PERTEMUAN 2

TIPE DATA DAN EKSPLORASI DATA MENGGUNAKAN WEKA DAN R

TUJUAN PRAKTIKUM

Mahasiswa akan dapat memahami Tipe data, Eksplorasi Data, Statistika ringkasan, Visualisasi menggunakan Weka dan R.

TEORI PENUNJANG

Eksplorasi Data

Eksplorasi data merupakan langkah untuk memahami data sebelum dilakukan praproses. Pemahaman data yang akan di-*mining* dapat membantu dalam menentukan teknik-teknik praproses dan analisa data terhadap data sebelum dilakukan tugas data mining. Dalam eksplorasi data, hal yang harus diperhatikan yaitu tipe data. Tipe data terdiri atas empat jenis, yaitu nominal, ordinal, interval, dan rasio. Penjelasan dan perbedaan setiap tipe data tersebut dapat dilihat pada Tabel 1.

Eksplorasi data dilakukan sebagai langkah awal untuk mengetahui karakteristik dari data. Tahapan ini dilakukan untuk menyeleksi teknik pemrosesan dan analisis data yang sesuai untuk dataset yang kita miliki. Eksplorasi data menggunakan R dimulai dengan memeriksa dimensi data, mencari statistika ringkasan, serta menampilkan berbagai jenis grafik seperti pie chart dan histogram.

Tabel 1 Deskripsi tipe data dan contohnya

Tipe data		**Deskripsi**	**Contoh**
Kategorik (kualitatif)	Nominal	Nilai dari atribut nominal adalah nama-nama sebagai pembeda antara satu dengan yang lain. Nilai nominal menyediakan informasi yang cukup untuk membedakan satu objek dengan objek yang lain. Operator aritmatik yang dapat digunakan ialah sama dengan (=) atau tidak sama dengan (≠).	PIN kartu ATM , NIM Mahasiswa, warna RGB , kode biner.
	Ordinal	Nilai dari atribut ordinal adalah nama-nama yang selain sebagai pembeda juga dapat menjadi ukuran perbandingan satu dengan yang lain. Nilai ordinal menyediakan informasi yang cukup untuk mengurutkan objek. Operator aritmatik yang dapat digunakan ialah sama dengan (=),tidak sama dengan (≠)lebih besar (>) , lebih kecil (<).	Nomor antrian, *Grade*, kecepatan prosesor {lambat, cepat, sangat cepat}

Tipe data		Deskripsi	Contoh
Numerik (kuantitatif)	Interval	Dalam atribut interval perbedaan antarnilai merupakan sesuatu yang berarti, pada atribut interval terdapat unit pengukuran. Operator aritmatik yang dapat digunakan ialah sama dengan (=),tidak sama dengan (≠)lebih besar (>) , lebih kecil (<), penambahan (+), pengurangan (-).	Tanggal pada kalender, Temperatur.
	Rasio	Dalam atribut rasio, perbedaan rasio merupakan hal yang berarti. Operator aritmatik yang dapat digunakan ialah sama dengan (=),tidak sama dengan (≠)lebih besar (>) , lebih kecil (<), penambahan (+), pengurangan (-), perkalian (*), pembagian (/).	Umur, berat badan, tinggi badan.

	Nominal	Ordinal	Interval	Rasio
Bilangan menunjukkan perbedaan	√	√	√	√
Pengukuran dapat digunakan untuk membuat peringkat atau mengurutkan obyek		√	√	√
Perbedaan bilangan mempunyai arti			√	√
Mempunyai nol mutlak dan rasio antara dua bilangan mempunyai arti				√

Eksplorasi data menggunakan Weka

Weka dapat membantu tahap eksplorasi data, namun di Weka hanya dapat mengenali tipe data. Tipe data yang dikenali di weka hanya nominal dan numerik, sehingga kita harus teliti dalam mengkategorisasikan data untuk tipe data ordinal dan interval. Untuk mengetahui lebih lanjut, kita menggunakan data **weather.arff** yang telah tersedia di weka.

a. Statistika Ringkasan

Gambar 1 merupakan nilai statistika ringkasan dari atribut outlook yang memiliki tipe data nominal. Seperti yang kita tahu, **tipe data nilai outlook {sunny, overcast, rainy} tidak hanya dapat diliha dari nilai namun juga dapat diurutkan, sehingga atribut outlook juga bisa dikategorikan sebagai ordinal**. Gambar 1 hanya menampilkan hasil *count* pada Weka.

