MODUL
DATA MINING

PRIMA DINA ATIKA | WOWON PRIATNA

FAKULTAS ILMU KOMPUTER
UNIVERSITAS BHAYANGKARA JAKARTA RAYA

KATA PENGANTAR

Tujuan modul Data Mining ini disusun sebagai rujukan bagi mahasiswa pada program studi Informatika Fakultas Ilmu Komputer Universitas Bhayangkara Jakarta Raya yang mengikuti perkuliahan Data Mining. Pada modul Data Mining ini membahas topik utama yaitu Pengantar Data Mining, Proses Data Data Mining dan Latihan Algoritma.

Modul ini masih belum sempurna, sehingga perlu dikaji baik oleh dosen pengajar, mahasiswa, dan pemakai modul ini. Oleh karena itu penyusun berharap agar para pemakai modul ini dapat memberikan sumbangan saran untuk perbaikan modul Data Mining ini. Semoga modul ini dapat bermanfaat bagi para personil yang terlibat dalam perkuliahan Data Mining, serta dapat meningkatkan kemampuan mahasiswa dalam menguasai konsep dasar, peran utama, metode, proses, dan algoritma data mining.

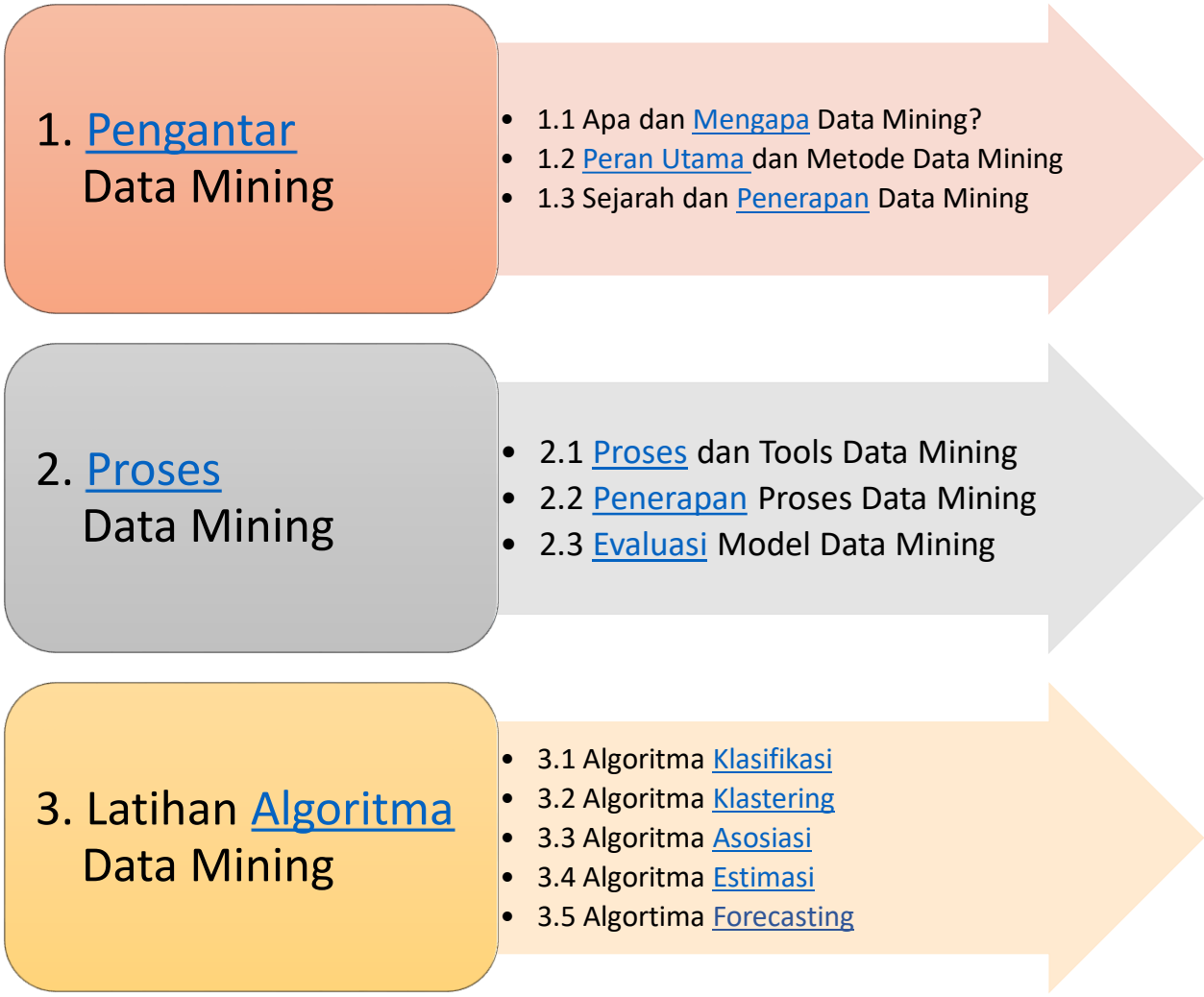
Jakarta, 19 Juli 2020

Tim Penulis

DAFTAR ISI

Modul 1 Kontrak Perkuliahan

1. Waktu perkuliahan sesuai dengan jadwal yang telah ditentukan
2. Perkuliahan Video Confrence, Google Classroom, Penugasan individu dan kelompok
3. Perkuliahan Video Confrence mahasiswa wajib menyalakan videonya dan hadir 15 menit sebelum waktu perkuliahan
4. Kehadiran dicatat jika mengikuti aturan yang ada
5. Semua komunikasi via group WA.



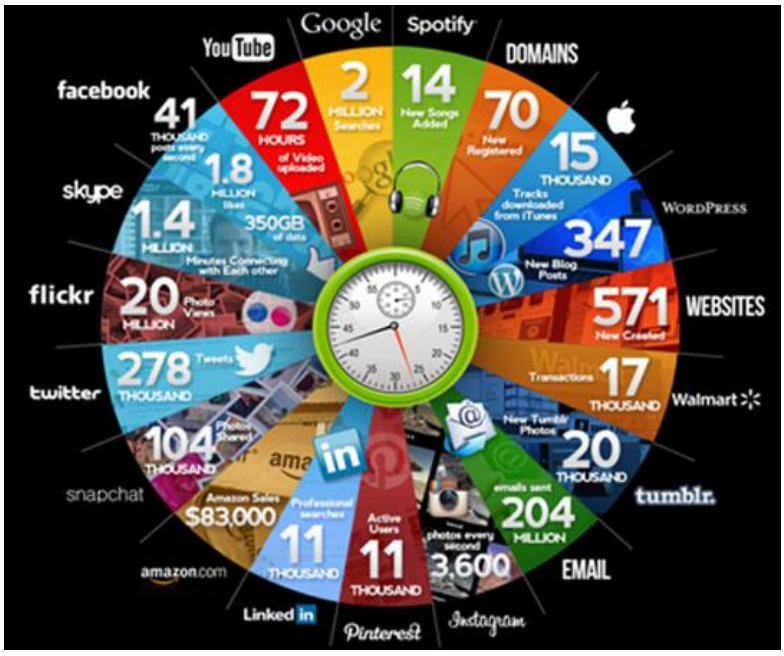
Modul 2 Pengantar Data Mining

2.1 Apa dan Mengapa Data Mining?

Manusia memerlukan melakukan data mining karena (1) manusia memproduksi data beragam data yang jumlah dan ukurannya sangat besar untuk bidang astronomi, bisnis, kedokteran, ekonomi, olahraga, cuaca, financial, dan masih banyak bidang. (2) Tejadinya pertumbuhan data seperti bidang astronomi berdasrakan survey yang dilakukan oleh perusahaan Sloan Digital Sky Survey di New Mexico pada 2000 bahwa data yang dihasilkan 140TB selama 10 tahun terakhir. Sedangkan survey dari perusahaan Large Synoptic Survey Telescope di Chile pada tahun 2016 bahwa setiap 5 hari dihasilkan data sebesar 140TB.

Pada bidang biologi dan kedokteran menurut European Bioinformatics Institute (EBI) dihasilkan data sebesar 20PB tiap tahun (genomic data doubles) dan Genom manusia berurutan tunggal bisa menghasilkan data sebesar 140GB.

Disamping itu terjadinya (3) perubahan kultur dan perilaku manusia bahwa tiap detik social media menghasilkan data yang fantastis seperti pada gambar di bawah ini:



Gambar 1 Sosial media menghasilkan data

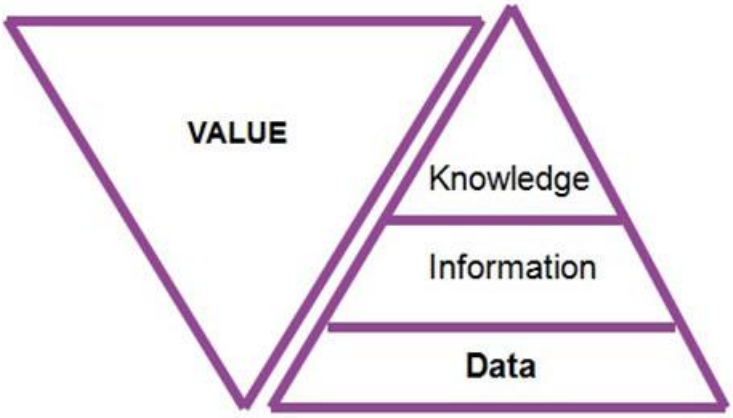
Alasan berikutnya adalah (4) terjadinya tsumani data seperti di bawah ini:

- Mobile Electronics market : 7B smartphone subscriptions in 2015
- Web & Social Networks generates amount of data
    - Google processes 100 PB per day, 3 million servers
    - Facebook has 300 PB of user data per day
    - Youtube has 1000PB video storage

Manusia kebanjiran data tetapi miskin pengetahuan untuk itu diperlukan bagaimana manusia bisa mengubah data menjadi pengetahuan. *(John Naisbitt, Megatrends, 1988)*

Data harus olah menjadi pengetahuan supaya bisa bermanfaat bagi manusia. Dengan pengetahuan tersebut, manusia dapat:

- Melakukan estimasi dan prediksi apa yang terjadi di depan
- Melakukan analisis tentang asosiasi, korelasi dan pengelompokan antar data dan atribut
- Membantu pengambilan keputusan dan pembuatan kebijakan



Gambar 2 Piramida Data - Informasi – Pengetahuan

Data - Informasi – Pengetahuan – Kebijakan

Table 1 Data kehadiran karyawan

NIP	TGL	DATANG	PULANG
1103	02/12/2004	07:20	15:40
1142	02/12/2004	07:45	15:33
1156	02/12/2004	07:51	16:00
1173	02/12/2004	08:00	15:15
1180	02/12/2004	07:01	16:31
1183	02/12/2004	07:49	17:00

Tabel 2 Informasi Akumulasi Bulanan Kehadiran Karyawan

NIP	Masuk	Alpa	Cuti	Sakit	Telat
1103	22				
1142	18	2		2	
1156	10	1	11		
1173	12	5			5
1180	10			12	

	Senin	Selasa	Rabu	Kamis	Jumat
Terlambat	7	0	1	0	5
Pulang Cepat	0	1	1	1	8
Izin	3	0	0	1	4
Alpa	1	0	2	0	2

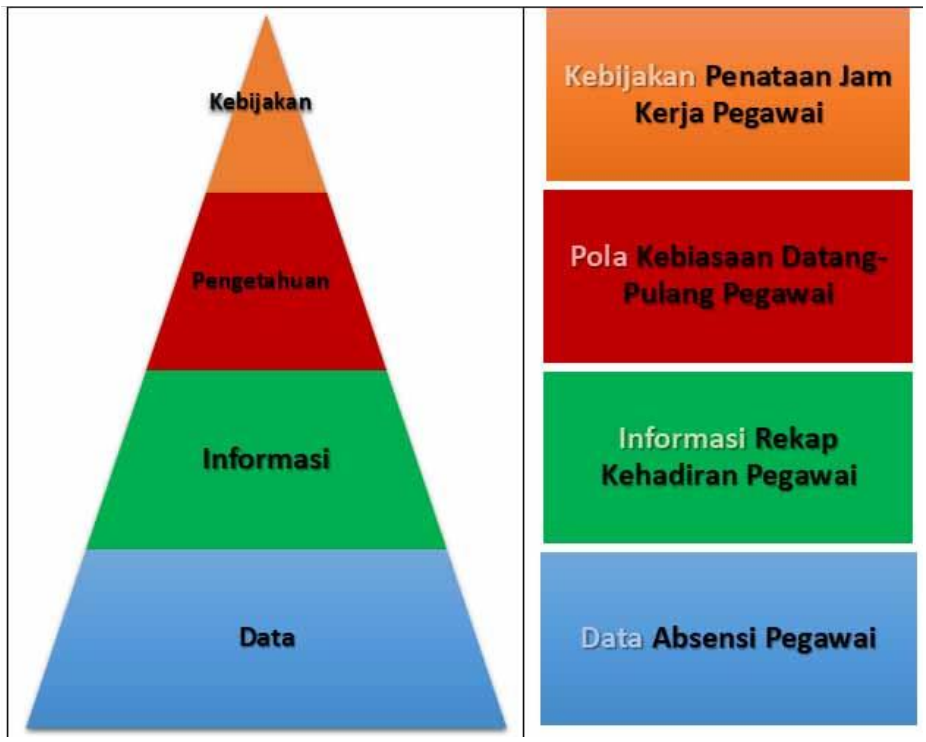
Table 3. Pola Kebiasaan Kehadiran Mingguan Karyawan

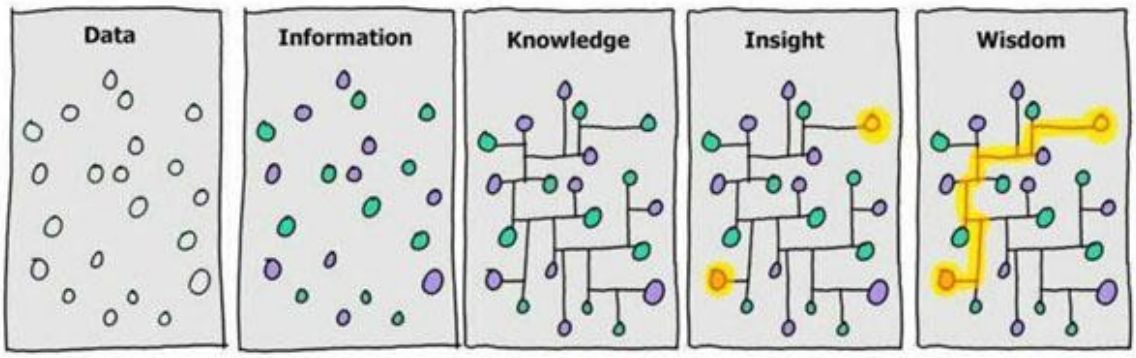
- Kebijakan penataan jam kerja karyawan khusus untuk hari senin dan jumat
- Peraturan jam kerja:
    - Hari Senin dimulai jam 10:00

- Hari Jumat diakhiri jam 14:00
- Sisa jam kerja dikompensasi ke hari lain

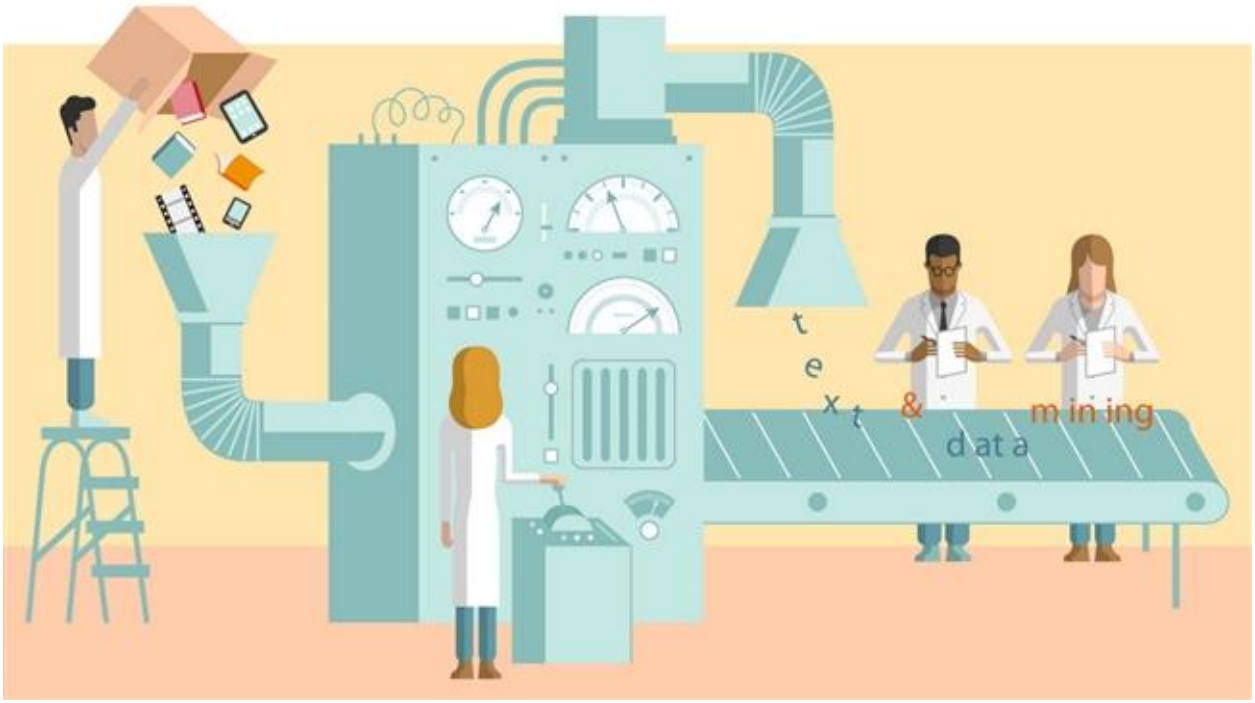
Gambar 3 Kebijakan penataan jam kerja karyawan

Data - Informasi – Pengetahuan – Kebijakan





Gambar 3. Data - Informasi – Pengetahuan – Kebijakan



Gambar 4. Data Mining

Data mining adalah disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar.

