NAMA	: SABRIELLA HAFIFAH
NIM	: 22110592
NO. ABSEN	: 32
KELAS	: 3SD1
NAMA DOSEN	: IBNU SANTOSO, SST, MT
MATA KULIAH	: DATA MINING AND KNOWLEDGE MANAGEMENT
HARI / TANGGAL UJIAN	: SELASA / 27-10-2020

" SAYA MENYATAKAN BAHWA UJIAN INI SAYA KERJAKAN DENGAN JUJUR SESUAI KEMAMPUAN SENDIRI DAN TIDAK MENGUTIP SEBAGIAN ATAU SELURUH PEKERJAAN ORANG LAIN. JIKA SUATU SAAT DITEMUKAN SAYA MELANGGAR KETENTUAN UJIAN, SAYA SIAP MENERIMA KONSEKUENSI YANG BERLAKU. "

(SABRIELLA HAFIFAH)

KIKY

NAMA : SABRIELLA HAFIFAH

NIM : 221810592

KELAS : 3SD1

UTS DATMIN

Sabriella Hafifah

October 27, 2020

Library

```
library(OneR)
```

```
## Warning: package 'OneR' was built under R version 4.0.3
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

Load Data

```
library(readxl)
data <- read_excel("C:/Users/SABRIELLA/Downloads/2014 and 2015 CSM dataset.xlsx")
str(data)
```

```
## tibble [231 x 14] (S3: tbl_df/tbl/data.frame)
##  $ Movie              : chr [1:231] "13 Sins" "22 Jump Street" "3 Days to Kill" "300: Rise of an Empire" ...
##  $ Year               : num [1:231] 2014 2014 2014 2014 2014 ...
##  $ Ratings            : num [1:231] 6.3 7.1 6.2 6.3 4.7 4.6 6.1 7.1 6.5 6.1 ...
##  $ Genre              : num [1:231] 8 1 1 1 8 3 8 1 10 8 ...
##  $ Gross              : num [1:231] 9.13e+03 1.92e+08 3.07e+07 1.06e+08 1.73e+07 2.90e+04 4.26e+07 5.75e+06 2.60e+07 4.86e+07 ...
```

```
##  $ Budget             : num [1:231] 4.00e+06 5.00e+07 2.80e+07 1.10e+08 3.
50e+06 5.00e+05 4.00e+07 2.00e+07 2.80e+07 1.25e+07 ...
##  $ Screens            : num [1:231] 45 3306 2872 3470 2310 ...
##  $ Sequel             : num [1:231] 1 2 1 2 2 1 1 1 1 1 ...
##  $ Sentiment          : num [1:231] 0 2 0 0 0 0 0 2 3 0 ...
##  $ Views              : num [1:231] 3280543 583289 304861 452917 3145573 .
..
##  $ Likes              : num [1:231] 4632 3465 328 2429 12163 ...
##  $ Dislikes           : num [1:231] 425 61 34 132 610 7 419 197 419 532 ..
.
##  $ Comments           : num [1:231] 636 186 47 590 1082 ...
##  $ Aggregate Followers: num [1:231] 1120000 12350000 483000 568000 1923800
...
```

