

# **MODUL DATA MINING**



**Disusun Oleh:**

**TIM DOSEN PRODI SARJANA TERAPAN MANAJEMEN INFORMASI KESEHATAN**

**UNIVERSITAS INDONESIA MAJU**

**PROGRAM STUDI SARJANA TERAPAN MANAJEMEN INFORMASIKESEHATAN**

**FAKULTAS VOKASI**

**UNIVERSITAS INDONESIA MAJU**

**JAKARTA**

**2022**



## **Modul Data Mining**

Nama Mahasiswa : \_\_\_\_\_  
NPM : \_\_\_\_\_

Program Studi Sarjana Terapan Manajemen Informasi Kesehatan

Fakultas Vokasi

Universitas Indonesia Maju

2022

## **KATA PENGANTAR**

Buku petunjuk praktikum ini disusun untuk memenuhi kebutuhan mahasiswa sebagai panduan dalam melaksanakan praktikum Data Mining, untuk mahasiswa program studi D4 Manajemen Informasi Kesehatan (MIK) UIMA. Dengan adanya buku petunjuk praktikum ini diharapkan akan membantu dan mempermudah mahasiswa dalam memahami dan melaksanakan praktikum Data Mining sehingga akan memperoleh hasil yang baik.

Materi yang dipraktikkan merupakan materi yang selaras dengan materi kuliah Data Mining. Untuk itu dasar teori yang didapatkan saat kuliah juga akan sangat membantu mahasiswa dalam melaksanakan praktikum ini.

Buku petunjuk ini masih dalam proses penyempurnaan. Insha Allah perbaikan akan terus dilakukan demi kesempurnaan buku petunjuk praktikum ini dan disesuaikan dengan perkembangan ilmu pengetahuan. Semoga buku petunjuk ini dapat dipergunakan sebagaimana mestinya.

Jakarta, September 2022

Penyusun

# DAFTAR ISI

HALAMAN SAMPUL.....	i
KATA PENGANTAR .....	iii
DAFTAR ISI.....	iv
BAB I DOWNLOAD DAN INSTAL R PADA WINDOWS .....	1
BAB II TEKNIK-TEKNIK PRAPROSES DATA .....	9
BAB III PROSES DATA MINING.....	15
BAB IV PENERAPAN DATA MINING.....	35
BAB V EVALUASI MODEL DATA MINING .....	44
DAFTAR PUSTAKA .....	56

# **BAB I**

## **DOWNLOAD DAN INSTAL R PADA WINDOWS**

### **A. Pendahuluan**

R adalah bahasa pemrograman *open source* yang biasa digunakan untuk komputasi dan pengolahan data statistik serta berhubungan dengan penampilan grafik menggunakan tools yang disediakan oleh paket-paketnya. R terdaftar dibawah GNU (*General Public License*). Versi awal R dibuat pada tahun 1992 di Universitas Auckland, New Zealand oleh Ross Ihaka dan Robert Gentleman.

R menyediakan beragam statistic, *machine learning* (pemodelan linier dan non linier, *classic statistic test*, *time series analyst*, klasifikasi, clustering). R memiliki berbagai fungsi *built-in* dan juga fungsi *extended* untuk tugas statistic, *machine learning* dan visualisasi, seperti:

1. Data extraction
2. Data cleaning
3. Data loading
4. Data transformation
5. Statistic analysis
6. Predictive modeling
7. Data visualization

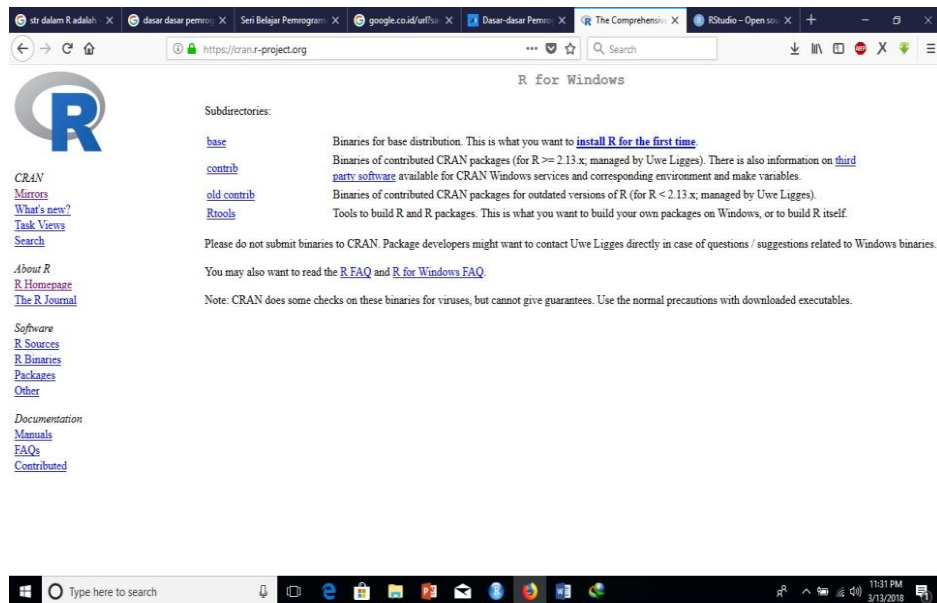
### **B. Understanding Features of R**

Beberapa fitur yang terdapat pada R,

1. Effective programming language
2. Relational database support
3. Data analytics
4. Data visualization
5. Package pada R sebagai penghubung dengan database yang besar.

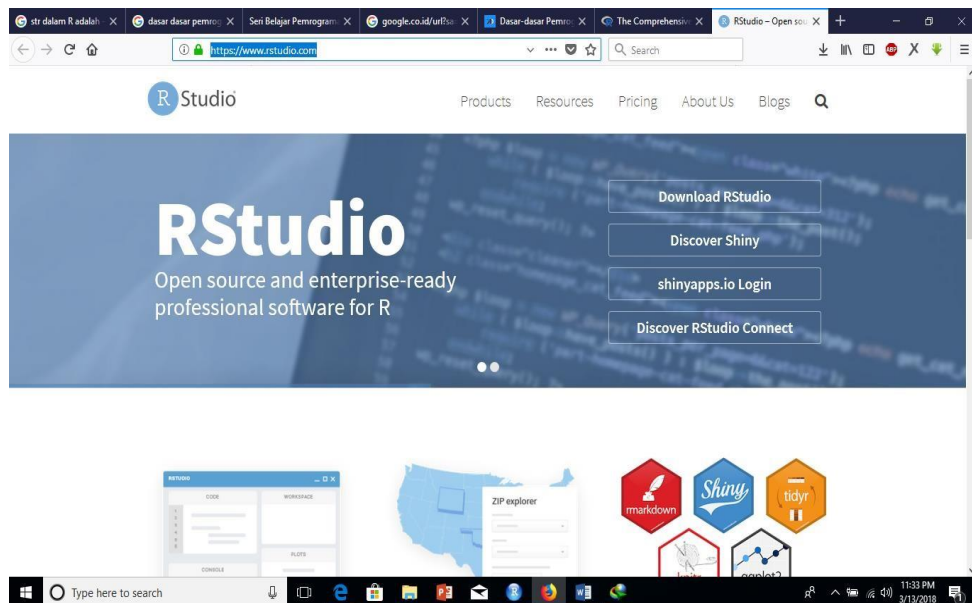
Cara install R adalah:

1. Download paket R di <https://cran.r-project.org/>



Gambar 1.1 Situs resmi dan pengunduhan R

Setelah menginstal paket R dasar, disarankan untuk menginstal RStudio, yang merupakan Integrated Development Environment (IDE) yang hebat dan intuitif untuk R di <https://www.rstudio.com/>



Gambar 1.2 Situs resmi dan pengunduhan RStudio

2. Klik pada CRAN *section*, pilih CRAN mirror dan pilih sesuai dengan sistem operasi yang digunakan (windows).
3. Download R version dari mirror
4. Install R dengan ekstention .exe

#### Cara install RStudio

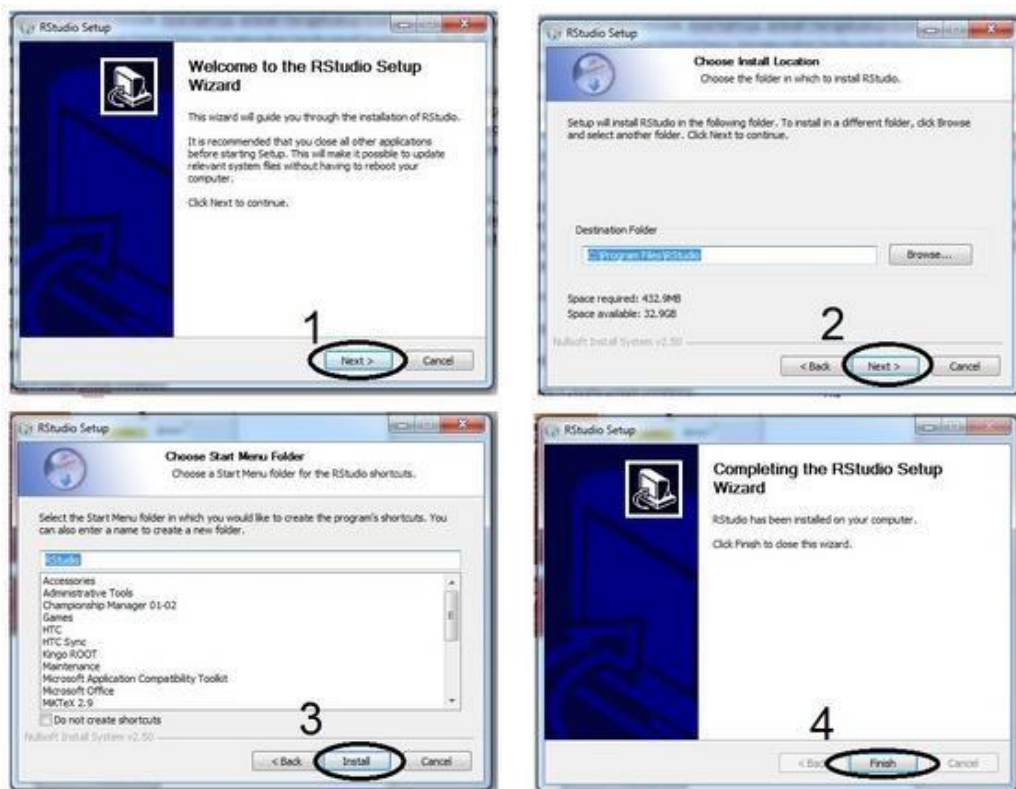
Rstudio diinstall setelah R diinstall. R dapat digunakan langsung, dengan cara mengetikkan kode-kode perintah pada jendela console. Kelemahannya adalah saat melakukan editing pada perintah yang sudah di-*create*. Hal ini dapat diatasi dengan menggunakan jendela R editor. R editor dapat dibuka dengan membuka file → new script, kemudian akan muncul jendela R editor.

Berikut cara mendownload Rstudio untuk windows:

1. Buka URL <https://www.rstudio.com/products/rstudio/download/>
2. Klik pada Rstudio untuk windows

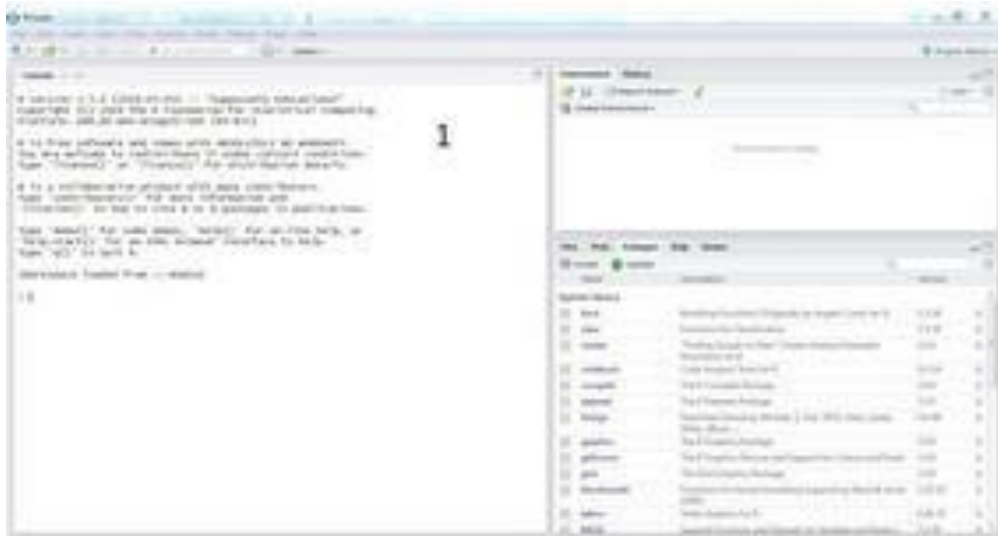
#### Cara menginstall

1. Klik kanan pada icon Rstudio dan pilih run as administrator.
2. Kemudian pilih next → next → install → finish



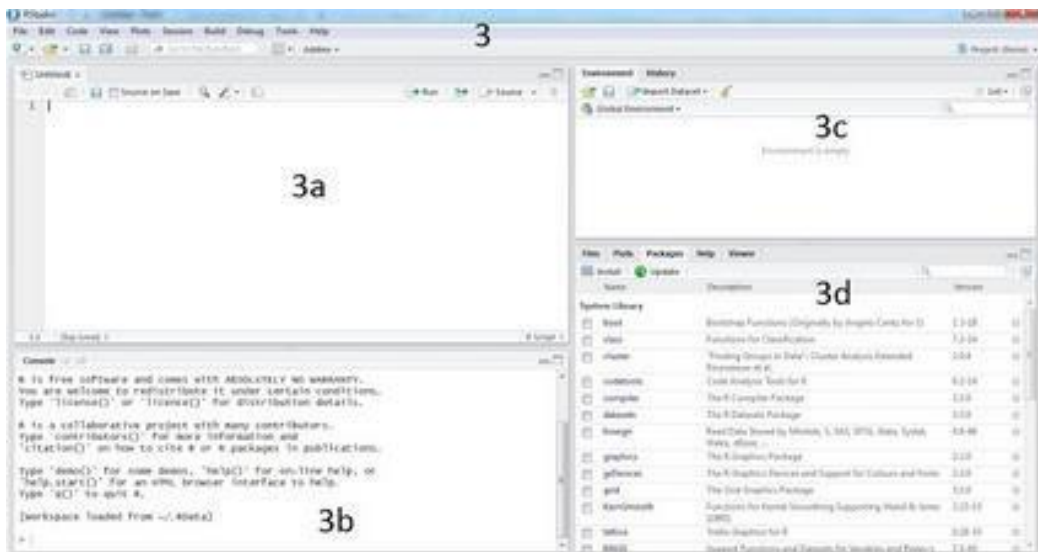
Gambar 1.3 proses install Rstudio pada windows

Jika proses install sudah selesai, maka Rstudio dapat dijalankan. Saat Rstudio pertama dijalankan, maka akan muncul jendela seperti Gambar 4.



Gambar 4 jendela awal Rstudio

Klik new file → new script, sehingga muncul jendela Rstudio yang baru. Didalam Rstudio ada 4 jendela seperti yang terlihat pada Gambar 5, yaitu:



Gambar 5 Jendela editor pada Rstudio

Keterangan:

- 3a. = Untuk menuliskan kode perintah (jendela console)
- 3b. = Tempat kode perintah dieksekusi (jendela environment)
- 3c. = Tempat daftar objek-objek yang sedang aktif, jendelafiles, plots, packages, help dan viewer.
- 3d. = Tempat melihat root files, hasil plot, paket yang diinstal dan help (bantuan)



Untuk melihat kode perintah pada jendela editor dapat dieksekusi dengan menekan alt+ enter atau menggunakan tombol Run.