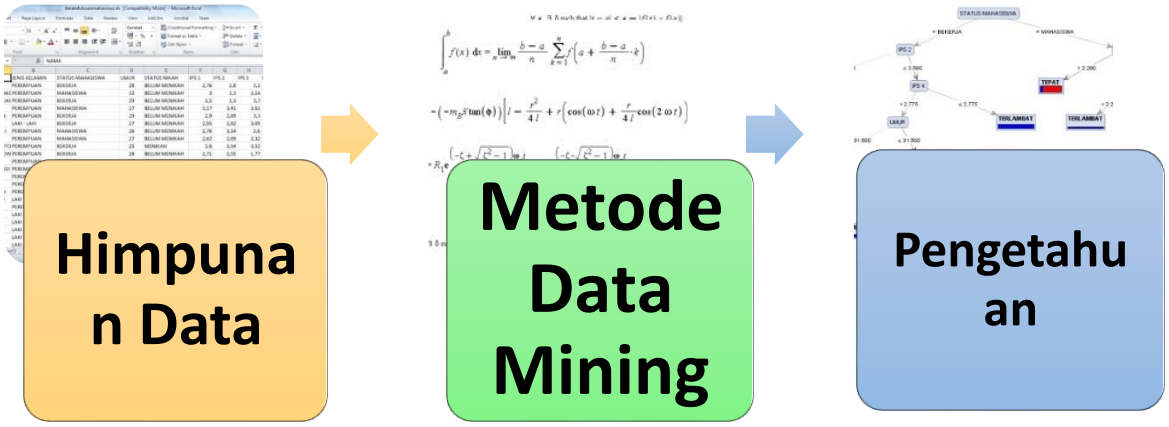
Ekstraksi dari data ke pengetahuan:

1. data: fakta yang terekam dan tidak membawa arti
2. informasi: rekap, rangkuman, penjelasan dan statistik dari data
3. pengetahuan: pola, rumus, aturan atau model yang muncul dari data

Nama lain data mining:

a. Knowledge Discovery in Database (KDD)
b. Big data
c. Business intelligence
d. Knowledge extraction
e. Pattern analysis
f. Information harvestin

2.2 Konsep Proses Data Mining



Gambar 5. Konsep Proses Data Mining

Definisi data mining

- Melakukan ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan sebelumnya tidak diketahui, dari suatu data *(Witten et al., 2011)*
- Kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola dan hubungan dalam set data berukuran besar *(Santosa, 2007)*
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data *(Han et al., 2011)*

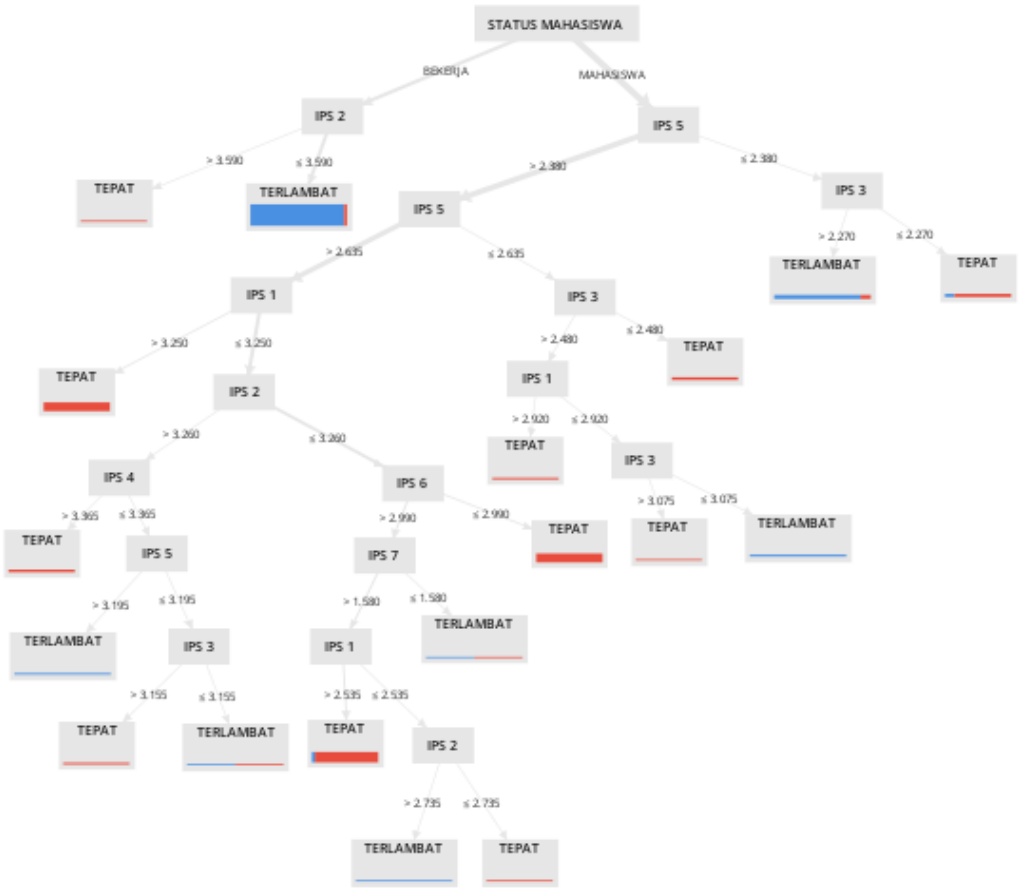
Contoh data mining di kampus

- Puluhan ribu data mahasiswa di kampus yang diambil dari sistem informasi akademik
- Apakah pernah kita ubah menjadi pengetahuan yang lebih bermanfaat? TIDAK!
- Seperti apa pengetahuan itu? Rumus, Pola, Aturan

Table 4. Daftar Kelulusan Mahasiswa

NIM	Gender	Nilai UN	Asal Sekolah	IPS1	IPS2	IPS3	IPS 4	...	Lulus Tepat Waktu
10001	L	28	SMAN 2	3.3	3.6	2.89	2.9		Ya
10002	P	27	SMA DK	4.0	3.2	3.8	3.7		Tidak
10003	P	24	SMAN 1	2.7	3.4	4.0	3.5		Tidak
10004	L	26.4	SMAN 3	3.2	2.7	3.6	3.4		Ya
...									
...									
11000	L	23.4	SMAN 5	3.3	2.8	3.1	3.2		Ya

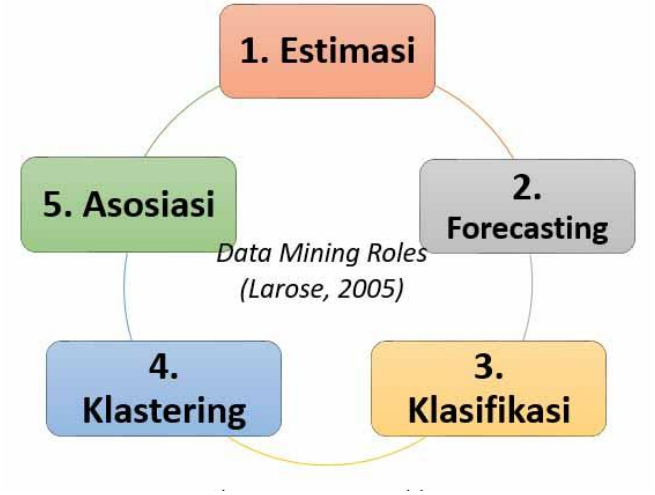
Prediksi Kelulusan Mahasiswa Pembelajaran dengan Metode Klasifikasi (*C4.5*)



Gambar 6. *Decision Tree* Kelulusan Mahasiswa.

MODUL 3. Peran Utama dan Metode Data Mining

3.1 Peran Utama Data Mining



Gambar 7. Peran Utama Data Mining

1. Estimation (estimasi), untuk menerka sebuah nilai yang belum diketahui, misal menerka penghasilan seseorang ketika informasi mengenai orang tersebut diketahui.

2. Forecasting (prediksi), untuk memperkirakan nilai masa mendatang, misal memprediksi stok barang satu tahun ke depan.

3. Classification (klasifikasi), merupakan proses penemuan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui.

4. Clustering (pengelompokan), yaitu pengelompokan mengidentifikasi data yang memiliki karakteristik tertentu.

5. Association (asosiasi), dinamakan juga analisis keranjang pasar dimana fungsi ini mengidentifikasi item-item produk yang kemungkinan dibeli konsumen bersamaan dengan produk lain.

3.2 Metode dalam data mining

1. Estimation (Estimasi):
   Linear Regression (LR), Neural Network (NN), Deep Learning (DL), Support Vector Machine (SVM), Generalized Linear Model (GLM), etc

2. Forecasting (Prediksi/Peramalan):
   Linear Regression (LR), Neural Network (NN), Deep Learning (DL), Support Vector Machine (SVM), Generalized Linear Model (GLM), etc
3. Classification (Klasifikasi):
   Decision Tree (CART, ID3, C4.5, Credal DT, Credal C4.5, Adaptative Credal C4.5), Naive Bayes (NB), K-Nearest Neighbor (kNN), Linear Discriminant Analysis (LDA), Logistic Regression (LogR), etc
4. Clustering (Klastering):
   K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means (FCM), etc
5. Association (Asosiasi):
   FP-Growth, A Priori, Coefficient of Correlation, Chi Square, etc

Output / Pola / Model / Knowledge adalah sebagai berikut:

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)

Contoh 1: Estimasi Waktu Pengiriman Pizza

Label

Customer	Jumlah Pesanan (P)	Jumlah Traffic Light (TL)	Jarak (J)	Waktu Tempuh (T)
1	3	3	3	16
2	1	7	4	20
3	2	4	6	18
4	4	6	8	36
...				
1000	2	4	2	12

Pembelajaran dengan Metode Estimasi (Regresi Linier)

Waktu Tempuh (T) = 0.48P + 0.23TL + 0.5J

Pengetahuan

Contoh 2: Forecasting Harga Saham

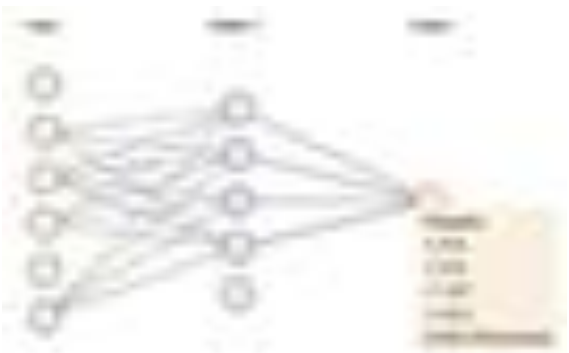
Label Time Series

Row No.	Close	Date	Open	High	Low	Volume
1	1286.570	Apr 11, 2006	1296.600	1300.710	1282.960	223288000(
2	1288.120	Apr 12, 2006	1286.570	1290.930	1286.450	193810000(
3	1289.120	Apr 13, 2006	1288.120	1292.090	1283.370	189194000(
4	1285.330	Apr 17, 2006	1289.120	1292.450	1280.740	179465000(
5	1307.280	Apr 18, 2006	1285.330	1309.020	1285.330	259544000(
6	1309.930	Apr 19, 2006	1307.650	1310.390	1302.790	244731000(
7	1311.460	Apr 20, 2006	1309.930	1318.160	1306.380	251292000(
8	1311.280	Apr 21, 2006	1311.460	1317.670	1306.590	239263000(
9	1308.110	Apr 24, 2006	1311.280	1311.280	1303.790	211733000(
10	1301.740	Apr 25, 2006	1308.110	1310.790	1299.170	236638000(
11	1305.410	Apr 26, 2006	1301.740	1310.970	1301.740	250269000(
12	1309.720	Apr 27, 2006	1305.410	1315	1295.570	277201000(
13	1310.610	Apr 28, 2006	1309.720	1316.040	1306.160	241992000(

Dataset harga saham dal
bentuk time series (rente
waktu)

Pembelajaran dengan
Metode Forecasting (*Neural Network*)

Pengetahuan berupa Rumus Neural Network



Predictio
Plot

Contoh 3: Klasifikasi Kelulusan Mahasiswa

Label

NIM	Gender	Nilai UN	Asal Sekolah	IPS1	IPS2	IPS3	IPS 4	...	Lulus Tepat Waktu
10001	L	28	SMAN 2	3.3	3.6	2.89	2.9		Ya
10002	P	27	SMA DK	4.0	3.2	3.8	3.7		Tidak
10003	P	24	SMAN 1	2.7	3.4	4.0	3.5		Tidak
10004	L	26.4	SMAN 3	3.2	2.7	3.6	3.4		Ya
...									
...									
11000	L	23.4	SMAN 5	3.3	2.8	3.1	3.2		Ya

Pembelajaran dengan
Metode Klasifikasi (*C4.5*)

Pengetahuan Berupa Pohon Keputusan

Gambar 8. Pohon Keputusan Kelulusan Mahasiswa

Contoh: Rekomendasi Main Golf

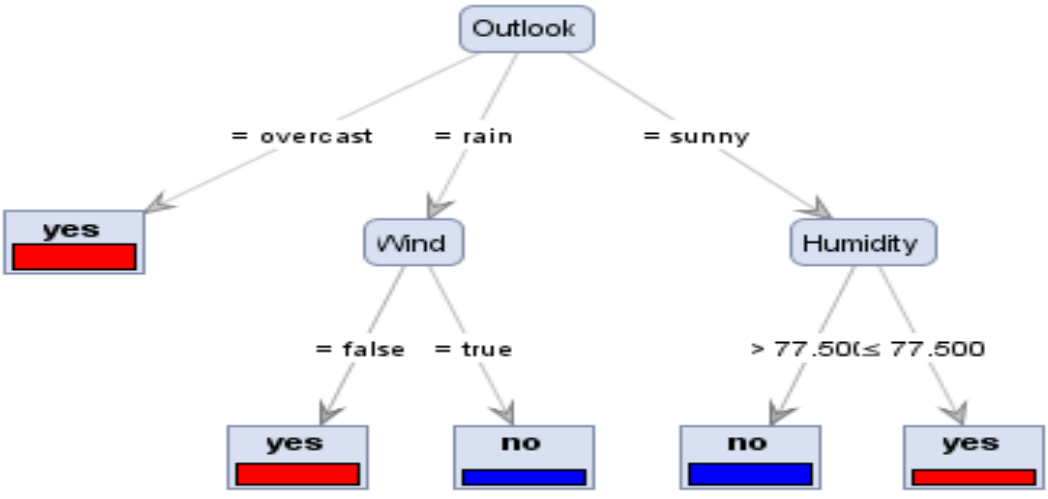
- **Input:**

Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

Output (Rules):

If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above then play = yes

Output (Tree):



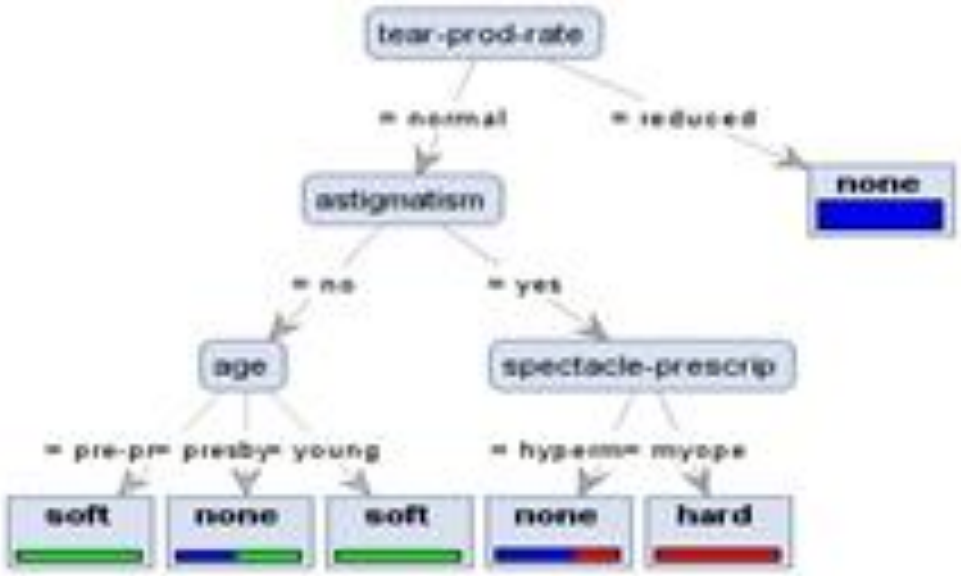
Gambar 10. Pohon Keputusan Main Golf

Contoh: Rekomendasi Contact Lens

Input:

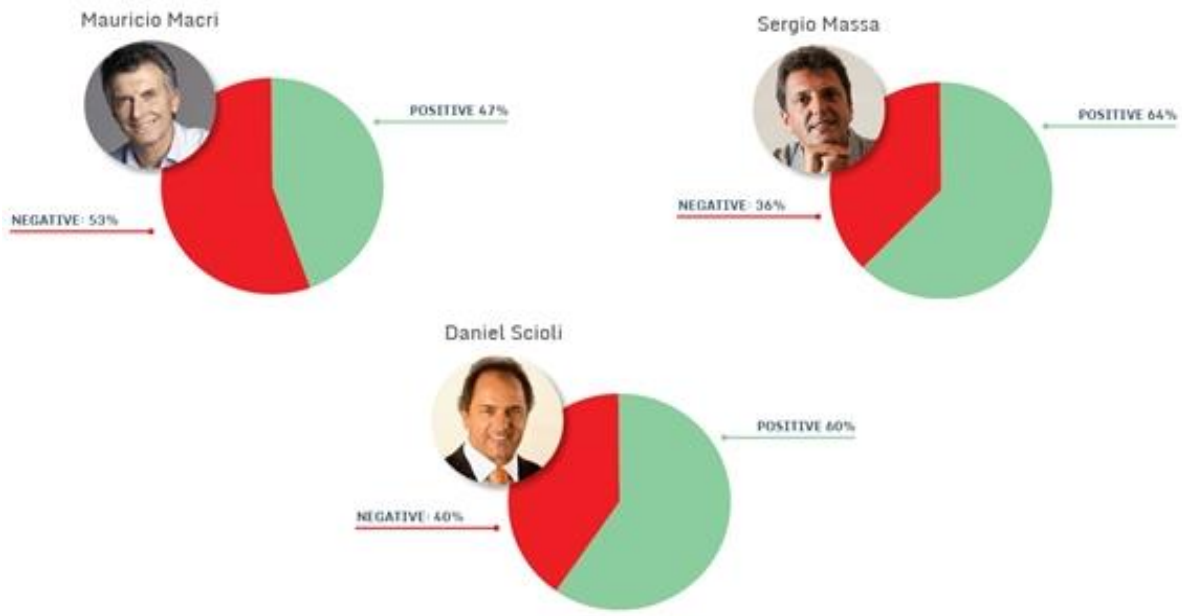
Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft

Output/Model (Tree):



Gambar 11 Pohon Keputusan Contact Lens

Klasifikasi Sentimen Analisis



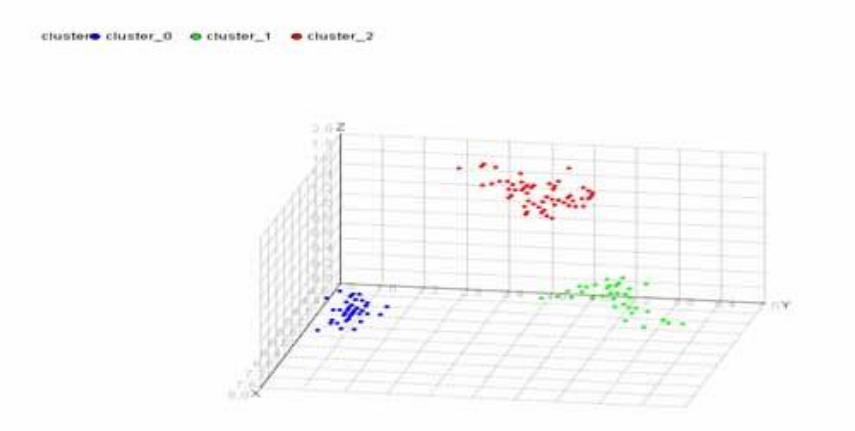
Contoh 4 : Klastering Bunga Iris

Dataset Tanpa Label

Row No.	id	a1	a2	a3	a4
1	id_1	5.100	3.500	1.400	0.200
2	id_2	4.900	3	1.400	0.200
3	id_3	4.700	3.200	1.300	0.200
4	id_4	4.600	3.100	1.500	0.200
5	id_5	5	3.600	1.400	0.200
6	id_6	5.400	3.900	1.700	0.400
7	id_7	4.600	3.400	1.400	0.300
8	id_8	5	3.400	1.500	0.200
9	id_9	4.400	2.900	1.400	0.200
10	id_10	4.900	3.100	1.500	0.100
11	id_11	5.400	3.700	1.500	0.200

Pembelajaran dengan
Metode Klastering (*K-Means*)

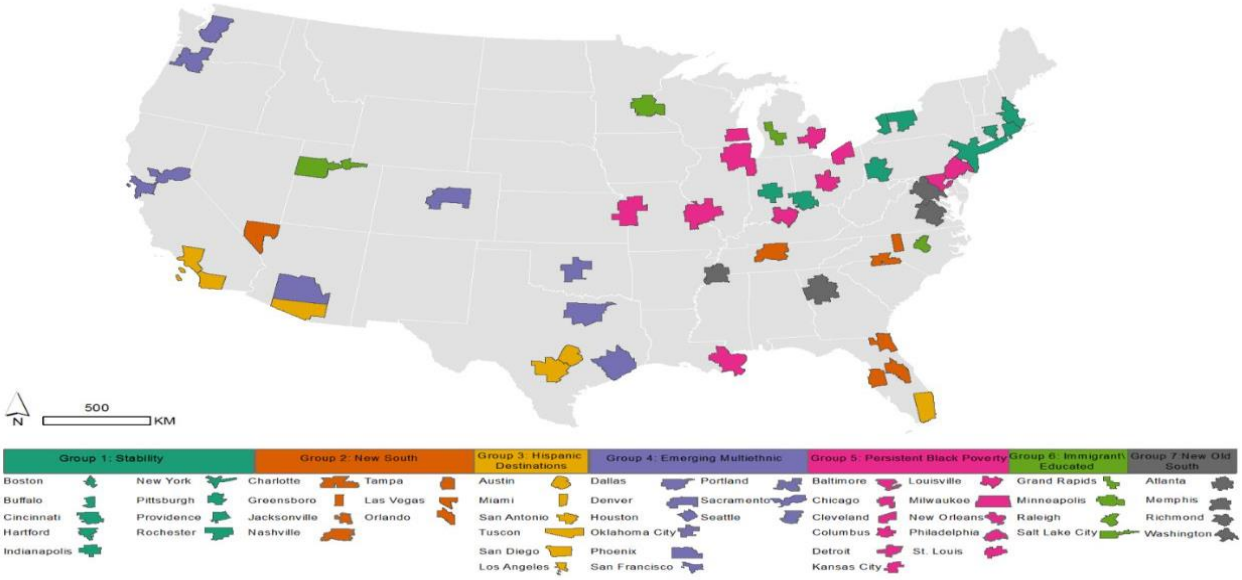
Pengetahuan (Model) Berupa Klaster



Klastering Jenis Pelanggan

Klastering Sentimen Warga

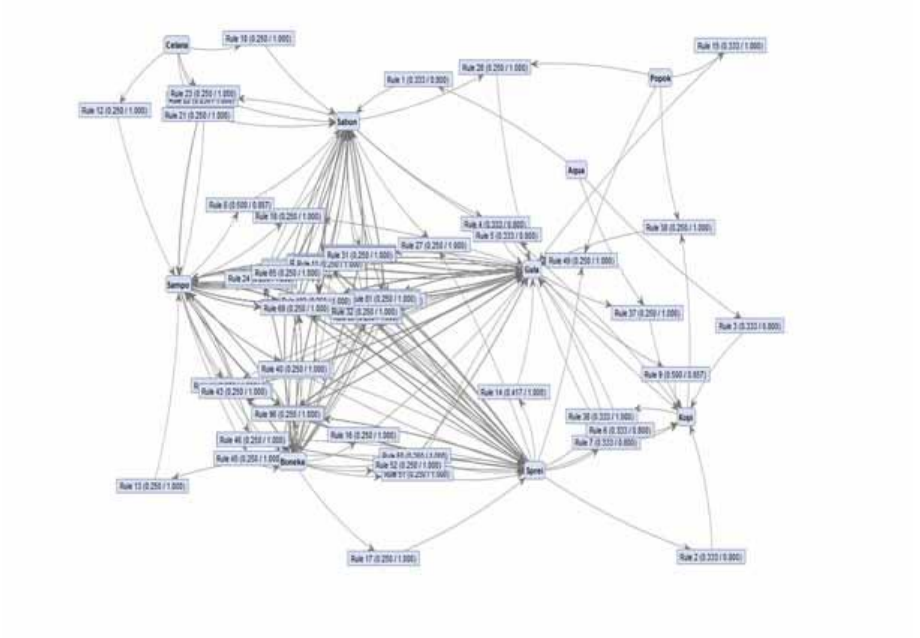
Poverty Rate Clustering



Contoh 5: Aturan Asosiasi Pembelian Barang

Pembelajaran dengan
Metode Asosiasi (*FP-Growth*)

Pengetahuan Berupa Aturan Asosiasi

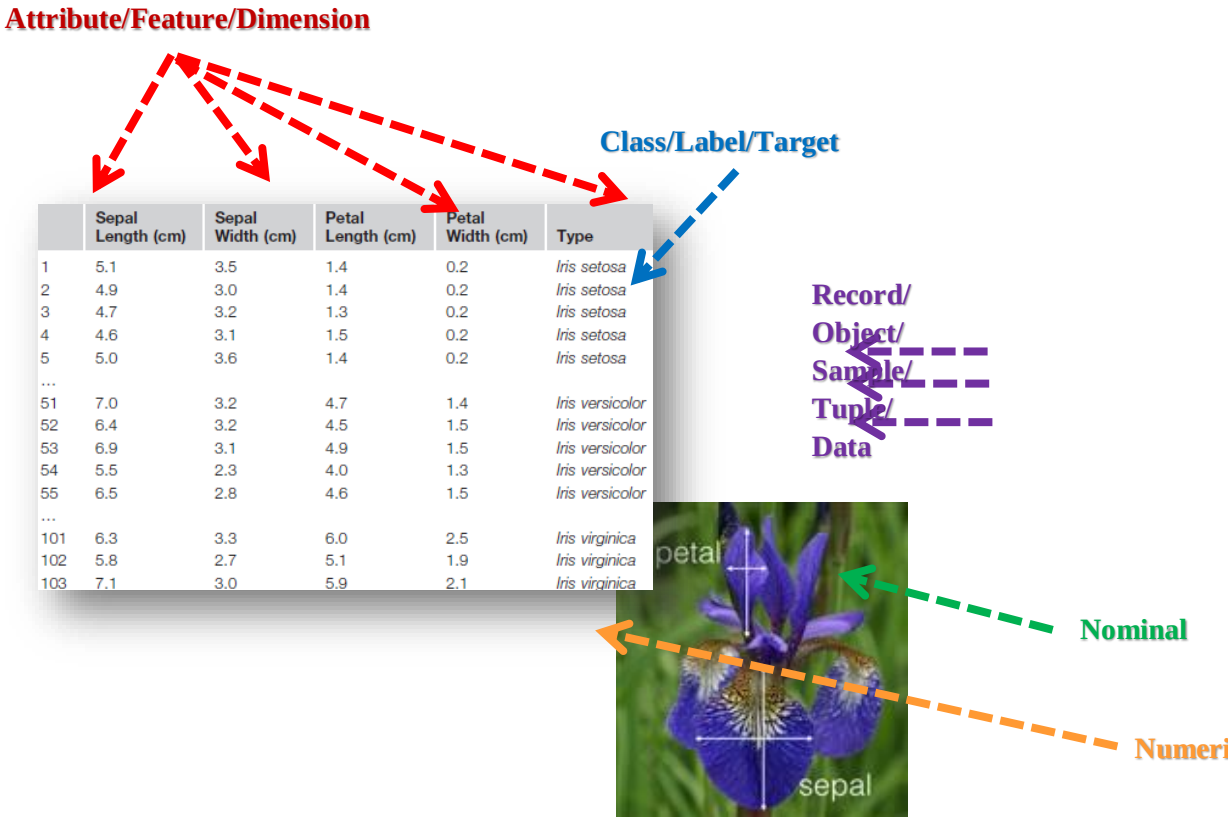


5. Tugas

Contoh Aturan Asosiasi

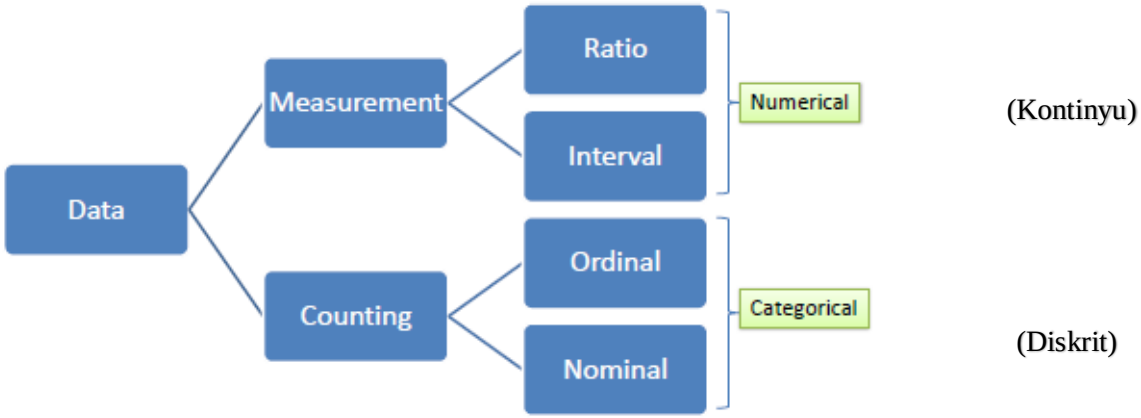
- Algoritma *association rule* (aturan asosiasi) adalah algoritma yang menemukan atribut yang "muncul bersamaan"
- Contoh, pada hari kamis malam, 1000 pelanggan telah melakukan belanja di supermaket ABC, dimana:
    - 200 orang membeli Sabun Mandi
    - dari 200 orang yang membeli sabun mandi, 50 orangnya membeli Fanta
- Jadi, association rule menjadi, "Jika membeli sabun mandi, maka membeli Fanta", dengan nilai support = 200/1000 = 20% dan nilai confidence = 50/200 = 25%

Dataset (Himpunan Data) dengan Class



	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
...					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris versicolor*
53	6.9	3.1	4.9	1.5	*Iris versicolor*
54	5.5	2.3	4.0	1.3	*Iris versicolor*
55	6.5	2.8	4.6	1.5	*Iris versicolor*
...					
101	6.3	3.3	6.0	2.5	*Iris virginica*
102	5.8	2.7	5.1	1.9	*Iris virginica*
103	7.1	3.0	5.9	2.1	*Iris virginica*

Tipe Data



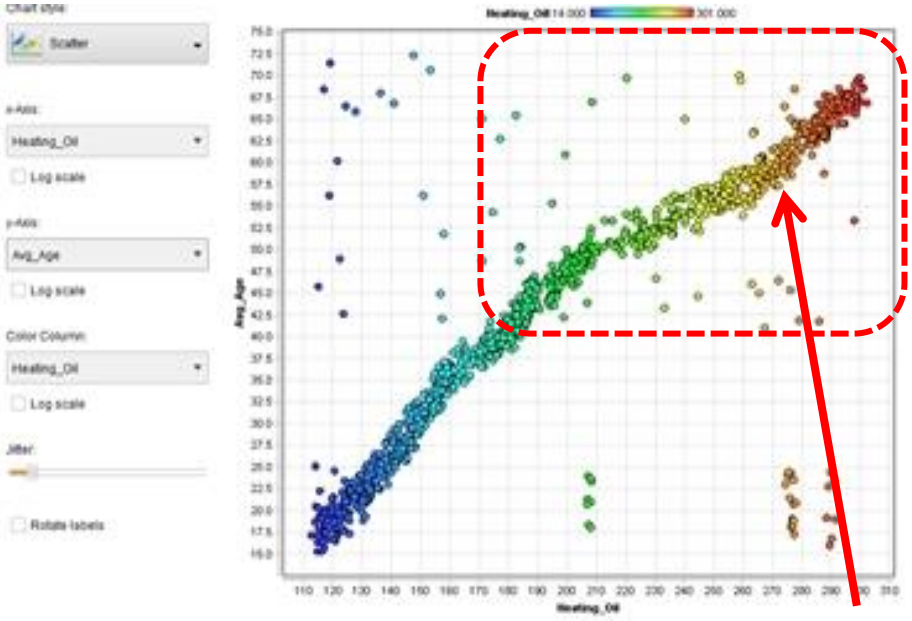
Tipe Data	Deskripsi	Contoh	Operasi
Ratio (Mutlak)	• Data yang diperoleh dengan cara pengukuran, dimana jarak dua titik pada skala sudah diketahui • Mempunyai titik nol yang absolut (*, /)	• Umur • Berat badan • Tinggi badan • Jumlah uang	geometric mean, harmonic mean, percent variation
Interval (Jarak)	• Data yang diperoleh dengan cara pengukuran, dimana jarak dua titik pada skala sudah diketahui • Tidak mempunyai titik nol yang absolut (+, -)	• Suhu 0°c-100°c, • Umur 20-30 tahun	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ordinal (Peringkat)	• Data yang diperoleh dengan cara kategorisasi atau klasifikasi • Tetapi diantara data tersebut terdapat hubungan atau berurutan (<, >)	• Tingkat kepuasan pelanggan (puas, sedang, tidak puas)	median, percentiles, rank correlation, run tests, sign tests
Nominal (Label)	• Data yang diperoleh dengan cara kategorisasi atau klasifikasi • Menunjukkan beberapa object yang berbeda (=, ≠)	• Kode pos • Jenis kelamin • Nomer id karyawan • Nama kota	mode, entropy, contingency correlation, χ^2 test

Aturan Asosiasi di Amazon.com

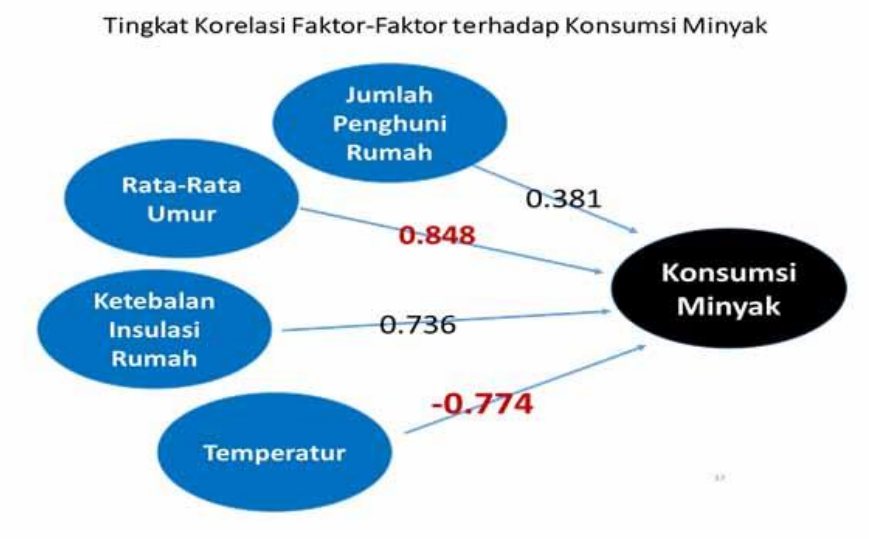
Heating Oil Consumption

Korelasi antara jumlah konsumsi minyak pemanas dengan faktor-faktor di bawah:

1. Insulation: Ketebalan insulasi rumah
2. Temperatur: Suhu udara sekitar rumah
3. Heating Oil: Jumlah konsumsi minyak pertahun perrumah
4. Number of Occupant: Jumlah penghuni rumah
5. Average Age: Rata-rata umur penghuni rumah
6. Home Size: Ukuran rumah



