

## Pokok Bahasan VII

### Naïve Bayes

**Kode Pokok Bahasan:** TIK.RPL03.005.00.01

**Deskripsi Pokok Bahasan:**

Membahas bagaimana penerapan Algoritma Naïve Bayes pada dataset titanic.

No	Elemen Kompetensi	Indikator Kinerja	Jml Jam	Hal
1.	Menampilkan peluang dari kasus yang diberikan.	Mampu melakukan analisis terhadap peluang atas kejadian yang ditentukan.	1	12
2.	Menggunakan fungsi naiveBayes dalam memprediksi data	Mengimplementasikan fungsi naive bayes pada prediksi data		

#### TUGAS PENDAHULUAN

Hal yang harus dilakukan dan acuan yang harus dibaca sebelum praktikum :

1. Menginstal R pada PC masing-masing praktikan.
2. Menginstal R Studio pada PC masing-masing praktikan.

#### DAFTAR PERTANYAAN

1. Apa itu algoritma Naïve Bayes?

Algoritma Naive Bayes adalah algoritma yang mempelajari probabilitas suatu objek dengan ciri-ciri tertentu yang termasuk dalam kelompok/kelas tertentu. Singkatnya, ini adalah pengklasifikasi probabilistik.

2. Apa kegunaan Naïve Bayes?

Prediksi multi-kelas: Algoritma klasifikasi Naive Bayes dapat digunakan untuk memprediksi probabilitas posterior dari beberapa kelas variabel target. Klasifikasi teks: Karena fitur prediksi multi-kelas, Naive Bayes algoritma klasifikasi sangat cocok untuk klasifikasi teks.

3. Sebutkan tahapan dari proses algoritma Naïve Bayes!

Menghitung jumlah kelas/label. Menghitung jumlah kasus perkelas. Mengalikan semua hasil variable kelas. Membandingkan hasil perkelas.

#### TEORI SINGKAT

Algoritma Naive Bayes merupakan sebuah metoda klasifikasi menggunakan metode probabilitas dan statistik yg dikemukakan oleh ilmuwan Inggris Thomas Bayes. Algoritma Naive Bayes memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Ciri

utama dr Naïve Bayes Classifier ini adalah asumsi yg sangat kuat (naïf) akan independensi dari masing-masing kondisi / kejadian.

Naive Bayes Classifier bekerja sangat baik dibanding dengan model classifier lainnya. Hal ini dibuktikan pada jurnal Xhemali, Daniela, Chris J. Hinde, and Roger G. Stone. "Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages." (2009), mengatakan bahwa "Naïve Bayes Classifier memiliki tingkat akurasi yg lebih baik dibanding model classifier lainnya".

Keuntungan penggunaan adalah bahwa metoda ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menentukan estimasi parameter yg diperlukan dalam proses pengklasifikasian. Karena yg diasumsikan sebagai variabel independent, maka hanya varians dari suatu variabel dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarians.

## **LAB SETUP**

Hal yang harus disiapkan dan dilakukan oleh praktikan untuk menjalankan praktikum modul ini.

1. Menginstall library yang dibutuhkan untuk mengerjakan modul.
2. Menjalankan R Studio.

## **ELEMEN KOMPETENSI I**

### **Deskripsi:**

Menampilkan peluang dari kasus yang diberikan.

### **Kompetensi Dasar:**

Mampu melakukan analisis terhadap peluang atas kejadian yang ditentukan.

### **Latihan**

### **Penjelasan Singkat :**

Pada latihan ini anda akan diminta untuk menampilkan summarize dari data menggunakan library yang disediakan oleh R.