Fungsi-fungsi dasar pada R:

**1. Visualisasi data:**

Sintax yang digunakan antaranya: barplot, pie, dotchart, dan histogram.

**2. Manipulasi data:**

Sintax yang digunakan diantaranya: sample, stack, unstack dan omit

Beberapa *package* dan fungsi pada R yang digunakan, yaitu:

1. *Clustering*

*Package* yang digunakan untuk *clustering* diantaranya, **fpc, cluster, pvclust, mclust**

2. *Classification*

*Package* yang digunakan untuk *classification* diantaranya, **rpart, tree, marginTree, party, randomForest, maptree**

3. Asosiasi

*Package* yang digunakan untuk asosiasi diantaranya arules dan drm.

4. *Sequential pattern*

*package* yang digunakan untuk *sequential pattern* diantaranya arulesSequences

5. *Time series*

*Package* yang digunakan untuk *time series* diantaranya timsac.

6. Statistik

*package* yang digunakan untuk statistik diantaranya BaseR dan nlme

Contoh kasus 1:

1. Buatlah file produksi.Rdata dan produksi.csv dari tabel berikut:

Bulan	Jawa	Sumatra	Sulawesi	Bali
Januari	12	6	4	2
Maret	9	5	3	5
Juni	8	4	9	7
September	12	5	2	5
Desember	10	8	6	6

2. Buatlah pie chart dari file produksi.Rdata
3. Buatlah histogram dari file produksi.Rdata.

Langkah-langkah menyelesaikan kasus 1:

1. Klik start → pilih Rstudio sehingga tampil layar editor seperti Gambar 5
2. Pada jendela console ketikkan perintah dibawah:

```
> Bulan <- c("Januari", "Maret", "Juni", "September", "Desember")
> Jawa <- c(12, 9, 8, 12, 10)
> Sumatra <- c(6, 5, 4, 5, 8)
> Sulawesi <- c(4, 3, 9, 2, 6)
> Bali <- c(2, 5, 7, 5, 6)
> a <- data.frame(Bulan, Jawa, Sumatra, Sulawesi, Bali)
> save(a, file = "produksi.Rdata")
> write.csv(a, "produksi.csv", row.names = FALSE)
> read.csv("produksi.csv")
```

Perintah read.csv("produksi.csv") berfungsi untuk menampilkan data produksi.csv sehingga tampil data tersebut pada jendela console, sebagai berikut.

```
> read.csv("produksi.csv")
  Bulan Jawa Sumatra Sulawesi Bali
1 Januari 12      6      4      2
2 Maret   9      5      3      5
3 Juni    8      4      9      7
4 September 12     5      2      5
5 Desember 10     8      6      6
```

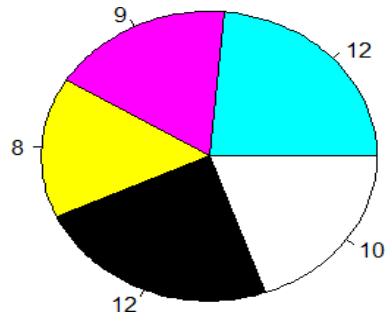
3. Tentukan warna setiap variabel dan jumlah produksi padi setiap wilayah dengan menggunakan perintah dibawah ini:

```
> color <- c("cyan", "magenta", "yellow", "black", "white")
> jawa_labels <- round(Jawa/sum(Jawa)*100, 1)
> jawa_labels <- paste(jawa_labels, "%", sep="")
> sumatra_labels <- round(Sumatra/sum(Sumatra)*100, 1)
> sumatra_labels <- paste(sumatra_labels, "%", sep="")
> sulawesi_labels <- round(Sulawesi/sum(Sulawesi)*100, 1)
> sulawesi_labels <- paste(sulawesi_labels, "%", sep="")
> bali_labels <- round(Bali/sum(Bali)*100, 1)
> bali_labels <- paste(bali_labels, "%", sep="")
```

4. Tampilkan bentuk pie chart dan histogram produksi padi pada jendela plot untuk masing-masing wilayah dengan menetikkan perintah dibawah ini:

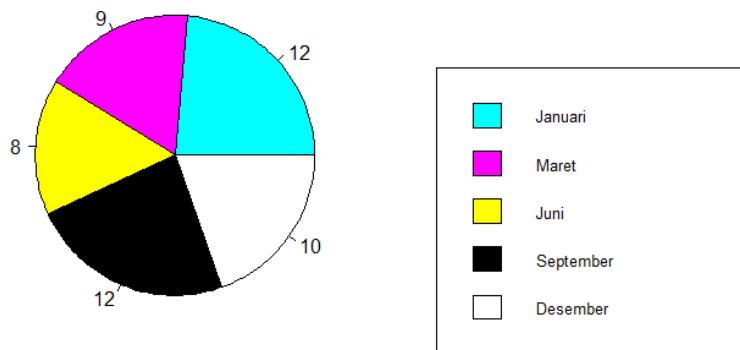
```
> pie(Jawa, main="Jumlah Prediksi Produksi Padi di Jawa Tahun 2015\ndalam satuan (ribu ton)", col=color, label=Jawa)
```

**Jumlah Prediksi Produksi Padi di Jawa Tahun 2014  
dalam satuan (ribu ton)**



```
> legend(1.5, 0.5, c("Januari","Maret","Juni","September","Desember"),cex=0.8,fill=color)
```

**Jumlah Prediksi Produksi Padi di Jawa Tahun 2014  
dalam satuan (ribu ton)**



```
> pie(Sumatra,main="Jumlah Prediksi Produksi Padi di Sumatra
Tahun 2014\ndalam satuan (ribu ton)",col=color,label=Sumat
ra)
> legend(1.5, 0.5, c("Januari","Maret","Juni","September","D
esember"),cex=0.8,fill=color)
> pie(Sulawesi,main="Jumlah Prediksi Produksi Padi di Sulawe
si Tahun 2014\ndalam satuan (ribu ton)",col=color,label=Su
lawesi)
> legend(1.5, 0.5, c("Januari","Maret","Juni","September","D
esember"),cex=0.8,fill=color)
> pie(Bali,main="Jumlah Prediksi Produksi Padi di Bali Tahun
2014\ndalam satuan (ribu ton)",col=color,label=Bali)
> legend(1.5, 0.5, c("Januari","Maret","Juni","September","D
esember"),cex=0.,fill=color)
```

5. Menampilkan histogram untuk data produksi padi dengan mengetikkan perintah dibawah ini:

a. Panggil data produksi padi setiap wilayah

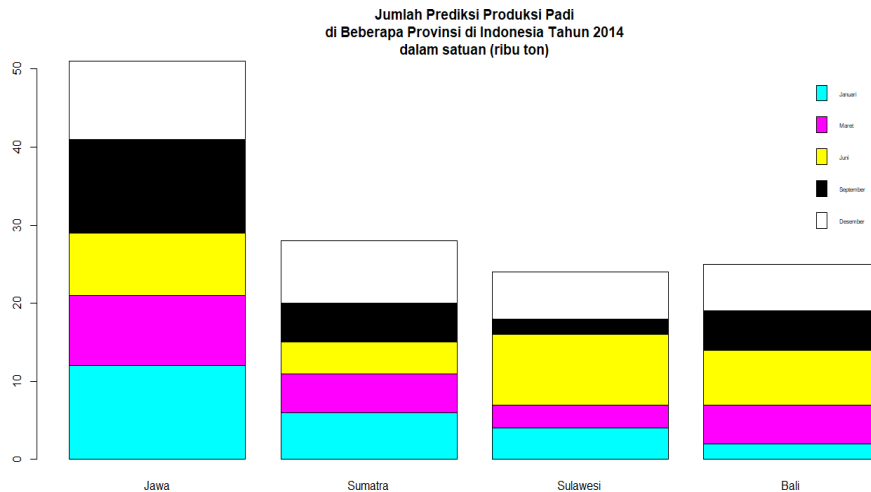
```
> b <- data.frame(Jawa,Sumatra,Sulawesi,Bali)
```

b. Tampilkan histogram untuk data produksi padi

```
> barplot(as.matrix(b), main="Jumlah Prediksi  
Produksi Padi\ndi Beberapa Provinsi di Indonesia  
Tahun 2014\ndalam satuan (ribu ton)",col=color)
```

c. Tampilkan legend histogram

```
> legend("topright",c("Januari","Maret","Juni","Septe  
mber","Desember"),cex=0.6,bty="n",fill=color)
```



## BAB II

### TEKNIK-TEKNIK PRAPROSES DATA

#### A. Pra-proses Data

Pra-proses dilakukan karena dimungkinkan *data set* yang tidak lengkap, mengandung *noise* atau *outlier*, data tidak konsisten, atau ada data yang berulang. Tujuan penting dari pra-proses data adalah untuk meningkatkan kualitas data, sehingga proses data mining juga menghasilkan pengetahuan baru yang lebih baik. Tugas utama dalam pra-proses data adalah pembersihan data, integrasi data, transformasi data, reduksi data dan diskretisasi data.

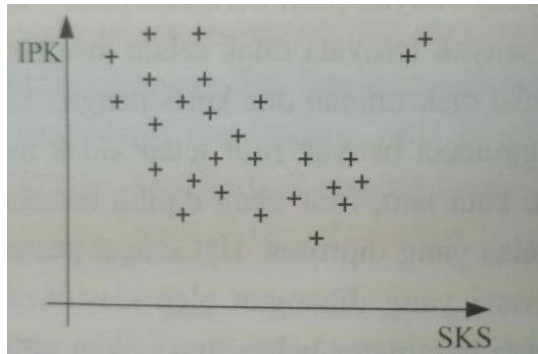
*Outlier* disebut juga *noise* didefinisikan sebagai titik yang terletak sangat jauh dari rata-rata variabel random pada umumnya yang berkorelasi dengan titik tersebut. Jumlah *outlier* biasanya sedikit, dan *outlier* biasanya dibuang dari data yang diproses. Pendeteksian *outlier* dapat dilakukan dengan menggunakan metode-metode seperti:

- a. Pendekatan statistic
- b. K-Nearest Neighbor
- c. Pemeriksaan kerapatan
- d. DBSCAN
- e. *Outlier Removal Clustering*
- f. Dan lain-lain

*Outlier* tidak selalu merupakan data dengan perilaku menyimpang yang akhirnya harus dibuang. Adakalanya *outlier* adalah data yang memang akan dicari karena keistimewaan perilakunya.

#### **Contoh:**

Kasus data akademik mahasiswa tingkat akhir (SKS banyak) dengan IPK tinggi.



Gambar 2.1 Pola data dengan keberadaan *outlier*

## B. Normalisasi Data

Normalisasi dalam kegiatan data mining merupakan proses penskalaan nilai atribut dari data sehingga bisa jatuh pada range tersebut. Ada beberapa metode yang digunakan untuk proses normalisasi, yaitu:

1. Min-Max
2. Z-Score
3. Decimal Scaling
4. Sigmoidal
5. DII

## C. Min-Max

Metode min-max merupakan metode normalisasi dengan melakukan transformasi linier terhadap data asli.

Rumus:

$$\text{Newdata} = (\text{data-min}) * (\text{newmax-newmin}) / (\text{max-min}) + \text{newmin}$$

Newdata = Data hasil normalisasi

Min = Nilai minimum dari data per kolom

Max = Nilai maximum dari data per kolom

Newmin = adalah batas minimum yang kita berikan

Newmax = adalah batas maximum yang kita berikan

#### D. Z-Score

Metode Z-score merupakan metode normalisasi yang berdasarkan mean (nilai rata-rata) dan standard deviation (deviasi standar) dari data.

Rumus

$$\text{newdata} = (\text{data} - \text{mean}) / \text{std}$$

newdata = Data hasil normalisasi

Mean = Nilai rata-rata dari data per kolom

std = Nilai dari standard deviasi

#### E. Decimal Scaling

Metode Decimal Scaling merupakan metode normalisasi dengan menggerakkan nilai desimal dari data ke arah yang diinginkan.

Rumus

$$\text{newdata} = \text{data} / 10^i$$

newdata = Data hasil normalisasi

i = Adalah nilai scaling yang kita inginkan

#### Sigmoidal

Sigmoidal merupakan metode normalization melakukan normalisasi data secara nonlinier ke dalam range -1 - 1 dengan menggunakan fungsi sigmoid.

Rumus

$$\text{newdata} = (1 - e^{-x}) / (1 + e^{-x})$$

x = (data - mean) / std

e = nilai eksponensial (2,718281828)

Metode ini sangat berguna pada saat data-data yang ada melibatkan data-data *outlier*. Data *outlier* data yang keluar jauh dari jangkauan data lainnya.

Seperti bahasa pemrograman pada umumnya, R memiliki nilai-nilai khusus yang merepresentasikan pengecualian-kecualian untuk tipe data normal lainnya. Nilai tersebut yaitu:

- NA, *not available*.

NA biasanya digunakan untuk menggantikan nilai *missing*. Pada R, operasi dasar yang ada dapat memproses dataset yang berisikan nilai NA. Perintah dibawah ini menjelaskan cara mengembalikan nilai NA sebagai hasil dari suatu operasi walaupun *input* dari *argument* tersebut tidak terdapat NA.

```
> NA + 1
[1] NA
> sum(c(NA, 1, 2))
[1] NA
> median(c(NA,1,2,3), na.rm = TRUE)
[1] 2
> length(c(NA, 2, 3, 4))
[1] 4
> 3==NA
[1] NA
> NA==NA
[1] NA
> TRUE | NA
[1] TRUE
```

- NULL

Berarti nilai yang kosong dan memiliki panjang 0.

```
> length(c(1,2,NULL,4))
[1] 3
> sum(c(1,2,NULL,4))
[1] 7
> x <- NULL
> c(x,2)
[1] 2
```

## 1. Eksplorasi data

Dalam R terdapat beberapa cara untuk melakukan eksplorasi data, yaitu dengan mengetahui **tipe data dari setiap atribut** dan mengetahui **persebaran data setiap atribut**.

```
> data<-airquality
> str(data)
'data.frame': 153 obs. of 6 variables:
 $ Ozone : int 41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int 190 118 149 313 NA NA 299 99 19 194 ...
 $ wind : num 7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6
 ...
```



```

$ Temp : int 67 72 74 62 56 66 65 59 61 69 ...
$ Month : int 5 5 5 5 5 5 5 5 5 5 ...
$ Day : int 1 2 3 4 5 6 7 8 9 10 ...

> summary(data)
      Ozone          Solar.R          Wind          Temp
Month
Min.   : 1.00   Min.   : 7.0    Min.   : 1.700   Min.   : 5
6.00   Min.   :5.000
1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:7
2.00   1st Qu.:6.000
Median : 31.50   Median :205.0   Median : 9.700   Median :7
9.00   Median :7.000
Mean   : 42.13   Mean    :185.9   Mean    : 9.958   Mean    :7
7.88   Mean    :6.993
3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:8
5.00   3rd Qu.:8.000
Max.   :168.00   Max.    :334.0   Max.    :20.700   Max.    :9
7.00   Max.    :9.000
NA's   :37      NA's    :7

      Day
Min.   : 1.0
1st Qu.: 8.0
Median :16.0
Mean   :15.8
3rd Qu.:23.0
Max.   :31.0

> view(data)

```

Untuk mengetahui jumlah data yang missing, dapat dilakukan dengan cara berikut:

- a. Install paket **mice**
- b. Gunakan fungsi `md.pattern(dataset)`

```

> library(mice)
> data <- airquality
> md.pattern(airquality)
      wind Temp Month Day Solar.R Ozone
111    1    1    1    1    1    1  0
35     1    1    1    1    1    0  1
5      1    1    1    1    0    1  1
2      1    1    1    1    0    0  2
      0    0    0    0    7    37 44

```

Dari hasil diatas dapat dilihat distribusi dari nilai *missing* disetiap atribut.