Attributes	Heating_Oil	Insulation	Temperature	Num_Occupants	Avg_Age	Home_Size
Heating_Oil	1	0.735	-0.774	-0.042	0.848	0.381
Insulation	0.735	1	-0.794	-0.013	0.643	0.201
Temperature	-0.774	-0.794	1	0.013	-0.671	-0.214
Num_Occupants	-0.042	-0.013	0.013	1	-0.048	-0.023
Avg_Age	0.848	0.643	-0.671	-0.048	1	0.307
Home_Size	0.381	0.201	-0.214	-0.023	0.307	1

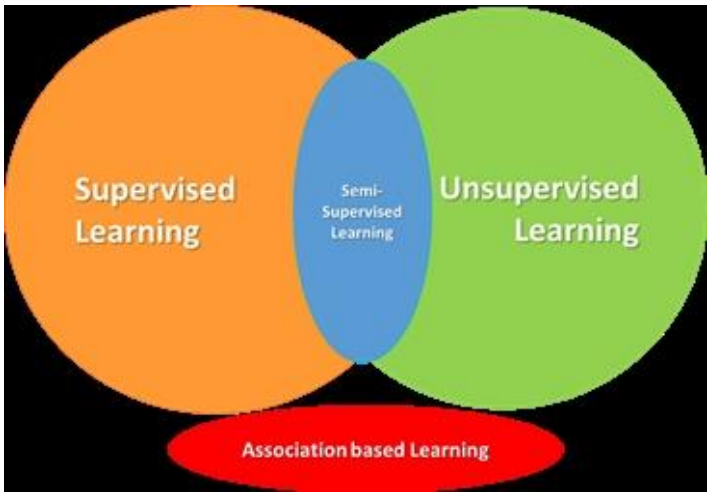


Insight Law (Data Mining Law 6)

Data mining amplifies perception in the business domain

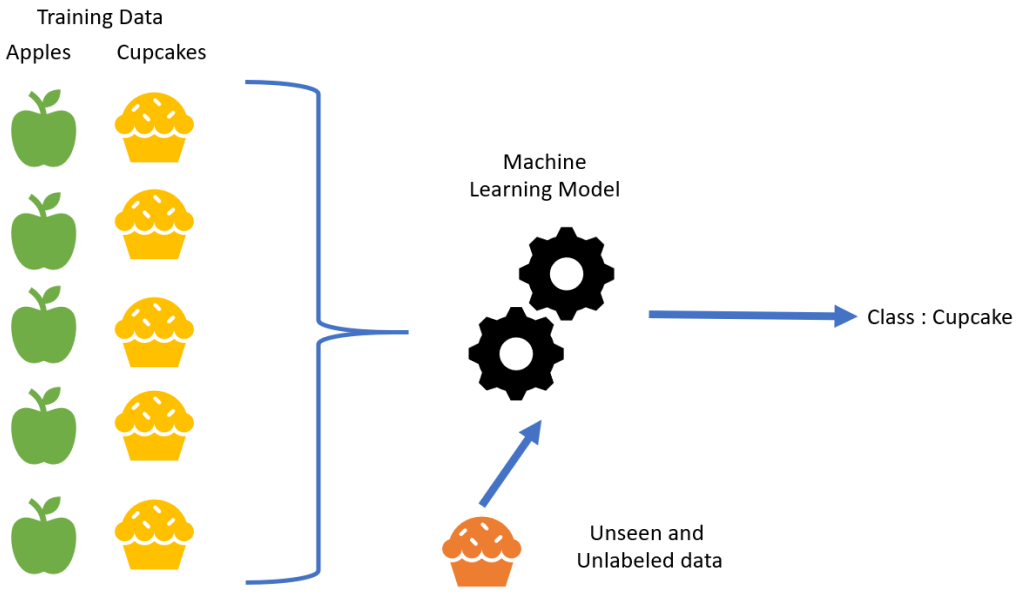
1. How does data mining produce insight? This law approaches the heart of data mining – why it must be a business process and not a technical one
   - Business problems are solved by people, not by algorithms
2. The data miner and the business expert "see" the solution to a problem, that is the patterns in the domain that allow the business objective to be achieved
   a. Thus data mining is, or assists as part of, a perceptual process
   b. Data mining algorithms reveal patterns that are not normally visible to human perception
3. Within the data mining process, the human problem solver interprets the results of data mining algorithms and integrates them into their business understanding

Metode Learning Algoritma Data Mining

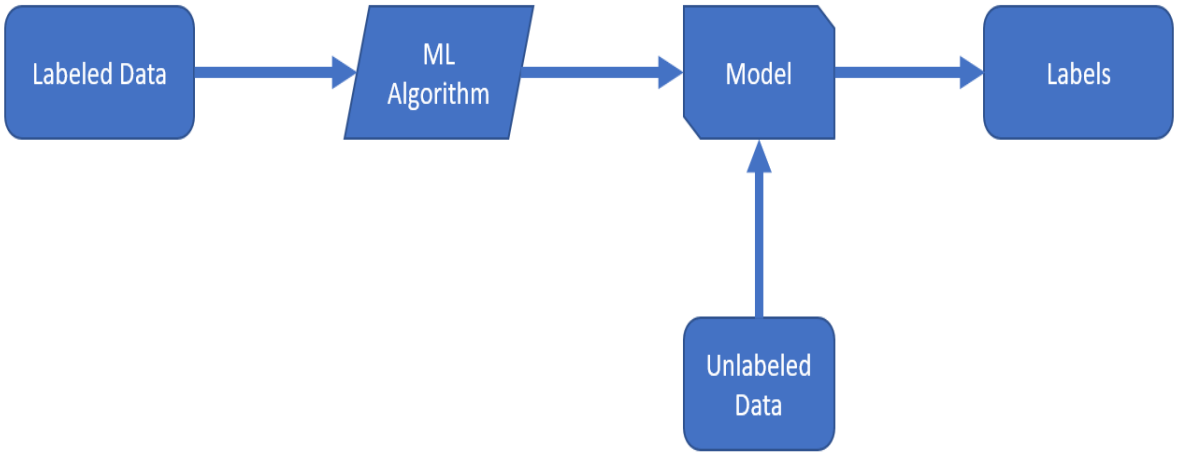


1. Supervised Learning

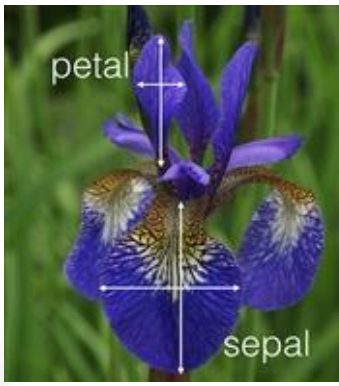
- Pembelajaran dengan guru, data set memiliki target/label/class
- Sebagian besar algoritma data mining (estimation, prediction/forecasting, classification) adalah supervised learning
- Algoritma melakukan proses belajar berdasarkan nilai dari variabel target yang terasosiasi dengan nilai dari variable prediktor



https://medium.com/@canburaktumer/machine-learning-basics-with-examples-part-2-supervised-learning-e2b740ff014c

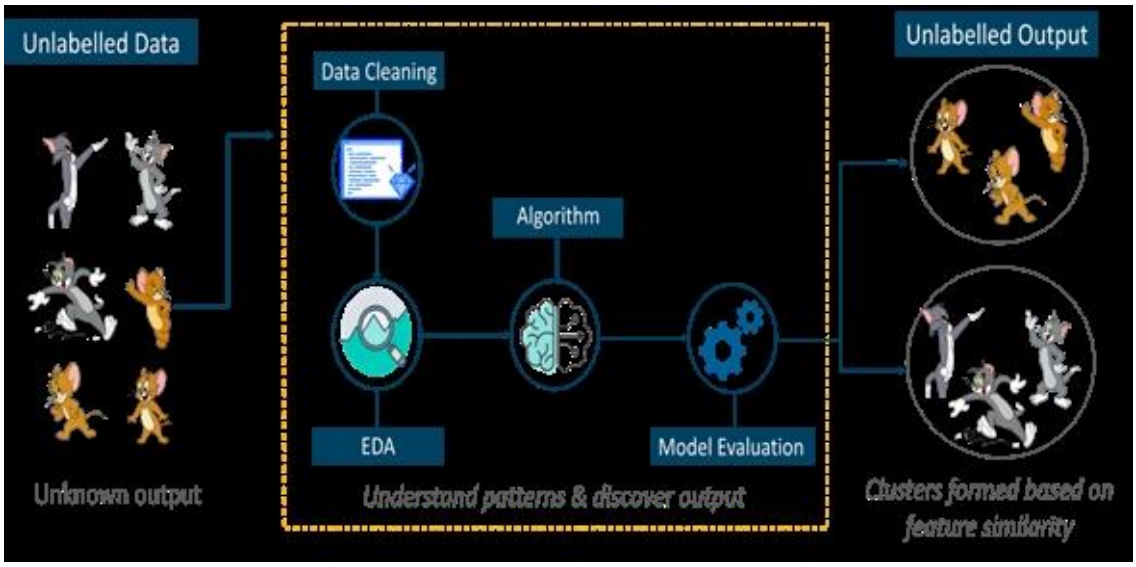


Dataset (Himpunan Data) dengan Class

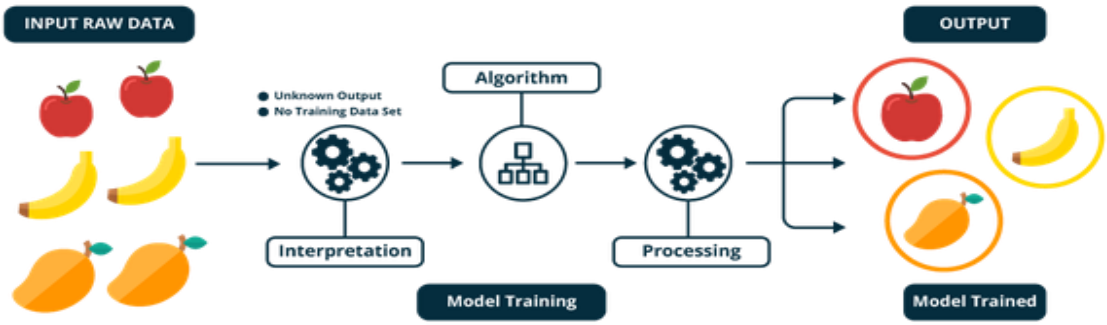


2. Unsupervised Learning

- Algoritma data mining mencari pola dari semua variable (atribut)
- Variable (atribut) yang menjadi target/label/class tidak ditentukan (tidak ada)
- Algoritma clustering adalah algoritma unsupervised learning

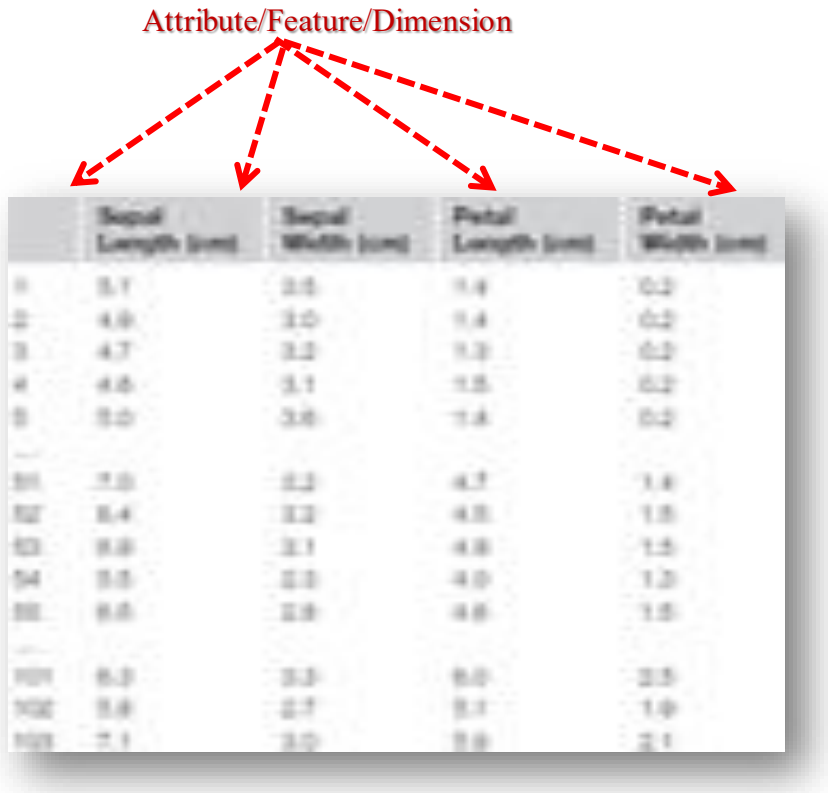


https://www.edureka.co/blog/introduction-to-machine-learning/

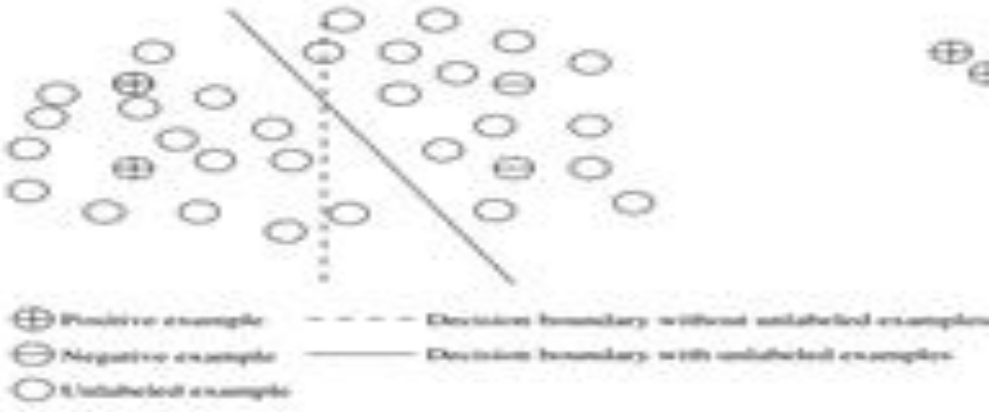


https://www.tes.com/lessons/mfiu9IDljkeE7A/it605d-l20-unsupervised-learning-winners-take-

Dataset (Himpunan Data) tanpa Class



3. Semi-Supervised Learning
   - Semi-supervised learning adalah metode data mining yang menggunakan data dengan label dan tidak berlabel sekaligus dalam proses pembelajarannya
   - Data yang memiliki kelas digunakan untuk membentuk model (pengetahuan), data tanpa label digunakan untuk membuat batasan antara kelas

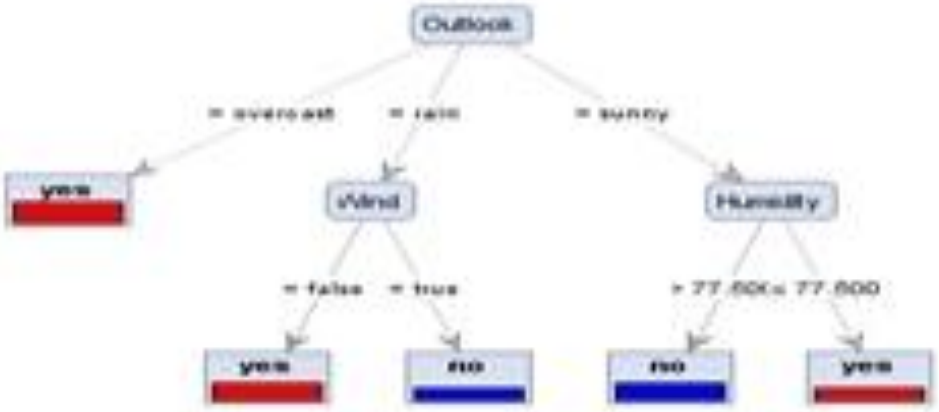


Algoritma Data Mining

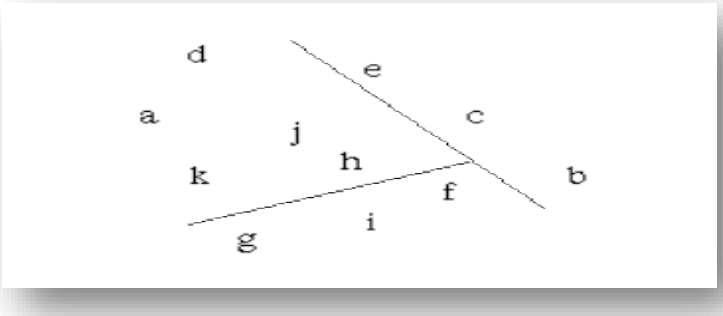
1. Estimation (Estimasi):
    - Linear Regression, Neural Network, Support Vector Machine, etc
2. Prediction/Forecasting (Prediksi/Peramalan):
    - Linear Regression, Neural Network, Support Vector Machine, etc
3. Classification (Klasifikasi):
    - Naive Bayes, K-Nearest Neighbor, C4.5, ID3, CART, 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.

Output/Pola/Model/Knowledge

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)



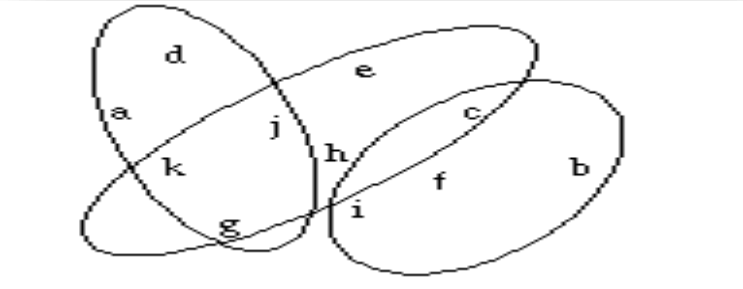
3. Tingkat Korelasi



4. Rule (Aturan)

IF ips3=2.8 THEN lulus tepat waktu

5. Cluster (Klaster)



Latihan

1. Sebutkan 5 peran utama data mining!
2. Jelaskan perbedaan estimasi dan forecasting!
3. Jelaskan perbedaan forecasting dan klasifikasi!
4. Jelaskan perbedaan klasifikasi dan klastering!
5. Jelaskan perbedaan klastering dan association!
6. Jelaskan perbedaan estimasi dan klasifikasi!
7. Jelaskan perbedaan supervised dan unsupervised learning!
8. Sebutkan tahapan utama proses data mining!