Selected attribute

Name: outlook Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

Gambar 1 Statistika ringkasan dari atribut *outlook*

Selected attribute

Name: temperature Type: Numeric
Missing: 0 (0%) Distinct: 12 Unique: 10 (71%)

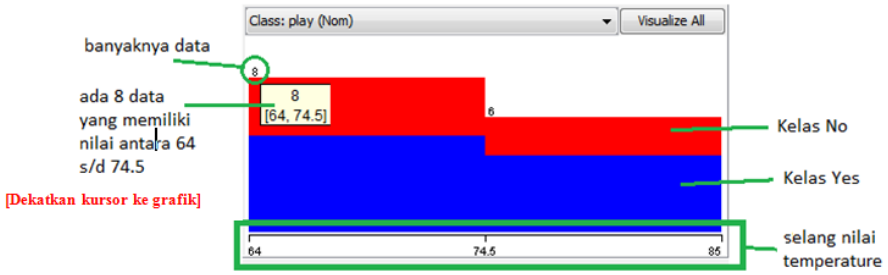
Statistic	Value
Minimum	64
Maximum	85
Mean	73.571
StdDev	6.572

Gambar 2 Statistika ringkasan dari atribut *temperature*

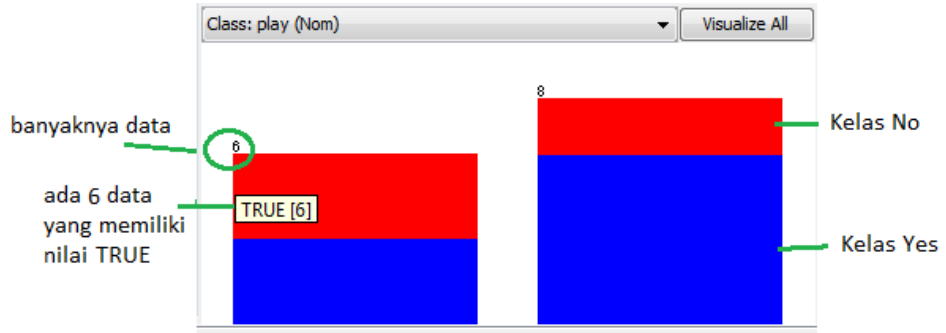
Gambar 2 merupakan nilai statistika ringkasan dari atribut temperature yang memiliki tipe data numeric. Berdasarkan dari penjelasan sebelumnya, tipe data numeric memiliki dua macam yaitu interval dan rasio. **Untuk kasus ini tipe data atribut temperature masuk kedalam kategori interval**. Nilai statistika yang muncul berupa minimum, maximum, mean, dan StdDev.

b. Visualisasi

Visualisai yang tersedia pada Weka yaitu histogram dan scatter plot. Histogram pada weka terdapat pada tab *preprocess* sedangkan scatter plot terdapat pada tab *Visualize.*

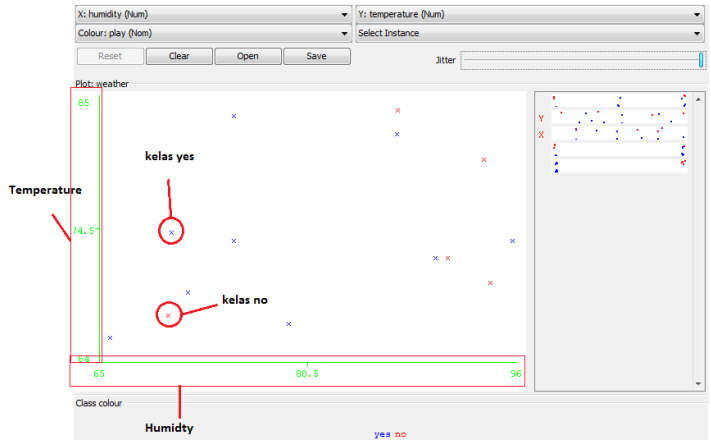


Gambar 3 Histogram atribut *temperature* terhadap kelas *play*

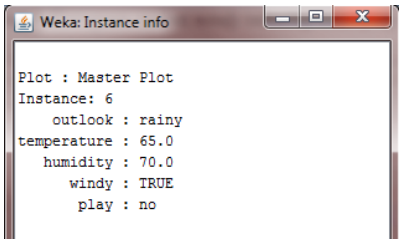


Gambar 4 Histogram atribut *windy* terhadap kelas *play*

Dalam *scatter plot*, setiap objek data diplot sebagai titik dalam bidang 2 dimensi dengan menggunakan nilai-nilai dari dua atribut sebagai koordinat x dan y. Gambar 5 merupakan contoh visualisasi *scatter plot* untuk atribut *temperature* dan *humidity*.



