**PERTEMUAN 2 DAN 3**

# KLASIFIKASI BAYES

## TUJUAN PRAKTIKUM

1. Mahasiswa mampu mengekplorasi Data lebih lanjut (dalam bentuk citra)
2. Mahasiswa mampu memahami dan menjelaskan mengenai konsep Bayes dalam pengenalan Pola
3. Mahasiswa mampu mengimplementasikan Algoritme Bayes untuk klasifikasi Data sederhana menggunakan R

## TEORI PENUNJANG DAN MATERI PRAKTIKUM

**Eksplorasi dan Analisis Data Lanjut**

Berikutnya kita membahas data dalam bentuk Citra. Untuk pengolahan data citra, kita membutuhkan library EBImage. Install library EBImage terlebih dahulu (link https://www.bioconductor.org/packages/release/bioc/html/EBImage.html), kemudian ikuti instruksi di bawah ini.

```
library(EBImage)
x <- readImage(system.file('images','shapes.png', package='EBImage'))
x <- x[110:312,1:130]
y <- bwlabel(x)
display(y, title='Binary')
```

Memanggil fungsi library EBImage , gambar dibaca menggunakan fungsi readImage dengan mengambil file shapes.png sebagai masukan, memilih posisi gambar pada pixels dengan koordinat (110:310, 1:30) kemudian objek gambar x dianggap sebagai gambar biner dimana piksel nilai 0 dianggap sebagai piksel latar belakang dan piksel lain sebagai latar depan. kemudian ditampilkan hasilnya.

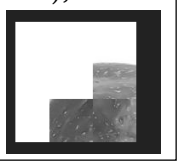

```
#Unduh citra buah (cth:alpukat) dari internet,dan baca melalui R

original_image  <-  readImage(file.choose())
display(original_image)
r = channel(original_image,"r")
g = channel(original_image,"g")
b = channel(original_image,"b")
new_image = 0.2126*r+0.7152*g+0.0722*b
display(new_image)
```

Membaca gambar dengan cara memilih file image pada direktori, kemudian ditampilkan, mengekstrak nilai dari channel R,G,B kedalam variabel baru yaitu r,g dan b kemudian mengubah nilai setiap channel red, green dan blue dengan nilai tertentu (luminance) kemudian ditampilkan

```
Dataimage <- new_image@.Data
Subdata1 <- Dataimage[110:312,130:200]
display(Subdata1)
Subdata2<- Dataimage[c(1:40, 100:150, 350:400 ), c(1:40, 100:150, 250:30
0 )]
display(Subdata2)
```

Answer 3.
Mengekstrak data dari citra kemudian mengambil subset dari citra tersebut pada koordinat piksel (110:312, 130:200) kemudian tampilkan
mengambil subset dari citra sebanyak sembilan segment yang koordinat pikselnya ditentukan melalui operasi matriks sehingga menghasilkan segementasi (1:40,1:40), (1:40,100:150 ), (1:40,250:300), (100:150,1:40),(100:150,100:150),(100:150,250:300), (350:400,1:40), (350:400,100:150), (350:400,250:300) kemudian tampilkan



```
# Unduh citra buah lain, dan lakukakan langkah yang sama dengan sebelumnya
# Ekstrak nilai citra dengan nama DataImage2
Dataimage2 <- Dataimage2[1:dim(Dataimage)[1], 1:dim(Dataimage)[2]]
obs1 <- as.vector(t(Dataimage))
obs2 <- as.vector(t(Dataimage2))
obs_gabung <- rbind(obs1,obs2)
dataset_buah <- as.data.frame(obs_gabung)
klas<- c("alpokat", "apel")
dataset_buah_baru<-cbind(dataset_buah, klas)
dim(dataset_buah_baru)
dataset_buah_baru[1,640001]
dataset_buah_baru[2,640001]
```