MODUL 4 Sejarah dan Penerapan Data Mining

4.1 Evolution of Sciences

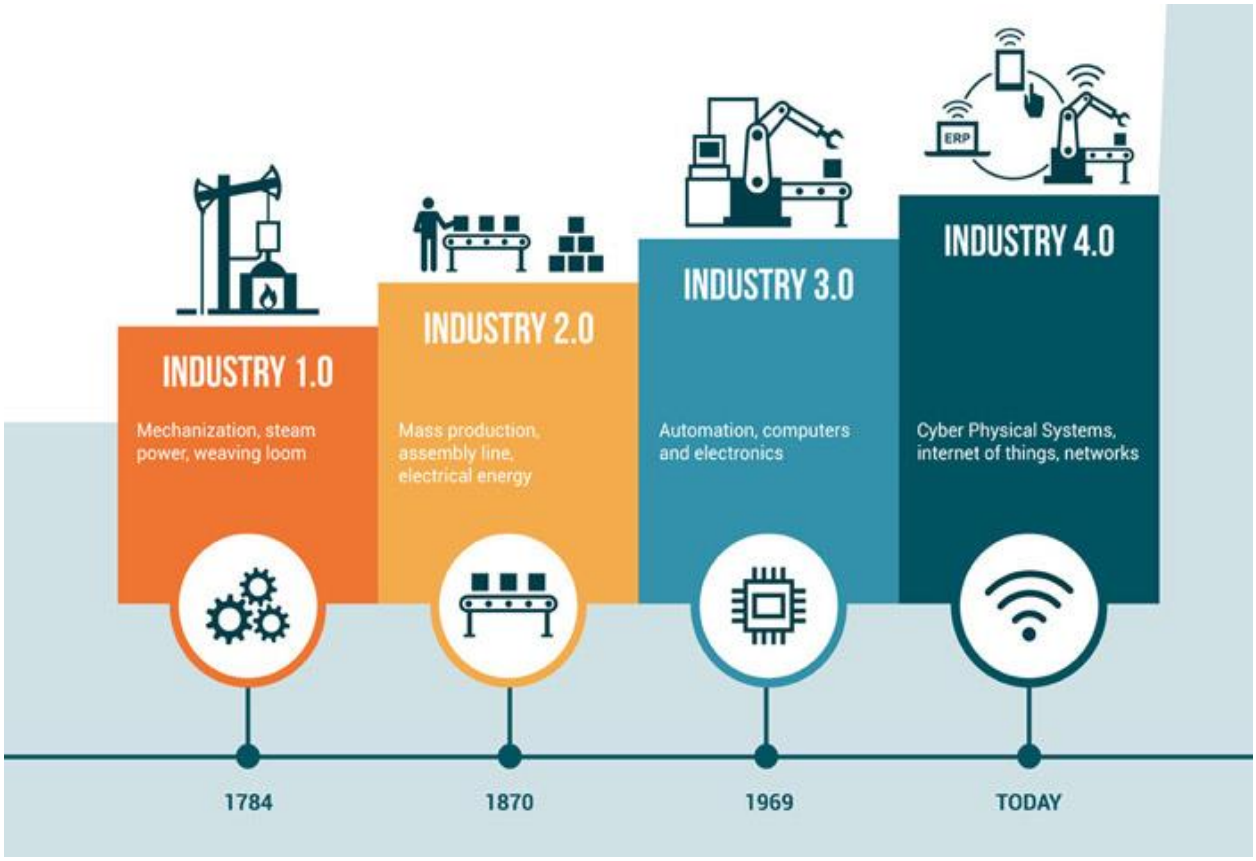
- Sebelum 1600: Empirical science
    - Disebut sains kalau bentuknya kasat mata
- 1600-1950: Theoretical science
    - Disebut sains kalau bisa dibuktikan secara matematis atau eksperimen

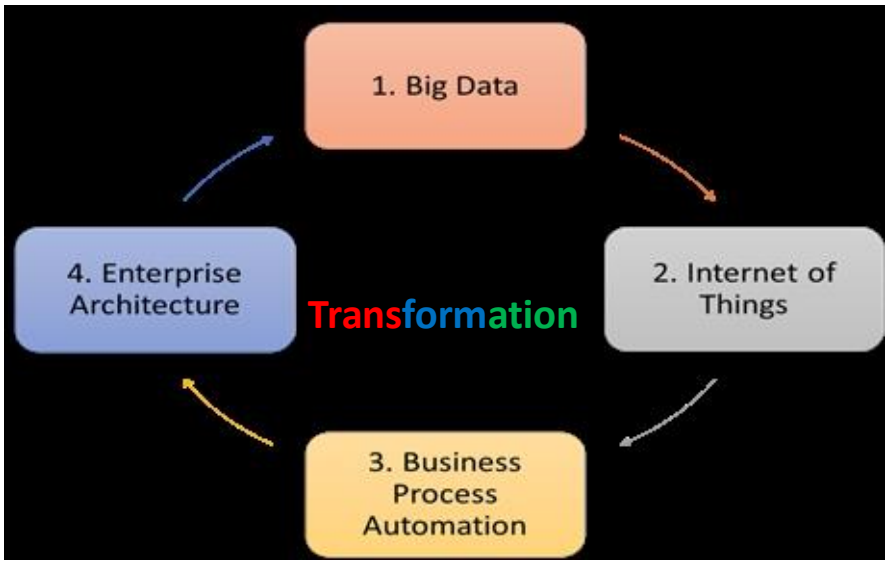
Jim Gray and Alex Szalay, The World Wide Telescope:

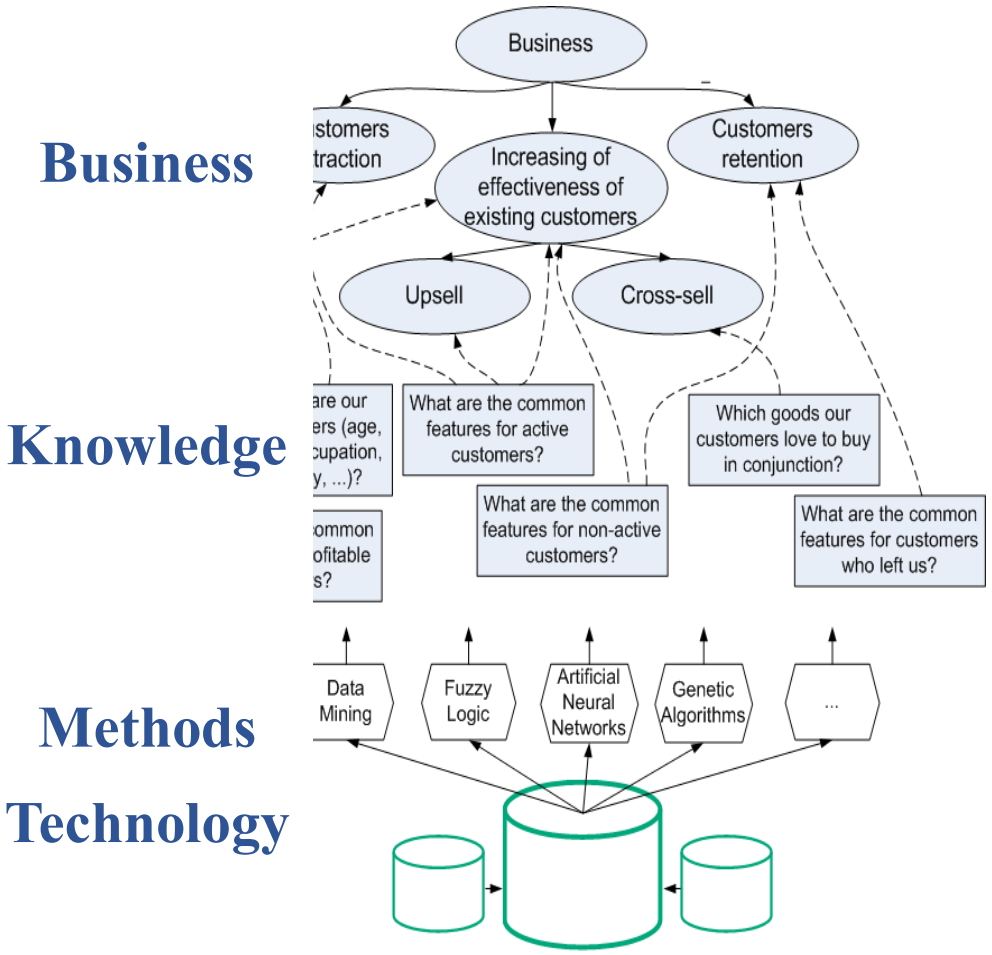
An Archetype for Online Science, Comm. ACM, 45(11): 50-54, Nov. 2002

- 1950s-1990: Computational science
    - Seluruh disiplin ilmu bergerak ke komputasi
    - Lahirnya banyak model komputasi
- 1990-sekarang: Data science
    - Kultur manusia menghasilkan data besar
    - Kemampuan komputer untuk mengolah data besar
    - Datangnya data mining sebagai arus utama sains

Jim Gray and Alex Szalay, The World Wide Telescope:An Archetype for Online Science, Comm. ACM, 45(11): 50-54, Nov. 2002


INDUSTRY 1.0
Mechanization, steam power, weaving loom
INDUSTRY 2.0
Mass production, assembly line, electrical energy
INDUSTRY 3.0
Automation, computers and electronics
INDUSTRY 4.0
Cyber Physical Systems, internet of things, networks
ERP
1784
1870
1969
TODAY


1. Big Data
2. Internet of Things
3. Business Process Automation
4. Enterprise Architecture
Transformation



4.2 Data Mining Law

Business Goals Law (Data Mining Law 1)

Business objectives are the origin of every data mining solution

- This defines the field of data mining: data mining is concerned with solving business problems and achieving business goals
- Data mining is not primarily a technology; it is a process, which has one or more business objectives at its heart
- Without a business objective, there is no data mining

The maxim: "Data Mining is a Business Process"

Business Knowledge Law (Data Mining Law 2)

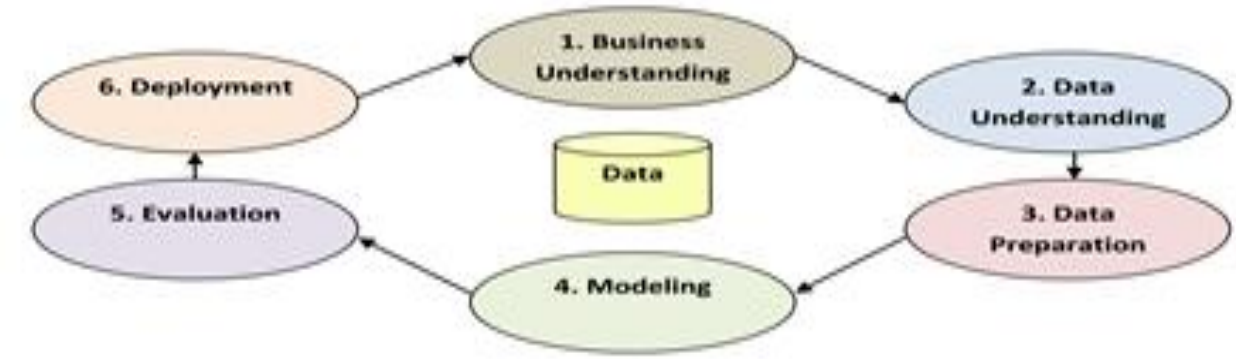
Business knowledge is central to every step of the data mining process

- A naive reading of CRISP-DM would see business knowledge used at the start of the process in defining goals, and at the end of the process in guiding deployment of results

Business Knowledge Law (Data Mining Law 3)

Business knowledge is central to every step of the data mining process

- This would be to miss a key property of the data mining process, that business knowledge has a central role in every step

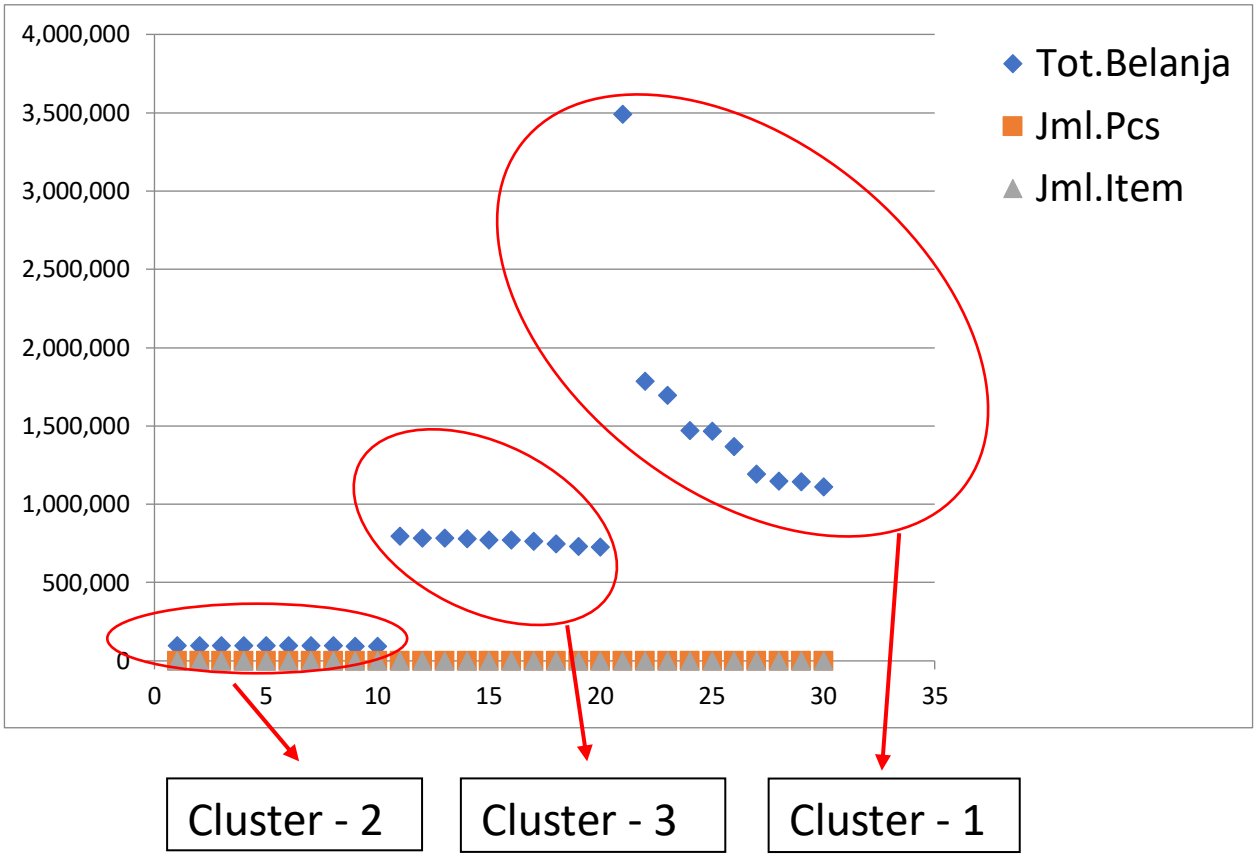


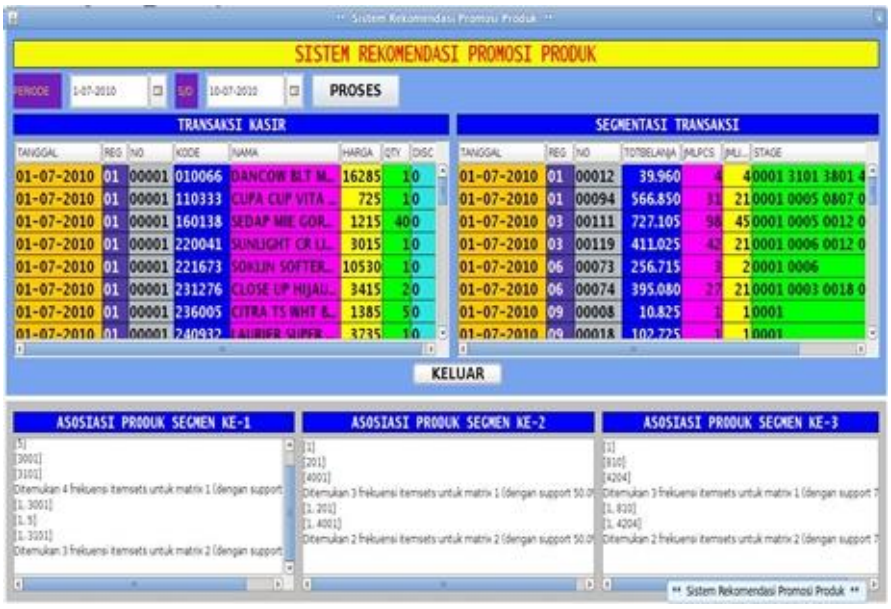
4.3 Penerapan Data Mining

Private and Commercial Sector

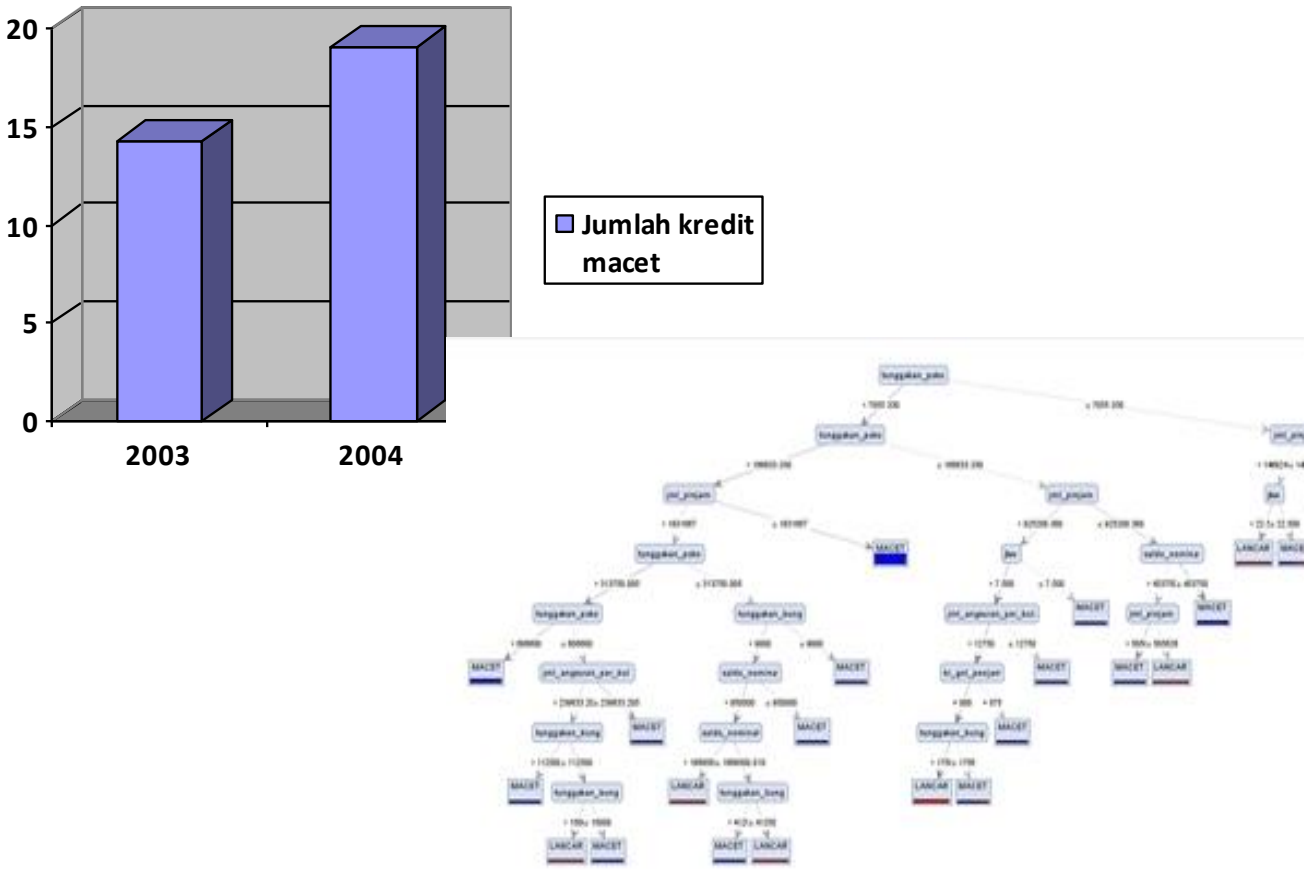
1. Marketing: product recommendation, market basket analysis, product targeting, customer retention
2. Finance: investment support, portfolio management, price forecasting
3. Banking and Insurance: credit and policy approval, money laundry detection
4. Security: fraud detection, access control, intrusion detection, virus detection
5. Manufacturing: process modeling, quality control, resource allocation
6. Web and Internet: smart search engines, web marketing
7. Software Engineering: effort estimation, fault prediction
8. Telecommunication: network monitoring, customer churn prediction, user behavior analysis