```
summary(data)
```

```
##     Movie                Year         Ratings          Genre
##  Length:231         Min.   :2014   Min.   :3.100   Min.   : 1.000
##  Class :character   1st Qu.:2014   1st Qu.:5.800   1st Qu.: 1.000
##  Mode  :character   Median :2014   Median :6.500   Median : 3.000
##                     Mean   :2014   Mean   :6.442   Mean   : 5.359
##                     3rd Qu.:2015   3rd Qu.:7.100   3rd Qu.: 8.000
##                     Max.   :2015   Max.   :8.700   Max.   :15.000
##
##      Gross              Budget             Screens         Sequel
##  Min.   :     2470   Min.   :    70000   Min.   :   2   Min.   :1.000
##  1st Qu.: 10300000   1st Qu.:  9000000   1st Qu.: 449   1st Qu.:1.000
##  Median : 37400000   Median : 28000000   Median :2777   Median :1.000
##  Mean   : 68066033   Mean   : 47921730   Mean   :2209   Mean   :1.359
##  3rd Qu.: 89350000   3rd Qu.: 65000000   3rd Qu.:3372   3rd Qu.:1.000
##  Max.   :643000000   Max.   :250000000   Max.   :4324   Max.   :7.000
##                      NA's   :1           NA's   :10
##    Sentiment          Views              Likes           Dislikes
##  Min.   :-38.00   Min.   :     698   Min.   :     1   Min.   :    0.0
##  1st Qu.:  0.00   1st Qu.:  623302   1st Qu.:  1776   1st Qu.:  105.5
##  Median :  0.00   Median : 2409338   Median :  6096   Median :  341.0
##  Mean   :  2.81   Mean   : 3712851   Mean   : 12732   Mean   :  679.1
##  3rd Qu.:  5.50   3rd Qu.: 5217380   3rd Qu.: 15248   3rd Qu.:  697.5
##  Max.   : 29.00   Max.   :32626778   Max.   :370552   Max.   :13960.0
##
##     Comments       Aggregate Followers
##  Min.   :    0.0   Min.   :    1066
##  1st Qu.:  248.5   1st Qu.:  183025
##  Median :  837.0   Median : 1052600
##  Mean   : 1825.7   Mean   : 3038193
##  3rd Qu.: 2137.0   3rd Qu.: 3694500
##  Max.   :38363.0   Max.   :31030000
##                    NA's   :35
```

Preprocessing

```
#Buat kategori untuk beberapa variabel yang akan digunakan
data$grup_rate <- cut(data$Ratings, breaks = c(0, 3, 7, 10), labels = c("Rend
ah", "Sedang", "Tinggi"))
data$Genre <- as.factor(data$Genre)
data$Sequel <- as.factor(data$Sequel)
data$grp_budget <- ifelse(data$Budget > 28000000, 1, 0)
data$grp_screens <- ifelse(data$Screens >2777, 1, 0)
data$grp_views <- ifelse(data$Views > 2409338, 1, 0)
data$grp_likes <- ifelse(data$Likes > 6096, 1, 0)
data$grp_dislikes <- ifelse(data$Dislikes > 342, 1, 0)
data$grp_komen <- ifelse(data$Comments > 837, 1, 0)

myData <- data.frame(data$grup_rate, data$Genre, data$Sequel, data$grp_budget
, data$grp_screens, data$grp_views, data$grp_dislikes, data$grp_likes, data$g
rp_komen)
str(myData)
```

```
## 'data.frame':    231 obs. of  9 variables:
##  $ data.grup_rate   : Factor w/ 3 levels "Rendah","Sedang",..: 2 3 2 2 2 2
2 3 2 2 ...
##  $ data.Genre       : Factor w/ 11 levels "1","2","3","4",..: 7 1 1 1 7 3
7 1 9 7 ...
##  $ data.Sequel      : Factor w/ 7 levels "1","2","3","4",..: 1 2 1 2 2 1 1
1 1 1 ...
##  $ data.grp_budget  : num  0 1 0 1 0 0 1 0 0 0 ...
##  $ data.grp_screens : num  0 1 1 1 0 NA 1 0 0 0 ...
##  $ data.grp_views   : num  1 0 0 0 1 0 1 0 0 1 ...
##  $ data.grp_dislikes: num  1 0 0 0 1 0 1 0 1 1 ...
##  $ data.grp_likes   : num  0 0 0 0 1 0 1 0 0 1 ...
##  $ data.grp_komen   : num  0 0 0 0 1 0 1 0 0 0 ...
```

Mengubah tipe variabel menjadi factor

```
for(i in names(myData)){
  myData[ ,i] <- as.factor(myData[ ,i])
}
str(myData)
```