**Question 4. Jelaskan apa maksud dari potongan kode di atas? Copy hasil display ke dalam box di bawah ini**

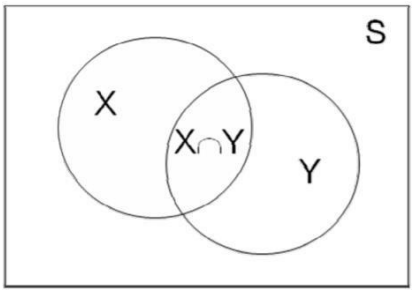
**Konsep dan Definisi Metode Bayes**

Pada Pengenalan Pola, beberapa metode sederhana contohnya metode find-S (berbasis hipotesa) dapat digunakan untuk mengklasifikasikan target suatu data baru, jika *dataset* yang dipakai dalam tahap pembelajaran adalah data dengan sifat konsisten dan tidak bias. Namun pada kenyataanya sulit untuk menemukan bentuk data yang konsisten dan tidak bias. Dalam praktiknya *dataset* selalu memiliki ketidakkonsistenan dan memiliki bias karena variasi yang terjadi pada data. Untuk mengatasi masalah tersebut salah satu metode dinamakan metode bayes diperkenalkan. Motede ini merupakan pendekatan statistik untuk melakukan inferensi

induksi pada persoalan klasifikasi tanpa harus melihat sifat datanya tetapi hubungan sebab akibat antara target dan atributnya (probabilitas bersyarat).

Metode Bayes menggunakan propabilitas bersyarat sebagai dasarnya. Dalam ilmu probabilitas bersyarat dinyatakan sebagai:

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$




Probabilitas X di dalam Y adalah probabilitas inteseksi X dan Y dari probabilitas Y, atau dengan bahasa lain P(X|Y) adalah prosentase banyaknya X di dalam Y. Probabilitas bersyarat dalam data diilustrasikan pada contoh berikut.

| Day | Cuaca | Temperatur | Kecepatan Angin | Berolah-raga |
|---|---|---|---|---|
| D1 | Cerah | Normal | Pelan | Ya |
| D2 | Cerah | Normal | Pelan | Ya |
| D3 | Hujan | Tinggi | Pelan | Tidak |
| D4 | Cerah | Normal | Kencang | Ya |
| D5 | Hujan | Tinggi | Kencang | Tidak |
| D6 | Cerah | Normal | Pelan | Ya |

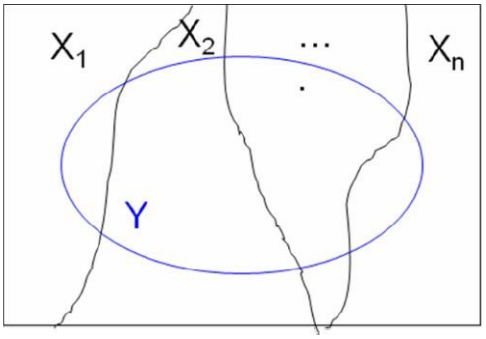
Banyaknya data "Berolah-raga = Ya" adalah 4 dari 6 data maka dituliskan P(Berolah-raga)= 4/6. Banyaknya data dengan atribut Cuaca Cerah dan target Berolah-raga Ya, adalah 4 dari 6 data, maka dituliskan P(cuaca=cerah dan Olahraga=ya)= 4/6.

**Answer . 5 Probabilitas cuca cerah pada saat olahraga adalah**

**P ( cuaca = cerah | olahraga = ya ) = (4/6) /(4/6) = 1**

**Naive Bayes /Hypothesis Maximum Appropri Probability (HMAP)**

Terminologi dari HMAP menyatakan hipotesa (keadaan Posterior /Probabilitas Xk di dalam Y) yang diambil/diputuskan didasarkan pada nilai probabilitas kondisi prior yang diketahui (Probabilitas Y di dalam Xk dibagi dengan jumlah probabilitas Y dalam semua Xi).




$$P(S \mid X) = \underset{x \in X}{\operatorname{argmax}} \frac{P(Y \mid X)\, P(X)}{\sum_{'} P(X)}$$

$$= \underset{x \in X}{\operatorname{argmax}} \; P(Y \mid X)\;\; P(X)$$

HMAP adalah model penyederhanaan dari metode bayes yang sering disebut sebagai metode ***Naive Bayes.*** HMAP dapat digunakan sebagai metode untuk mendapatkan hipotesis dari suatu keputusan. HMAP dapat diartikan untuk mencari probabilitas terbesar dari semua *instance* pada atribut, target atau semua kemungkinan keputusan.

