Nama : Mardiningtyas Bayu Sulthani

NIM : 221810412

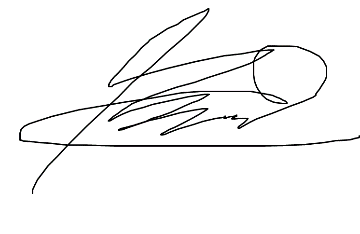
No. Absen : 19

Dosen : Ibnu Santoso SST, MT

Mata Kuliah : DMKM

Hari/Tanggal : Selasa, 27 Oktober 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 INI, SAYA SIAP MENERIMA KONSEKUENSI YANG BERLAKU.”



(Mardiningtyas Bayu Sulthani)

221810412-Mardiningtyas Bayu Sulthani

10/27/2020

## Pembacaan Data

Data dibaca menggunakan fungsi read.csv dari direktori yang sudah ditentukan lalu dilihat class tipe data tiap kolomnya

dataHep <- read.csv("C:\\Users\\moonc\\Desktop\\PJJ\\Datmin\\hepatitis.data")  
sapply(dataHep, class)

## Class AGE SEX STEROID ANTIVIRALS   
## "integer" "integer" "integer" "integer" "integer"   
## FATIGUE MALAISE ANOREXIA LIVERBIG LIVERFIRM   
## "integer" "integer" "integer" "integer" "integer"   
## SPLEENPALPABLE SPIDERS ASCITES VARICES BILIRUBIN   
## "integer" "integer" "integer" "integer" "numeric"   
## ALKPHOSPHATE SGOT ALBUMIN PROTIME HISTOLOGY   
## "integer" "integer" "numeric" "integer" "integer"

## Perbaikan Tipe data

Karena ada beberapa tipe data yang tidak sesuai, maka harus diperbaiki terlebih dahulu menggunakan fungsi transform satu per satu kemudian baris yang mengandung NA akan dihapus sehingga data menjadi bersih menggunakan fungsi remove.na dari package rgr

dataHep <- transform(  
 dataHep,  
 Class=as.factor(Class),  
 AGE=as.numeric(AGE),  
 SEX=as.factor(SEX),  
 STEROID=as.factor(STEROID),  
 ANTIVIRALS=as.factor(ANTIVIRALS),  
 FATIGUE=as.factor(FATIGUE),  
 MALAISE=as.factor(MALAISE),  
 ANOREXIA=as.factor(ANOREXIA),  
 LIVERBIG=as.factor(LIVERBIG),  
 LIVERFIRM=as.factor(LIVERFIRM),  
 SPLEENPALPABLE=as.factor(SPLEENPALPABLE),  
 SPIDERS=as.factor(SPIDERS),  
 ASCITES=as.factor(ASCITES),  
 VARICES=as.factor(VARICES),  
 BILIRUBIN=as.numeric(BILIRUBIN),  
 ALKPHOSPATE=as.numeric(ALKPHOSPHATE),  
 SGOT=as.numeric(SGOT),  
 ALBUMIN=as.numeric(ALBUMIN),  
 PROTIME=as.numeric(PROTIME),  
 HISTOLOGY=as.factor(HISTOLOGY)  
)  
  
dataHep <- dataHep[,1:20] #karena kolom ALKHOSPAHTE entah kenapa ada dua  
rem <- remove.na(dataHep, iftell = TRUE)

##   
## 75 row(s) with NA(s) removed from matrix

dataRem <- rem$x  
head(dataRem)

## Class AGE SEX STEROID ANTIVIRALS FATIGUE MALAISE ANOREXIA LIVERBIG LIVERFIRM  
## 6 2 34 1 2 2 2 2 2 2 2  
## 11 2 39 1 1 1 2 2 2 1 1  
## 12 2 32 1 2 1 1 2 2 2 1  
## 13 2 41 1 2 1 1 2 2 2 1  
## 14 2 30 1 2 2 1 2 2 2 1  
## 16 2 38 1 1 2 1 1 1 2 2  
## SPLEENPALPABLE SPIDERS ASCITES VARICES BILIRUBIN ALKPHOSPHATE SGOT ALBUMIN  
## 6 2 2 2 2 0.9 95 28 4.0  
## 11 2 2 2 2 1.3 78 30 4.4  
## 12 2 1 2 2 1.0 59 249 3.7  
## 13 2 2 2 2 0.9 81 60 3.9  
## 14 2 2 2 2 2.2 57 144 4.9  
## 16 2 2 1 2 2.0 72 89 2.9  
## PROTIME HISTOLOGY  
## 6 75 1  
## 11 85 1  
## 12 54 1  
## 13 52 1  
## 14 78 1  
## 16 46 1

