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 dipraktikumkan 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

KATA PENGANTAR	HALAMAN SAMPUL	i
DAFTAR ISIiv BAB I DOWNLOAD DAN INSTAL R PADA WINDOWS	KATA PENGANTAR	iii
BAB I DOWNLOAD DAN INSTAL R PADA WINDOWS	DAFTAR ISI	iv
BAB II TEKNIK-TEKNIK PRAPROSES DATA	BAB I DOWNLOAD DAN INSTAL R PADA WINDOWS	1
BAB III PROSES DATA MINING	BAB II TEKNIK-TEKNIK PRAPROSES DATA	9
BAB IV PENERAPAN DATA MINING	BAB III PROSES DATA MINING	15
BAB V EVALUASI MODEL DATA MINING	BAB IV PENERAPAN DATA MINING	35
DAFTAR PUSTAKA	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 analysist*, klasifikasi, clustering). R memiliki berbagai fungsi *built-in* dan juga fungsi *extended* untuk tugas statistic, *machine learning* dn 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 <u>https://www.rstudio.com/</u>



Gambar 1.2 Situs resmi dan pengunduhan RStudio

- 2. Clik pada CRAN *section*, pilih CRAN mirror dan pilih sesuai dengan sistem operasi yang digunakan (windows).
- 3. Download R version dari mirror
- 4. Istall 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 \rightarrow 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 kana pada icon Rstudio dan pilih run as administrator.
- 2. Kemudian pilih next \rightarrow next \rightarrow install \rightarrow 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.

the local data and the second s				1.8
Contra a second de an			-	_
1000 () ()				-
Lange 12-100 million for the second second	B inclusion		1.17	
BARRADSTRUCT				
A in a sufficient of a state of the set of t				
ten and the set of the		- 10 000		-
descendent basing that stating			-	
18	target land			
	and the second	- Andrew Collection (Collection of Street Collection of St.	1.1.8	100
	10. mm	Summarily Institution	111	
	10 mm	Name Lind of the One-Industry		- 10
	C. H. wanted	Case Realist Rescip-1	100	- 61
	10 maple	Test Constitution	1.00	
	11 million	The Convertings	- 14	
	11. Trees.	familiar involve more 1 ing (MS) that (see		
	The second	In J. Courts, And Mr.	1.00	1.1
	11 attent	The Printer Security Security Security	A	161
	10.00	Territoring to head	1.00	1.0
	10 million (from the second se	a 1000 -	1
	E. M. Law	San Associated	14.4	1.5
	Total Mark	And investory of the state of t	1 Aug. 1	1.0

Gambar 4 jendela awal Rstudio

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

) Realer - a commer frank					1 Sector	- Apr., 10
File Bill Colle View Flats Insuine Build Denug Taxe. Phys.	3				-	
With the state of					S from	1 (Sec.)
Eliment a	-7	Training (Holey)				-01
C B Deventer S Z + D	CARL TRECEMENT IN	if il iPhanthes	ate 🖌			e+119
11		d interterent.		-		
				3c		
			- Photos			
3a						
54						
		In the later	And America	_	_	-
		States Street		and the second s		1000
		Same Provent	(Income in case of the local section of the local s	34	-	-10
		System Likeway		Ju		1.14
		E best	Roman Purchase (2	Ingenality by the good Caring Acri (1	1148.	- 43
Add Deliver I	a longe 1	(C) 4444	Autolices for Casalicy		3.3-10.2	- 11
Constitution of the		0.000	Training Groups in Data	Contaction designs (Interceded	358	- 12
& it free officers and comet with Abdocativ we wantery.		C unteres	Caste Strations Tests for	4	6214	- A.U.
Type "license()' or 'license()' for door theirs details.		C sargete	Darb Compile Factor		-13.8	- 49
		C Alexen	The # Dationals Publicage		118	
rype contributors) for more toferaulton and		E here	Read Cale Strend by Marries, efforts -	house 5, 567, 2015, Harry Syntol,	-14-10	- *1
contraction of a state of a state of a particular of particular state.		EL grates	Datingto, Peter	£	218	- 11
Type "deno()" for some denon, "he'p()" for on-live he'p, of The's start()" for an own broater interface to he's.		21 planet	Da Chahia Perce	and Support for Calcura and Socio-	3.88	- 4
YyBe 'BC1' to april #,		10 ml	The Dire Insultion Part		3,1.9	
(serkspace loaded from -/. 45eta) 3h		E Ardense	Apagent he form in 2003	making Laportry Hand & Some	10.11	
	10	E 164	Tradis-Depther fold		3128-44	
		C IN BRID	thereast furniture and	Names for Social and Passars	11.00	