Use Case: Product Recommendation



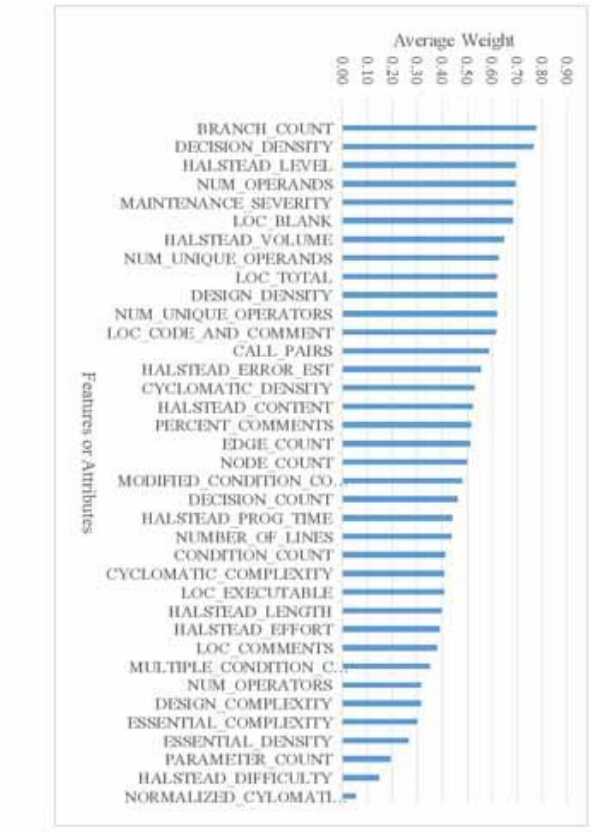


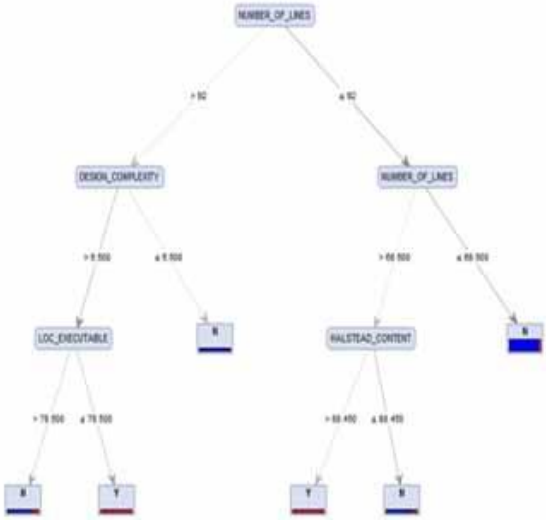
Use Case: Penentuan Kelayakan Kredit



Use Case: Software Fault Prediction

- The cost of capturing and correcting defects is expensive
  - $14,102 per defect in post-release phase *(Boehm & Basili 2008)*
  - $60 billion per year (NIST 2002)
- Industrial methods of manual software reviews activities can find only 60% of defects
  (Shull et al. 2002)
- The probability of detection of software fault prediction models is higher (71%) than software reviews (60%)

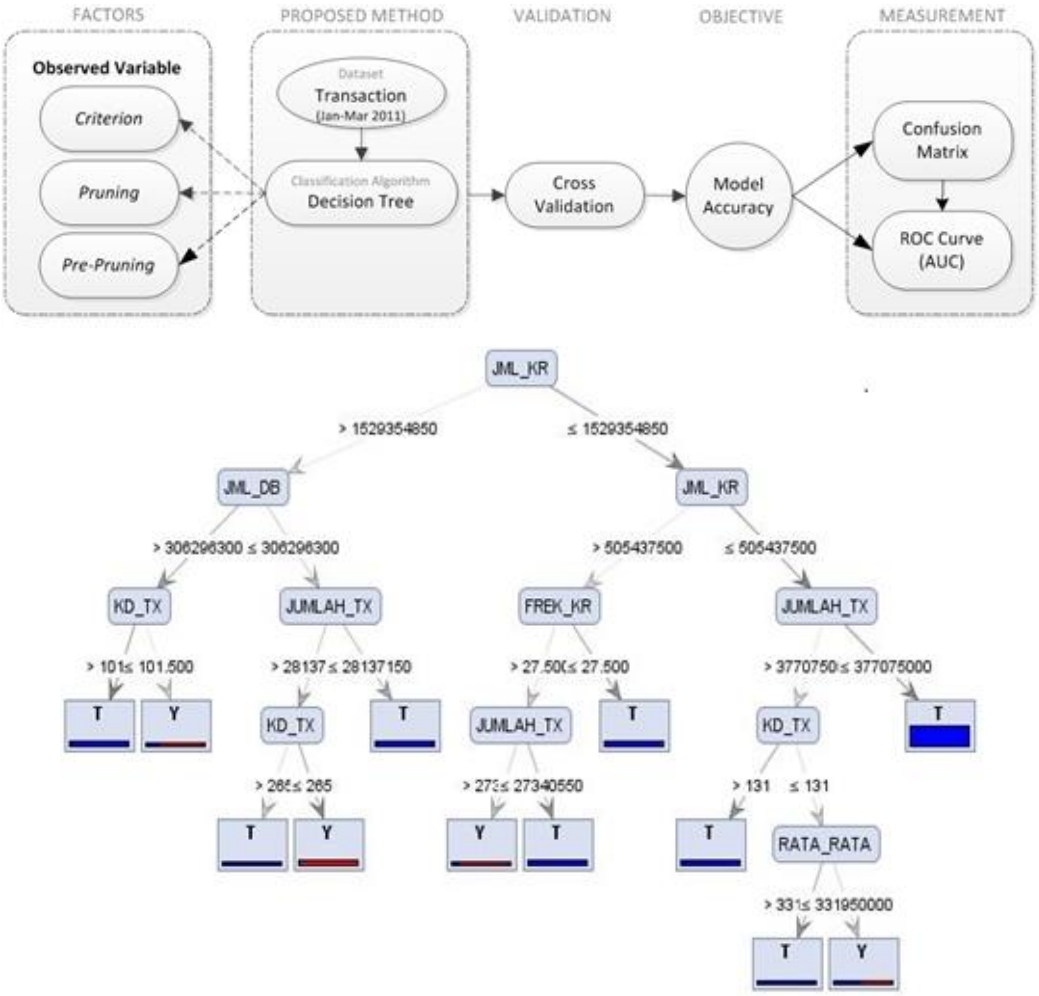




Public and Government Sector

1. Finance: exchange rate forecasting, sentiment analysis
2. Taxation: adaptive monitoring, fraud detection
3. Medicine and Healt Care: hypothesis discovery, disease prediction and classification, medical diagnosis
4. Education: student allocation, resource forecasting
5. Insurance: worker's compensation analysis
6. Security: bomb, iceberg detection
7. Transportation: simulation and analysis, load estimation
8. Law: legal patent analysis, law and rule analysis
9. Politic: election prediction

Use Case: Deteksi Pencucian Uang

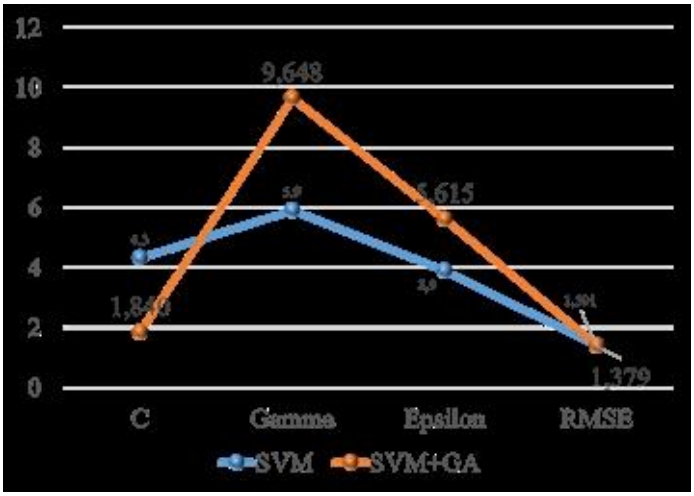


Use Case: Prediksi Kebakaran Hutan

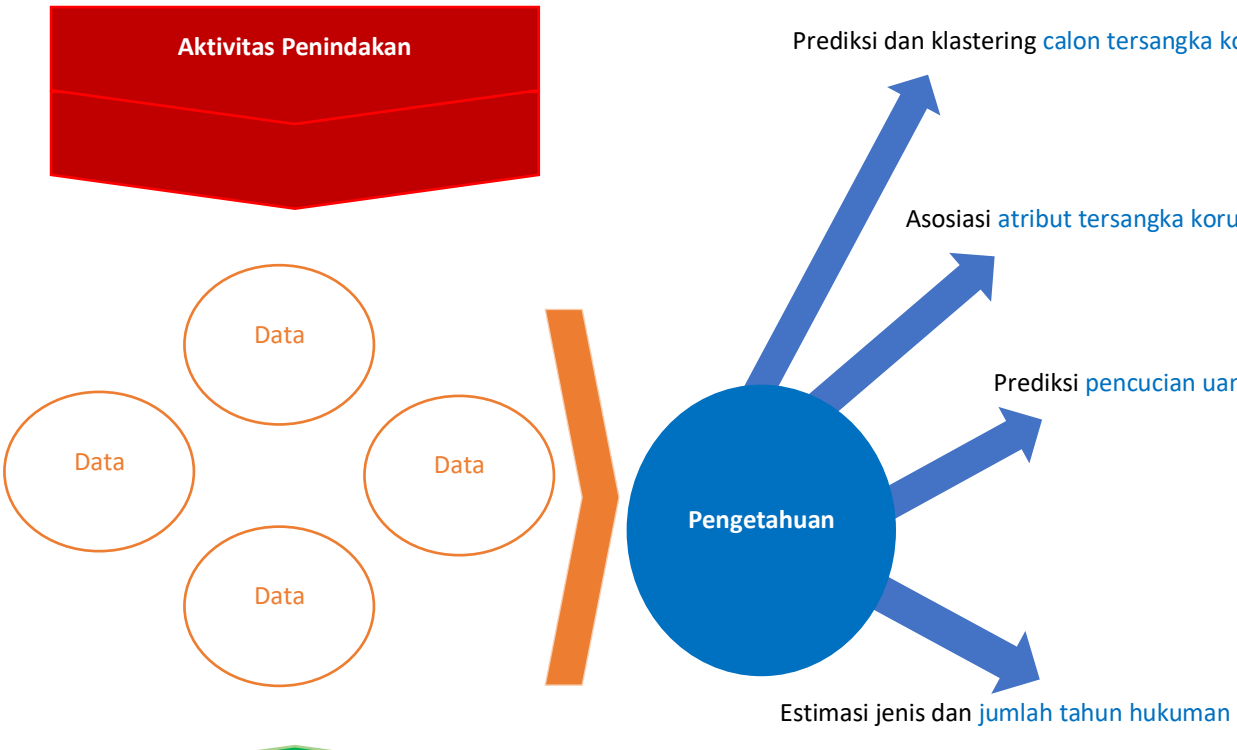
FFMC	DMC	DC	ISI	temp	RH	wind	rain	ln(area+1)
93.5	139.4	594.2	20.3	17.6	52	5.8	0	0
92.4	124.1	680.7	8.5	17.2	58	1.3	0	0
90.9	126.5	686.5	7	15.6	66	3.1	0	0
85.8	48.3	313.4	3.9	18	42	2.7	0	0.307485
91	129.5	692.6	7	21.7	38	2.2	0	0.357674
90.9	126.5	686.5	7	21.9	39	1.8	0	0.385262

95.5	99.9	513.3	13.2	23.3	31	4.5	0	0.438255

	SVM	SVM+GA
C	4.3	1,840
Gamma (γ)	5.9	9,648
Epsilon (ε)	3.9	5,615
RMSE	1.391	1.379



Use Case: Profiling dan Prediksi Koruptor



Contoh Penerapan Data Mining

1. Penentuan kelayakan kredit pemilihan rumah di bank
2. Penentuan pasokan listrik PLN untuk wilayah Jakarta
3. Prediksi profile tersangka koruptor dari data pengadilan
4. Perkiraan harga saham dan tingkat inflasi
5. Analisis pola belanja pelanggan
6. Memisahkan minyak mentah dan gas alam
7. Penentuan pola pelanggan yang loyal pada perusahaan operator telepon
8. Deteksi pencucian uang dari transaksi perbankan
9. Deteksi serangan (intrusion) pada suatu jaringan

Data Mining Society

1. 1989 IJCAI Workshop on Knowledge Discovery in Databases
   - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
2. 1991-1994 Workshops on Knowledge Discovery in Databases
   - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
3. 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
   - Journal of Data Mining and Knowledge Discovery (1997)
4. ACM SIGKDD conferences since 1998 and SIGKDD Explorations
5. More conferences on data mining
   - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), WSDM (2008), etc.
6. ACM Transactions on KDD (2007)

Conferences Data Mining

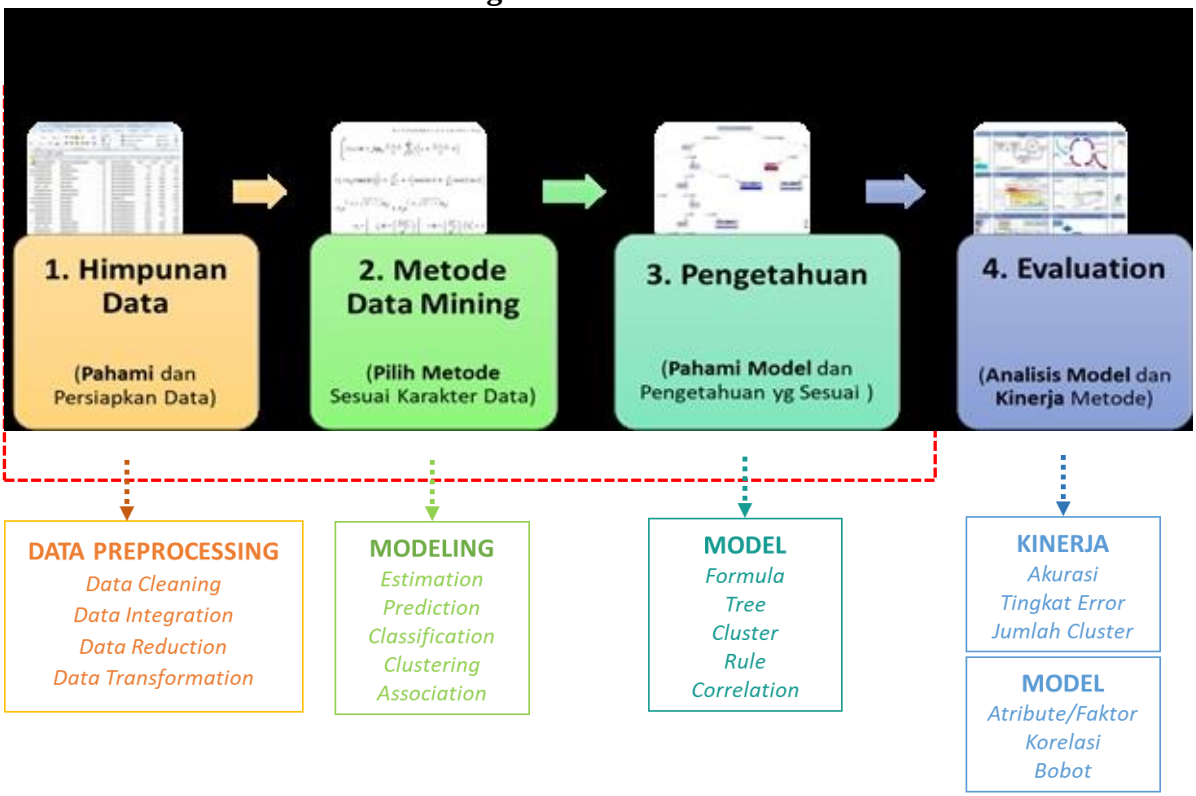
- ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
- SIAM Data Mining Conf. (SDM)
- (IEEE) Int. Conf. on Data Mining (ICDM)
- European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
- Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- Int. Conf. on Web Search and Data Mining (WSDM)