Gambar 5 Scatter plot *humidty* terhadap *temperature*



Gambar 6 *Instance* info

Pada instance info, terdapat keterangan seperti nama atribut beserta nilainya. Hal ini akan memudahkan pengguna untuk membaca dan menganalisa hasil scatter plot. Misalkan outlook: rainy, temperature: 65.0, humidity: 70.0, windy: TRUE, maka masuk ke kelas no.

Eksplorasi data menggunakan R

Eksplorasi pada R tidak jauh berbeda dengan weka, hanya saja di R memiliki tipe data dan visualisasi data yang lebih lengkap dibandingkan weka. Dengan menggunakan R kita juga bisa melakukan eksplorasi pada multiple variables. Eksplorasi pada multiple variables dilakukan untuk mengetahui hubungan antara dua variabel atau lebih. Pada R, hal ini dapat dilihat melalui bar chart, boxplot, scatter plot, pairs plot, serta dengan menghitung nilai covariance dan correlation antar variabel. Untuk lebih jelasnya kita import data **insurance.csv** (tersedia di http://bit.ly/insurance_csv) dan simpan pada variabel **data**.

a. Statistika Ringkasan

Untuk melihat tipe data tiap atribut ketikkan perintah :

```
> str(data)
'data.frame': 1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3
2 1 2 ...
 $ charges : num  16885 1726 4449 21984 3867 ...
```

Jika dibandingkan dengan Weka, R memiliki tipe data yang lebih lengkap seperti int, Factor, num. Untuk melihat statistika ringkasan (misal pada atribut age dan sex) ketikkan perintah:

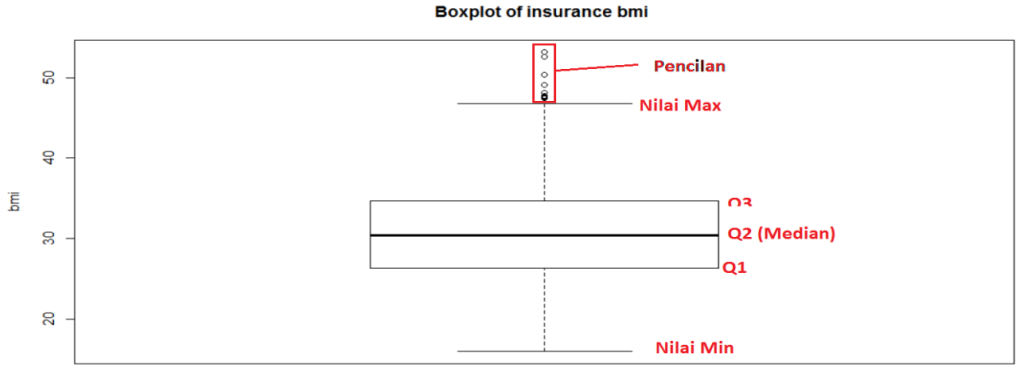
```
> summary(data$sex)
female   male
   662    676
> summary(data$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   27.00   39.00   39.21   51.00   64.00
```

Jika dibandingkan dengan Weka, R memiliki hasil statistika ringkasan lebih lengkap yaitu Min, Q1, Median, Mean, Q3, dan Max.

a. Visualisasi

Beberapa visualisasi yang tersedia pada R yaitu histogram, pie chart, line chart, boxplot, scatter plot**.** Untuk membuat boxplot ketikkan perintah :