Gambar 5 Jendela editor pada Rstudio

Keterangan:

- 3a. = Untuk menuliskan kode perintah (jendela console)
- 3b. = Tempat kode perintah dieksekusi (jendela environtment)
- 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. Assosiasi

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:

- Bulan Jawa Sumatra Sulawesi Bali 2 Januari 12 4 6 9 5 5 Maret 3 7 8 4 9 Juni 5 September 12 5 2 10 8 6 Desember 6
- 1. Buatlah file produksi.Rdata dan produksi.csv dari tabel berikut:

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

Langkah-langkah menyelesaikan kasus 1:

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

>	<pre>Bulan <- c("Januari", "Maret", "Juni", "September", "De</pre>
	sember")
>	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)
>	<pre>save(a,file = "produksi.Rdata")</pre>
>	<pre>write.csv(a,"produksi.csv", row.names = FALSE)</pre>
>	<pre>read.csv("produksi.csv")</pre>

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

>	read.csv('	'produ	uksi.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 <- paste(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="")</pre>
```

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)</pre>

- 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. Dll

C. Min-Max

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

Rumus:

Newdata = (data-min) * (newmax-newmin) / (max-min) + 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

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

newdata = 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

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 nlai-nilai khusus yang merepresentasikan pengecualian-pengecualian 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
```

<pre>\$ Temp : int \$ Month : int \$ Day : int</pre>	67 72 74 62 56 0 5 5 5 5 5 5 5 5 5 1 2 3 4 5 6 7 8	66 65 59 61 69 5 5 9 10	
ş Day . Inc	12343070	9 10	
<pre>> summary(data)</pre>	Solar R	Wind	Temp
Month	Solutin	WING	remp
Min. : 1.00	Min. : 7.0	Min. : 1.700	Min. :5
1st Qu.: 18.00	1st Qu.:115.8	1st Qu.: 7.400	1st Qu.:7
Median : 31.50	Median :205.0	Median : 9.700	Median :7
Mean : 42.13	Mean :185.9	Mean : 9.958	Mean :7
3rd Qu.: 63.25	3rd Qu.:258.8	3rd Qu.:11.500	3rd Qu.:8
Max. :168.00	Max. :334.0	Max. :20.700	Max. :9
NA's :37	NA'S :7		
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)

> 1ib	rary(mice)						
> dat	a <-	airqua	ality					
> md.	patte	rn(ai)	rquali	ty)				
W	ind т	emp Mo	onth D	ay	Solar.R (Dzone		
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	Ŏ	Ō	2	
	ō	Ō	ō	0	7	37	44	
	Ū	Ū	Ũ	Ũ			•••	

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 vind Temp Month Day Ozone
116 1 1 1 1 1 0
37 1 1 1 1 0 1
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 pubik 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)

7

	Cases			•	kependary 🔍 V		Search
chine Le	earning Repository				Vie	WALL Dat	ta Se
se Through	360 Data Sets					Table View	List Vie
terask Scaton (262)	Name	Data Types	Default Task	Attribute	# Instances	Attributes	Year
15(0) (63) 1000 (54) (52)	Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
osical (37) lical (213) (56)	Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Vpe ariate (281) 1ate (16)		Multivariate	Classification	Categorical, Integer, Real	798	38	
eries (05) 2) In-Theory (22)	UCI Anonymous Microsoft Web Data		Recommender- Systems	Categorical	37711	294	1998
sences (82)	Arrhythmia	Multivariate	Classification	Categorical. Integer, Real	452	279	1998
Sciences (23)	Artificial Characters	Multivariate	Classification	Categorical, Integer, Real	6000	7	1992
67) butes	Audiology (Original)	Muttivariate	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)

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

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. Kehandalan
 - 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)
- Perspektif Desain
 (Design perspective)
- Perspektif Hasil (Result perspective)



View Operator



- 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

2	F	Parameters 💥 💱 🗢 🔯 🔿				
- Frank - Fran	K1	- 2 -	-		😼 📑 🗝	
logverbosity	init		-	Nom	inal to Date	
logfile		date typ	e name e	•	date	
resultfile		date for	nat			
random seed	2001	time zo	ne		SYSTEM	
send mail	never	locale	en old s	dirihu	English (Unite	
encoding	SYSTEM	=	p 010 0			