## 2. Pembersihan data

Proses mendeteksi dan mengoreksi (menghapus) *record* yang tidak akurat dari *set record*, tabel atau database yang tidak komplit, *incorrect*, *inaccurate* kemudian menggantikan, memodifikasi atau menghapus data tersebut.

```

> data <- airquality
> data$Solar.R[is.na(data$Solar.R)] <- mean(data$Solar.R, na
.rm = TRUE)
> md.pattern(data)

```

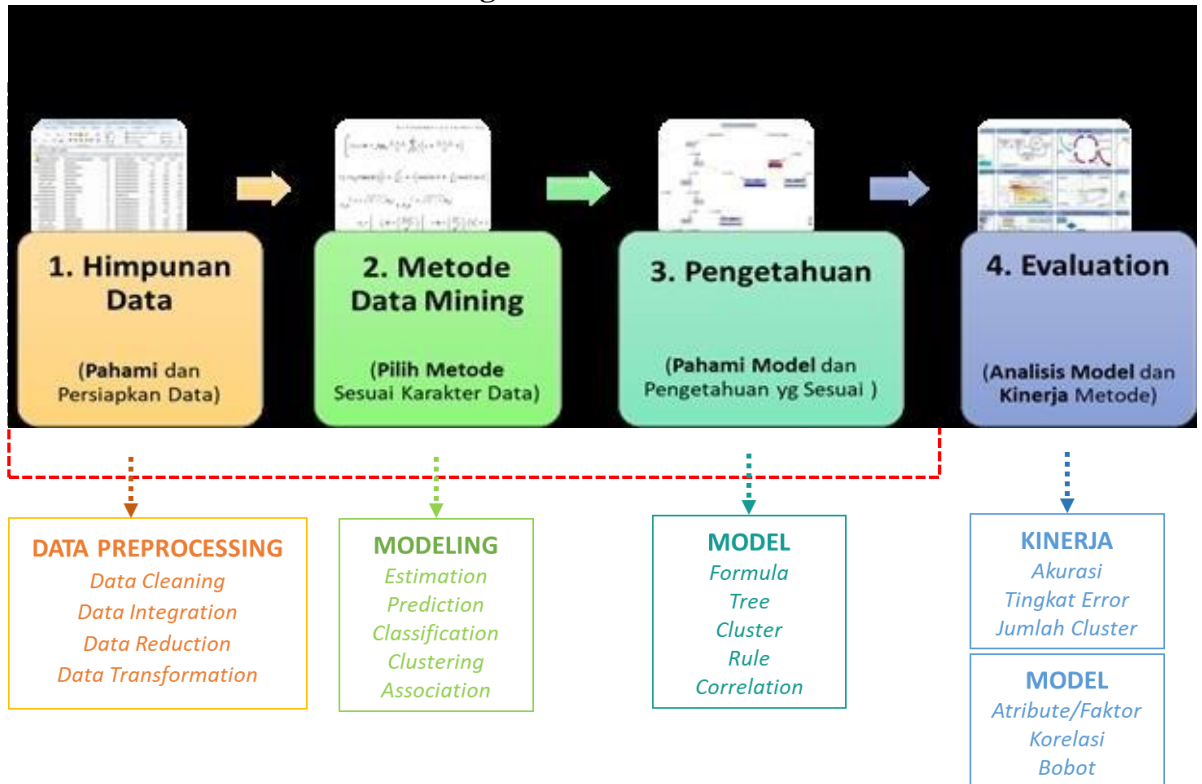
	Solar.R	Wind	Temp	Month	Day	Ozone	
116	1	1	1	1	1	1	0
37	1	1	1	1	1	0	1
	0	0	0	0	0	37	37

Untuk atribut dengan tipe data kategorikal dapat menggunakan fungsi modus, yaitu:

```
names(sort(-table(x)))[1]
```

# BAB III PROSES DATA MINING

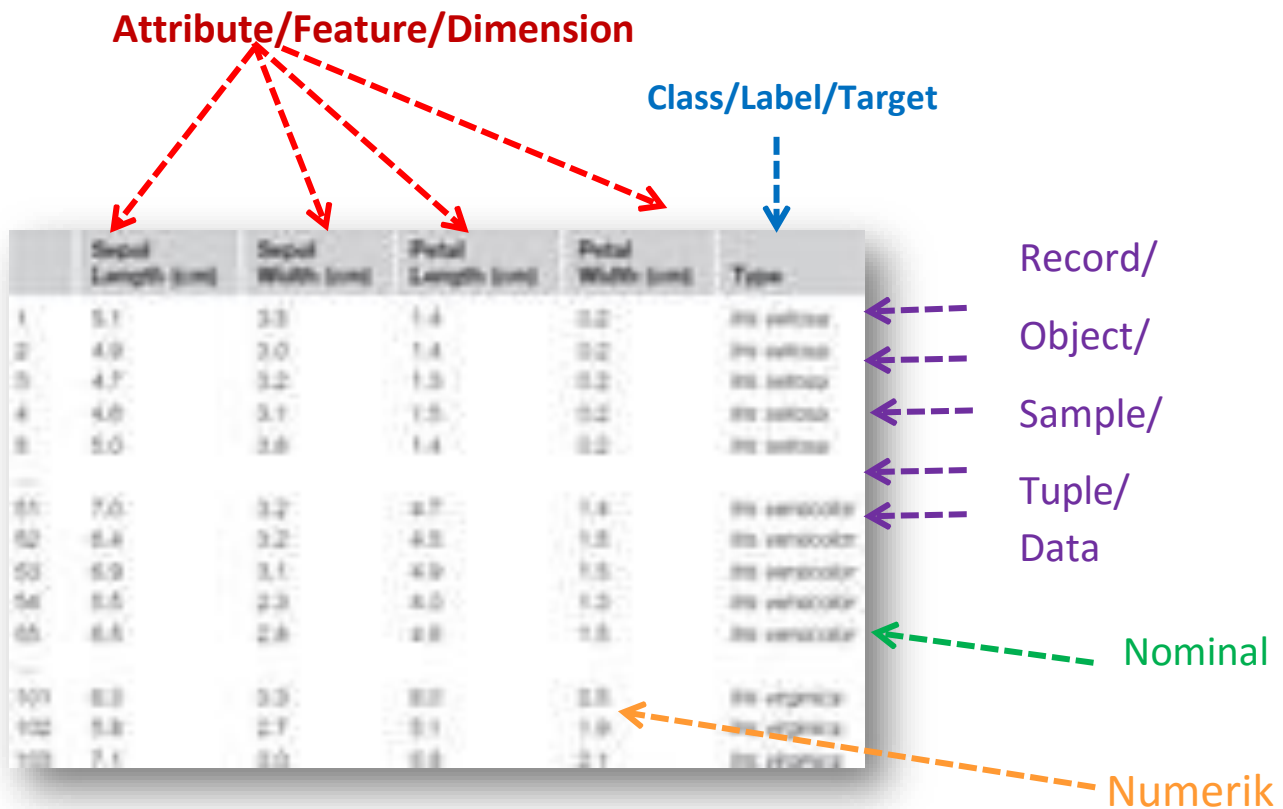
## A. Proses dan Tools Data Mining




### 1. Himpunan Data (Dataset)

- Atribut adalah faktor atau parameter yang menyebabkan class/label/target terjadi
- Jenis dataset ada dua: Private dan Public
- Private Dataset: data set dapat diambil dari organisasi yang kita jadikan obyek penelitian
  - Bank, Rumah Sakit, Industri, Pabrik, Perusahaan Jasa, etc
- Public Dataset: data set dapat diambil dari repositori publik yang disepakati oleh para peneliti data mining
  - UCI Repository  
(<http://www.ics.uci.edu/~mlearn/MLRepository.html>)
  - ACM KDD Cup (<http://www.sigkdd.org/kddcup/>)
  - PredictionIO (<http://docs.prediction.io/datacollection/sample/>)

- Trend penelitian data mining saat ini adalah menguji metode yang dikembangkan oleh peneliti dengan public dataset, sehingga penelitian dapat bersifat: comparable, repeatable dan verifiable.



Public Data Set (UCI Repository)










UCI Machine Learning Repository  
Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

Repository View  Search

[View ALL Data Sets](#)

Browse Through: **360 Data Sets** Table View [ListView](#)

Default Task	Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Classification (262) Regression (93) Clustering (64) Other (62)	 Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
<b>Attribute Type</b> Categorical (137) Numerical (213) Mixed (106)	 Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
<b>Data Type</b> Multivariate (231) Univariate (16) Sequential (38) Time-Series (95) Text (32) Domain Theory (22) Other (21)	 Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
<b>Area</b> Life Sciences (62) Physical Sciences (43) CS / Engineering (111) Social Sciences (23) Biology (11) Game (19) Other (67)	 Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	204	1998
<b># Attributes</b> Less Than 10 (86) 10 to 100 (162)	 Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998
	 Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
	 Audiology (Original)	Multivariate	Classification	Categorical	226		1987

## 2. Metode Data Mining (DM)

### 1. Estimation (Estimasi):

- Linear Regression, Neural Network, Support Vector Machine, Deep Learning, etc

### 2. Prediction/Forecasting (Prediksi/Peramalan):

- Linear Regression, Neural Network, Support Vector Machine, Deep Learning, etc

### 3. Classification (Klasifikasi):

- Decision Tree (CART, ID3, C4.5, Credal DT, Credal C4.5, DynamicCC4.5), Naive Bayes, K-Nearest Neighbor, Linear Discriminant Analysis, Logistic Regression, etc

### 4. Clustering (Klastering):

- K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means, etc

### 5. Association (Asosiasi):

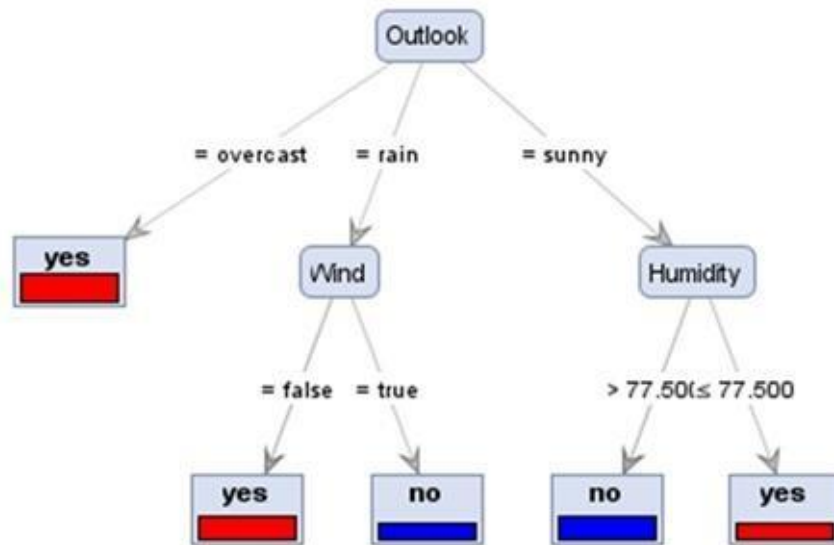
- FP-Growth, A Priori, Coefficient of Correlation, Chi Square, etc

## 3. Pengetahuan (Pola/Model)

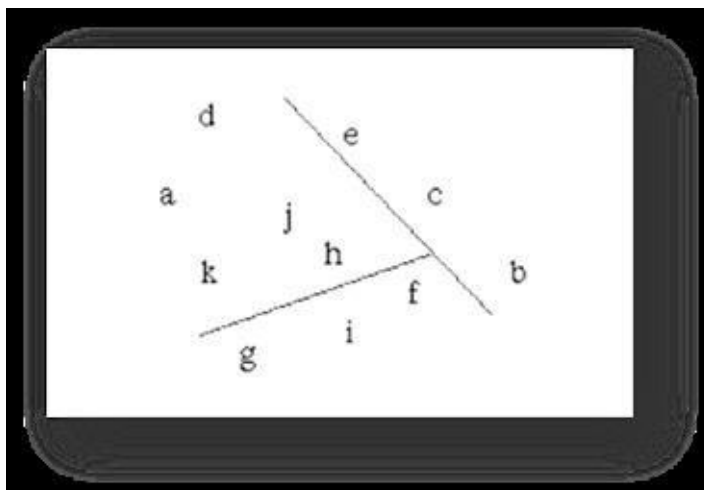
### 1. Formula/Function (Rumus atau Fungsi Regresi)

$$\text{WAKTU TEMPUH} = 0.48 + 0.6 \text{ JARAK} + 0.34 \text{ LAMPU} + 0.2 \text{ PESANAN}$$

### 2. Decision Tree (Pohon Keputusan)



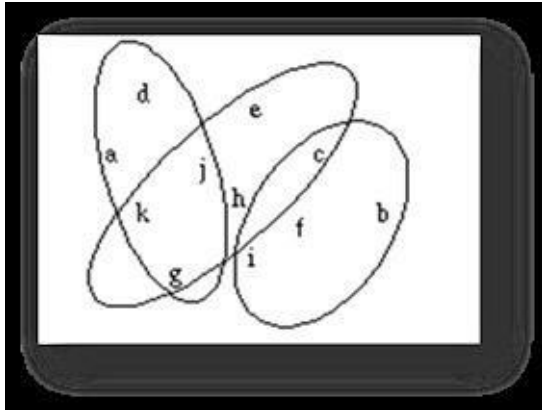
### 3. Korelasi dan Asosiasi



### 4. Rule (Aturan)

IF ips3=2.8 THEN lulus tepat waktu

### 5. Cluster (Klaster)



#### 4. Evaluasi Model Data Mining

Evaluasi Model Data Mining merupakan tahap ke empat dari proses data mining. Berikut ini evaluasi model data mining sesuai dengan peran utama:

##### 1. Estimation:

- **Error:** Root Mean Square Error (RMSE), MSE, MAPE, etc

##### 2. Prediction/Forecasting (Prediksi/Peramalan):

- **Error:** Root Mean Square Error (RMSE) , MSE, MAPE, etc

##### 3. Classification:

- **Confusion Matrix:** Accuracy
- **ROC Curve:** Area Under Curve (AUC)

##### 4. Clustering:

- **Internal Evaluation:** Davies–Bouldin index, Dunn index,
- **External Evaluation:** Rand measure, F-measure, Jaccard index, Fowlkes–Mallows index, Confusion matrix

##### 5. Association:

- **Lift Charts:** Lift Ratio
- **Precision and Recall** (F-measure)

#### Kriteria Evaluasi dan Validasi Model

##### 1. Akurasi

- a. Ukuran dari seberapa baik model mengkorelasikan antara hasil dengan atribut dalam data yang telah disediakan

- b. Terdapat berbagai model akurasi, tetapi semua model akurasi tergantung pada data yang digunakan
- 2. Keandalan
  - a. Ukuran di mana model data mining diterapkan pada dataset yang berbeda
  - b. Model data mining dapat diandalkan jika menghasilkan pola umum yang sama terlepas dari data testing yang disediakan
- 3. Kegunaan
  - Mencakup berbagai metrik yang mengukur apakah model tersebut memberikan informasi yang berguna.

