**Praktikum 6**

**Naïve Bayes**

|  |
| --- |
| install.packages("tidyverse")  library("tidyverse") #for data wrangling tools  install.packages("titanic")  library("titanic")  tdf <- titanic\_train #training set of Titanic data  head(tdf)  # Compute the probability that a randomly selected passenger on the Titanic was female given that the passenger was at least 35 years old.  tdf %>%  summarize(prob = sum(Age >= 35 & Sex == "female", na.rm = TRUE)/sum(Age >= 35, na.rm = TRUE)) |

Output :

|  |
| --- |
|  |

Lakukan perhitungan kasus di atas dengan menggunakan Excel. Pisahkan age dengan batas 35 tahun. Output :

|  |
| --- |
|  |

Gunakan [titanic.csv](http://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv) yang berisi data 887 penumpang Titanic passengers. Kolom data menggambarkan survived (*S*), age (*A*), passenger-class (*C*), sex (*G*) and the fare paid (*X*).

Hitung peluang bersyarat (conditional probability) di bawah ini   
*P*(*S*= true | *G*=female)

*P*(*S*= true | *G*=male)

*P*(*S*= true | *C*=1)

*P*(*S*= true | *C*=2)

*P*(*S*= true | *C*=3)

*P*(*S*= true | *G*=female,*C*=1) =

*P*(*S*= true | *G*=female,*C*=2) =  
*P*(*S*= true | *G*=female,*C*=3) =  
*P*(*S*= true | *G*=male,*C*=1) =  
*P*(*S*= true | *G*=male,*C*=2) =  
*P*(*S*= true | *G*=male,*C*=3) =

Gunakan R dan excel :

|  |
| --- |
| **Excel**    **R**  > tdf %>%summarize(prob = sum(Survived == 1 & Sex == "female", na.rm = TRUE)/sum(Survived==1, na.rm = TRUE)\*(sum(Survived==1,na.rm=TRUE)/sum(Survived==1 | Survived==0 ,na.rm = TRUE)))  prob  1 0.2615039  > tdf %>%summarize(prob = sum(Survived == 1 & Sex == "male", na.rm = TRUE)/sum(Survived==1, na.rm = TRUE)\*(sum(Survived==1,na.rm=TRUE)/sum(Survived==1 | Survived==0 ,na.rm = TRUE)))  prob  1 0.1223345  > tdf %>%summarize(prob = sum(Survived == 1 & Pclass == "1", na.rm = TRUE)/sum(Survived==1, na.rm = TRUE)\*(sum(Survived==1,na.rm=TRUE)/sum(Survived==1 | Survived==0 ,na.rm = TRUE)))  prob  1 0.1526375  > tdf %>%summarize(prob = sum(Survived == 1 & Pclass == "2", na.rm = TRUE)/sum(Survived==1, na.rm = TRUE)\*(sum(Survived==1,na.rm=TRUE)/sum(Survived==1 | Survived==0 ,na.rm = TRUE)))  prob  1 0.0976431  > tdf %>%summarize(prob = sum(Survived == 1 & Pclass == "3", na.rm = TRUE)/sum(Survived==1, na.rm = TRUE)\*(sum(Survived==1,na.rm=TRUE)/sum(Survived==1 | Survived==0 ,na.rm = TRUE)))  prob  1 0.1335578  > tdf %>%summarize(prob = sum(Survived == 1 & Sex=="female"& Pclass == "1", na.rm = TRUE)/sum(Survived==1, na.rm = TRUE)\*(sum(Survived==1,na.rm=TRUE)/sum(Survived==1 | Survived==0 ,na.rm = TRUE)))  prob  1 0.1021324  > tdf %>%summarize(prob = sum(Survived == 1 & Sex=="female"& Pclass == "2", na.rm = TRUE)/sum(Survived==1, na.rm = TRUE)\*(sum(Survived==1,na.rm=TRUE)/sum(Survived==1 | Survived==0 ,na.rm = TRUE)))  prob  1 0.07856341  > tdf %>%summarize(prob = sum(Survived == 1 & Sex=="female"& Pclass == "3", na.rm = TRUE)/sum(Survived==1, na.rm = TRUE)\*(sum(Survived==1,na.rm=TRUE)/sum(Survived==1 | Survived==0 ,na.rm = TRUE)))  prob  1 0.08080808  > tdf %>%summarize(prob = sum(Survived == 1 & Sex=="male"& Pclass == "1", na.rm = TRUE)/sum(Survived==1, na.rm = TRUE)\*(sum(Survived==1,na.rm=TRUE)/sum(Survived==1 | Survived==0 ,na.rm = TRUE)))  prob  1 0.05050505  > tdf %>%summarize(prob = sum(Survived == 1 & Sex=="male"& Pclass == "2", na.rm = TRUE)/sum(Survived==1, na.rm = TRUE)\*(sum(Survived==1,na.rm=TRUE)/sum(Survived==1 | Survived==0 ,na.rm = TRUE)))  prob  1 0.01907969  > tdf %>%summarize(prob = sum(Survived == 1 & Sex=="male"& Pclass == "3", na.rm = TRUE)/sum(Survived==1, na.rm = TRUE)\*(sum(Survived==1,na.rm=TRUE)/sum(Survived==1 | Survived==0 ,na.rm = TRUE)))  prob  1 0.05274972 |

Jalankan perintah R di bawah ini :

|  |
| --- |
| # https://www.kaggle.com/brirush/naive-bayes-for-titanic  # This R script will run on our backend. You can write arbitrary code here!  # Many standard libraries are already installed, such as e1071, which has naiveBayes  library(e1071)  # The train and test data is stored in the ../input directory  > train <- read.csv("D:/Semester 5/Data mining/Praktikum/prak4/train.csv")  > test <- read.csv("D:/Semester 5/Data mining/Praktikum/prak4/test.csv")  # We can inspect the train data. The results of this are printed in the log tab below  #summary(train)  BayesTitanicModel<-naiveBayes(as.factor(Survived)~., train)  str(BayesTitanicModel)  BayesPrediction<-predict(BayesTitanicModel, test)  str(BayesPrediction)  summary(BayesPrediction)  output<-data.frame(test$PassengerId, BayesPrediction)  str(output)  colnames(output)<-cbind("PassengerId","Survived")  write.csv(output, file = 'Rushton\_Solution.csv', row.names = F) |

Output :

|  |
| --- |
| Output file csv |

Berikan penjelasan terhadap output di atas

|  |
| --- |
| Hasil output dari data perhitungan prediksi BayesPrediction diekspor dalam excel Rusthon\_Solution yang hanya diambil data Training dan test.  Data ini menunjukkan prediksi untuk Id penumpang selanjutnya dengan status Survivednya masing-masing. |

Sumber :

<http://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html>

<http://rstudio-pubs-static.s3.amazonaws.com/6595_b57093a21dfc46a4b3338cfee29ec61e.html>

<https://community.rstudio.com/t/conditional-probability-with-dplyr/5117>

<https://www.kaggle.com/brirush/naive-bayes-for-titanic>