View Help dan View Comment

View Help menampilkan deskripsi dari operator

View Comment menampilkan komentar yang dapat diedit terhadap operator



View Problems

🎽 📁 🛏 🔹 🕨 •	100 C	views	Design	Results	Turbo Prep	Auto Model	Deployments		And data spectros, et	2	All Studio *
Repository ×	Process					0	0		Parameters	× [
Compositions II ·	() Process					-	10 10 10	10 M	logwitosity	nt	
data data data	Refere	ne databung	at. Decis	ine free					logtie		
databungaite label (torce + in datavelukusanmahapidwa () co	2	1							resultire		
datapemikukpu (a calco a to mu -									random seed	2001	
harpasaham-training (action 7.1				Missing la	abel				send mail	never	*
Koroder i sizzizi tri di Auria tri al 🗸				input Examp attribute.	AeSet does not ha	ve a label	2		encoding	SYSTEM	
Operators 🛛				Got it!	Mara Infa				Z 16de advances	1 parameters	
decisi X					0.000000 N				✓ Change come	100 CT (0 100	20
Modeling (8)											
Trees (8) Decision Tree Random Forest									Process RapedMee	r Studio Core	
Gradient Boosted Trees									Synopsis The root operato	r which is the	outer
<	Leverage the Wisdom	n of Crowds 1	lo gell operator n	ecommendations	based on your pro	was design			most operator of	every proces	5
Get more operators from the Marketplace	-			A varia	e waracen of Crowd	•.:			Description		

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)

1	Apply	/ Moc	lel
	mod	0	lab D
0	unl	U	mod D
	• /	2	_



- 4. Indikator status dari operator:
 - Lampu status: merah (tak tersambung), kuning (lengkap tetapi belum dijalankan), hijau (sudah behasil 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 \rightarrow Run
- Menekan kunci F11



Melihat Hasil



Panduan Install Rapid Miner







	Registrasi Lisensi	10	
	Lengkapi semua field. Konten dapat mengacu pada histori perkuliahan anda Ceklis "have read and accept the end- user license agreement" dan "l hereby confirm that I am eligible and that I agree to meet the requirements." Kik Apply for license		
1	Registrasi Lisensi 9 Juka Kembali aplikasi Rapidminer 9 Astikan bahwa lisensi anda telah aktif seperti pada gamber berikut Fridemonda gamber berikut Fr		

BAB IV PENERAPAN DATA MINING

A. Proses Data Mining

	$ \begin{array}{c} & & \\ \left\{ \begin{array}{c} (\alpha + \alpha + \beta + \alpha + \beta + \beta + \beta + \beta + \beta + \beta + $	H ¹ H ²	
1. Himpunan Data (Pemahaman dan Pengolahan Data)	2. Metode Data Mining (Pilih Metode Sesuai Karakter Data)	3. Pengetahuan (Pola/Model/Rumus/ Tree/Rule/Cluster)	4. Evaluation (Akurasi, AUC, RMSE, Lift Ratio,)
•	•		
Data PreProcesing Data Cleaning Data Integration Data Reduction Data Transformation	Estimation Prediction Classification Clustering Association		

1. Latihan: Rekomendasi Main Golf

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



Buka RapidMiner



Muncul Lembar Kerja



Membaca Data dan menampilkan data

	C ► • ■	Views F	Design	Results	Ovestor
Reposition X Add Data C Add Data C Add Data C Add Data C Coll Testset (+7) C Coll T	Process × Process Proc	و هر هر ۲۵۵۳ و	> <mark> 2</mark>	Par C repo	ameters × Retrieve Golf (Retrieve) siltory entry pieardata/Golf
Data Access (46)	Problems × % No problems found Message		Fixes	Hel	Retrieve RapidMiner Studio Core NapidMiner Studio Core

