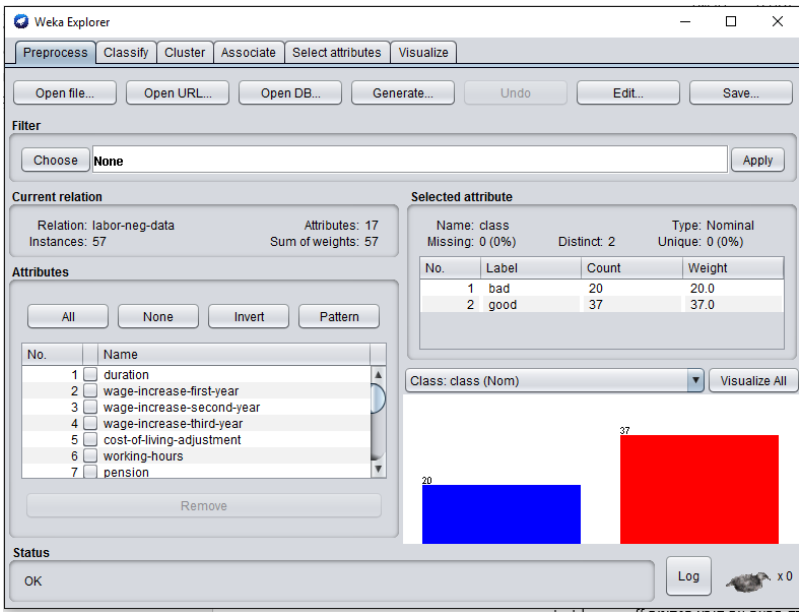
מטלה 1 בבטיחות תוכנה

מגיש: עוז לוי מעתוק 305181158

שאלה 1

.1



.2 התפלגות המחלקות היא 20 למחלקה הכחולה (bad) ו 37 למחלקה האדומה (good)

.3

@ 'duration' numeric (min = 1, max = 3, Mean = 2.161, stdDev = 0.708)

@ 'wage-increase-first-year' numeric (min = 2, max = 7, Mean = 3.804, stdDev = 1371)

@ 'wage-increase-second-year' numeric (min = 2, max = 7, Mean = 3.972, stdDev = 1.164)

@ 'wage-increase-third-year' numeric (min = 2, max = 5.1, Mean = 3.913, stdDev = 1.304)

@ 'cost-of-living-adjustment' {'none','tcf','tc'} non numeric

@ 'working-hours' numeric (min = 27, max = 40, Mean = 38.039, stdDev = 2.506)

@ 'pension' {'none','ret_allw','empl_contr'} non numeric

@ 'standby-pay' numeric (min = 2, max = 14, Mean = 7.444, stdDev = 5.028)

@ 'shift-differential' numeric (min = 0, max = 25, Mean = 4.871, stdDev = 4.544)

@ 'education-allowance' {'yes','no'} non numeric

@ 'statutory-holidays' numeric (min = 9, max = 15, Mean = 11.094, stdDev = 1.26)

@ 'vacation' {'below_average','average','generous'} (min = 1, max = 3, Mean = 2.161, stdDev = 0.708)

@ 'longterm-disability-assistance' {'yes','no'} non numeric

@ 'contribution-to-dental-plan' {'none','half','full'} non numeric

@ 'bereavement-assistance' {'yes','no'} non numeric

@ 'contribution-to-health-plan' {'none','half','full'} non numeric

@ 'class' {'bad','good'} non numeric

שאלה 2

1. [אינטואיציה: נקבל עץ מינימאלי בעל שתי איברים (עלים) אם"ם ערכי ה X שלנו יהיו קטנים שווים ל 2.5 בשגיאת מסווג של 15.27/2.27.]
   ז"א כי כאשר נגדיר בנתוני בניית העץ כי numOfObj > 10 אזי נקבל עץ מינימאלי בעל 2 עלים. אחוז השגיאה הוא 31.5789%
2. [אינטואיציה: נקבל עץ מקסימלי בעל חמש איברים (שלושה עלים) אם"ם כלל הערכים שלנו יהיו גדולים מ 2.5 בשגיאת מסווג של 30.96/1.0 עבור המחלקה good ו 10.77/4.77 עבור המחלקה bad.]
   ז"א כי כאשר נגדיר בנתוני בניית העץ כי numOfObj > 2 אזי נקבל עץ מקסימאלי בעל 8 עלים. אחוז השגיאה הוא 15.7895%