**Keseimbangan diantaranya** ketiganya diperlukan karena belum tentu model yang akurat adalah handal, dan yang handal atau akurat belum tentu berguna

## B. Tools Data Mining

### Magic Quadrant for Data Science Platform (*Gartner, 2017*)





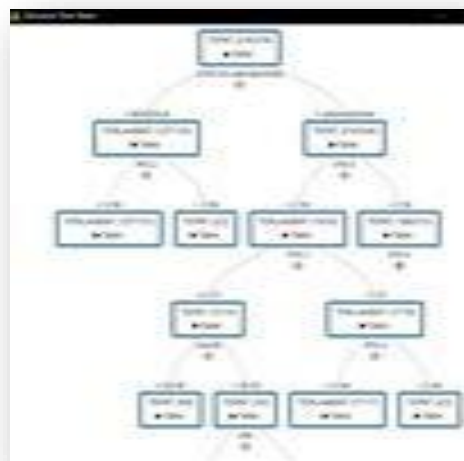
## Magic Quadrant for Data Science Platform (Gartner, 2018)



## KNIME



- KNIME (Konstanz Information Miner) adalah platform data mining untuk analisis, pelaporan, dan integrasi data yang termasuk perangkat lunak bebas dan sumber terbuka
- KNIME mulai dikembangkan tahun 2004 oleh tim pengembang perangkat lunak dari Universitas Konstanz, yang dipimpin oleh Michael Berthold, yang awalnya digunakan untuk penelitian di industri farmasi
- Mulai banyak digunakan orang sejak tahun 2006, dan setelah itu berkembang pesat sehingga tahun 2017 masuk ke Magic Quadrant for Data Science Platform (Gartner Group)



## Sejarah Rapidminer

Pengembangan dimulai pada 2001 oleh Ralf Klinkenberg, Ingo Mierswa, dan Simon Fischer di Artificial Intelligence Unit dari University of Dortmund, ditulis dalam bahasa Java



Open source berlisensi AGPL (GNU Affero General Public License) versi 3.  
Meraih penghargaan sebagai software data mining dan data analytics terbaik di berbagai lembaga kajian, termasuk IDC, Gartner, KDnuggets, dsb

### **Fitur Rapidminer**

Menyediakan prosedur data mining dan machine learning termasuk: ETL (extraction, transformation, loading), data preprocessing, visualisasi, modelling dan evaluasi. Proses data mining tersusun atas operator-operator yang nestable, dideskripsikan dengan XML, dan dibuat dengan GUI. Mengintegrasikan proyek data mining Weka dan statistika R.

### **Atribut Pada Rapidminer**

1. Atribut: karakteristik atau fitur dari data yang menggambarkan sebuah proses atau situasi
  - ID, atribut biasa
2. Atribut target: atribut yang menjadi tujuan untuk diisi oleh proses data mining
  - Label, cluster, weight

### **Tipe Nilai Atribut pada Rapidminer**

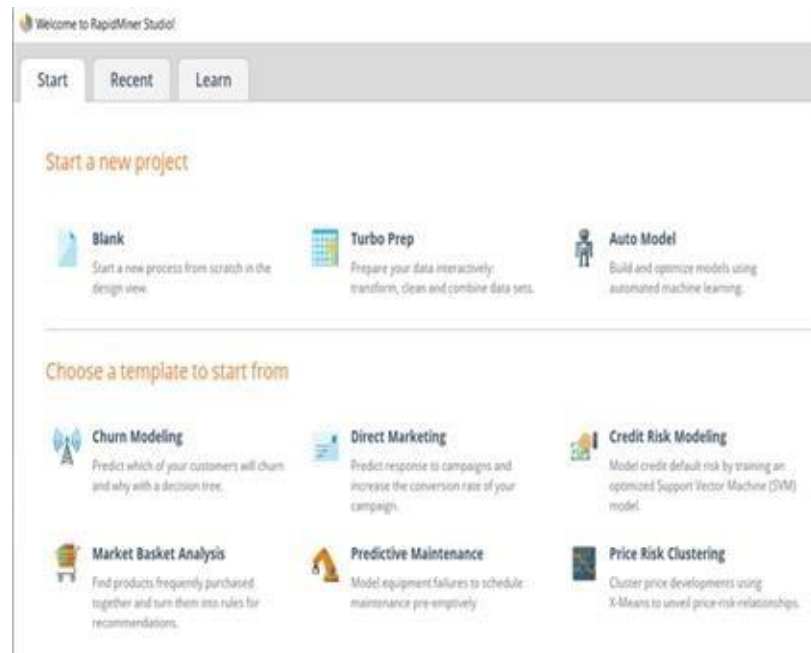
1. nominal: nilai secara kategori
2. binominal: nominal dua nilai
3. polynominal: nominal lebih dari dua nilai
4. numeric: nilai numerik secara umum
5. integer: bilangan bulat
6. real: bilangan nyata
7. text: teks bebas tanpa struktur
8. date\_time: tanggal dan waktu
9. date: hanya tanggal
10. time: hanya waktu

### **Data dan Format Data**

1. Data menyebutkan obyek-obyek dari sebuah konsep, ditunjukkan sebagai baris dari tabel
2. Metadata menggambarkan karakteristik dari konsep tersebut, ditunjukkan sebagai kolom dari tabel
3. Dukungan Format data
  - Oracle, IBM DB2, Microsoft SQL Server, MySQL, PostgreSQL, Ingres, Excel, Access, SPSS, CSV files dan berbagai format lain.

### **Perspektif dan View**

1. Perspektif Selamat Datang (Welcome perspective)
2. Perspektif Desain  
(Design perspective)
3. Perspektif Hasil  
(Result perspective)



## View Operator

### 1. Repository Access

Untuk membaca dan menulis repository

### 2. Import Data

Untuk membaca data dari berbagai format



### 3. Data Access

Untuk membaca dan menulis repositori

### 3. Blending

Untuk menggabungkan data dari berbagai format

### 4. Cleansing

Untuk memberisihkan data

### 5. Modelling

Untuk proses data mining yang sesungguhnya seperti klasifikasi, regresi, clustering, aturan asosiasi dll

### 5. Scoring

Untuk menghitung confidence, apply model

### 6. Validation

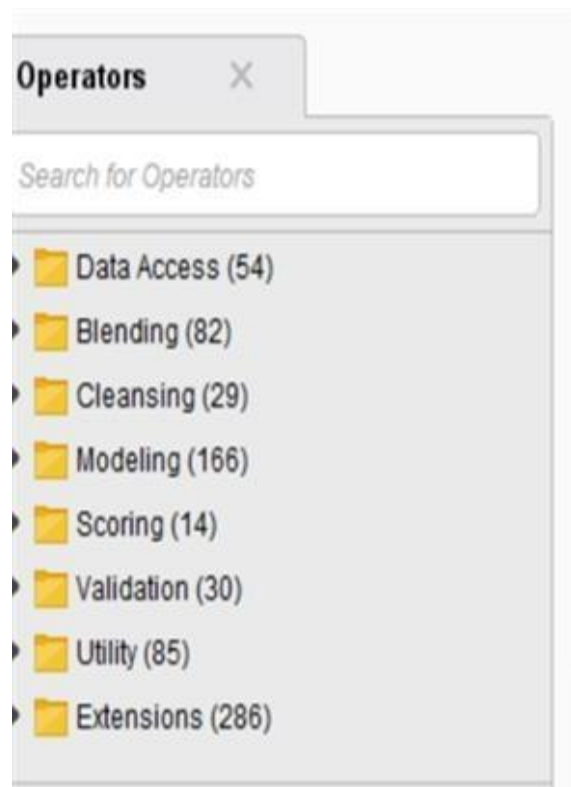
Untuk menghitung kualitas dan perfomansi dan validasi dari model

### 7. Utility

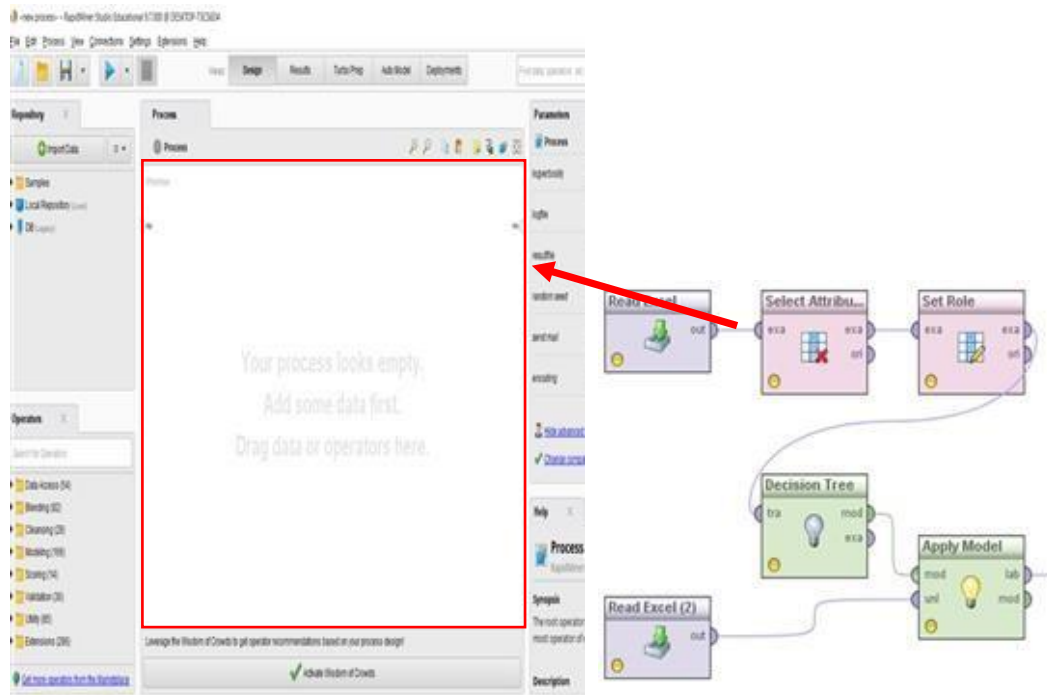
Untuk mengelompokkan subprocess, juga macro dan logger eksternal

### 8. Extentions

Fasilitas tambahan seperti Text Mining, Web Mining, dll

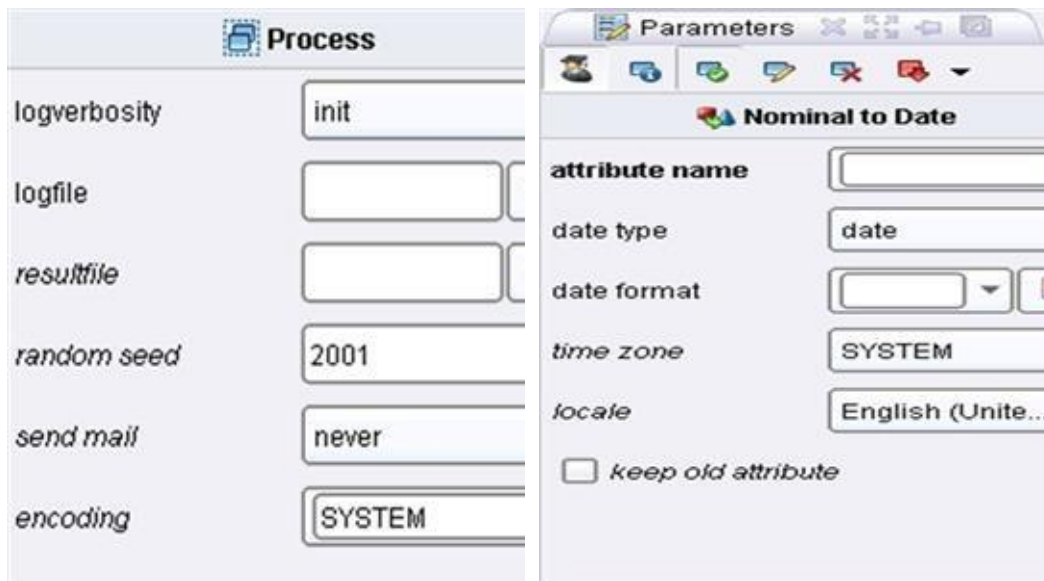


## View Proses



## View Parameter

- Operator kadang memerlukan parameter untuk bisa berfungsi
- Setelah operator dipilih di view Proses, parameternya ditampilkan di view ini



### View Help dan View Comment

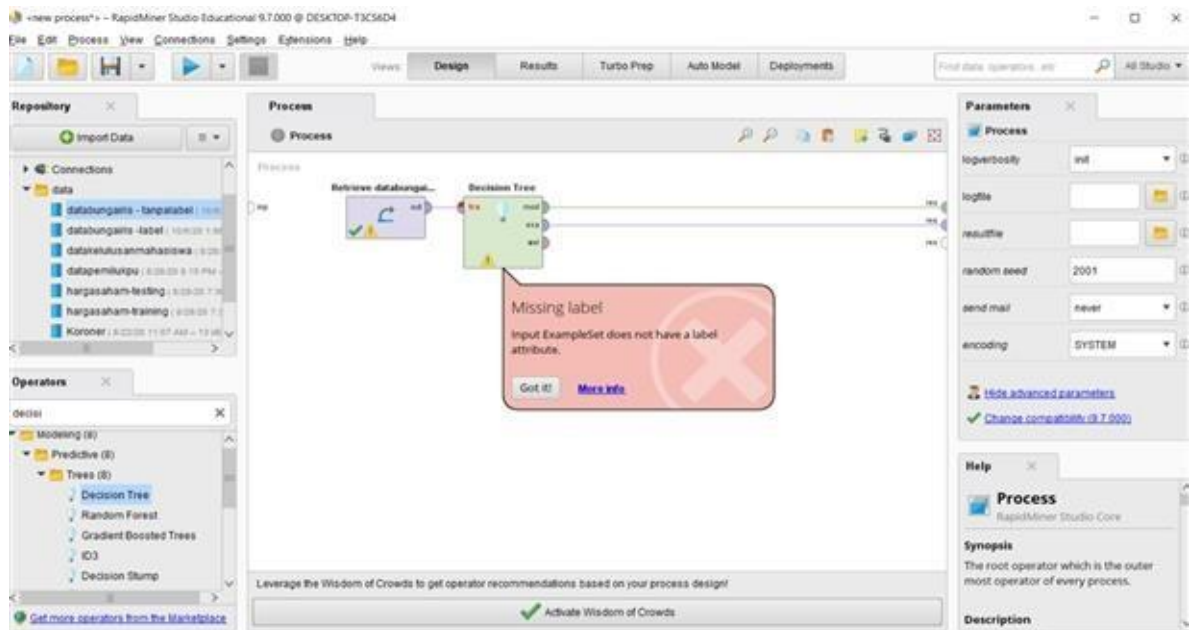
View Help menampilkan deskripsi dari operator

View Comment menampilkan komentar yang dapat diedit terhadap operator



### View Problems





## Operator dan Proses

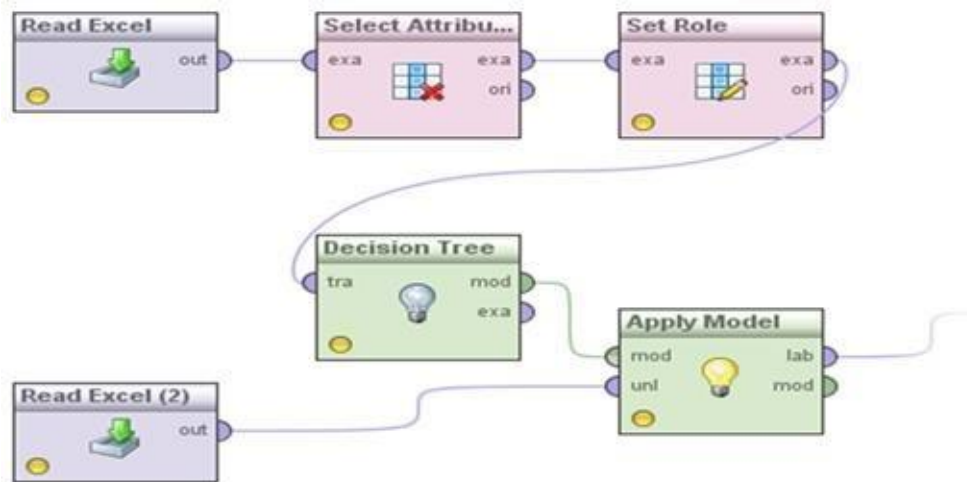
1. Proses data mining pada dasarnya adalah proses analisa yang berisi alur kerja dari komponen data mining
2. Komponen dari proses ini disebut operator, yang didefinisikan dengan:
  - a. Deskripsi input
  - b. Deskripsi output
  - c. Aksi yang dilakukan
  - d. Parameter yang diperlukan
3. Sebuah operator bisa disambungkan melalui port masukan (kiri) dan port keluaran (kanan)



