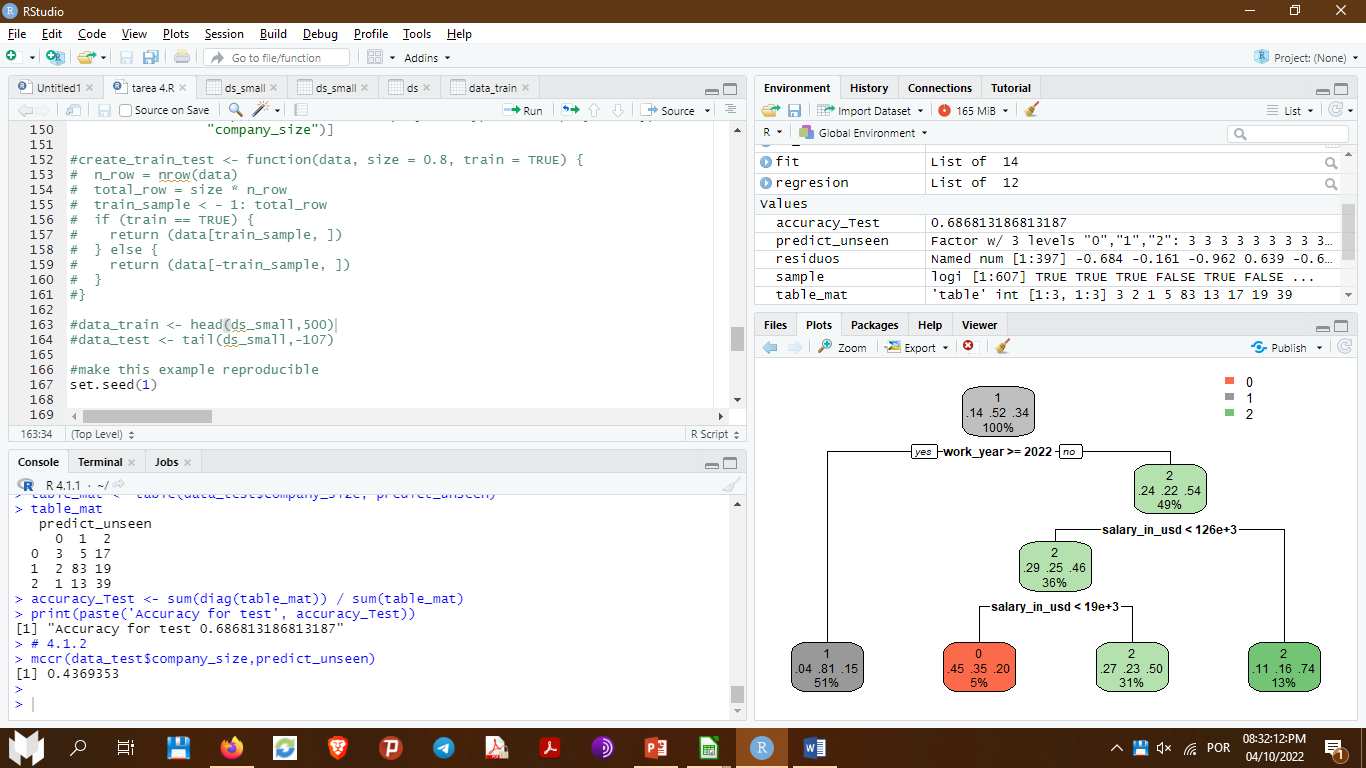
* fit <- rpart(company\_size~., data = data\_train, method = 'class', maxdepth=3)

dim(data\_train)