## 

## Pecah data ke train dan test

Data kemudian dipecat menggunakan fungsi sample.split dari package caTools menjadi set training dan set test

sample = sample.split(dataRem$Class, SplitRatio = .75)  
  
train = subset(dataRem, sample == TRUE)  
test = subset(dataRem, sample == FALSE)  
  
dim(train)

## [1] 60 20

dim(test)

## [1] 20 20

## Fungsi fungsi yang diperlukan

Kemudian beberapa fungsi yang diperlukan didefinisikan

IsPure <- function(data) {  
 length(unique(data[,ncol(data)])) == 1  
}  
  
Entropy <- function( vls ) {  
 res <- vls/sum(vls) \* log2(vls/sum(vls))  
 res[vls == 0] <- 0  
 -sum(res)  
}  
  
InformationGain <- function( tble ) {  
 tble <- as.data.frame.matrix(tble)  
 entropyBefore <- Entropy(colSums(tble))  
 s <- rowSums(tble)  
 entropyAfter <- sum (s / sum(s) \* apply(tble, MARGIN = 1, FUN = Entropy ))  
 informationGain <- entropyBefore - entropyAfter  
 return (informationGain)  
}  
  
TrainID3 <- function(node, data) {  
   
 node$obsCount <- nrow(data)  
   
   
 if (IsPure(data)) {  
   
 child <- node$AddChild(unique(data[,ncol(data)]))  
 node$feature <- tail(names(data), 1)  
 child$obsCount <- nrow(data)  
 child$feature <- ''  
 } else {  
   
 ig <- sapply(colnames(data)[-ncol(data)],   
 function(x) InformationGain(  
 table(data[,x], data[,ncol(data)])  
 )  
 )  
 feature <- names(ig)[ig == max(ig)][1]  
   
 node$feature <- feature  
   
   
 childObs <- split(data[,!(names(data) %in% feature)], data[,feature], drop = TRUE)  
   
 for(i in 1:length(childObs)) {  
   
 child <- node$AddChild(names(childObs)[i])  
   
   
 TrainID3(child, childObs[[i]])  
 }  
   
 }  
   
}

## Training Model

Model mulai dilatih menggunakan fungsi yang ada pada package data.tree

tree <- Node$new("train")  
TrainID3(tree, train)  
print(tree, "feature", "obsCount")