4. Indikator status dari operator:
  - a. Lampu status: merah (tak tersambung), kuning (lengkap tetapi belum dijalankan), hijau (sudah berhasil dijalankan)
  - b. Segitiga warning: bila ada pesan status

- c. Breakpoint: bila ada breakpoint sebelum/sesudahnya
- d. Comment: bila ada komentar
- e. Subprocess: bila mempunyai subprocess

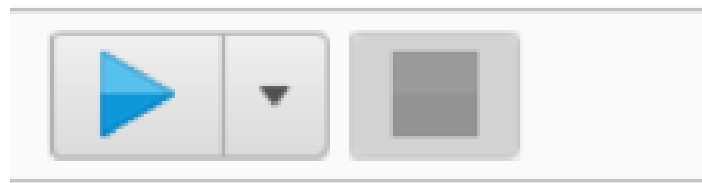
### Mendesain Proses



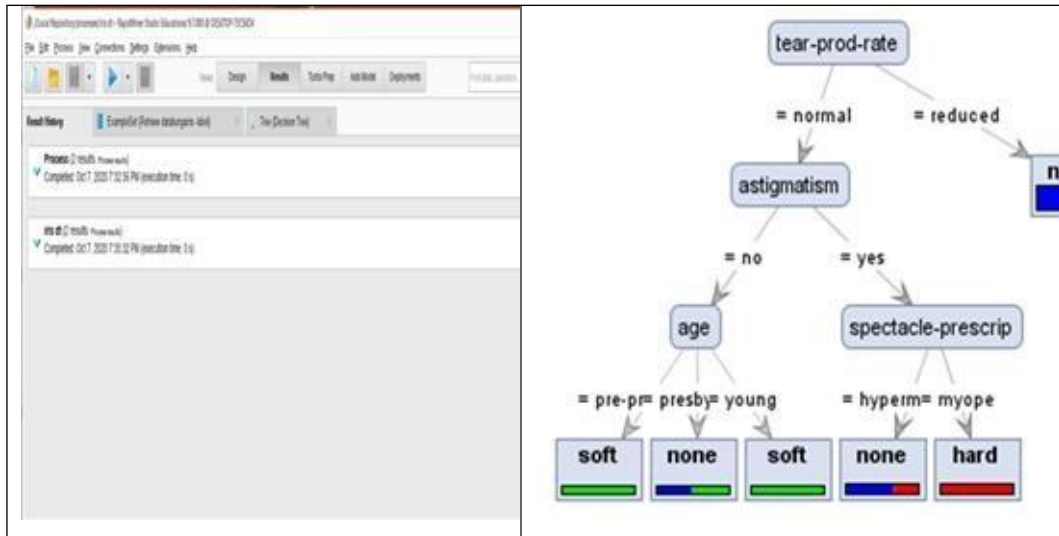
### Menjalankan Proses

Proses dapat dijalankan dengan:




- Menekan tombol Play
- Memilih menu Process → Run
- Menekan kunci F11





### Melihat Hasil







## Panduan Install Rapid Miner





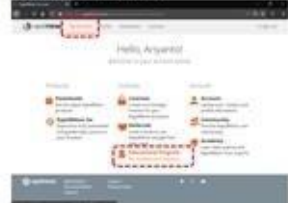
I. Instalasi JDK	
1	<p><b>Instalasi JDK</b></p> <ul style="list-style-type: none"> <li>• Double-click pada file installer JDK yang akan diinstall</li> <li>• Klik "Next"</li> </ul> 
2	<p><b>Instalasi JDK</b></p> <ul style="list-style-type: none"> <li>• Klik "Next"</li> </ul> 
3	<p><b>Instalasi JDK</b></p> <ul style="list-style-type: none"> <li>• Instalasi JDK selesai</li> <li>• Klik "Close"</li> </ul> 
II. Downlaod RapidMiner	



1	<p>Download RapidMiner</p> <ul style="list-style-type: none"> <li>• <a href="https://rapidminer.com">https://rapidminer.com</a></li> <li>• Klik "Get Started"</li> </ul> 	2
3	<p>Download RapidMiner</p>  <p>• Kemudian Klik Download</p>	4

### III. Instalasi RapidMiner

1	<p>Instalasi RapidMiner</p> <ul style="list-style-type: none"> <li>• Double-click pada file installer Rapidminer yang akan diinstall</li> <li>• Klik "Next"</li> </ul> 	2
3	<p>Instalasi RapidMiner</p> <ul style="list-style-type: none"> <li>• Pilih lokasi penginstalan</li> <li>• Klik "Install"</li> </ul> 	4
2	<p>Instalasi RapidMiner</p> <ul style="list-style-type: none"> <li>• Klik "I Agree"</li> </ul> 	4
4	<p>Instalasi RapidMiner</p> <ul style="list-style-type: none"> <li>• Klik "Finish"</li> </ul> 	4

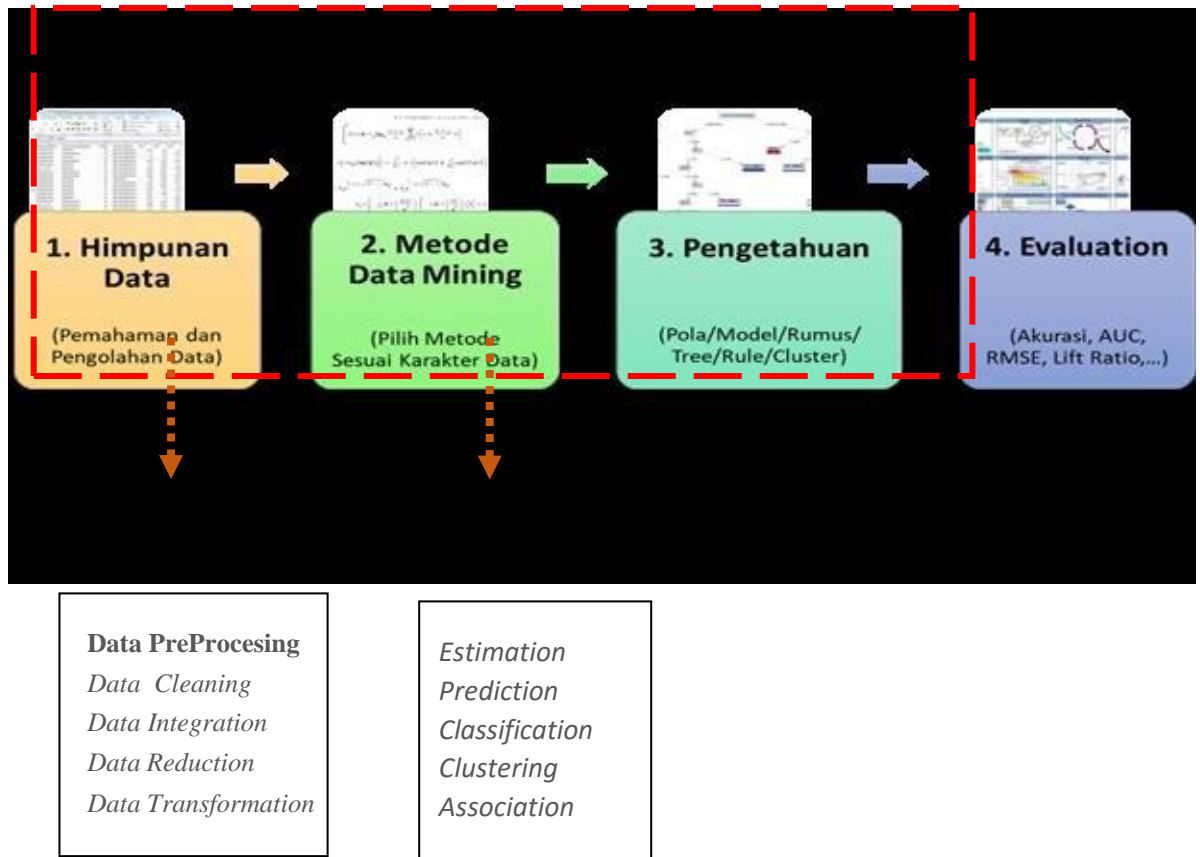
## IV. Registrasi Akun

1	<p><b>Registrasi Akun</b></p> <ul style="list-style-type: none"> <li>• Checklist "I have read and understand the terms..."</li> <li>• Klik "I Accept"</li> </ul> 	2	
3	<p><b>Registrasi Akun</b></p> <p>Setelah membuat akun, periksa email aktivasi</p> 	4	
5	<p><b>Registrasi Akun</b></p> <ul style="list-style-type: none"> <li>• Buka email verify dari RapidMiner</li> </ul> 	6	
7	<p><b>Registrasi Akun</b></p> <ul style="list-style-type: none"> <li>• Klik "Refresh", akun anda akan terverifikasi</li> </ul> 	8	<p><b>Registrasi Lisensi</b></p> <p>Apabila saat ini anda telah terdaftar di RapidMiner namun belum memiliki lisensi:</p> <ul style="list-style-type: none"> <li>• Akses My Account Rapidminer</li> <li>• Klik Educational Program</li> </ul> 

9	<p><b>Registrasi Lisensi</b></p> <ul style="list-style-type: none"> <li>• Lengkapi semua field. Konten dapat mengacu pada histori perkuliahan anda</li> <li>• Ceklis "I have read and accept the end-user license agreement" dan "I hereby confirm that I am eligible and that I agree to meet the requirements."</li> <li>• Klik Apply for licensi</li> </ul> 	10	
11	<p><b>Registrasi Lisensi</b></p> <ul style="list-style-type: none"> <li>• Buka Kembali aplikasi Rapidminer</li> <li>• Pastikan bahwa lisensi anda telah aktif seperti pada gambar berikut</li> </ul> 		

## BAB IV PENERAPAN DATA MINING

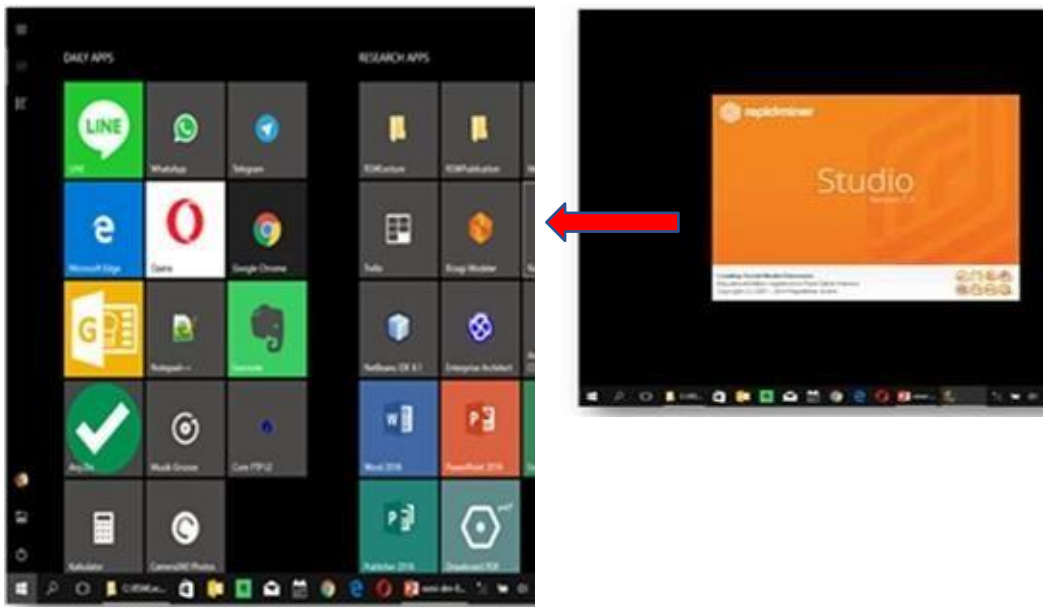
### A. Proses Data Mining



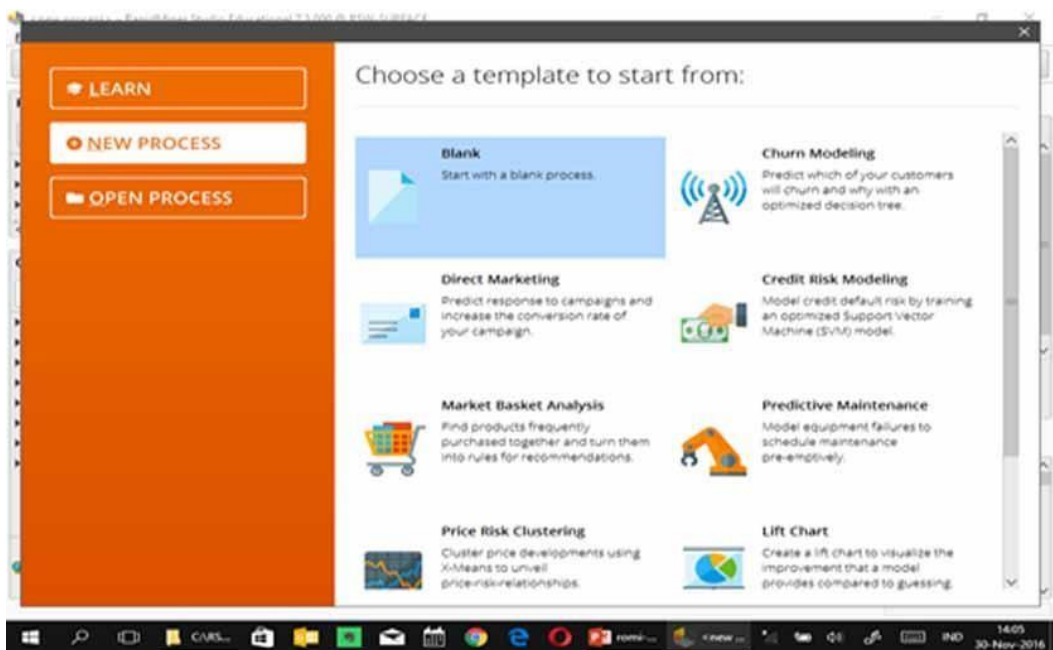
#### 1. Latihan: Rekomendasi Main Golf

- Lakukan training pada data golf (maingolf.xls) dengan menggunakan algoritma decision tree
- Tampilkan himpunan data (dataset) dan pengetahuan (model tree) yang terbentuk