```
## 'data.frame':    231 obs. of  9 variables:
##  $ data.grup_rate   : Factor w/ 3 levels "Rendah","Sedang",..: 2 3 2 2 2 2
2 3 2 2 ...
##  $ data.Genre       : Factor w/ 11 levels "1","2","3","4",..: 7 1 1 1 7 3
7 1 9 7 ...
##  $ data.Sequel      : Factor w/ 7 levels "1","2","3","4",..: 1 2 1 2 2 1 1
1 1 1 ...
##  $ data.grp_budget  : Factor w/ 2 levels "0","1": 1 2 1 2 1 1 2 1 1 1 ...
##  $ data.grp_screens : Factor w/ 2 levels "0","1": 1 2 2 2 1 NA 2 1 1 1 ...
##  $ data.grp_views   : Factor w/ 2 levels "0","1": 2 1 1 1 2 1 2 1 1 2 ...
```

```
##  $ data.grp_dislikes: Factor w/ 2 levels "0","1": 2 1 1 1 2 1 2 1 2 2 ...
##  $ data.grp_likes   : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 2 1 1 2 ...
##  $ data.grp_komen   : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 2 1 1 1 ...
```

Split Data

Memecah data menjadi data training(80% dari data awal) dan data test (20% dari data awal)

```
set.seed(1233)
sampel <- sample(2,nrow(myData), replace = TRUE, prob = c(0.8,0.2))
trainData <- myData[sampel==1, ]
testingData<- myData[sampel==2, ]
print(paste("Jumlah Train Data: ", nrow(trainData), "| Jumlah Test Data: ", n
row(testingData)))
```

```
## [1] "Jumlah Train Data:  193 | Jumlah Test Data:  38"
```

Membuat Model

```
modelOneR <- OneR(data.grup_rate~., data = trainData, verbose = TRUE)
```

```
## Warning in bin(data): 8 instance(s) removed due to missing values
```

```
## Warning in OneR.data.frame(x = data, ties.method = ties.method, verbose =
## verbose, : data contains unused factor levels
```

```
##
##     Attribute         Accuracy
## 1 * data.Sequel       74.05%
## 2   data.Genre        71.89%
## 2   data.grp_budget   71.89%
## 2   data.grp_screens  71.89%
## 2   data.grp_views    71.89%
## 2   data.grp_dislikes 71.89%
## 2   data.grp_likes    71.89%
## 2   data.grp_komen    71.89%
## ---
## Chosen attribute due to accuracy
## and ties method (if applicable): '*'
```

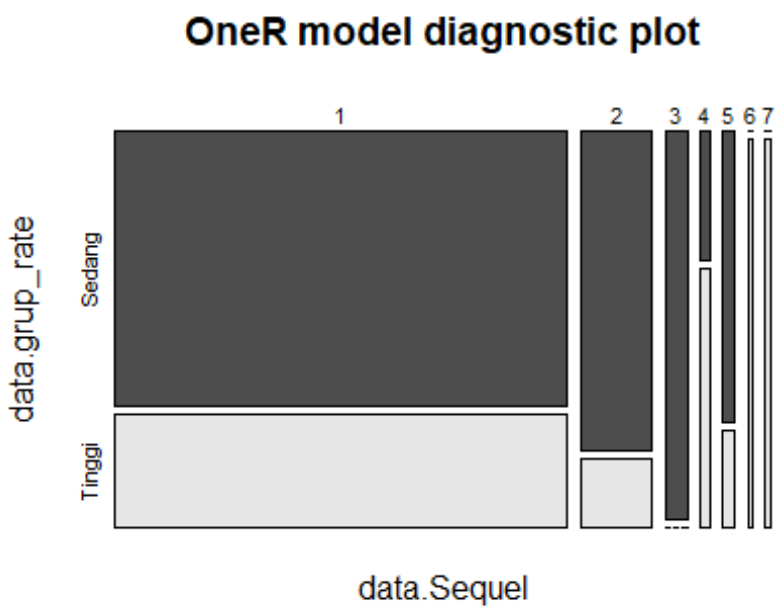
```
summary(modelOneR)
```