Normalize particular partecolar particular particular particular particular		14 .	2 6		0	These states	Cwords	Results		Carolina Carolina	100
Descripted (54 exercipted, 1 space ablode, 6 registra ablode) Pilter (14) 4 stamples)	Result History	× 🖬 (nampleSet (Rot	rieve Golf) — IC							
Rev Ru. Pary Oblick Temperature MemBit Prior 1 ni. survey 85 50 50 2 ni. survey 85 50 50 2 ni. survey 85 50 50 3 prior survey 87 98 50 3 prior survey 87 98 50 3 prior survey 87 98 50 6 prior survey 87 98 50 6 prior survey 88 50 50 7 prior survey 16 50 50 6 prior survey 72 50 50 6 prior survey 73 50 50 6 prior survey 73 50 50 6 prior survey 73 50 50	-	Danskebet (5	4 exempted, 1 top	coar affritude, 4 regula	(anticides)			Filter (14/154 assertgiles)	10	*	18
1 no nonv 45 66 Nohe 2 no surve 83 90 Train 2 no surve 83 90 Train 3 pro overcard 83 90 None 3 pro overcard 83 90 None 6 pro overcard 83 90 None 6 pro overcard 83 50 None 6 pro overcard 84 85 Store 6 pro overcard 84 85 Store 7 pro overcard 84 85 Store 7 pro overcard 72 94 Note 9 pro overcard 73 Store Note 11 pro overcard 72 93 Note 12 pro overcard 72 94 Note	Orde	Now No.	Flay	Cuttook	Temperature	Hamilty	Wed				
2 m sum 88 90 true 3 98 6vercarl 80 78 Mite 4 960 can 70 Mite 6 76 can 70 Mite 6 76 can 70 Mite 6 76 can 70 Mite 7 169 can 70 Mite 7 res can 70 Mite 7 res can 70 Mite 6 res can 70 Mite 7 res can 70 Mite 6 res can 71 Mite 71 res can 72 Mite 71 res res res res 71 res res res res 72 res res res res 72 res			84	survy :	85	- 85	taise				
Description 3 yes evercast 63 78 Mate 4 yes tain 79 Mate 4 yes tain 79 Mate 6 yes tain 79 Mate 6 yes tain 69 Mate 6 tein cast 69 Mate 7 tein tein tein Mate 6 tein central 68 79 Mate 6 tein central 68 79 Mate 6 tein central 71 Mate Mate 7 tein tein tein 71 Mate 6 tein tein 71 89 Mate 7 tein tein 71 89 Mate 7 tein tein 72 79 Mate 7 tein tein 75 M	1 mm	2	-	samy	80	90	THE				
Bankess 4 pe ean 70 90 5000 6 98 1000 69 500 5000 6 60 600 60 70 Fore 70 98 000 70 Fore 70 98 000 70 Fore 8 00 0000 70 5060 9 98 0000 70 5060 10 98 0000 70 5060 11 98 0000 72 50 5000 11 98 0000 72 50 5000 Anneet 11 98 0000 72 50 5000 Anneet 11 98 0000 6000 50 5000 5000 Anneet 11 98 0000 50 5000 5000	Σ	3	198	overcast	40	28	Table				
6 yes ran 48 50 64a 6 6a 6a 6a 73 Rue 7 yes owcat# 64 65 Sue 7 yes owcat# 64 65 Sue 8 me owcat# 64 65 Sue 9 me owcat# 72 66 Nue 13 me owcat# 75 50 Sue 14 yes owcat# 72 50 Sue 14 yes owcat# 72 50 Sue 14 yes owcat# 72 50 Sue 14 me owcat# 72 50 Sue 14 me owcat# 72 50 Sue	Distance	4	199	1901	78	99	Talaxe				
6 60 60 60 60 70 Inde 7 190 outstaft 64 65 Inde 9 10 100 100 100 Inde Abanced 10 190 00000 17 100 Inde Abanced 13 140 00000 75 20 Inde Abanced 13 140 000000 75 20 Inde Abanced 13 140 0000000 72 20 Inde Abanced 14 0000000 17 20 1000000000000000000000000000000000000			144	1971	44	80	false				
Chains 7 yes constant 64 65 fore 0 res europ 72 16 hate 0 res europ 72 16 hate 0 res europ 73 50 hate 0 res control 73 50 hate 0 res control 73 23 hate 11 yes control 72 30 hote 22 yes oversatt 81 75 bits Annethine 14 res control 73 bits	100		-	1991	60	79	Put .				
0 eve evere 72 95 Mate Advanced 0 ref evere 13 70 Mate 13 ref ref ref 35 Mate 14 ref evere 72 20 Mate 22 ref evere 72 30 Mate Anneet ref evere 72 30 Mate Anneet ref ref 73 Mate ref Anneet ref ref ref ref ref ref	Charts	7	199	overcast	84	65	the				
B yes summy 68 79 false Advances 13 yes rass 75 80 bites 11 yes summy 72 73 bites 12 yes oversistit 72 29 bites Associations 14 ma oversistit 72 39 bites			**	0.0712	72	95	Talan				
Advanced Device 13 yes rgs rgs 20 Main 11 yes survey 75 25 Inse 12 yes oversatil 72 35 Inse Annual yes oversatil 81 55 Main Annual yes oversatil 81 95 Main	100		198	aures -	48	79	taise				
Obsets 11 yes surfly 75 79 Isse 12 yes oversait 72 90 Pose 13 yes oversait 81 75 talse Annufatives 14 ms tan 71 80 Pose	Abanad	-10	143	1941	78	80	Talan.				
12 pilo overlash 72 30 Pue 13 pilo overlash 81 75 talsa 14 ma tam 71 80 Pue	Churts	99	785	samy	75	79	the .				
13 yes outstate 81 75 take Annotations 14 ms tam 71 80 two	100	-12	195	overcast	72	30	848				
Accolutions 14 ms cam 11 80 true		13	198	overcast	81	75	Tales				
	Acculations	34	**	1911	.79	80	914				