## Buka RapidMiner

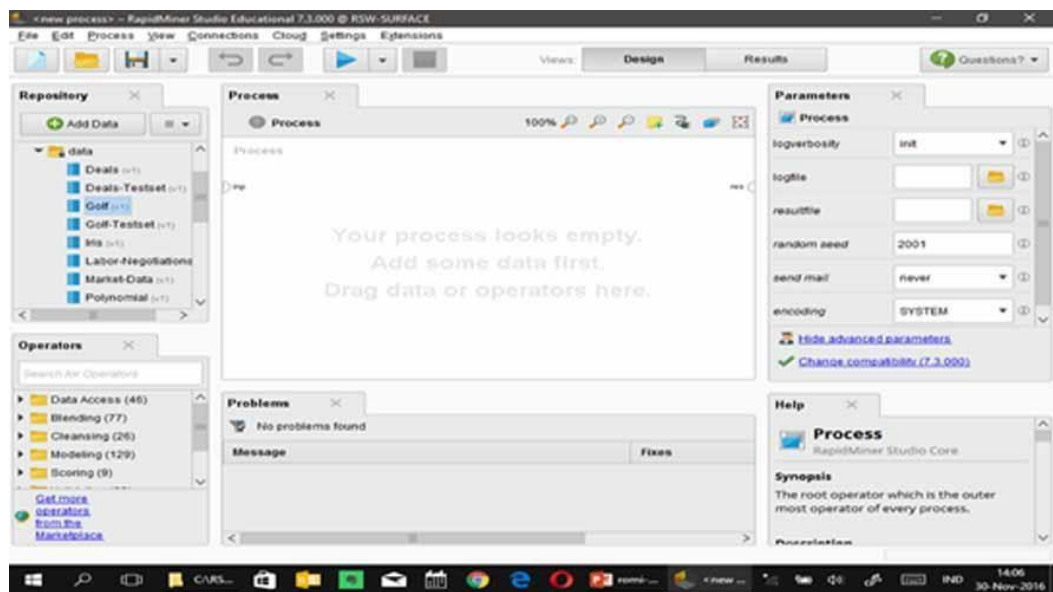


Klik Blank

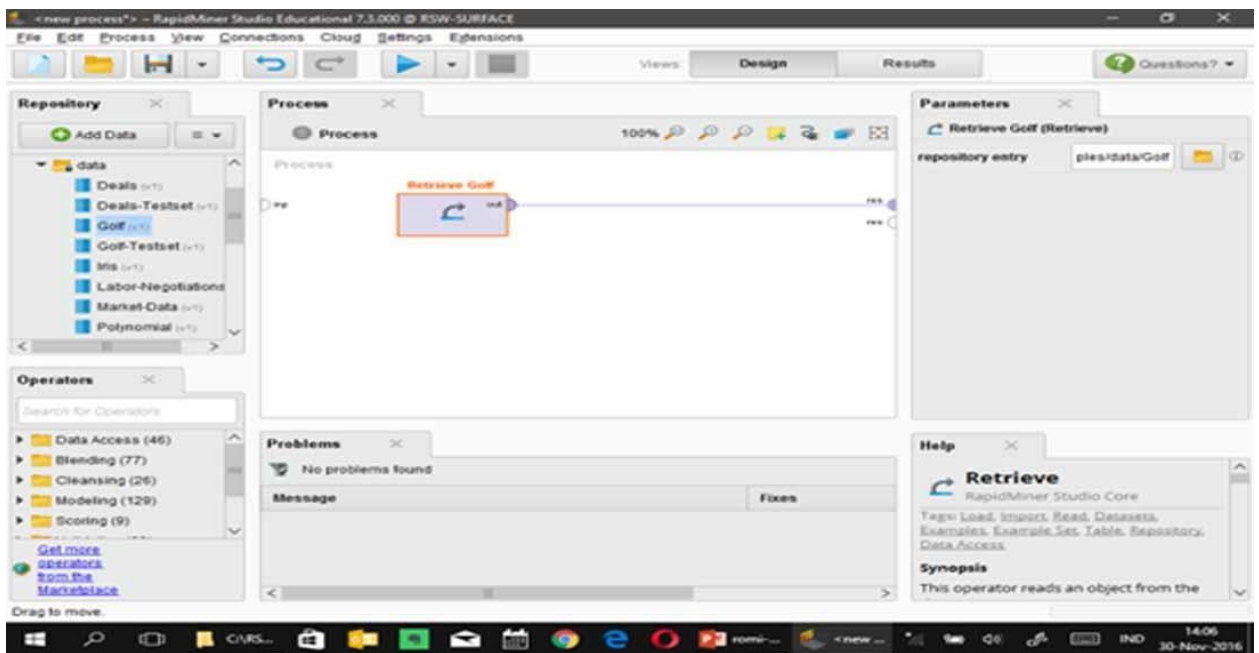




## Muncul Lembar Kerja



## Membaca Data dan menampilkan data



ExampleSet (14 examples, 1 special attribute, 4 regular attributes) Filter (14 / 14 examples): all

Row No.	Play	Outlook	Temperature	Humidity	Wind
1	no	sunny	85	85	false
2	no	sunny	85	90	true
3	yes	overcast	83	78	false
4	yes	rain	76	96	false
5	yes	rain	68	80	false
6	no	rain	85	70	true
7	yes	overcast	64	65	true
8	no	sunny	72	95	false
9	yes	sunny	69	70	false
10	yes	rain	75	80	false
11	yes	sunny	75	70	true
12	yes	overcast	72	30	true
13	yes	overcast	81	75	false
14	no	rain	71	80	true

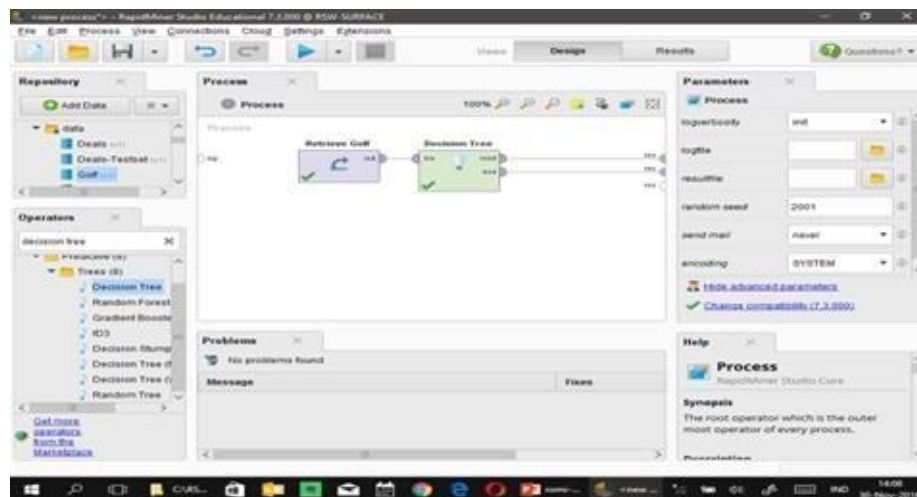
## Menampilkan Statistik Data

ExampleSet (Retrieve Golf) Filter (5 / 5 attributes): Search for attributes

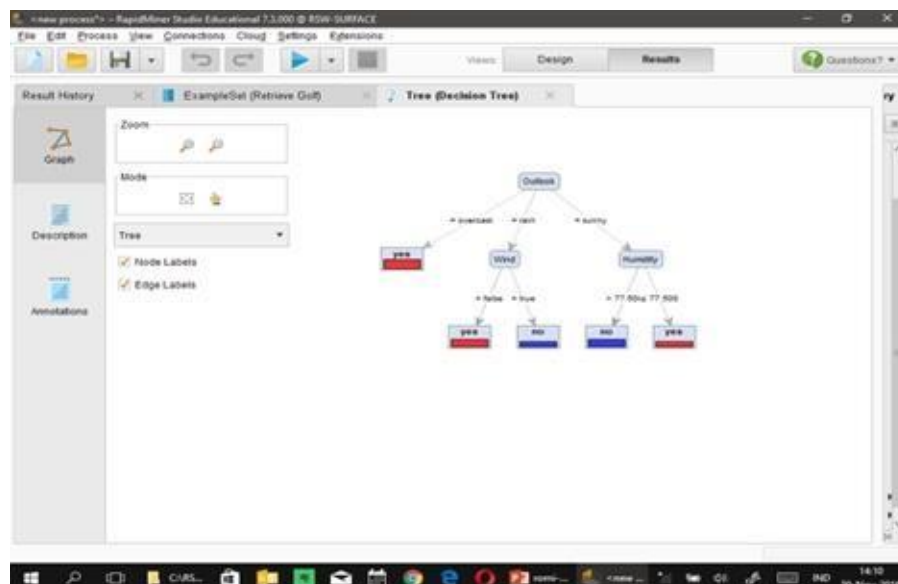
Name	Type	Missing	Statistics								
Play	Nominal	0	<table border="1"> <tr><th>Value</th><th>Count</th></tr> <tr><td>no</td><td>5</td></tr> <tr><td>yes</td><td>9</td></tr> </table>	Value	Count	no	5	yes	9		
Value	Count										
no	5										
yes	9										
Outlook	Nominal	0	<table border="1"> <tr><th>Value</th><th>Count</th></tr> <tr><td>overcast</td><td>4</td></tr> <tr><td>rain</td><td>5</td></tr> <tr><td>sunny</td><td>5</td></tr> </table>	Value	Count	overcast	4	rain	5	sunny	5
Value	Count										
overcast	4										
rain	5										
sunny	5										
Temperature	Integer	0	<table border="1"> <tr><th>Min</th><th>Max</th><th>Average</th></tr> <tr><td>64</td><td>85</td><td>73.571</td></tr> </table>	Min	Max	Average	64	85	73.571		
Min	Max	Average									
64	85	73.571									
Humidity	Integer	0	<table border="1"> <tr><th>Min</th><th>Max</th><th>Average</th></tr> <tr><td>65</td><td>96</td><td>80.286</td></tr> </table>	Min	Max	Average	65	96	80.286		
Min	Max	Average									
65	96	80.286									
Wind	Nominal	0	<table border="1"> <tr><th>Value</th><th>Count</th></tr> <tr><td>true</td><td>8</td></tr> <tr><td>false</td><td>6</td></tr> </table>	Value	Count	true	8	false	6		
Value	Count										
true	8										
false	6										

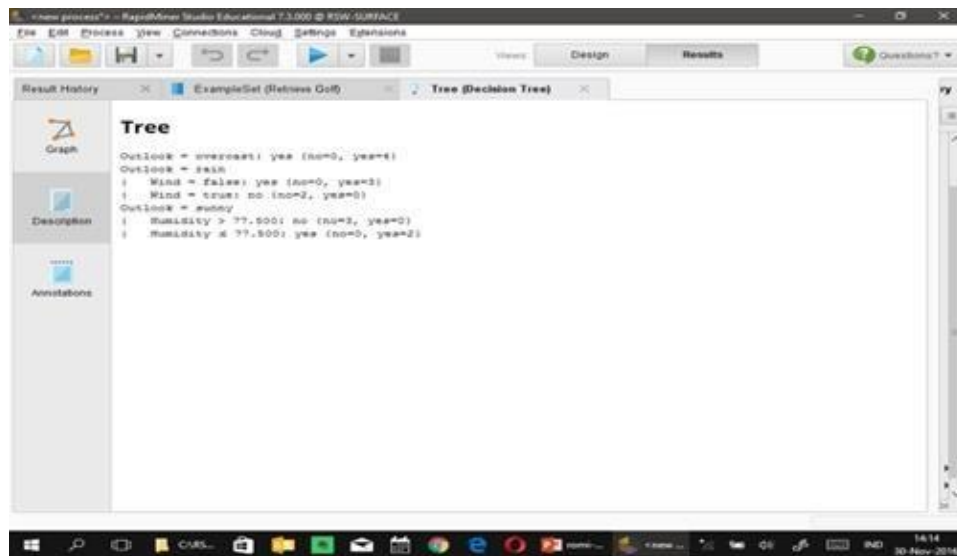
Showing attributes 1 - 5 Examples: 14 Special Attributes: 1 Regular Attributes: 4

## Membuat Model



## Menampilkan Hasil



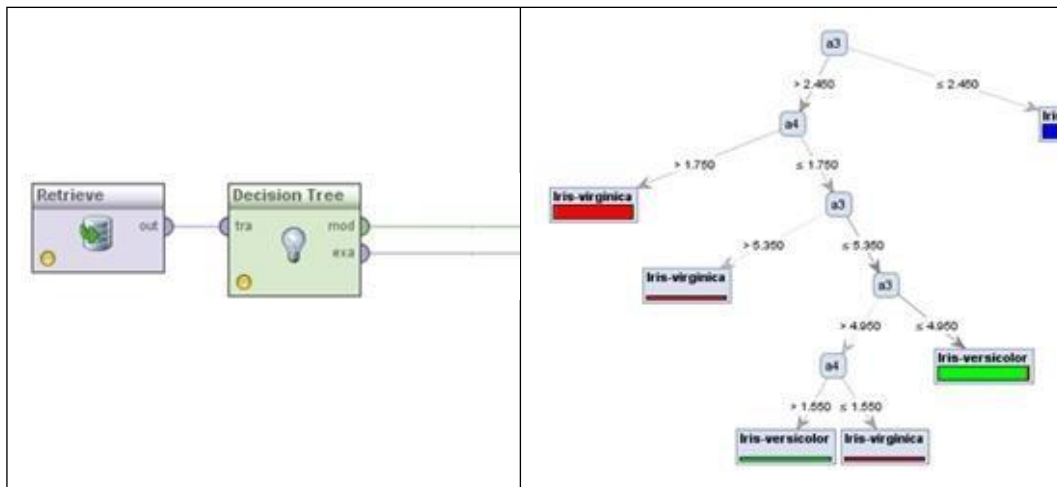


## 2. Latihan: Rekomendasi Main Tenis

1. Lakukan training pada data tenis (tenis.xls) dengan menggunakan algoritma decision tree
2. Tampilkan himpunan data (dataset) dan pengetahuan (model tree) yang terbentuk

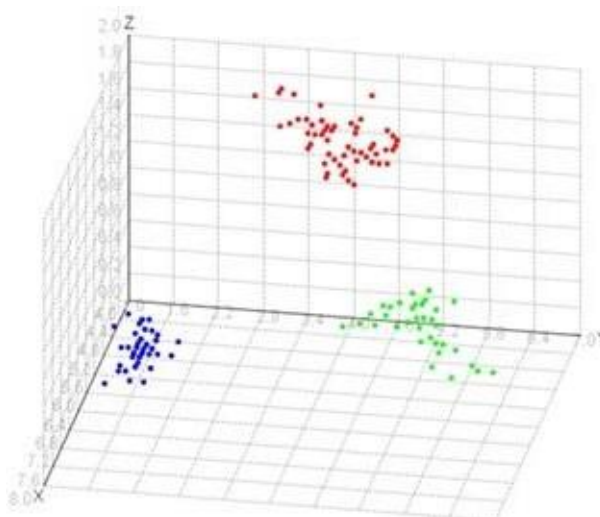
## 3. Latihan: Penentuan Jenis Bunga Iris

1. Lakukan training pada data Bunga Iris (ambil dari repositories rapidminer) dengan menggunakan algoritma decision tree
2. Tampilkan himpunan data (dataset) dan pengetahuan (model tree) yang terbentuk



#### 4. Latihan: Klastering Jenis Bunga Iris

1. Lakukan training pada data Bunga Iris (ambil dari repositories rapidminer) dengan menggunakan algoritma k-Means
2. Tampilkan himpunan data (dataset) dan pengetahuan (model tree) yang terbentuk
3. Tampilkan grafik dari cluster yang terbentuk seperti di bawah ini.