שאלה 3

1. [אינטואיציה: נקבל עץ מינימאלי בעל שתי איברים (עלים) אם"ם כלל הערכים שלנו עבור tear prob rate יהיו reduced בעל סטייה עבור none של 12.0]
   קל לראות כי כאשר נגדיר בנתוני בניית העץ כי numOfObj =< 7 אזי נקבל עץ מינימאלי בעל 2 עלים. אחוז השגיאה הוא 62.5%
2. [אינטואיציה: נקבל עץ מקסימלי בעל שיבעה איברים (ארבעה עלים) אם"ם קיים לנו ערך normal עבור tear prob rate וגם yes עבור astigmatism. נקבל סטיות של none = 3.0, hard = 3.0, soft = 6.0/1.0]
   ז"א כי כאשר נגדיר בנתוני בניית העץ כי numOfObj < 4 אזי נקבל עץ מינימאלי בעל 2 עלים. אחוז השגיאה הוא 62.5%

שאלה 4

Code:

```
install.packages("RWeka")
install.packages("xlsx")
install.packages("caret")
update.packages()
library("RWeka")
library("xlsx")
library("caret")

table <- read.csv(file=file.choose())

smp_size <- floor(0.7 * nrow(table))
train_ind <- sample(seq_len(nrow(table)), size = smp_size)
train <- table[train_ind, ]
test <- table[-train_ind, ]

J48Model <- J48(Contact.lenses ~ ., data = table)
cMatrixJ48 <- summary(J48Model)

LRModel <- glm(Contact.lenses ~.,family=binomial(link='logit'),data=train)
LRpredict <- predict(LRModel, type = 'response')
cMatrixLR <- table(train$Contact.lenses, LRpredict > 0.7)
LRaccuracy <- sum(diag(cMatrixLR))/nrow(train)*100
```

- כמו שביקשת, מצורף קובץ txt לקוד.
  בנוסף הוספתי עמודת אינדקסים לטבלה ובזכות זה החישוב של LOGISTIC REGRESSION מתבצע כהלכה (ניתן לראות בפלט של הטבלה/המודל/המטריצה)

Output:

```
# CONFUSION MATRIX OF LOGISTIC REGRESSION = cMatrixLR
#       FALSE TRUE
#  hard     2    0
#  none     0    7
#  soft     0    5
#
# CONFUSION MATRIX OF J48 = cMatrixJ48
#  a  b  c   <-- classified as
#  3  0  0 |  a = hard
#  0 11  1 |  b = none
#  0  0  5 |  c = soft
#
# LOGISTIC REGRESSION MODEL accuracy = LRaccuracy
# LRaccuracy = 64.28571
#
# J48 MODEL accuracy = summary(J48Model)
# Correctly Classified Instances          19               95      %
# Incorrectly Classified Instances         1                5      %
# Kappa statistic                          0.9127
# Mean absolute error                      0.0556
# Root mean squared error                  0.1667
# Relative absolute error                 14.631  %
# Root relative squared error             38.6831 %
# Coverage of cases (0.95 level)          100      %
# Mean rel. region size (0.95 level)      43.3333 %
# Total Number of Instances               20
```

<u>ההבדל בתוצאות המודלים נוצר משתי סיבות:</u>

הראשונה כי שהם אינם עובדים על אותה צורת חישוב.

השנייה כי מודל LR בתכליתו אינו אמור להתמודד עם יותר משתי ערכים, ואחרי מודולציה של נתוני המודל הצלחתי להוציא תוצאת חישוב עבור שלושת הערכים (אם כי קצת לא מדוייקת....).