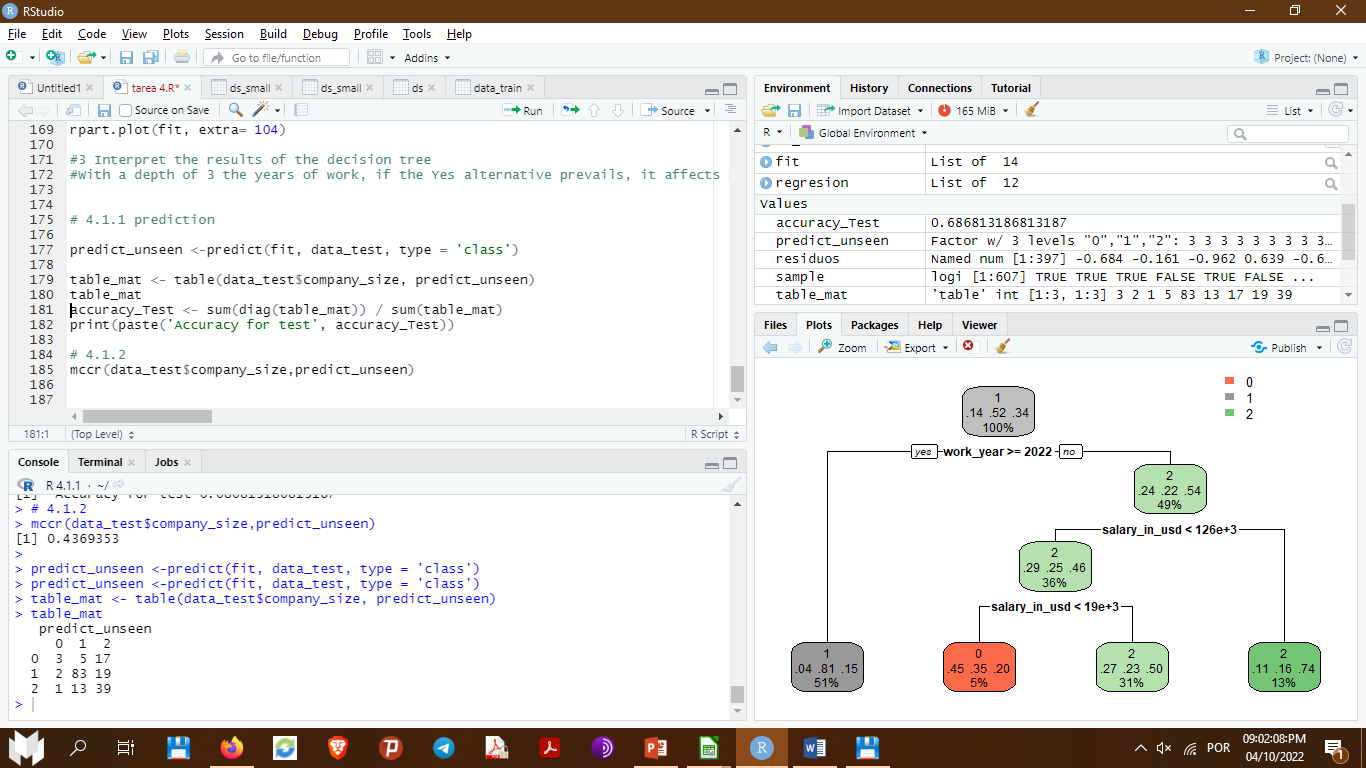
R/[1] 425 14

2.2. Plot the decision tree. (2 point)

* rpart.plot(fit, extra= 104)



> predict\_unseen <-predict(fit, data\_test, type = 'class')