Menampilkan Statistik Data

Result History	ExampleSe	c (Rettleve Gott) 🔅				
	112-TH) < . Tate	Maxing	Sulvice	Filer (5+5 abrovles): Unerth to	aleman Tra
Data				1		
	A Play	Nominat			no (5)	yes (0)
Σ				- Lawrence	ter.	
SURVICE .		and the second		Look (d)	trial (Th)	100 million (15) augusta
-	Outlook	ALC: NO.		Controller (4)	nam (o)	
Charts	Temperature	Integer		64	85	73.571
				100	-	And and a second se
	Y Humidity	bileger		65	96	80.296
Charts	-	history	1	true (D)	tions (E)	Table (E), true (I
-	wind	10000	÷	a de fait	and (a)	
Antoitationa						
	<					5
	Showing attributes 1-5				Examples 14 Opecial Attributes	1 Regular Attributes: 4

Membuat Model

- I-I -	*> C*		 10001 	Veren	Design	14	auto	60	Questions?
Add Date .	Process © Process	×		100%	P P - 3	- 131	Parameters	× []	
geraters Gat into Gat into Gat into Present (R) Present (R) Present (R) Present (R) Present (R) Present (R)	Pressie D re		<i>C</i>	Proban Iron		110	Inguistanty Register Resultive Resultive Resultive Resultive Resultive Resultive Resultive Change comp	Inst 2003 nevel bystem dd pacateleta astroky (7.3.0	
2 603 Decision filore Decision filore Decision filore Randem filer Section file Section file Sec	Problems To protein Nessage	ini In Frankel			Frees	,	Help X Proces Paperster Synapsis The root operator most operator	s or shalls Corr or which is the d every proce	outer IS.

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

- 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	(Chronich Territory)
1	none	young	myope	no	(sea through the
2	soft	young	туоре	no	
3	none	young	туоре	yes	a kontak di sekan
4	hard	young	myope	yes	W. A.
5	none	young	hypermetrop	no	adignation
6	soft	young	hypermetrop	no	and the second second
7	none	young	hypermetrop	yes	* 22 * 22 million
8	hard	young	hypermetrop	yes	
9	none	pre-presbyoj	myope	no	Tank Constants in the
10	soft	pre-presbyor	myope	no	(and (and and a second
11	none	pre-presbyoj	myope	yes	a second design of the second second
12	hard	pre-presbyoj	myope	yes	- to to bagh bound - planae and
13	none	pre-presbyoj	hypermetrop	no	Track I Frank Track Track Track
14	soft	pre-presbyoj	hypermetrop	no	A DESCRIPTION OF THE OWNER OWNE

