Excercise\_6\_Data\_Mining

Duncan Ferguson

10/20/2021

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.5 v dplyr 1.0.7  
## v tidyr 1.1.4 v stringr 1.4.0  
## v readr 2.0.2 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

# Importing Table

df <- read.table("Table\_8\_1.csv", header=TRUE, sep=",")  
df

## RID Age income student credit\_rating Class..buys\_computer  
## 1 1 youth high no fair no  
## 2 2 youth high no excellent no  
## 3 3 middle\_aged high no fair yes  
## 4 4 senior medium no fair yes  
## 5 5 senior low yes fair yes  
## 6 6 senior low yes excellent no  
## 7 7 middle\_aged low yes excellent yes  
## 8 8 youth medium no fair no  
## 9 9 youth low yes fair yes  
## 10 10 senior medium yes fair yes  
## 11 11 youth medium yes excellent yes  
## 12 12 middle\_aged medium no excellent yes  
## 13 13 middle\_aged high yes fair yes  
## 14 14 senior medium no excellent no

subset <- df[df$Age =="youth",]  
subset

## RID Age income student credit\_rating Class..buys\_computer  
## 1 1 youth high no fair no  
## 2 2 youth high no excellent no  
## 8 8 youth medium no fair no  
## 9 9 youth low yes fair yes  
## 11 11 youth medium yes excellent yes

Info\_i\_D <- -((3/5)\*log2(3/5))-((2/5)\*log2(2/5))  
Info\_i\_D

## [1] 0.9709506

*Classifying Credit* # Splitting on Credit Rating

subset\_credit\_e <- subset[subset$credit\_rating =="excellent",]  
subset\_credit\_f <- subset[subset$credit\_rating =="fair",]  
subset\_credit\_e

## RID Age income student credit\_rating Class..buys\_computer  
## 2 2 youth high no excellent no  
## 11 11 youth medium yes excellent yes

subset\_credit\_f

## RID Age income student credit\_rating Class..buys\_computer  
## 1 1 youth high no fair no  
## 8 8 youth medium no fair no  
## 9 9 youth low yes fair yes

Info\_Credit\_Fair <- -((1/3)\*log2(1/3))-((2/3)\*log2(2/3))  
Info\_Credit\_Fair

## [1] 0.9182958

Info\_Credit\_Excellent <- -((1/2)\*log2(1/2))-((1/2)\*log2(1/2))  
Info\_Credit\_Excellent

## [1] 1

Info\_Credit <- ((2/5)\*Info\_Credit\_Excellent) + ((3/5)\*Info\_Credit\_Fair)  
Info\_Credit

## [1] 0.9509775

Gain\_Credit <- Info\_i\_D-Info\_Credit  
Gain\_Credit

## [1] 0.01997309

*Classifying Student* # Splitting on Student

subset\_student\_y <- subset[subset$student == "yes",]  
subset\_student\_n <- subset[subset$student == "no",]  
subset\_student\_y

## RID Age income student credit\_rating Class..buys\_computer  
## 9 9 youth low yes fair yes  
## 11 11 youth medium yes excellent yes

subset\_student\_n

## RID Age income student credit\_rating Class..buys\_computer  
## 1 1 youth high no fair no  
## 2 2 youth high no excellent no  
## 8 8 youth medium no fair no

Info\_Student\_Yes <- -((2/2)\*log(2/2))-((0/2)\*log2(0/2))  
Info\_Student\_Yes

## [1] NaN

Info\_Student\_No <- -((0/3)\*log(0/3))-((3/3)\*log2(3/3))  
Info\_Student\_No

## [1] NaN

Info\_Student <- ((2/5)\*0)+((3/5)\*0)  
Info\_Student

## [1] 0

Gain\_Student <- Info\_i\_D - Info\_Student  
Gain\_Student

## [1] 0.9709506

*Classifying Income*

# Splitting on Income

subset\_income\_high <- subset[subset$income == "high",]  
subset\_income\_medium <- subset[subset$income == "medium",]  
subset\_income\_low <- subset[subset$income == "low",]  
subset\_income\_low

## RID Age income student credit\_rating Class..buys\_computer  
## 9 9 youth low yes fair yes

subset\_income\_medium

## RID Age income student credit\_rating Class..buys\_computer  
## 8 8 youth medium no fair no  
## 11 11 youth medium yes excellent yes

subset\_income\_high

## RID Age income student credit\_rating Class..buys\_computer  
## 1 1 youth high no fair no  
## 2 2 youth high no excellent no

Info\_Income\_Low <- -(1/1)\*log2(1/1) - 0  
Info\_Income\_Low

## [1] 0

Info\_Income\_Medium <- -((1/2)\*log2(1/2))-((1/2)\*log2(1/2))  
Info\_Income\_Medium

## [1] 1

Info\_Income\_High <- 0-((2/2)\*log2(2/2))  
Info\_Income\_High

## [1] 0

Info\_Income <- ((1/5)\*Info\_Income\_Low) + ((2/5)\*Info\_Income\_Medium) + ((2/5)\*Info\_Income\_High)  
Info\_Income

## [1] 0.4

Gain\_Income <- Info\_i\_D - Info\_Income  
Gain\_Income

## [1] 0.5709506