### **Langkah-Langkah Praktikum:**

1. Instal dan panggil package berikut jika belum terinstall

```
install.packages("tidyverse")  
library("tidyverse") #for data wrangling tools  
install.packages("titanic")  
library("titanic")
```

2. Inisialisasi titanic\_train ke dalam variabel tdf.

```
tdf <- titanic_train #training set of Titanic data
head(tdf)
```

3. Lakukan perintah berikut :

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 :

```
> head(tdf)
  PassengerId Survived Pclass Name Sex Age SibSp Parch
1           1         0      3 Braund, Mr. Owen Harris male 22     1     0
2           2         1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38     1     0
3           3         1      3 Heikkinen, Miss. Laina female 26     0     0
4           4         1      1 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35     1     0
5           5         0      3 Allen, Mr. William Henry male 35     0     0
6           6         0      3 Moran, Mr. James male NA     0     0

  Ticket Fare Cabin Embarked
1  A/5 21171  7.2500      S
2  PC 17599 71.2833      C85
3 STON/O2. 3101282  7.9250      S
4  113803 53.1000    C123
5  373450  8.0500      S
6  330877  8.4583      Q

> tdf %>%
+   summarize(prob = sum(Age >= 35 & Sex == "female", na.rm = TRUE)/sum(Age >= 35, na.rm = TRUE))
# A tibble: 1 x 2
#   prob
#   <dbl>
1 0.3446809
>
```

## ELEMEN KOMPETENSI II

### Deskripsi:

Menggunakan fungsi naïve bayes dalam memprediksi data.

### Kompetensi Dasar:

Mengimplementasikan fungsi naive bayes pada prediksi data

### Latihan 1.2.1

#### Penjelasan Singkat :

Pada latihan ini anda akan diminta untuk mengimplementasikan naïve bayes pada kasus yang diberikan.

#### Langkah-Langkah Praktikum:

Gunakan [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

$$P(S = \text{true} \mid G = \text{female})$$

$P(S = \text{true} \mid G = \text{male})$

$P(S = \text{true} \mid C = 1)$

$P(S = \text{true} \mid C = 2)$

$P(S = \text{true} \mid C = 3)$

$P(S = \text{true} \mid G = \text{female}, C = 1) =$

$P(S = \text{true} \mid G = \text{female}, C = 2) =$

$P(S = \text{true} \mid G = \text{female}, C = 3) =$

$P(S = \text{true} \mid G = \text{male}, C = 1) =$

$P(S = \text{true} \mid G = \text{male}, C = 2) =$

$P(S = \text{true} \mid G = \text{male}, C = 3) =$

Gunakan R :

```
> #S=True/G=Female
> tdf %>%
+ summarize(prob = sum(Survived == "1" & Sex == "female", na.rm = TRUE)/sum(Survived == "1", na.rm = TRUE))
      prob
1 0.6812865
> #S=True/G=Male
> tdf %>%
+ summarize(prob = sum(Survived == "1" & Sex == "male", na.rm = TRUE)/sum(Survived == "1", na.rm = TRUE))
      prob
1 0.3187135
> #S=True/C=1
> tdf %>%
+ summarize(prob = sum(Survived == "1" & Pclass == "1", na.rm = TRUE)/sum(Survived == "1", na.rm = TRUE))
      prob
1 0.3976608
> #S=True/C=2
> tdf %>%
+ summarize(prob = sum(Survived == "1" & Pclass == "2", na.rm = TRUE)/sum(Survived == "1", na.rm = TRUE))
      prob
1 0.254386
> #S=True/C=3
> tdf %>%
+ summarize(prob = sum(Survived == "1" & Pclass == "3", na.rm = TRUE)/sum(Survived == "1", na.rm = TRUE))
      prob
1 0.3479532
> #S=True/G=Female/C=1
> tdf %>%
+ summarize(prob = sum(Survived == "1" & Sex == "female" & Pclass == "1", na.rm = TRUE)/sum(Survived == "1", na.rm = TRUE))
      prob
1 0.2660819
> #S=True/G=Female/C=2
> tdf %>%
+ summarize(prob = sum(Survived == "1" & Sex == "female" & Pclass == "2", na.rm = TRUE)/sum(Survived == "1", na.rm = TRUE))
      prob
1 0.2046784
> #S=True/G=Female/C=3
> tdf %>%
+ summarize(prob = sum(Survived == "1" & Sex == "female" & Pclass == "3", na.rm = TRUE)/sum(Survived == "1", na.rm = TRUE))
      prob
1 0.2105263
> #S=True/G=Male/C=1
> tdf %>%
+ summarize(prob = sum(Survived == "1" & Sex == "male" & Pclass == "1", na.rm = TRUE)/sum(Survived == "1", na.rm = TRUE))
      prob
1 0.1315789
> #S=True/G=Male/C=2
> tdf %>%
+ summarize(prob = sum(Survived == "1" & Sex == "male" & Pclass == "2", na.rm = TRUE)/sum(Survived == "1", na.rm = TRUE))
      prob
1 0.0497076
> #S=True/G=Male/C=3
> tdf %>%
+ summarize(prob = sum(Survived == "1" & Sex == "male" & Pclass == "3", na.rm = TRUE)/sum(Survived == "1", na.rm = TRUE))
      prob
1 0.1374269
> |
```