```
##
## Call:
## OneR.formula(formula = data.grup_rate ~ ., data = trainData,
##     verbose = TRUE)
##
## Rules:
## If data.Sequel = 1 then data.grup_rate = Sedang
## If data.Sequel = 2 then data.grup_rate = Sedang
```

```
## If data.Sequel = 3 then data.grup_rate = Sedang
## If data.Sequel = 4 then data.grup_rate = Tinggi
## If data.Sequel = 5 then data.grup_rate = Sedang
## If data.Sequel = 6 then data.grup_rate = Tinggi
## If data.Sequel = 7 then data.grup_rate = Tinggi
##
## Accuracy:
## 137 of 185 instances classified correctly (74.05%)
##
## Contingency table:
##                data.Sequel
## data.grup_rate     1    2   3   4   5   6   7 Sum
##         Sedang * 103 * 19 * 7   1 * 3   0   0 133
##         Tinggi    42    4   0 * 2   1 * 1 * 2  52
##         Sum      145   23   7   3   4   1   2 185
## ---
## Maximum in each column: '*'
##
## Pearson's Chi-squared test:
## X-squared = 13.996, df = 6, p-value = 0.02968
```

OneR Model

```
plot(modelOneR)
```



Model Evaluation

```r
p1 <- predict(modelOneR, testingData, type = "class")
eval_model(p1, testingData)
```

```
##
## Confusion matrix (absolute):
##           Actual
## Prediction  0  1 Sedang Sum
##     0       0  0      0   0
##     1       0  0      0   0
##     Sedang 19 19      0  38
##     Sum    19 19      0  38
##
## Confusion matrix (relative):
##           Actual
## Prediction   0   1 Sedang Sum
##     0      0.0 0.0    0.0 0.0
##     1      0.0 0.0    0.0 0.0
##     Sedang 0.5 0.5    0.0 1.0
##     Sum    0.5 0.5    0.0 1.0
##
## Accuracy:
## 0 (0/38)
##
## Error rate:
## 1 (38/38)
##
## Error rate reduction (vs. base rate):
## -1 (p-value = 1)
```

```r
confusionMatrix(p1, testingData$data.grup_rate)
```

```
## Warning in confusionMatrix.default(p1, testingData$data.grup_rate): Levels
are
## not in the same order for reference and data. Refactoring data to match.
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Rendah Sedang Tinggi
##     Rendah      0      0      0
##     Sedang      0     28     10
##     Tinggi      0      0      0
##
## Overall Statistics
##
##                Accuracy : 0.7368
##                  95% CI : (0.569, 0.866)
##     No Information Rate : 0.7368
##     P-Value [Acc > NIR] : 0.584
##
##                   Kappa : 0
```

```
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Rendah Class: Sedang Class: Tinggi
## Sensitivity                     NA        1.0000        0.0000
## Specificity                      1        0.0000        1.0000
## Pos Pred Value                  NA        0.7368           NaN
## Neg Pred Value                  NA           NaN        0.7368
## Prevalence                       0        0.7368        0.2632
## Detection Rate                   0        0.7368        0.0000
## Detection Prevalence             0        1.0000        0.0000
## Balanced Accuracy               NA        0.5000        0.5000
```

ZeroR

```
# Hanya mengambil target class saja
head(trainData)
```

```
##   data.grup_rate data.Genre data.Sequel data.grp_budget data.grp_screens
## 1         Sedang          8           1               0                0
## 2         Tinggi          1           2               1                1
## 3         Sedang          1           1               0                1
## 4         Sedang          1           2               1                1
## 5         Sedang          8           2               0                0
## 6         Sedang          3           1               0             <NA>
##   data.grp_views data.grp_dislikes data.grp_likes data.grp_komen
## 1              1                 1              0             0
## 2              0                 0              0             0
## 3              0                 0              0             0
## 4              0                 0              0             0
## 5              1                 1              1             1
## 6              0                 0              0             0
```