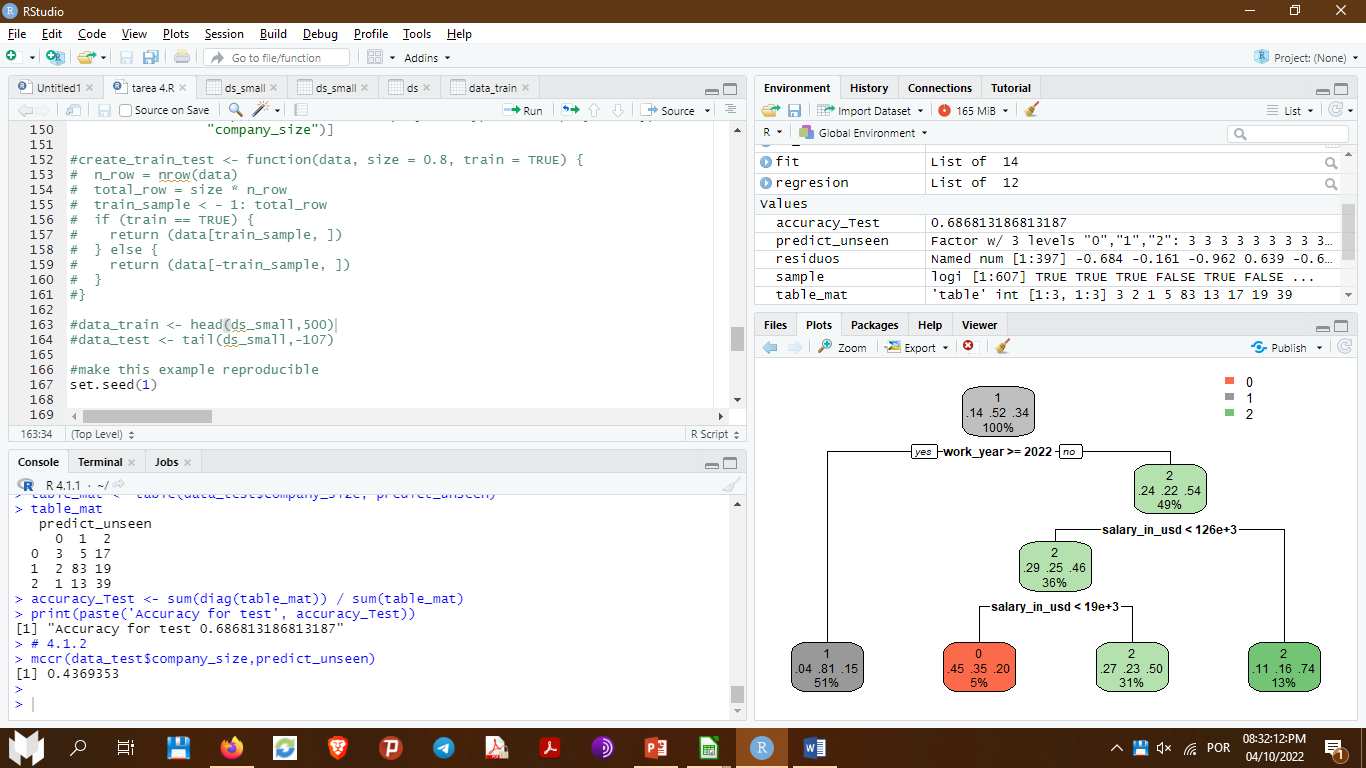
> table\_mat <- table(data\_test$company\_size, predict\_unseen)

> table\_mat

3. Interpret the results of the decision tree (3 points)

With a depth of 3 the years of work, if the Yes alternative prevails, it affects 51%, if the No prevails, it affects 49%, which allows the approach of salary in usd <133, if this value is greater, it affects by 13%, if the opposite happens, the value takes 36%, so this value allows having two alternatives salary in usd<19, if this decision is fulfilled, it generates 5%, otherwise it generates 31%