Jalankan perintah R di bawah ini :

```
# https://www.kaggle.com/brirush/naive-bayes-for-titanic
library(e1071)
train <- read.csv("F:/Kuliah Data Mining gasal 1819/Kaggle/Titanic/train.csv")
```

```

test <- read.csv("F:/Kuliah Data Mining gasal 1819/Kaggle/Titanic/test.csv")
BayesTitanicModel<-naiveBayes(as.factor(Survived)~., train)
BayesPrediction<-predict(BayesTitanicModel, test)
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 :

```

> # https://www.kaggle.com/brirush/naive-bayes-for-titanic
> library(e1071)
> ibnu_train <- read.csv("D:/File Kuliah Semester 5/Penambahan Data/Prak-7/train_prak7.csv")
> ibnu_test  <- read.csv("D:/File Kuliah Semester 5/Penambahan Data/Prak-7/test_prak7.csv")
> BayesTitanicModel<-naiveBayes(as.factor(Survived)~., ibnu_train)
> BayesPrediction<-predict(BayesTitanicModel, ibnu_test)
> summary(BayesPrediction)
 0    1
301 117
> output<-data.frame(ibnu_test$PassengerId, BayesPrediction)
> str(output)
'data.frame':  418 obs. of  2 variables:
 $ ibnu_test.PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
 $ BayesPrediction      : Factor w/ 2 levels "0","1": 1 1 1 2 1 2 1 2 1 2 ...
> colnames(output)<-cbind("PassengerId","Survived")
> write.csv(output, file = 'Rushton_Solution.csv', row.names = F)
> |

```

PassengerId	Survived
892	0
893	0
894	0
895	0
896	1
897	0
898	1
899	0
900	1
901	0
902	0
903	0
904	1
905	0
906	1
907	1
908	0
909	0
910	0
911	1
912	1
913	0
914	1
915	1
916	1
917	0
918	1

Berikan penjelasan terhadap output di atas

Mengeluarkan data dari File, mengecek data dari penumpang dan status nya survived atau tidak.

## Tugas : Kasus “playing golf”

Data [excelNaive](#)

id	Outlook	Temp	Humidity	Windy	PlayGolf
1	Rainy	Hot	High	FALSE	No
2	Rainy	Hot	High	TRUE	No
3	Overcast	Hot	High	FALSE	Yes
4	Sunny	Mild	High	FALSE	Yes
5	Sunny	Cool	Normal	FALSE	Yes
6	Sunny	Cool	Normal	TRUE	No
7	Overcast	Cool	Normal	TRUE	Yes
8	Rainy	Mild	High	FALSE	No
9	Rainy	Cool	Normal	FALSE	Yes
10	Sunny	Mild	Normal	FALSE	Yes
11	Rainy	Mild	Normal	TRUE	Yes
12	Overcast	Mild	High	TRUE	Yes
13	Overcast	Hot	Normal	FALSE	Yes
14	Sunny	Mild	High	TRUE	No

Data [excelNaiveTest](#)

id	Outlook	Temp	Humidity	Windy
16	Rainy	Mild	Normal	TRUE

Script R :