**Contoh 1:**

Diketahui hasil survey yang dilakukan sebuah lembaga kesehatan menyatakan bahwa 30% penduduk di dunia menderita sakit paru-paru. Dari 90% penduduk yang sakit paru-paru 60% adalah perokok, dan dari penduduk yang tidak sakit paru-paru 20% adalah perokok.

Fakta hasil survey didefinisikan sebagai berikut:

| | | | |
|---|---|---|---|
| Y | : Sakit paru-paru | X | : Perokok |
| ~Y | : Tidak sakit Paru | ~X | : Bukan Perokok |

**Answer. 6**

Peluang orang sakit paru: P(Y) = 0.9
Peluang orang tidak sakit paru: P(-Y) = 0.1
P(X|Y)=0.6 P(~X|Y) 0.4
P(X|~Y)=0.2 P(~X|~Y) = 0.8
P(S|Y) = argmax (x Y) P({X}|Y) P(Y)
= max( P(X|Y) P(Y), P(X|~Y)P(~Y))
= max (0.6 x 0.9, 0.2 x 0.1 ) = max (0.54, 0.02)
karena 0.54 terpilih yang menyatakan seorang menderita sakit paru lebih besar dari seorang yang tidak menderita sakit paru maka disimpulkan seorang akan menderita sakit paru jika dia perokok

**Contoh 2:**

| # | Cuaca | Temperatur | Kecepatan Angin | Berolah-raga |
|---|---|---|---|---|
| 1 | Cerah | Normal | Pelan | Ya |
| 2 | Cerah | Normal | Pelan | Ya |
| 3 | Hujan | Tinggi | Pelan | Tidak |
| 4 | Cerah | Normal | Kencang | Ya |
| 5 | Hujan | Tinggi | Kencang | Tidak |
| 6 | Cerah | Normal | Pelan | Ya |

Answer 7.
Atribut : X1 = cuaca, X2 = Temperatur, X3 = Kecepatan angin.
Fakta menunjukkan
P(Y=Ya) = 4/6 P(Y=Tidak) = 2/6
P(X1= Cerah | Y=Ya)=1 P(X1= Cerah | Y=Tidak) = 0
P(X2= Tinggi | Y=Ya)= 0 P(X2 = Tinggi | Tidak ) = 1
P (X3=Kencang|Y=Ya)=1/4 P (X3=Kencang|Y=Tidak)=1/2
bila Cuaca cerah, Temperatur tinggi dan Kecepatan angin kencang maka:
P( X1=cerah,X2=tinggi dan X3=kencang | Y=ya )
= { P(X1=cerah|Y=ya).P(X2=tinggi|Y=ya).P(X3=kencang|Y=ya) } . P(Y=ya)
= { (1) . (0). (1/4) } . (4/6) = 0
P( X1=cerah,X2=Tinggi dan X3=kencang | Y=tidak )
= { P(X1=cerah|Y=tidak).P(X2=tinggi|Y=tidak)P(X3=kencang|Y=tidak) } . P(Y=tidak)
= { (0) . (1).(1/2) } . (2/6) = 0 Kesimpulan : kedua bernilai nol

**Naive Bayes untuk Data Numerik**

Contoh sebelumnya merupakan Data dengan atribut dengan bentuk data berupa data nominal. Sedangkan untuk data numerik (Kontinyu) maka perlu dilakukan diskretisasi data atau menggunakan asumsi distribusi dari data atributnya. Diskretisasi adalah mentransformasi nilai numerik dari data atribut ke bentuk interval (*categorical counterparts)*/ *binning* sebelum diolah dalam tabel frekuensi. Dan cara berikutnya adalah menghitung fungsi densitas peluang dari data atribut (dalam bentuk tebakan terhadap asumsi distribusi peluangnya), yang biasakan distribusi nya diasumsikan dalam bentuk sebaran normal dari atribut data Numerik.