Read Excel Operator

2 📩 너	Data Import wisar		Sep 4 of 4	12								×	200
Repository ×	Step- Furth- opera	4: R. erm dors	apidMiner Studio uses str re, RapidMiner Studio as . These roles can be also		gly typed attrib pris roles to the efficient here. Fil	utes in thi e attributes naity, you c	in 1 6, 6 300	dep, you can define tefning what they c rename attributes	an be u or dese	ta types of your altr sed for by the indu- ilect them entirely.	ibutes. idual		
Samples	C Beload data	16	Guess value types	4	ate format	Circularia							Witte
• 📕 DB	Preview uses on	dy 84	at 100 rows.										
E Local Repository	1		1		120			20		32			
· data mente	age		spectacle-prescrip		astigmatism			tear-prod-rate		contact-lenses			
CP 00 Test 0	polynominal	٠	binominal	٠	binominal		•	binominal		polynominal			
BlogGende	attribute	٠	attribute	•	attribute		•	attribute		label			
gr cpu - tr	presbyopsc		myope		0.0			normal		none		12	M
G Credit App	presbyopic		myope					reduced		none			
C 11 M 11	presbyopic		myope		1985			normal		hard			
	presbyopic		hypermetrope		no			reduced		none			
Operators ×	presbyopic		hypermetrope		no			normal		808		1	
readex	presbyopic		hypermetrope		yes			reduced		none		18	
Data Access (6)	presbyopic		hypermetrope		yes			normal		none		4	4.9
 Files (5) Read (4) 	0 errors.								- (c) 19	nore errors 📄 8	how only gr	> rors	
<	Row, Column		Error			Origi	-	il value		Message			
We found "MeaningCloud Te Analytics" and													Cor
"MLWizard" in the Marketplace Show							+	- Erevious -	+ int	Post Entern	X Can	cel	1
met		-		-			-			Synopsi		_	1

Import Data Function

	mpo	rt Data - Format your colum	ns.	Design	Beaulta	Hodoop Data X	P
Repository			Form	nat your columns	i.		1
Samples		Date format MMM d, yyyy h	n:mm:ss a z 🔹 🔻	Replace err	ors with missing values 🕧)	rati
Local Reposi		age 🔅 🔻	spectacle-presc • •	astigmatism 🔅 🔻 binominal	tear-prod-rate 🔹 🔻 binominal	contact-lenses & • polynominal label	
processes	1	young	myope	no	reduced	none	<
Cioud Repos	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	
101	9	pre-presbyopic	myope	no	reduced	none	het
	10	pre-presbyopic	myope	no	normal	soft	
perators	11	pre-presbyopic	myope	yes	reduced	none	
ris	12	pre-presbyopic	myope	yes	normal	hard	
Modeling (8)	47	nra nrashuania	hunarmatrana		raduand	💙 no problems.	e

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 tig acara 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
- Pisahkan data menjadi dua secara otomatis (Split Data): data training (90%) dan data testing (10%)
- 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
- Terapkan algoritma yang sesuai dan ukur performance dari model yang dibentuk

Proses Prediksi Kelulusan Mahasiswa



🔎 🐸 Hell -	5 C	•	100		Vere.	Design	Results			Questions?
Repository 🖂	Process	×								*arameters
O Add Data = +	tat Para	anar Line pian	Nora .					×	Pa 🖝 133	T Split Data
-	~									partito. 78.
Na Constant		Edt Parame	oter List parts na mat should	be created.						and an a
CeGroup Promitants - ef. 11111	-1									sauge, se. e
DateKetulusanMahasiswa	ratio								1.000	C use locer rande
MFCountry Aundania + et. 110	0.9									
Trataatal Avendaria - e1, 1111										
CPU (Novellance - v). 22218-114	0.1									
DataPersiukPU (feedbarre1										
HeatingOE monthere - 11.0000										
MubicGenre munchana - rt. 311	31 I									
	-									
Operators 🔍										
performance >										- Hits advances
-	4									** zacameterz
S Cluster Density Pe	P.									
S item Distribution P									- C	Help 🛪
Tenformance										Calls
B Combine Budience					-	(-	Sprit
R Partymore disas						Bernese Entry	✓ QK	X Cancel		Toata
R Performance (Min-Ma	Message			Fixes			Location		1	Studio
S Performance to Data	A Paramater	repository em	ty accesses a	. O teo o	uick fix available		rt Retrieve Da	takietutus antitatria	NVAS.	Core
Edenziona (8)	3									Synopsis
	(W)									Contraction and

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		93%
2		91%
3		90%
4		93%
5		93%
6		91%
7		94%
8		93%
9		91%
10		90%
	92%	

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



Confusion Matrix \rightarrow Accuracy



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 lancer $(\ensuremath{\textbf{FP}})$

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

Accuracy =
$$\frac{\mathbf{TP} + \mathbf{TN}}{\frac{\mathbf{TP} + \mathbf{TN} + \mathbf{FP} + \mathbf{FN}}{90}} = \frac{53 + 37}{53 + 37 + 4 + 6}$$

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

Precision and Recall, and F-measures

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

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

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

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

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

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

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

Fß: weighted measure of precision and recall

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

• assigns β 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)

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

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

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

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

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

• 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 Yaxis 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.