Journals Data Mining

- ACM Transactions on Knowledge Discovery from Data (TKDD)
- ACM Transactions on Information Systems (TOIS)
- IEEE Transactions on Knowledge and Data Engineering
- Springer Data Mining and Knowledge Discovery
- International Journal of Business Intelligence and Data Mining (IJBIDM)

Modul 5 Proses Data Mining

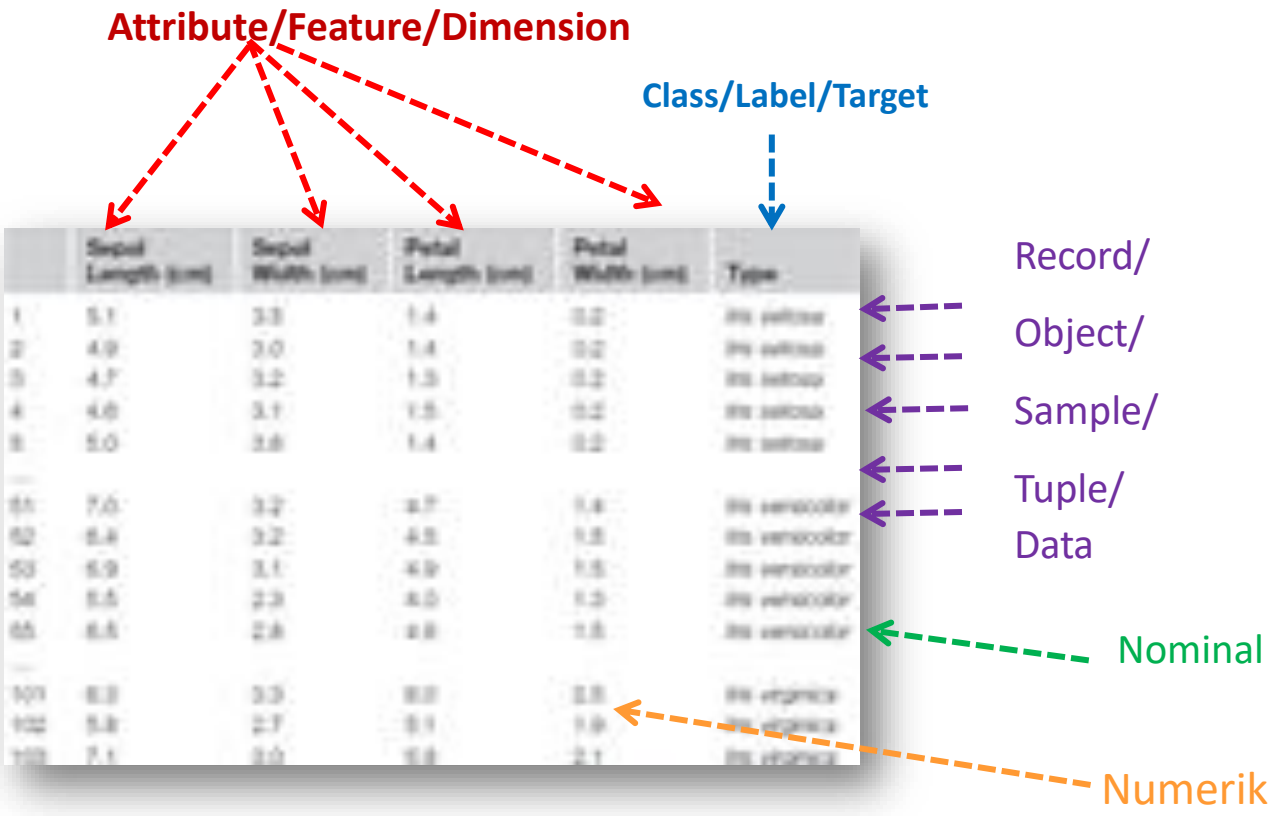
5.1 Proses dan Tools Data Mining



1. Himpunan Data (Dataset)

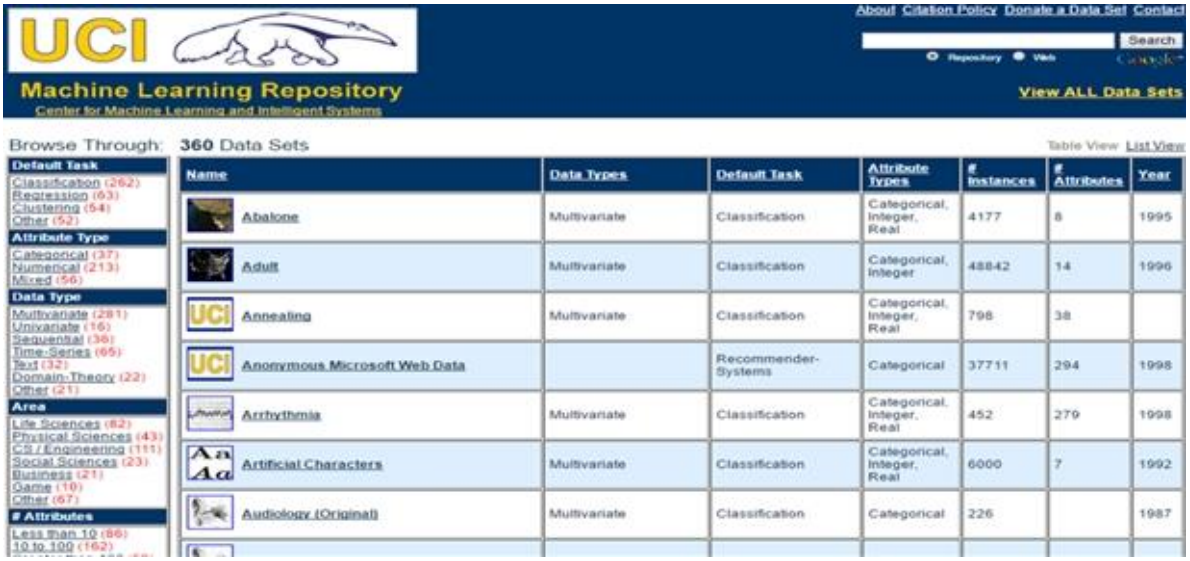
- Atribut adalah faktor atau parameter yang menyebabkan class/label/target terjadi
- Jenis dataset ada dua: Private dan Public
- Private Dataset: data set dapat diambil dari organisasi yang kita jadikan obyek penelitian
    - Bank, Rumah Sakit, Industri, Pabrik, Perusahaan Jasa, etc
- Public Dataset: 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



2. Metode Data Mining (DM)

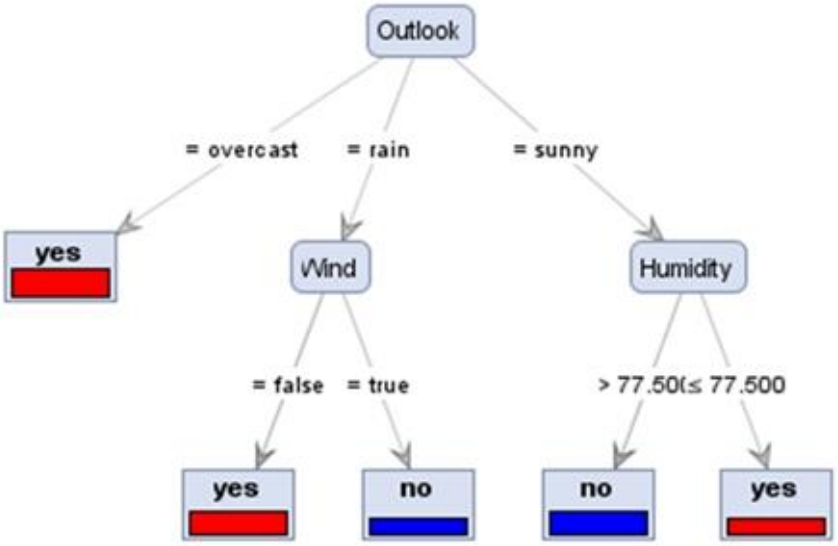
1. Estimation (Estimasi):
   - Linear Regression, Neural Network, Support Vector Machine, Deep Learning, etc
2. Prediction/Forecasting (Prediksi/Peramalan):
   - Linear Regression, Neural Network, Support Vector Machine, Deep Learning, etc
3. Classification (Klasifikasi):
   - Decision Tree (CART, ID3, 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)

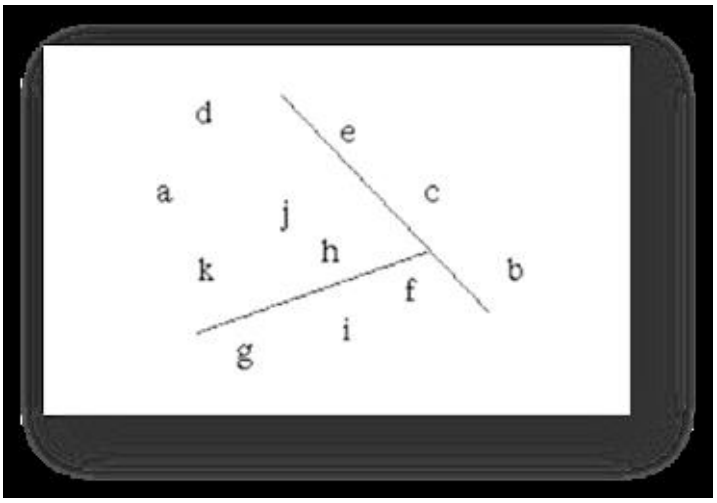
1. Formula/Function (Rumus atau Fungsi Regresi)

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

2. Decision Tree (Pohon Keputusan)



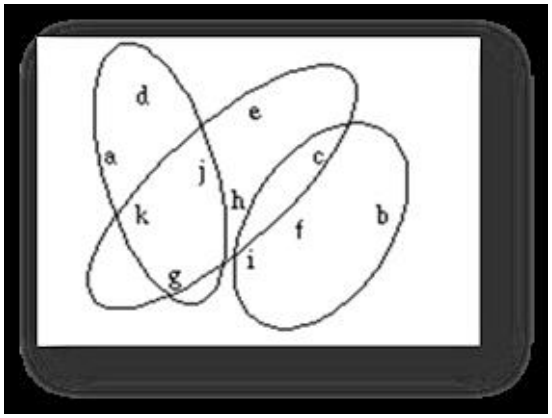
3. Korelasi dan Asosiasi



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

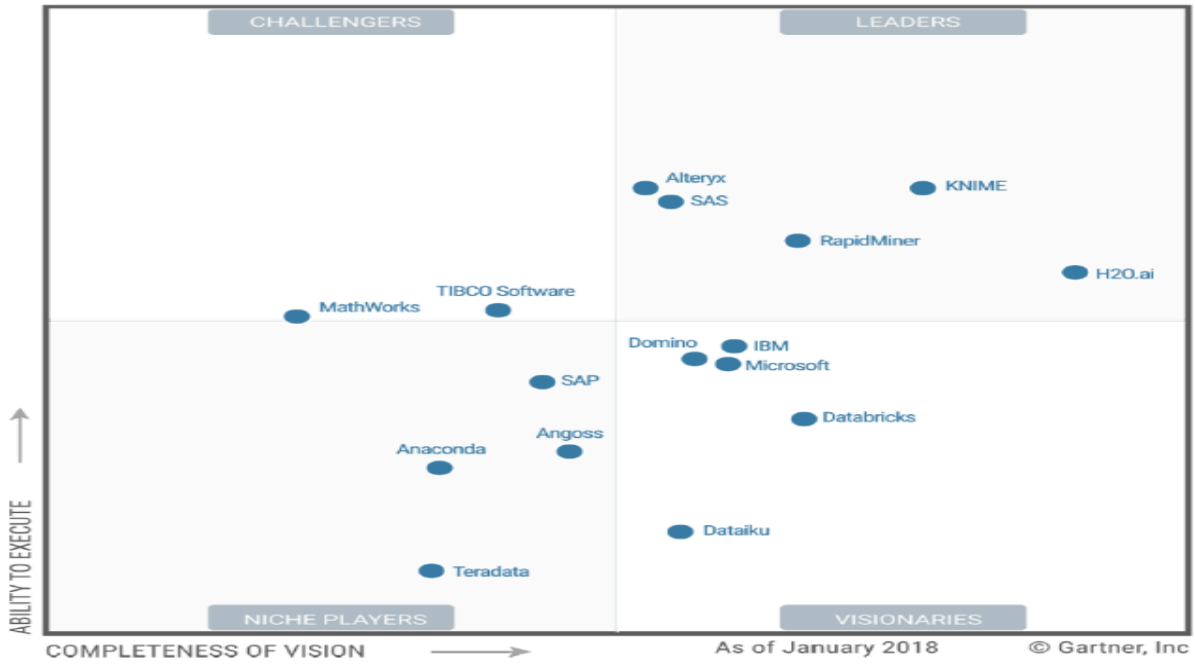
2. Kehandalan
   a. Ukuran di mana model data mining diterapkan pada dataset yang berbeda
   b. Model data mining dapat diandalkan jika menghasilkan pola umum yang sama terlepas dari data testing yang disediakan
3. Kegunaan
   - Mencakup berbagai metrik yang mengukur apakah model tersebut memberikan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

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

5.2 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

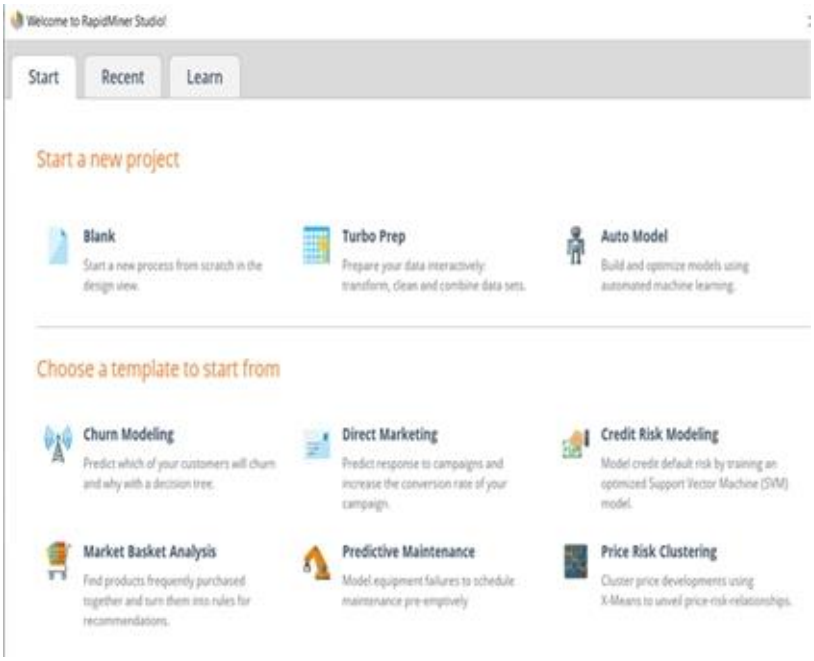
1. nominal: nilai secara kategori
2. binominal: nominal dua nilai
3. polynominal: nominal lebih dari dua nilai
4. numeric: nilai numerik secara umum
5. integer: bilangan bulat
6. real: bilangan nyata
7. text: teks bebas tanpa struktur
8. date_time: tanggal dan waktu
9. date: hanya tanggal
10. time: hanya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

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

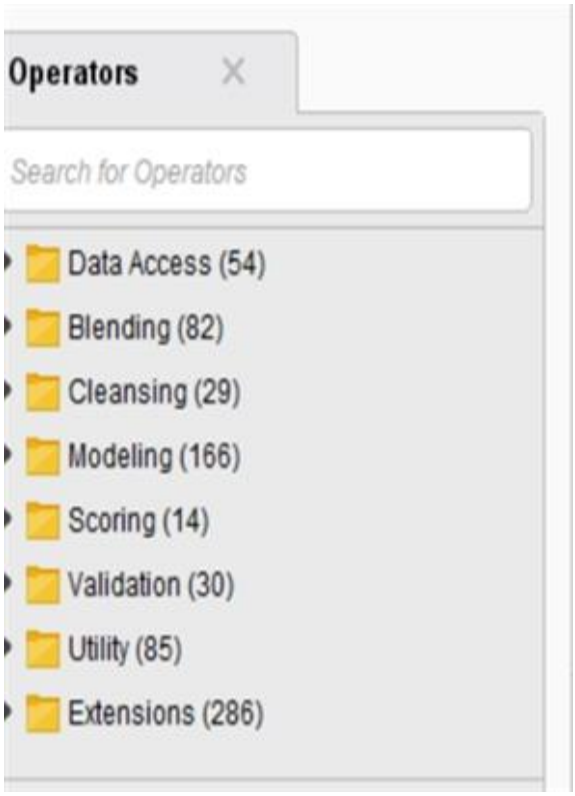
Untuk menghitung kualitas dan perfomansi dan validasi dari model