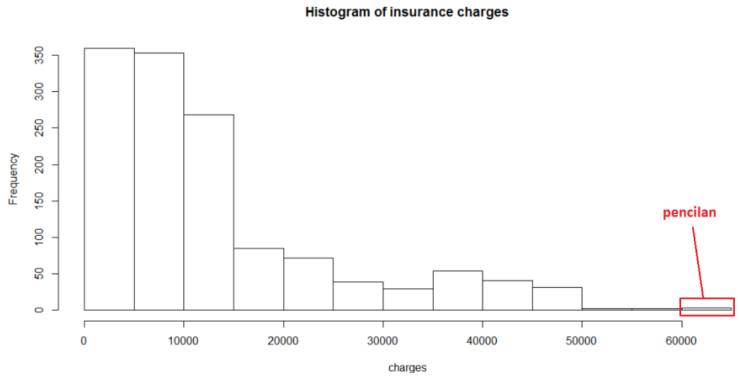
```
boxplot(data$bmi, main="Boxplot of insurance bmi",
        ylab="bmi")
```



Gambar 7 Box plot atribut bmi

Berdasarkan hasil boxplot yang diperoleh, nilai minimum dan maximum berturut-turut yaitu 15.96 dan 53.13. Nilai Q1, Q2, dan Q3 berturut-turut yaitu 26.30, 30.40, 34.69. Data disebut pencilan jika data tersebut lebih besar dari nilai maximum atau lebih kecil dari nilai minimum. Untuk membuat histogram ketikkan perintah:

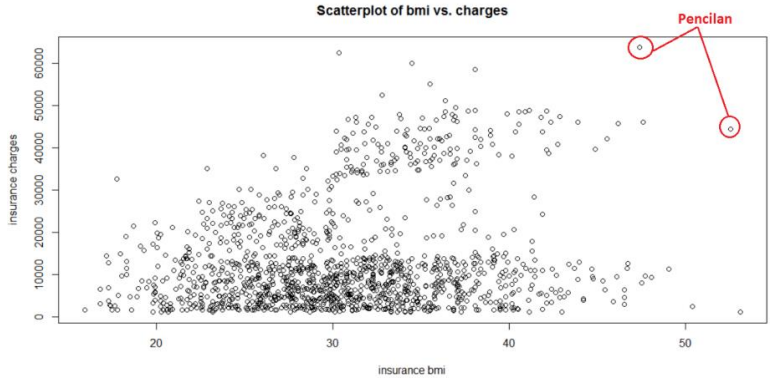
```
hist(data$charges, main = "Histogram of insurance charges",
     xlab = "charges")
```



Gambar 8 Histogram atribut charges

Bentuk histogram tersebut cenderung condong ke kiri dan terdapat pencilan pada data tersebut. Untuk membuat scatter plot ketikkan perintah:

```
plot(x = data$bmi, y = data$charges,
     main = "Scatterplot of bmi vs. charges",
     xlab = "insurance bmi",
     ylab = "insurance charges")
```



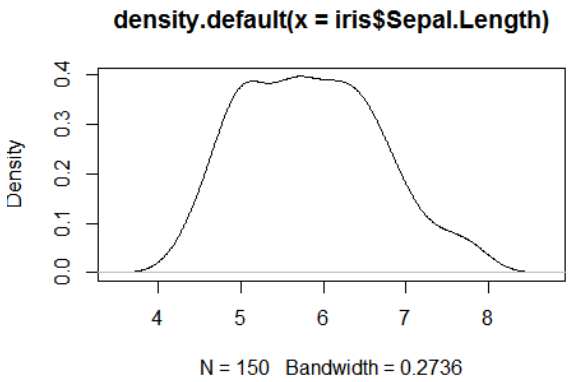
Gambar 9 Scatter plot bmi vs charges

Kita dapat melihat dua data yang dapat dikatakan pencilan. Hal ini dikarenakan kedua data tersebut sangat berbeda / jauh dengan data yang lainnya. Terdapat tiga jenis analisa yang dapat dilakukan dengan menggunakan *scatter plot*:

1 Scatter plot dapat menunjukkan hubungan (korelasi) antara dua variabel/atribut dan juga dapat digunakan untuk mendeteksi hubungan non linier antar dua variabel/atribut.
2. Ketika label dari kelas tersedia, *scatter plot* dapat digunakan untuk menyelidiki bidang pemisah antar kelas
3. Menganalisa pencilan/*outlier*.