3.1. Write a paragraph interpreting the structure of the decision tree.



**Post-processing**

4. Measure the performance of the decision tree to make accurate predictions.

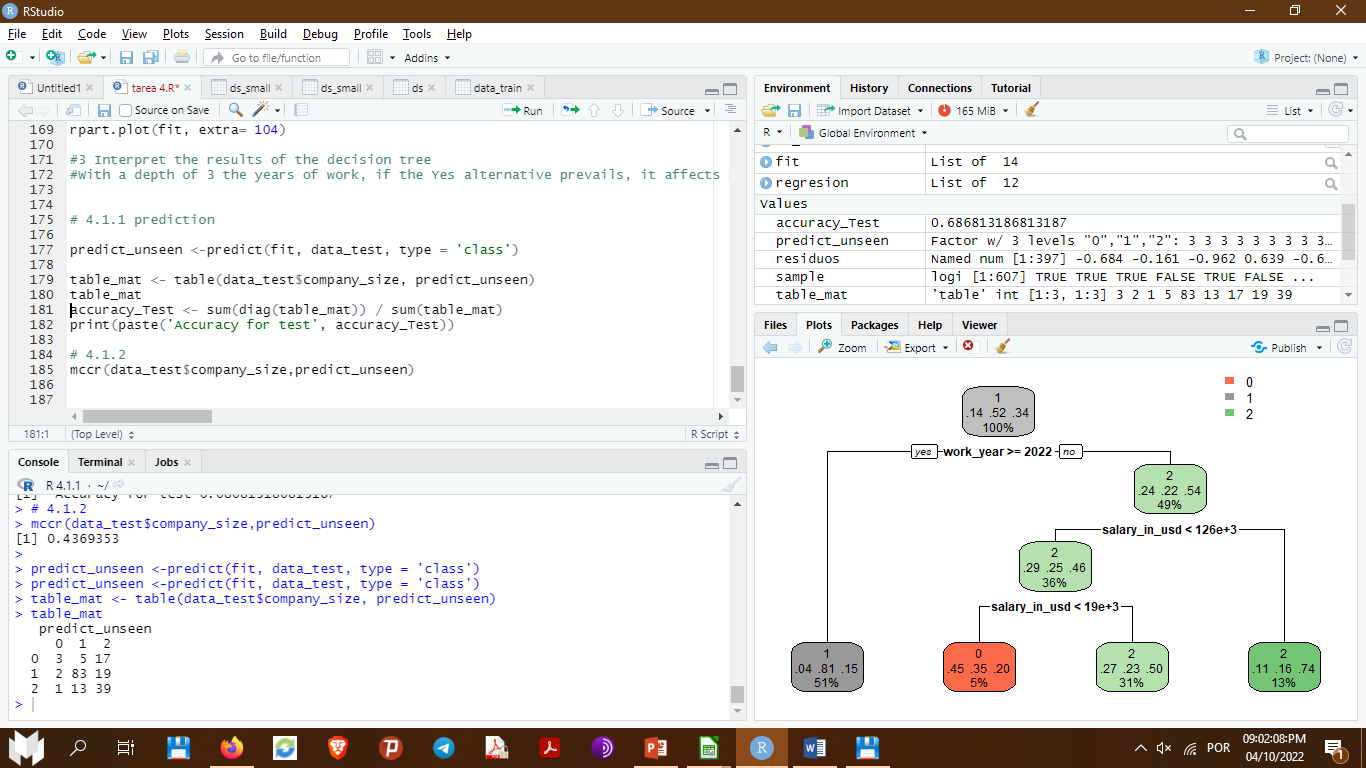
4.1. By hold-out cross-validation performance (3 points)

4.1.1. Estimate the accuracy, interpret the results. (1.5 points)

* predict\_unseen <-predict(fit, data\_test, type = 'class')

table\_mat <- table(data\_test$company\_size, predict\_unseen)

table\_mat

 R/

accuracy\_Test <- sum(diag(table\_mat)) / sum(table\_mat)

print(paste('Accuracy for test', accuracy\_Test))

R/ [1] "Accuracy for test 0.686813186813187"

4.1.2. Estimate the Matthews correlation coefficient, interpret the results (1.5 points)

* mccr(data\_test$company\_size,predict\_unseen)

R/[1] 0.4369353