Untuk Fungsi Likelihood dari data numerik (menggunakan asumsi distribusi normal) adalah:

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma^2 ij}}$$

Di mana :
$P$ : Peluang
$Xi$ : Atribut ke $i$
$xi$ : Nilai atribut ke $i$
$Y$ : Kelas yang dicari
$yi$ : Sub kelas $Y$ yang dicari
$\mu$ : *mean*, menyatakan rata – rata dari seluruh atribut
$\sigma$ :Deviasi standar, menyatakan varian dari seluruh atribut.

Contoh 3:

| | | Humidity | *Mean* | *StDev* |
|---|---|---|---|---|
| **Play** | yes | 86 96 80 65 70 80 70 90 75 | 79.1 | 10.2 |
| **Golf** | no | 85 90 70 95 91 | 86.2 | 9.7 |

**Answer 8.**
**P(humidity=74 | play=yes) = 1/√(2π(10,2)) e^( 〖-(74-79,1)〗 ^2/ 〖2(10,2)〗 ^2 ) = 0,0344**
**P(humidity=74 | play=no) = 1/√(2π(9,7)) e^( 〖-(74-86,2)〗 ^2/ 〖2(9,7)〗 ^2 ) = 0,0187**

## IMPLEMENTASI DALAM R

### Partisi dan pembagian Data

Data yang disiapkan pada tahap pembelajaran untuk membuat suatu model pembelajaran menggunakan metode yang ada di machine learning/data mining disebut sebagai data training, sedangkan data yang digunakan pada saat pengujian model dan validasi hasil model disebut sebagai data testing. Pembagaian dataset menjadi data training dan data testing merupakan hal yang penting. Untuk Dataset yang berukuran besar proporsi yang digunakan pada pembagian data tidak menjadi masalah dan validasi model dan juga uji data juga cukup mudah dilakukan. Untuk Data yang ukurannya tidak besar, proporsi data training untuk tahap pembalajaran dan validasi modelnya bisa diakali dengan penggunaan teknik *cross validation*.

Berikut ini adalah cara-cara sederhana untuk partisi data.

```
> #mengambil dataset dari url

> url =
"https://vincentarelbundock.github.io/Rdatasets/csv/MASS/Pima
.te.csv"
```

**Answer . 9**

```
#mengambil dataset dari url
url = "https://vincentarelbundock.github.io/Rdatasets/csv/MASS/
Pima.te.csv"
datadiabet <- read.csv(url, header= TRUE)
View (datadiabet)
head (datadiabet)
jmlbrs <- nrow(datadiabet)
# Split dataset menjadi dtraining dan dtesting
# Proporsi dtraining adalah 80%, dtesting adalah 20%
datasample <- sample(2, jmlbrs, replace = TRUE, prob= c(0.8,0.2))
dtraining <- datadiabet[datasample == 1,]
dtesting <- datadiabet[datasample == 1,]
```

**Partisi data dengan *cross validation***

Diperlukan untuk menginstal ***package* cvTools**

**Answer 10.**
**library(cvTools)**
**library(lattice)**
**library (robustbase)**
**fold <- cvFolds(nrow(datadiabet), K = 8, R = 1, type = "random")**
**# 8-fold cross validation**
**k=1**
**dtrain <- datadiabet[fold$subsets[fold$which != k], ]**
**dtes <- datadiabet[fold$subsets[fold$which == k], ]**
**View (dtrain)**

**Naive Bayes**

Untuk menerapkan Naive Bayes bisa digunakan package e1071 yang sudah menyediakan fungsi naiveBayes.

```
> library(e1071)
> data(iris)
> head(iris)
> datalatih<-iris[c(1:40,  51:90,  101:140),]
> datauji<-iris[-c(1:40,  51:90,  101:140),]
> # Membuat model Naïve Bayes menggunakan datalatih
>  model.nB <- naiveBayes(Species ~ ., datalatih)
> # Prediksi datauji / data baru
> predict(model.nB, datauji[,-5])
> predict(model.nB, datauji[,-5], type="raw")
> table(predict=predict(model.nB, datauji[, -5]), true=datauji[,5])
```

## TUGAS PRAKTIKUM

**Dikerjakaan dan dikumpulkan maksimal pekan depan**

## DAFTAR PUSTAKA

1. Victor A. Bloomfield. 2014. *Using R for Numerical Analysis in Science and Engineering*. 1 edition. Chapman and Hall/CRC