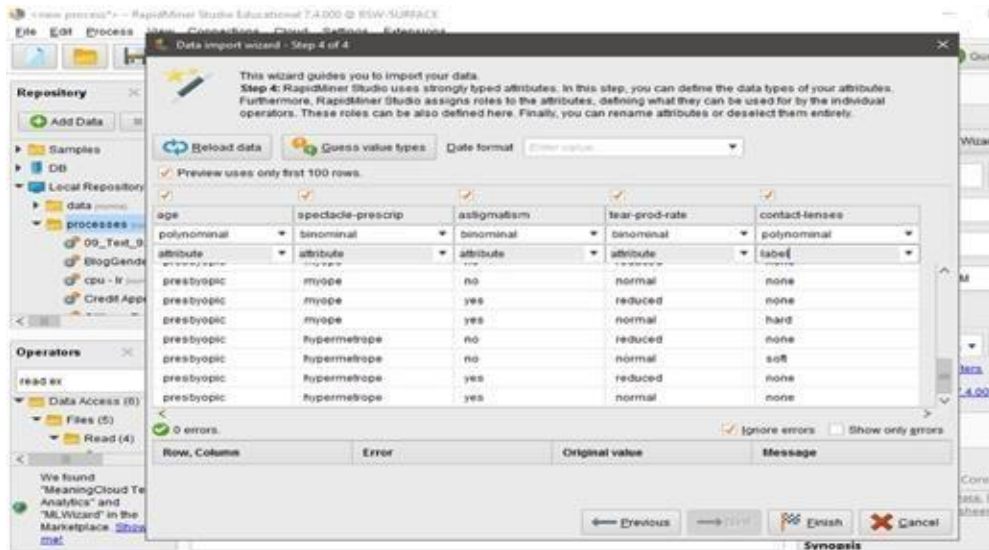


#### 5. Latihan: Rekomendasi Contact Lenses

1. Lakukan training pada data Contact Lenses (contact-lenses.xls) dengan menggunakan algoritma decision tree
2. Gunakan operator Read Excel (on the fly) atau langsung menggunakan fitur Import Data (persistent)
3. Tampilkan himpunan data (dataset) dan pengetahuan (model tree) yang terbentuk

Row No.	contact-len.	age	spectacle-p.	astigmatism
1	none	young	myope	no
2	soft	young	myope	no
3	none	young	myope	yes
4	hard	young	myope	yes
5	none	young	hypermetrop	no
6	soft	young	hypermetrop	no
7	none	young	hypermetrop	yes
8	hard	young	hypermetrop	yes
9	none	pre-presbyoi	myope	no
10	soft	pre-presbyoi	myope	no
11	none	pre-presbyoi	myope	yes
12	hard	pre-presbyoi	myope	yes
13	none	pre-presbyoi	hypermetrop	no
14	soft	pre-presbyoi	hypermetrop	no

## Read Excel Operator



## Import Data Function

<new process\*> - RapidMiner Studio Educational 7.4.000 © RSW-SURFACE

File Edit Process View Connections Cloud Settings Extensions

Import Data - Format your columns.

### Format your columns.

Date format: MMM d, yyyy h:mm:ss a z  Replace errors with missing values ⓘ

	age <i>polynomial</i>	spectacle-presc... <i>binominal</i>	astigmatism <i>binominal</i>	tear-prod-rate <i>binominal</i>	contact-lenses <i>polynomial label</i>
1	young	myope	no	reduced	none
2	young	myope	no	normal	soft
3	young	myope	yes	reduced	none
4	young	myope	yes	normal	hard
5	young	hypermetrope	no	reduced	none
6	young	hypermetrope	no	normal	soft
7	young	hypermetrope	yes	reduced	none
8	young	hypermetrope	yes	normal	hard
9	pre-presbyopic	myope	no	reduced	none
10	pre-presbyopic	myope	no	normal	soft
11	pre-presbyopic	myope	yes	reduced	none
12	pre-presbyopic	myope	yes	normal	hard
13	pre-presbyopic	hypermetrope	no	reduced	none

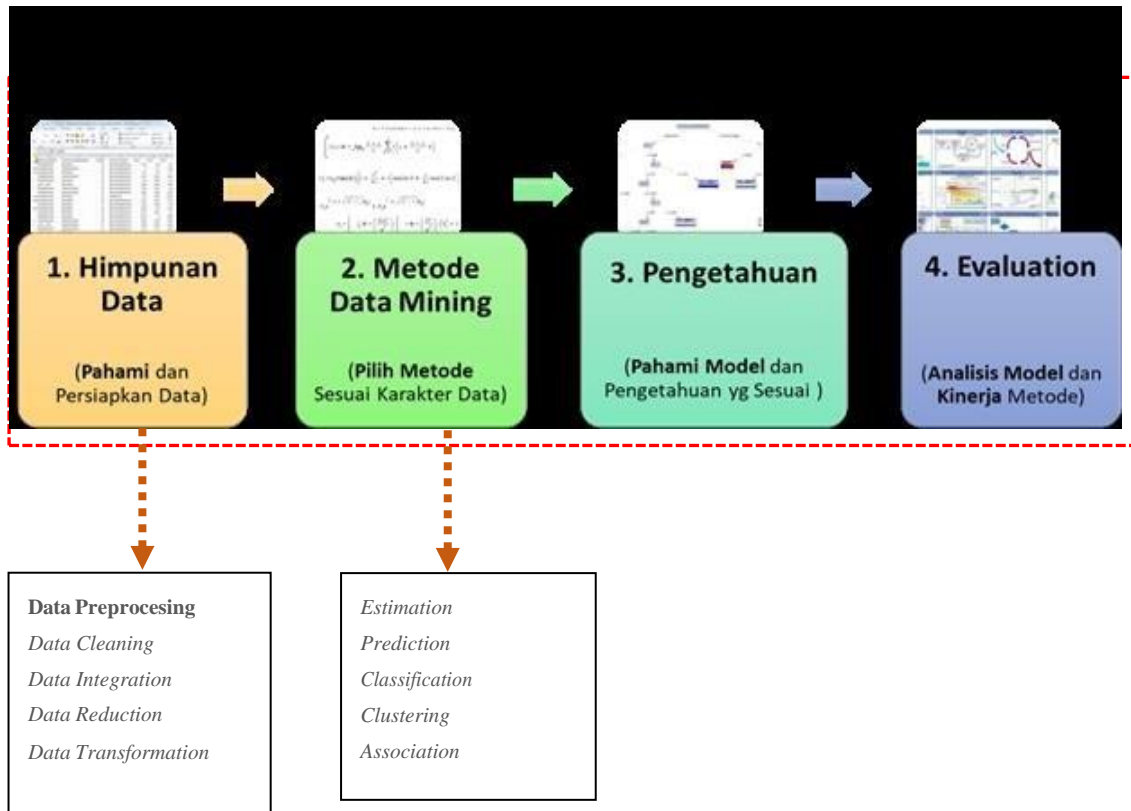
no problems.

Previous Next Cancel

We found "WhiBo" in the Chaid Trees

## BAB V EVALUASI MODEL DATA MINING

### A. Proses Data Mining



### B. Evaluasi Data Mining

1. Estimation:
  - Error: Root Mean Square Error (RMSE), MSE, MAPE, etc
2. Prediction/Forecasting (Prediksi/Peramalan):
  - Error: Root Mean Square Error (RMSE) , MSE, MAPE, etc
3. Classification:
  - Confusion Matrix: Accuracy
  - ROC Curve: Area Under Curve (AUC)
4. Clustering:
  - Internal Evaluation: Davies–Bouldin index, Dunn index,
  - External Evaluation: Rand measure, F-measure, Jaccard index, Fowlkes–Mallows index, Confusion matrix
5. Association:



- Lift Charts: Lift Ratio

Precision and Recall (F-measure)

Pembagian dataset, perbandingan 90:10 atau 80:20. Data training 90 dan data testing 10 atau Data training 80 dan data testing 20. Data training untuk pembentukan model, dan data testing digunakan untuk pengujian model. Pemisahan data training dan testing ada tiga cara yaitu:

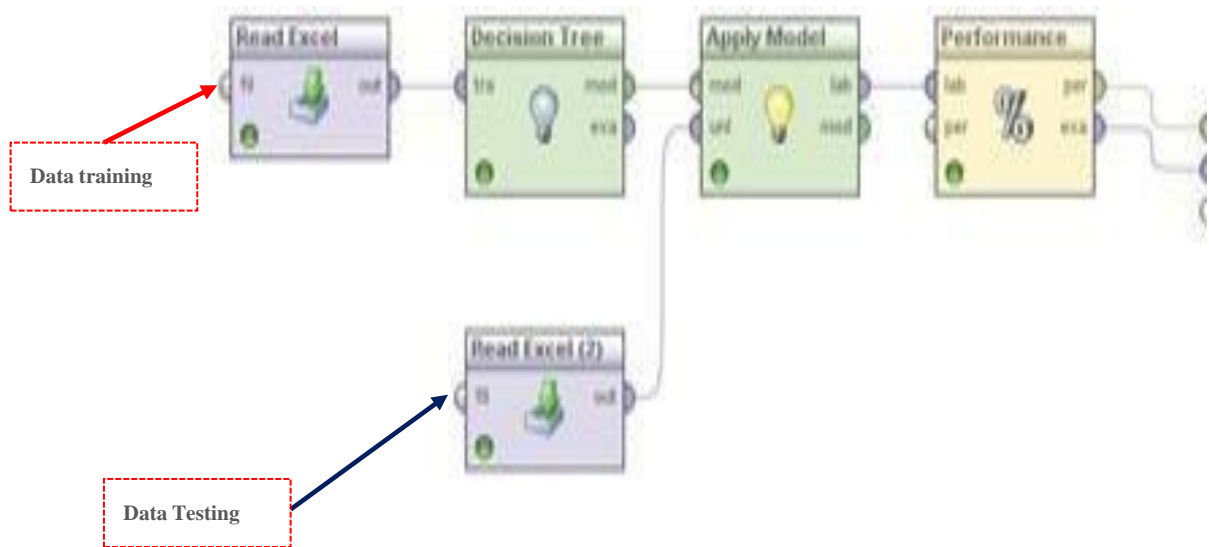
1. Data dipisahkan secara manual
2. Data dipisahkan otomatis dengan operator Split Data
3. Data dipisahkan otomatis dengan X Validation

### **Pemisahan Data Manual**

Pemisahan data manual adalah dataset dipisahkan secara fisik. Seperti contoh latihan di bawah ini.

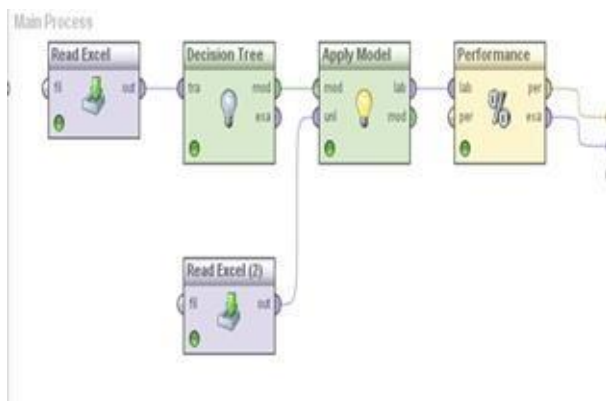
#### **Latihan: Penentuan Kelayakan Kredit**

- a. Gunakan dataset di bawah:
  - creditapproval-training.xls: untuk membuat model
  - creditapproval-testing.xls: untuk menguji model
- b. Data di atas terpisah dengan perbandingan:  
data training (90%) dan data testing (10%)
- c. Data training sebagai pembentuk model, dan data testing untuk pengujian model, ukur performancenya



### Latihan: Deteksi Serangan Jaringan

- Gunakan dataset di bawah:
  - intrusion-training.xls: untuk membuat model
  - intrusion-testing.xls: untuk menguji model
- Data di atas terpisah dengan perbandingan: data training (90%) dan data testing (10%)
- Jadikan data training sebagai pembentuk model/pola/knowledge, dan data testing untuk pengujian model
- Ukur performance (AUC dan Accuracy)

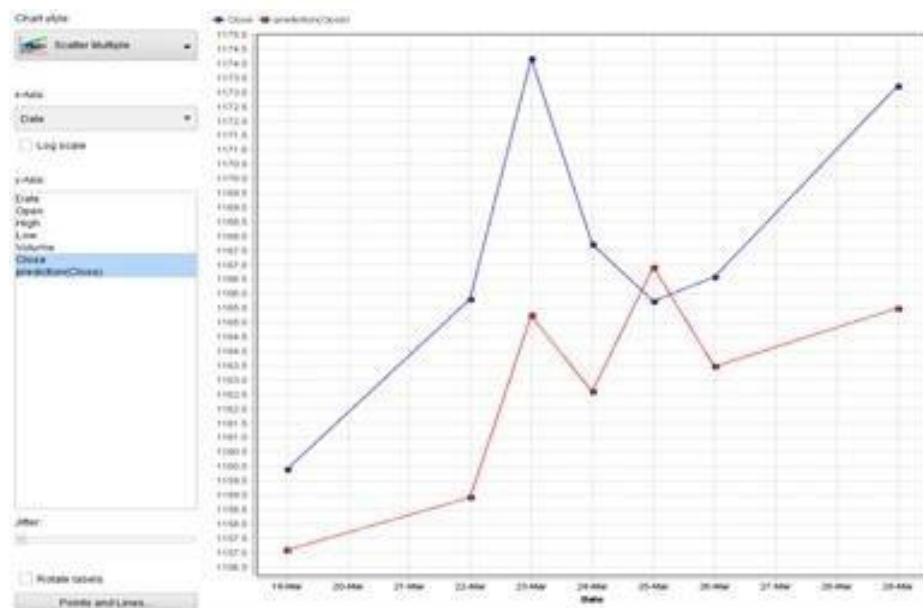
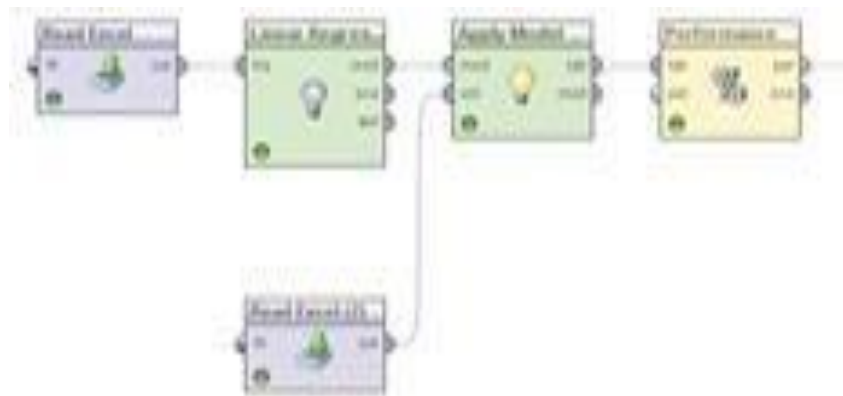


	C4.5
Accuracy	58%
AUC	0.86

### Latihan: Prediksi Harga Saham

- Gunakan dataset di bawah:

- hargasaham-training.xls: untuk membuat model
- hargasaham-testing.xls: untuk menguji model
- Data di atas terpisah dengan perbandingan: data training (90%) dan data testing (10%)
- Jadikan data training sebagai pembentuk model/pola/knowledge, dan data testing untuk pengujian model
- Ukur performance