```
> excelNaive <- read.delim('clipboard')
> excelNaiveTest <- read.delim('clipboard')
> excelNaiveModel <- naiveBayes(as.factor(PlayGolf)~., excelNaive)
> excelNaivePredic <- predict(excelNaiveModel, excelNaiveTest)
> summary(excelNaivePredic)
> excelNaiveOutput <- data.frame(excelNaiveTest$id, excelNaivePredic)
> str(excelNaiveOutput)
> colnames(excelNaiveOutput) <- cbind('id', 'PlayGolf')
> write.csv(excelNaiveOutput, file = "rainy.csv", row.names = F)
```

Output :

```
> #TUGAS
> excelNaive <- read.delim('clipboard')
> excelNaiveTest <- read.delim('clipboard')
> excelNaiveModel <- naiveBayes(as.factor(PlayGolf)~., excelNaive)
> excelNaivePredic <- predict(excelNaiveModel, excelNaiveTest)
> summary(excelNaivePredic)
  No  Yes
0    1
> excelNaiveOutput <- data.frame(excelNaiveTest$id, excelNaivePredic)
> str(excelNaiveOutput)
'data.frame':  1 obs. of  2 variables:
 $ excelNaiveTest.id: int 16
 $ excelNaivePredic : Factor w/ 2 levels "No","Yes": 2
> colnames(excelNaiveOutput) <- cbind('id', 'PlayGolf')
> write.csv(excelNaiveOutput, file = "rainy.csv", row.names = F)
> |
```

Hitungan manual data test :

	A	B
1	id,"PlayGolf"	
2	16,"Yes"	
3		

Setelah mengolah data dan survei terhadap data-data yang sudah ada, ketika id baru masuk dengan data yang ada maka terbukti lah data yang baru masuk seperti diatas

Sumber :

<http://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html>  
[http://rstudio-pubs-static.s3.amazonaws.com/6595\\_b57093a21dfc46a4b3338cfee29ec61e.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>  
<https://www.geeksforgeeks.org/naive-bayes-classifiers/>

## CEK LIST

Elemen Kompetensi	No Latihan	Penyelesaian	
		Selesai	Tidak selesai
1	1.1.1	✓	

2	1.2.1	✓	
---	-------	---	--

## FORM UMPAN BALIK

Elemen Kompetensi	Tingkat Kesulitan	Tingkat Ketertarikan	Waktu Penyelesaian dalam menit
Memahami data pre-processing.	<input type="checkbox"/> Sangat Mudah <input type="checkbox"/> Mudah <input checked="" type="checkbox"/> Biasa <input type="checkbox"/> Sulit <input type="checkbox"/> Sangat Sulit	<input type="checkbox"/> Tidak Tertarik <input type="checkbox"/> Cukup Tertarik <input type="checkbox"/> Tertarik <input checked="" type="checkbox"/> Sangat Tertarik	
Mengimplementasikan pre-processing data.	<input type="checkbox"/> Sangat Mudah <input type="checkbox"/> Mudah <input checked="" type="checkbox"/> Biasa <input type="checkbox"/> Sulit <input type="checkbox"/> Sangat Sulit	<input type="checkbox"/> Tidak Tertarik <input type="checkbox"/> Cukup Tertarik <input type="checkbox"/> Tertarik <input checked="" type="checkbox"/> Sangat Tertarik	

Elemen Kompetensi	Tingkat Kesulitan	Tingkat Ketertarikan	Waktu Penyelesaian dalam menit
Menggunakan fungsi naiveBayes dalam memprediksi data	<input type="checkbox"/> Sangat Mudah <input type="checkbox"/> Mudah <input checked="" type="checkbox"/> Biasa <input type="checkbox"/> Sulit <input type="checkbox"/> Sangat Sulit	<input type="checkbox"/> Tidak Tertarik <input type="checkbox"/> Cukup Tertarik <input type="checkbox"/> Tertarik <input checked="" type="checkbox"/> Sangat Tertarik	