## levelName feature obsCount  
## 1 train SGOT 60  
## 2 ¦--15 HISTOLOGY 1  
## 3 ¦ °--1 1  
## 4 ¦--18 AGE 3  
## 5 ¦ ¦--25 HISTOLOGY 1  
## 6 ¦ ¦ °--1 1  
## 7 ¦ ¦--38 HISTOLOGY 1  
## 8 ¦ ¦ °--2 1  
## 9 ¦ °--56 HISTOLOGY 1  
## 10 ¦ °--1 1  
## 11 ¦--19 HISTOLOGY 2  
## 12 ¦ °--2 2  
## 13 ¦--20 Class 2  
## 14 ¦ ¦--1 HISTOLOGY 1  
## 15 ¦ ¦ °--2 1  
## 16 ¦ °--2 HISTOLOGY 1  
## 17 ¦ °--1 1  
## 18 ¦--24 STEROID 2  
## 19 ¦ ¦--1 HISTOLOGY 1  
## 20 ¦ ¦ °--1 1  
## 21 ¦ °--2 HISTOLOGY 1  
## 22 ¦ °--2 1  
## 23 ¦--28 HISTOLOGY 2  
## 24 ¦ °--1 2  
## 25 ¦--29 HISTOLOGY 1  
## 26 ¦ °--1 1  
## 27 ¦--30 AGE 2  
## 28 ¦ ¦--39 HISTOLOGY 1  
## 29 ¦ ¦ °--1 1  
## 30 ¦ °--54 HISTOLOGY 1  
## 31 ¦ °--2 1  
## 32 ¦--31 HISTOLOGY 2  
## 33 ¦ °--1 2  
## 34 ¦--38 HISTOLOGY 2  
## 35 ¦ °--1 2  
## 36 ¦--42 HISTOLOGY 1  
## 37 ¦ °--2 1  
## 38 ¦--44 HISTOLOGY 2  
## 39 ¦ °--1 2  
## 40 ¦--46 HISTOLOGY 2  
## 41 ¦ °--1 2  
## 42 ¦--49 HISTOLOGY 1  
## 43 ¦ °--2 1  
## 44 ¦--52 HISTOLOGY 1  
## 45 ¦ °--2 1  
## 46 ¦--53 HISTOLOGY 1  
## 47 ¦ °--1 1  
## 48 ¦--54 HISTOLOGY 1  
## 49 ¦ °--1 1  
## 50 ¦--55 Class 2  
## 51 ¦ ¦--1 HISTOLOGY 1  
## 52 ¦ ¦ °--2 1  
## 53 ¦ °--2 HISTOLOGY 1  
## 54 ¦ °--1 1  
## 55 ¦--58 HISTOLOGY 2  
## 56 ¦ °--1 2  
## 57 ¦--60 HISTOLOGY 1  
## 58 ¦ °--1 1  
## 59 ¦--63 HISTOLOGY 1  
## 60 ¦ °--1 1  
## 61 ¦--64 HISTOLOGY 1  
## 62 ¦ °--1 1  
## 63 ¦--65 HISTOLOGY 1  
## 64 ¦ °--1 1  
## 65 ¦--69 HISTOLOGY 1  
## 66 ¦ °--2 1  
## 67 ¦--70 HISTOLOGY 1  
## 68 ¦ °--2 1  
## 69 ¦--75 HISTOLOGY 2  
## 70 ¦ °--2 2  
## 71 ¦--78 HISTOLOGY 1  
## 72 ¦ °--1 1  
## 73 ¦--80 HISTOLOGY 1  
## 74 ¦ °--2 1  
## 75 ¦--81 HISTOLOGY 1  
## 76 ¦ °--2 1  
## 77 ¦--89 HISTOLOGY 1  
## 78 ¦ °--1 1  
## 79 ¦--90 HISTOLOGY 1  
## 80 ¦ °--1 1  
## 81 ¦--92 AGE 2  
## 82 ¦ ¦--35 HISTOLOGY 1  
## 83 ¦ ¦ °--1 1  
## 84 ¦ °--54 HISTOLOGY 1  
## 85 ¦ °--2 1  
## 86 ¦--98 HISTOLOGY 2  
## 87 ¦ °--1 2  
## 88 ¦--117 HISTOLOGY 1  
## 89 ¦ °--2 1  
## 90 ¦--118 HISTOLOGY 1  
## 91 ¦ °--2 1  
## 92 ¦--140 HISTOLOGY 1  
## 93 ¦ °--2 1  
## 94 ¦--144 HISTOLOGY 1  
## 95 ¦ °--1 1  
## 96 ¦--145 HISTOLOGY 1  
## 97 ¦ °--1 1  
## 98 ¦--156 HISTOLOGY 1  
## 99 ¦ °--1 1  
## 100 °--... 4 nodes w/ 4 sub NA

## Terlihat bahwa sebuah decision tree telah terbentuk

## Prediksi

Predict <- function(tree, features) {  
 if (tree$children[[1]]$isLeaf) return (tree$children[[1]]$name)  
 child <- tree$children[[features[[tree$feature]]]]  
 return ( Predict(child, features))  
}

Predict(tree, test[2,2:20])

## [1] "1"

Maaf pak perhitungan akurasi, sensitivity, specificity, precision, recall, F-1 score, dan confusion matrix tidak sempat terhitung karena fungsi untuk mendapatkan vector prediksi belum sempat terbuat dan waktunya sudah habis terlebih dahulu dan yang bisa saya kerjakan hanya pembuatan model dan preprocessing data nya.