Density/densitas dapat digunakan untuk melihat pola distribusi data/peluang distribusi. Gambar 10 menunjukkan grafik densitas untuk atribut sepal length. Dapat dilihat bahwa atribut sepal length memiliki distribusi normal. Berikut adalah kode untuk menghasilkan grafik densitas:

```
> data("iris")
> plot(density(iris$Sepal.Length))
```



Gambar 10 Grafik densitas variabel Sepal Length

Eksplorasi data pada *multiple variables*

Eksplorasi data pada multiple variables dilakukan untuk melihat hubungan antara dua buah variabel atau lebih. Untuk melihat hubungan tersebut bisa dilakukan dengan menghitung nilai covariance dan correlation antara variabel tersebut.

```
> cov(iris$Sepal.Length, iris$Petal.Length)
> cov(iris[,1:4])
> cor(iris$Sepal.Length, iris$Petal.Length)
> cor (iris[,1:4])
> aggregate(Sepal.Length ~ Species, summary, data=iris)
```

Visualisasi data untuk *multiple variables* di R berupa boxplot, scatter plot, 3D scatter plot, pairs plot, heat map, level plot, contour, 3D surface, dan parallel coordinates.

```
#boxplot
> boxplot(Sepal.Length~Species, data=iris)

#scatter plot
>   with(iris,   plot(Sepal.Length,   Sepal.Width,
col=Species, pch=as.numeric(Species)))

#scatter plot with jitter
>     plot(jitter(iris$Sepal.Length,
jitter(iris$Sepal.Width))

#matrix of scatter plots (Pairs plot)
> pairs(iris)

#3D scatter plot
> library(scatterplot3d)
> scatterplot3d(iris$Petal.Width, iris$Sepal.Length,
iris$Sepal.Width)

#interactive 3D scatter plot
> library(rgl)
> plot3d(iris$Petal.Width, iris$Sepal.Length,
iris$Sepal.Width)

#heat map
> distMatrix <- as.matrix(dist(iris[,1:4])
> heatmap(distMatrix)

#level plot
> library(lattice)
> levelplot(Petal.Width~Sepal.Length*Sepal.Width, iris,
cuts=9, col.regions=grey.colors(10)[10:1])
#contour
```

```
> filled.contour(volcano, color=terrain.colors, asp=1,
plot.axes=contour(volcano, add=T))
#3D surface

> persp(volcano, theta = 25, phi = 30, expand = 0.5, col
= "lightblue")

#Parallel coordinates
> library(MASS)
> parcoord(iris[1:4],    col=iris$Species)

#Parallel Coordinates with lattice
> library(lattice)
> parallelplot(~iris[1:4] | Species, data=iris)

#Scatter Plot with ggplot2
> library(ggplot2)
> qplot(Sepal.Length,     Sepal.Width,     data=iris,
facets=Species ~.)

#save as a PDF file
> pdf("myPlot.pdf")
> x <- 1:50
> plot(x, log(x))
> graphics.off()

#Save as a postscript file
> postscript("myPlot2.ps")
> x <- -20:20
> plot(x, x^2)
> graphics.off()
```

Berikut perbedaan secara general terkait eksplorasi data dengan menggunakan Weka dan R

	R	**Weka**
Statistika Ringkasan	Min, Max, Q1, Median, Q3, Mean	Min, Max, StdDev, Mean
Visualisasi Data	Histogram, Scatter plot, pie chart, line chart, box plot	Scatter plot, Histogram
Analisis multi-variable	Ya	Tidak

LAPORAN PENDAHULUAN

1. Apa yang anda ketahui tentang Eksplorasi Data?
2. Sebutkan contoh dari tipe data nominal, ordinal, interval, dan rasio sesuai Prodi anda masing-masing?
3. Apa yang anda ketahui tentang *outlier*?

MATERI PRAKTIKUM

1. Penjelasan Eksplorasi Data
2. Perbedaan Tipe data (nominal, ordinal, interval, rasio) beserta contohnya
3. Statistika Ringkasan dan Visualisasi Data dengan Weka
4. Statistika Ringkasan dan Visualisasi Data dengan R
5. Eksplorasi dan visualisasi data untuk *multiple variables* di R.

DAFTAR PUSTAKA

1. Yanchang Zhao. 2012. R and Data Mining: Examples and Case Studies. http://www.rdatamining.com/docs/RDataMining.pdf
2. Bouckaert, R. R.. et al. 2013. *WEKA Manual for Version 3-6-9*. Edition of January 21, 2013. http://jaist.dl.sourceforge.net/project/weka/documentation/3.6.x/WekaManual-3-6-9.pdf . Accessed on 27 January 2013.