```
trainingdat <- trainData[,1]
testingdat <- testingData[,1]
trainingdat
```

```
##   [1] Sedang Tinggi Sedang Sedang Sedang Sedang Sedang Tinggi Sedang Sedan
g
##  [11] Tinggi Sedang Sedang Sedang Sedang Sedang Sedang Tinggi Sedang Sedan
g
##  [21] Sedang Sedang Tinggi Tinggi Sedang Sedang Sedang Sedang Sedang Sedan
g
##  [31] Sedang Sedang Tinggi Sedang Tinggi Sedang Sedang Tinggi Sedang Sedan
g
##  [41] Sedang Sedang Sedang Sedang Tinggi Sedang Sedang Sedang Sedang Sedan
g
##  [51] Sedang Tinggi Sedang Sedang Sedang Sedang Sedang Sedang Sedang Tingg
```

```
i
##  [61] Sedang Sedang Sedang Sedang Sedang Tinggi Sedang Sedang Sedang Tingg
i
##  [71] Sedang Sedang Sedang Sedang Sedang Sedang Sedang Sedang Tinggi Sedan
g
##  [81] Sedang Sedang Sedang Sedang Tinggi Sedang Sedang Sedang Sedang Tingg
i
##  [91] Sedang Sedang Sedang Tinggi Sedang Tinggi Sedang Sedang Sedang Sedan
g
## [101] Tinggi Tinggi Sedang Tinggi Sedang Tinggi Tinggi Sedang Tinggi Sedan
g
## [111] Sedang Sedang Sedang Tinggi Sedang Sedang Sedang Sedang Sedang Tingg
i
## [121] Sedang Sedang Sedang Sedang Sedang Tinggi Sedang Sedang Tinggi Tingg
i
## [131] Tinggi Sedang Sedang Tinggi Sedang Tinggi Tinggi Tinggi Tinggi Sedan
g
## [141] Sedang Tinggi Sedang Sedang Sedang Tinggi Tinggi Sedang Tinggi Sedan
g
## [151] Sedang Sedang Sedang Sedang Sedang Sedang Sedang Sedang Sedang Tingg
i
## [161] Sedang Tinggi Tinggi Sedang Sedang Sedang Tinggi Tinggi Sedang Sedan
g
## [171] Sedang Sedang Sedang Tinggi Sedang Sedang Sedang Sedang Sedang Tingg
i
## [181] Sedang Tinggi Tinggi Tinggi Tinggi Tinggi Sedang Sedang Sedang Sedan
g
## [191] Sedang Sedang Sedang
## Levels: Rendah Sedang Tinggi
```

```
testingdat
```

```
##  [1] Sedang Sedang Sedang Tinggi Sedang Tinggi Sedang Sedang Tinggi Sedang
## [11] Tinggi Tinggi Sedang Sedang Sedang Sedang Sedang Sedang Tinggi Sedang
## [21] Sedang Tinggi Tinggi Sedang Sedang Sedang Tinggi Sedang Sedang Sedang
## [31] Sedang Sedang Sedang Sedang Tinggi Sedang Sedang Sedang
## Levels: Rendah Sedang Tinggi
```

```
# Ambil banyak rendah, sedang, dan tinggi pada target class
banyakRendah <- sum(trainingdat == "Rendah")
banyakRendah
```

```
## [1] 0
```

```
banyakSedang <- sum(trainingdat == "Sedang")
banyakSedang
```

```
## [1] 141
```

```
banyakTinggi <- sum(trainingdat == "Tinggi")
banyakTinggi
```

```
## [1] 52
```

Hitung Peluang

```
probRendah <- banyakRendah/length(trainingdat)
probSedang <- banyakSedang/length(trainingdat)
probTinggi <- banyakTinggi/length(trainingdat)
print(paste("Peluang Rendah: ", probRendah, " | Peluang Sedang: ", probSedang
, " | Peluang Tinggi: ", probTinggi))
```

```
## [1] "Peluang Rendah:  0  | Peluang Sedang:  0.730569948186529  | Peluang T
inggi:  0.269430051813472"
```