### C. Pemisahan Data otomatis dengan operator Split Data

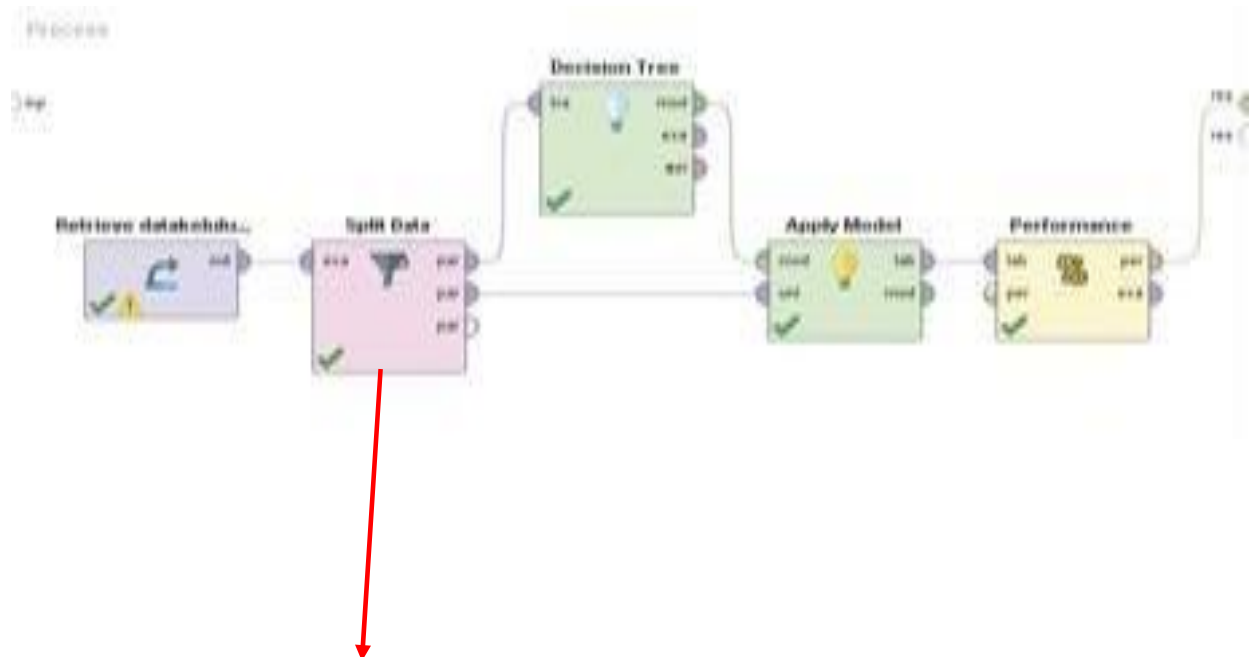
The Split Data operator takes a dataset as its input and delivers the subsets of that dataset through its output ports. The sampling type parameter decides how the examples should be shuffled in the resultant partitions:

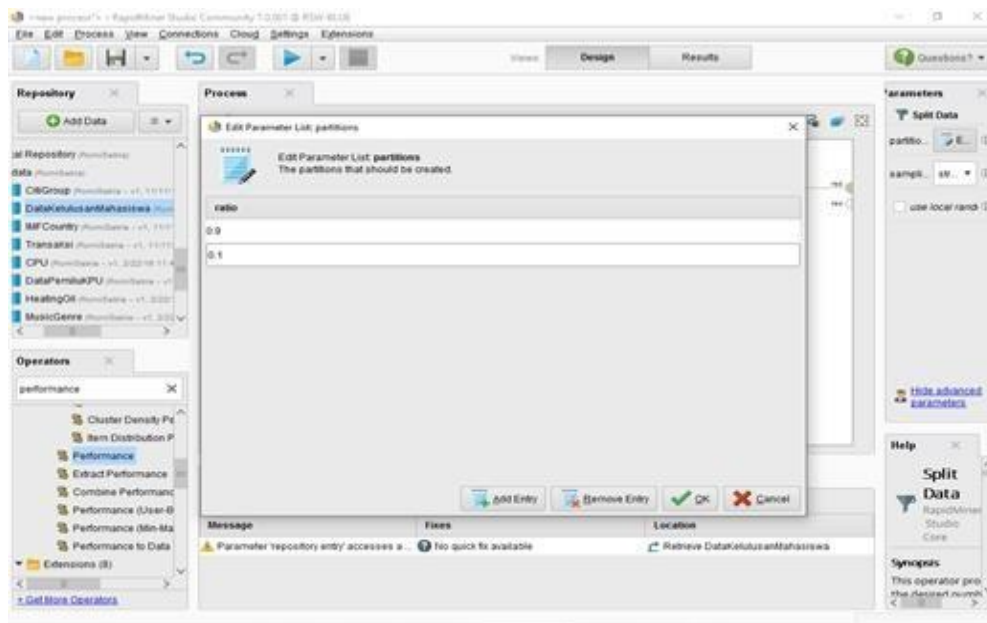
- a. Linear sampling: Divides the dataset into partitions without changing the order of the examples
- b. Shuffled sampling: Builds random subsets of the dataset
- c. Stratified sampling: Builds random subsets and ensures that the class distribution in the subsets is the same as in the whole dataset

### Latihan: Prediksi Kelulusan Mahasiswa

1. Dataset: datakelulusanmahasiswa.xls
2. Pisahkan data menjadi dua secara otomatis (Split Data): data training (90%) dan data testing (10%)
3. Ujicoba parameter pemisahan data baik menggunakan Linear Sampling, Shuffled Sampling dan Stratified Sampling
4. Jadikan data training sebagai pembentuk model/pola/knowledge, dan data testing untuk pengujian model
5. Terapkan algoritma yang sesuai dan ukur performance dari model yang dibentuk

### Proses Prediksi Kelulusan Mahasiswa





#### D. Pemisahan Data dan Evaluasi Model Otomatis dengan Cross-Validation

Metode cross-validation digunakan untuk menghindari overlapping pada data testing. Tahapan cross-validation:

1. Bagi data menjadi k subset yg berukuran sama
2. Gunakan setiap subset untuk data testing dan sisanya untuk data training

Disebut juga dengan k-fold cross-validation. Seringkali subset dibuat stratified (bertingkat) sebelum cross-validation dilakukan, karena stratifikasi akan mengurangi variansi dari estimasi

#### 10 Fold Cross-Validation

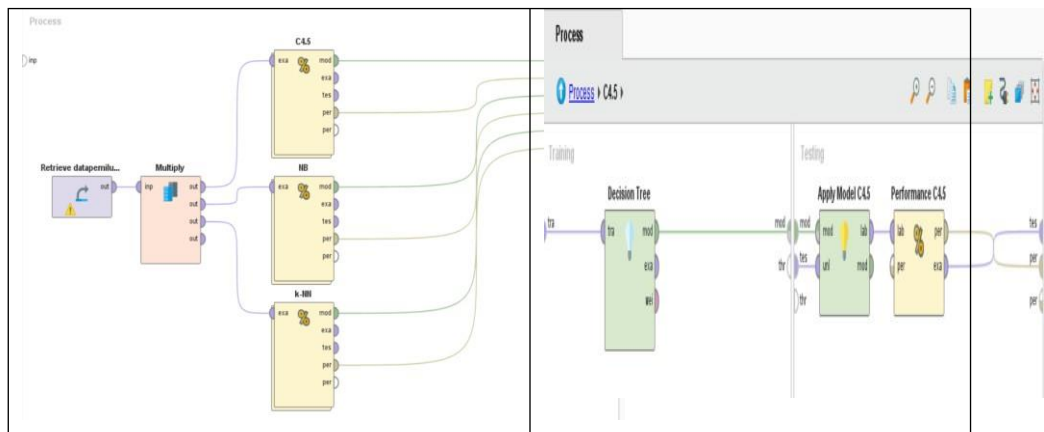
Metode evaluasi standard: stratified 10-fold cross-validation. Mengapa 10? Hasil dari berbagai percobaan yang ekstensif dan pembuktian teoritis, menunjukkan bahwa 10-fold cross-validation adalah pilihan terbaik untuk mendapatkan hasil validasi yang akurat. 10-fold cross-validation akan mengulang pengujian sebanyak 10 kali dan hasil pengukuran adalah nilai rata-rata dari 10 kali pengujian seperti pada gambar di bawah ini:

Eksperimen	Dataset	Akurasi
1	[Orange]	93%
2	[Orange]	91%
3	[Orange]	90%
4	[Orange]	93%
5	[Orange]	93%
6	[Orange]	91%
7	[Orange]	94%
8	[Orange]	93%
9	[Orange]	91%
10	[Orange]	90%
<b>Akurasi Rata-Rata</b>		<b>92%</b>

Gambar fold cross-validation

**Latihan: Prediksi Elektabilitas Caleg**

1. Lakukan training pada data pemilu (datapemilukpu.xls)
2. Lakukan pengujian dengan menggunakan 10-fold X Validation
3. Ukur performance-nya dengan confusion matrix dan ROC Curve
4. Lakukan ujicoba, ubah algoritma menjadi C4.5, Naive Bayes, dan k-NN, analisis mana algoritma yang menghasilkan model yang lebih baik (akurasi tinggi)



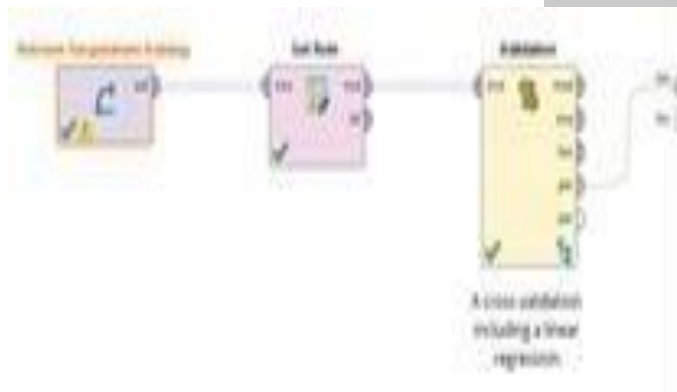
	C4.5	NB	k-NN
Accuracy	92.87%	79.34%	88.7%
AUC	0.934	0.849	0.5

### Latihan: Prediksi Harga Saham

1. Gunakan dataset harga saham (hargasaham-training.xls)
2. Lakukan pengujian dengan menggunakan 10-fold X Validation
3. Lakukan ujicoba dengan algoritma NN

Operator untuk menetapkan label data

	NN
RMSE	7.334



Confusion Matrix → Accuracy

accuracy: 90.00%

	true MACET	true LANCAR	class precision
pred. MACET	53	4	92.98%
pred. LANCAR	6	37	86.05%
class recall	89.83%	90.24%	

pred MACET- true MACET: Jumlah data yang diprediksi macet dan kenyataannya macet (**TP**)

pred LANCAR-true LANCAR: Jumlah data yang diprediksi lancar dan kenyataannya lancar (**TN**)

pred MACET-true LANCAR: Jumlah data yang diprediksi macet tapi kenyataannya lancar (**FP**)

pred LANCAR-true MACET: Jumlah data yang diprediksi lancar tapi kenyataannya macet (**FN**)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{53 + 37}{53 + 37 + 4 + 6}$$

$$= \frac{90}{100} = 90\%$$

### Precision and Recall, and F-measures

Precision: exactness – what % of tuples that the classifier labeled as positive are actually positive

$$\text{precision} = \frac{TP}{TP + FP}$$

Recall: completeness – what % of positive tuples did the classifier label as positive?

$$\text{recall} = \frac{TP}{TP + FN}$$

- Perfect score is 1.0
- Inverse relationship between precision & recall

F measure (F1 or F-score): harmonic mean of precision and recall,



$$F = \frac{2 \times \textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}}$$

F $\beta$ : weighted measure of precision and recall

$$F_{\beta} = \frac{(1 + \beta^2) \times \textit{precision} \times \textit{recall}}{\beta^2 \times \textit{precision} + \textit{recall}}$$

- assigns  $\beta$  times as much weight to recall as to precision

### Sensitivity and Specificity

Binary classification should be both sensitive and specific as much as possible:

1. Sensitivity measures the proportion of true 'positives' that are correctly identified (True Positive Rate (TP Rate) or Recall)

$$\textit{Sensitivity} = \frac{\textit{Number of 'True Positives'}}{\textit{Number of 'True Positives'} + \textit{Number of 'False Negatives'}}$$

2. Specificity measures the proportion of true 'negatives' that are correctly identified (False Negative Rate (FN Rate) or Precision)

$$\textit{Specificity} = \frac{\textit{Number of 'True Negatives'}}{\textit{Number of 'True Negatives'} + \textit{Number of 'False Positives'}}$$

### PPV and NPV

We need to know the probability that the classifier will give the correct diagnosis, but the sensitivity and specificity do not give us this information

- Positive Predictive Value (PPV) is the proportion of cases with 'positive' test results that are correctly diagnosed

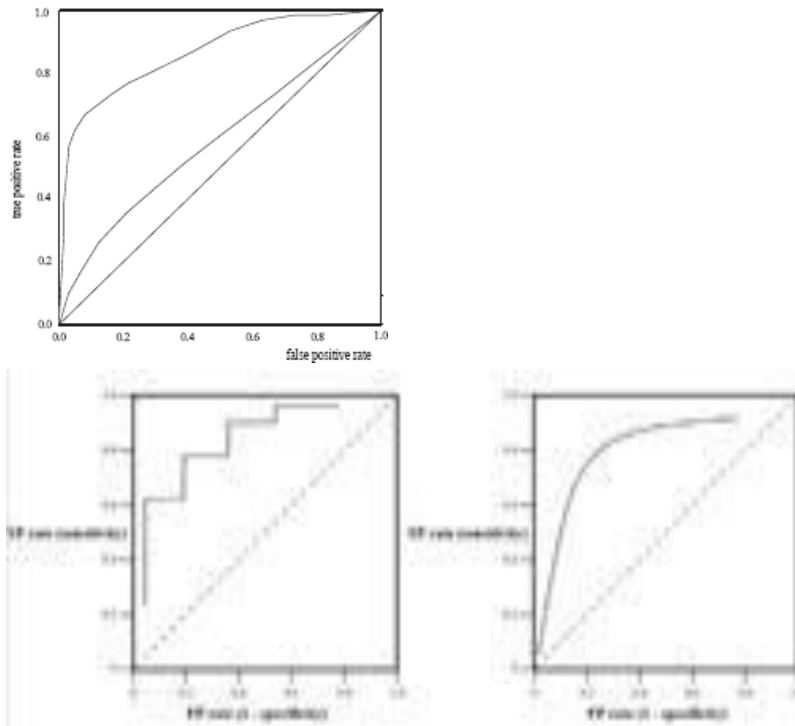
$$PPV = \frac{\textit{Number of 'True Positives'}}{\textit{Number of 'True Positives'} + \textit{Number of 'False Positives'}}$$

- Negative Predictive Value (NPV) is the proportion of cases with 'negative' test results that are correctly diagnosed

### Kurva ROC - AUC (Area Under Curve)

- ROC (Receiver Operating Characteristics) curves: for visual comparison of classification models
  - Originated from signal detection theory

- ROC curves are two-dimensional graphs in which the TP rate is plotted on the Y-axis and the FP rate is plotted on the X-axis
- ROC curve depicts relative trade-offs between benefits ('true positives') and costs ('false positives')
- Two types of ROC curves: discrete and continuous



### **Guide for Classifying the AUC**

1. 0.90 - 1.00 = excellent classification
2. 0.80 - 0.90 = good classification
3. 0.70 - 0.80 = fair classification
4. 0.60 - 0.70 = poor classification
5. 0.50 - 0.60 = failure

*(Gorunescu, 2011)*

## DAFTAR PUSTAKA

- Cirillo, A. (2017). *R Data Mining: Implement data mining techniques through practical use cases and real world datasets*. Packt Publishing Ltd.
- García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (Vol. 72). Cham, Switzerland: Springer International Publishing.
- King, R. S. (2015). *Cluster analysis and data mining: An introduction*. Stylus Publishing, LLC.
- Makhabel, B. (2015). *Learning data mining with R*. Packt Publishing Ltd.
- Patel, S., & Patel, H. (2016). Survey of data mining techniques used in healthcare domain. *International Journal of Information*, 6(1/2), 53-60.
- Roiger, R. J. (2017). *Data mining: a tutorial-based primer*. Chapman and Hall/CRC.