7. Utility

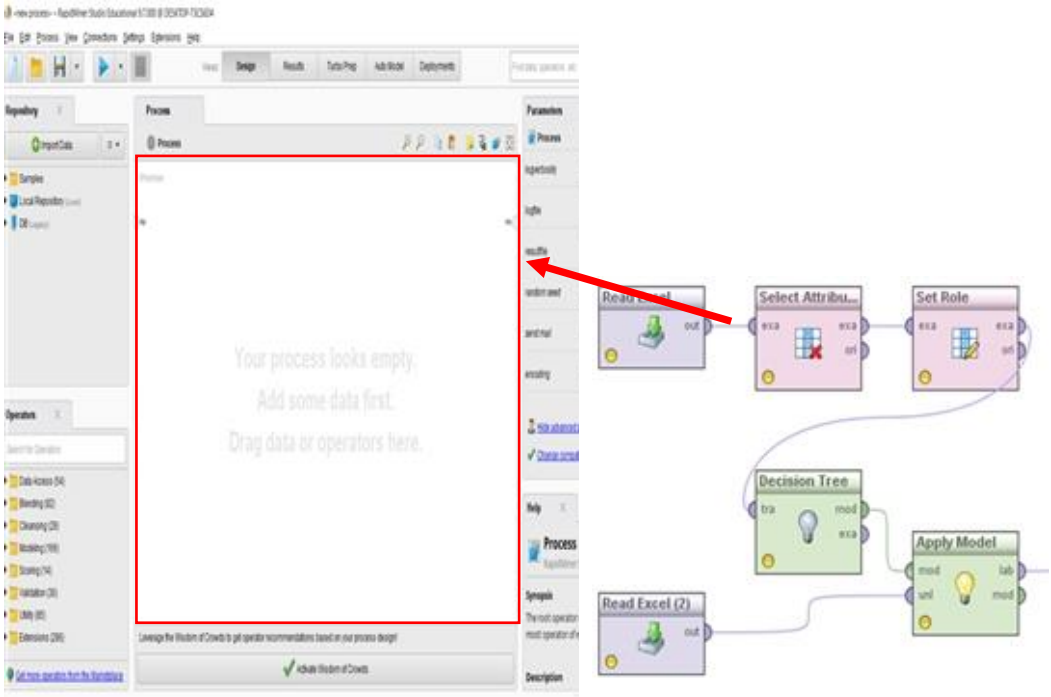
Untuk mengelompokkan subprocess, juga macro dan logger eksternal

8. Extentions

Fasilitas tambahan seperti Text Mining, Web Mining, dll

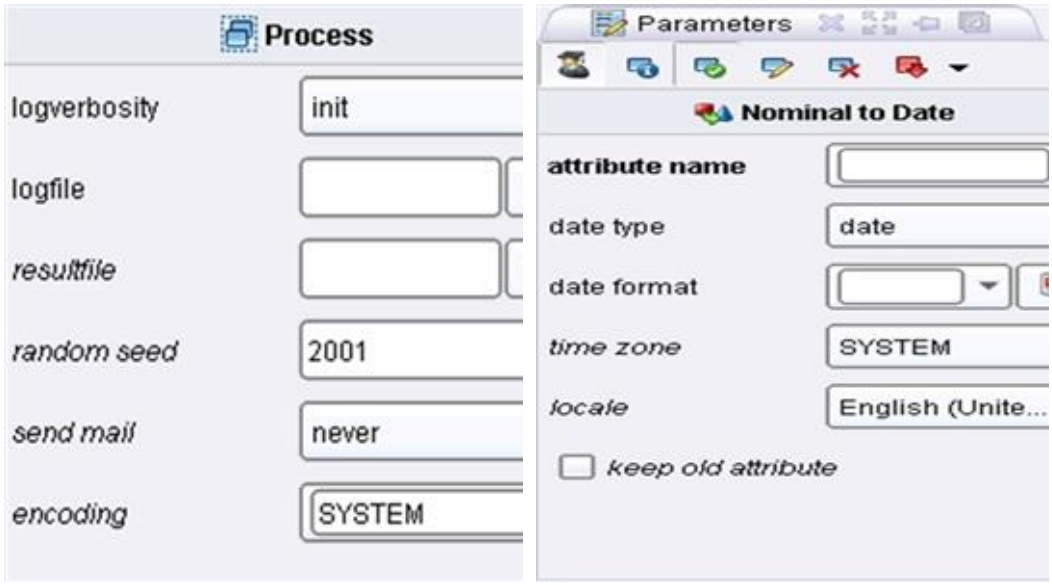


View Proses



View Parameter

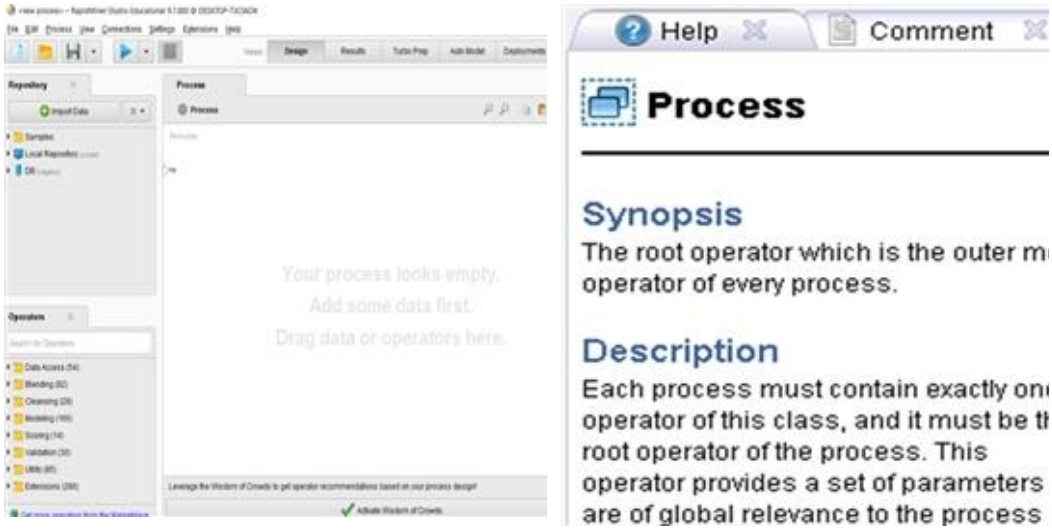
- Operator kadang memerlukan parameter untuk bisa berfungsi
- Setelah operator dipilih di view Proses, parameternya ditampilkan di view ini



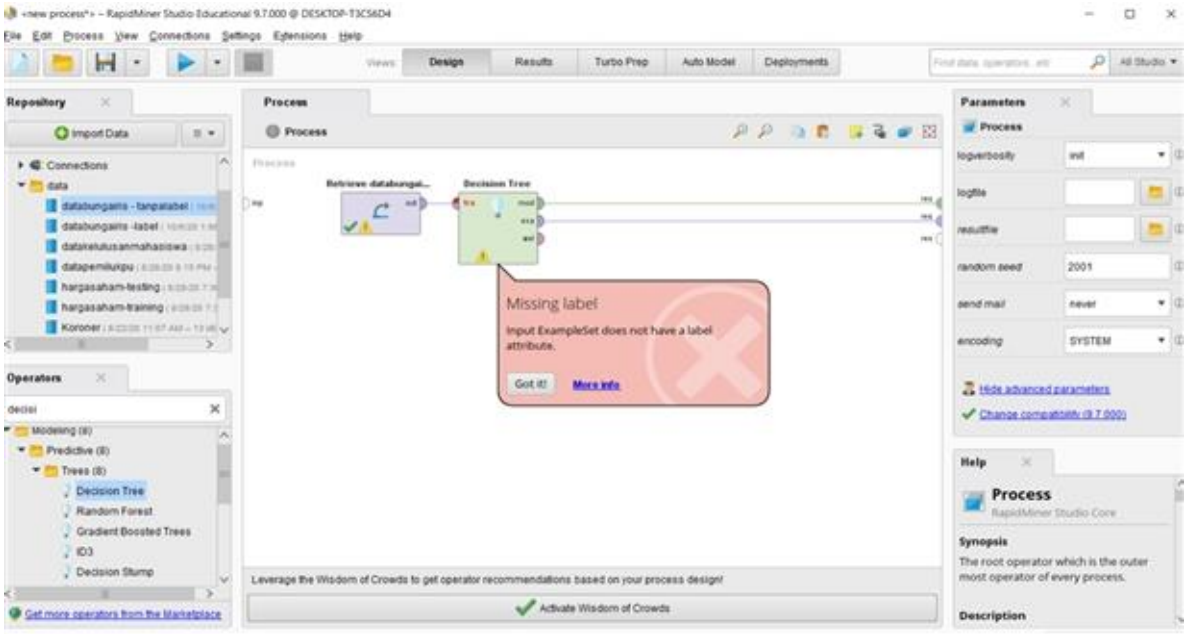
View Help dan View Comment

View Help menampilkan deskripsi dari operator

View Comment menampilkan komentar yang dapat diedit terhadap operator



View Problems



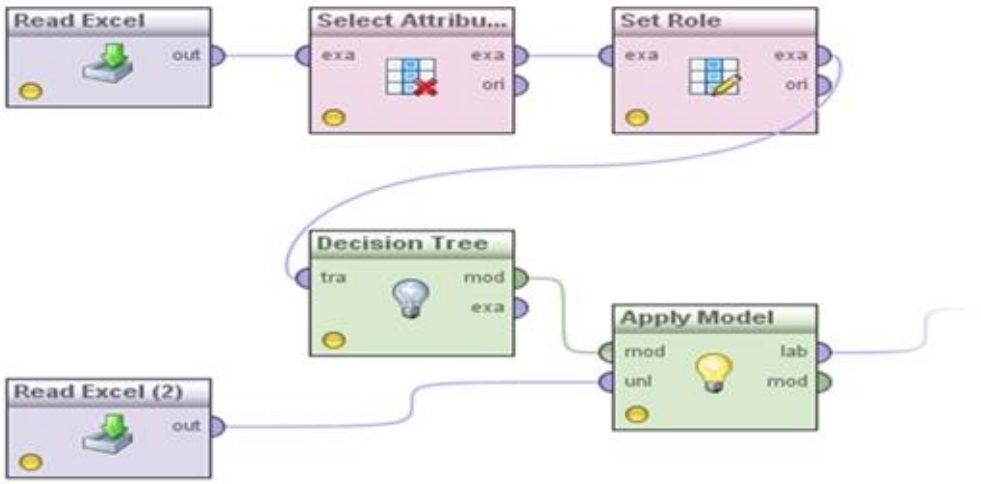
Operator dan Proses

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 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

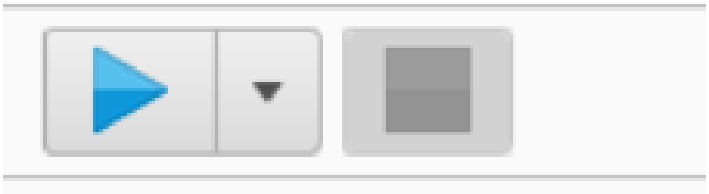
Mendesain Proses



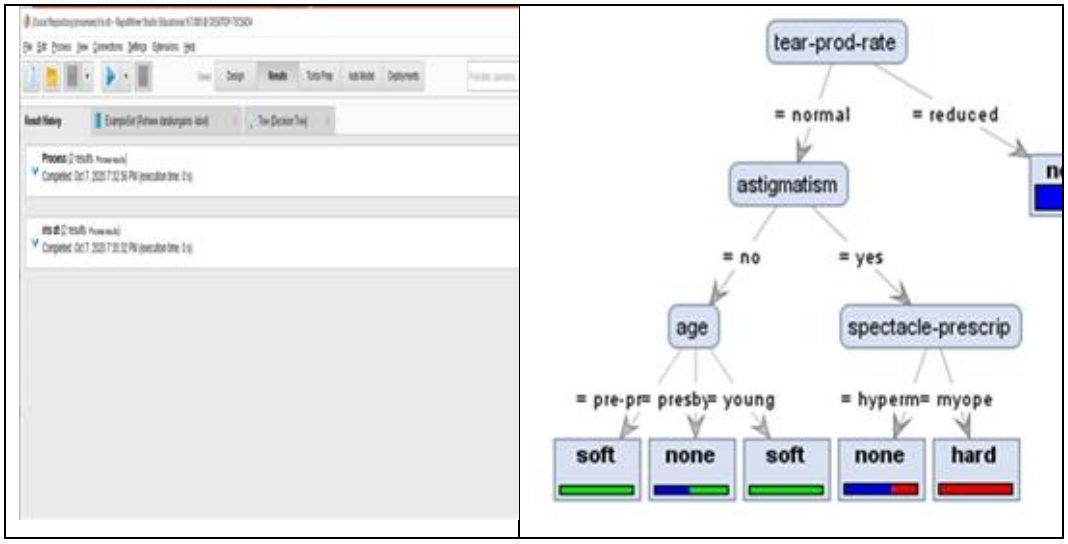
Menjalankan Proses

Proses dapat dijalankan dengan:

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



Melihat Hasil



Panduan Install Rapid Miner

I. Instalasi JDK			
1	Instalasi JDK • Double-click pada file instaler JDK yang akan diinstall • Klik "Next"	2	Instalasi JDK • Klik "Next"
3	Instalasi JDK • Instalasi JDK selesai • Klik "Close"		
II. Downlaod RapidMiner			

1	Download RapidMiner • https://rapidminer.com • Klik "Get Started" 	2	
3		4	

III. Instalasi RapidMiner

1	Instalasi RapidMiner • Double-click pada file instaler Rapidminer yang akan diinstall • Klik "Next" 	2	Instalasi RapidMiner • Klik "I Agree" 
3	Instalasi RapidMiner • Pilih lokasi penginstalan • Klik "Install" 	4	Instalasi RapidMiner • Klik "Finish" 

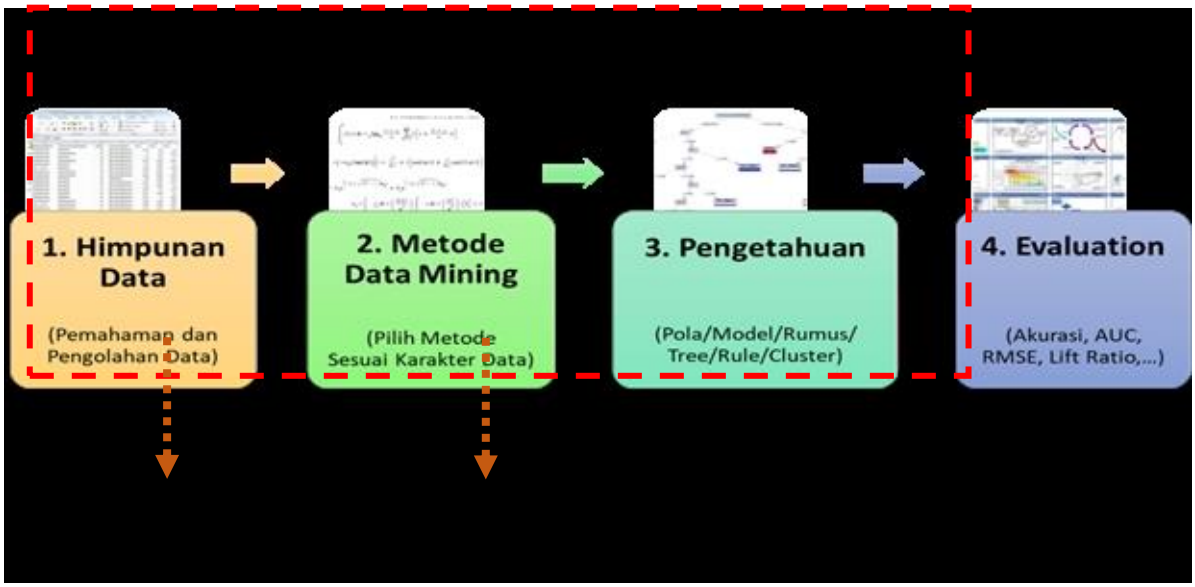
IV. Registrasi Akun

1	Registrasi Akun • Checklist "I have read and unserstand the terms..." • Klik "I Accept" 	2	Registrasi Akun • Pilih "Educational" • Lengkapi data yang diperlukan • Gunakan email yang aktif untuk verifikasi akun • Klik Create my Account! 
3	Registrasi Akun Setelah membuat akun, periksa email aktivasi 	4	Registrasi Akun • Buka email verify dari RapidMiner 
5	Registrasi Akun • Buka email verify dari RapidMiner 	6	Registrasi Akun • Klik "confirm your email address" 
7	Registrasi Akun • Klik "Refresh", akun anda akan terverifikasi 	8	Registrasi Lisensi Apabila saat ini anda telah terdaftar di RapidMiner namun belum memiliki lisensi: • Akses My Account Rapidminer • Klik Educational Program 

9	Registrasi Lisensi • Lengkapi semua field. Konten dapat mengacu pada histori perkuliahan anda • 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." • Klik Apply for licens 	10	
11	Registrasi Lisensi • Buka Kembali aplikasi Rapidminer • Pastikan bahwa lisensi anda telah aktif seperti pada gamber berikut 		

MODUL 6. Penerapan Proses Data Mining

Proses Data Mining



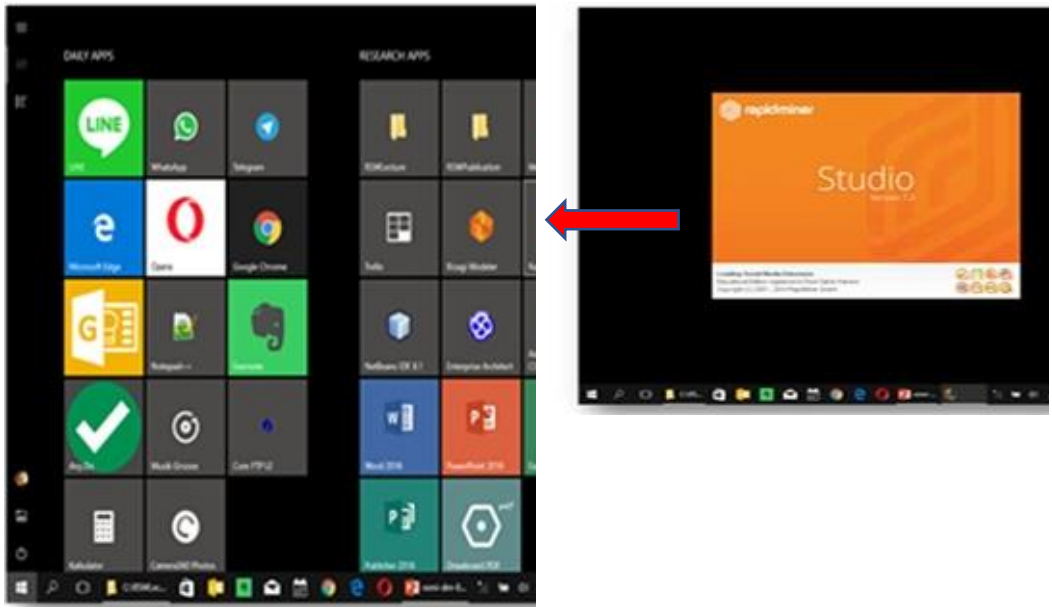
Data PreProcesing
Data Cleaning
Data Integration
Data Reduction
Data Transformation

Estimation
Prediction
Classification
Clustering
Association

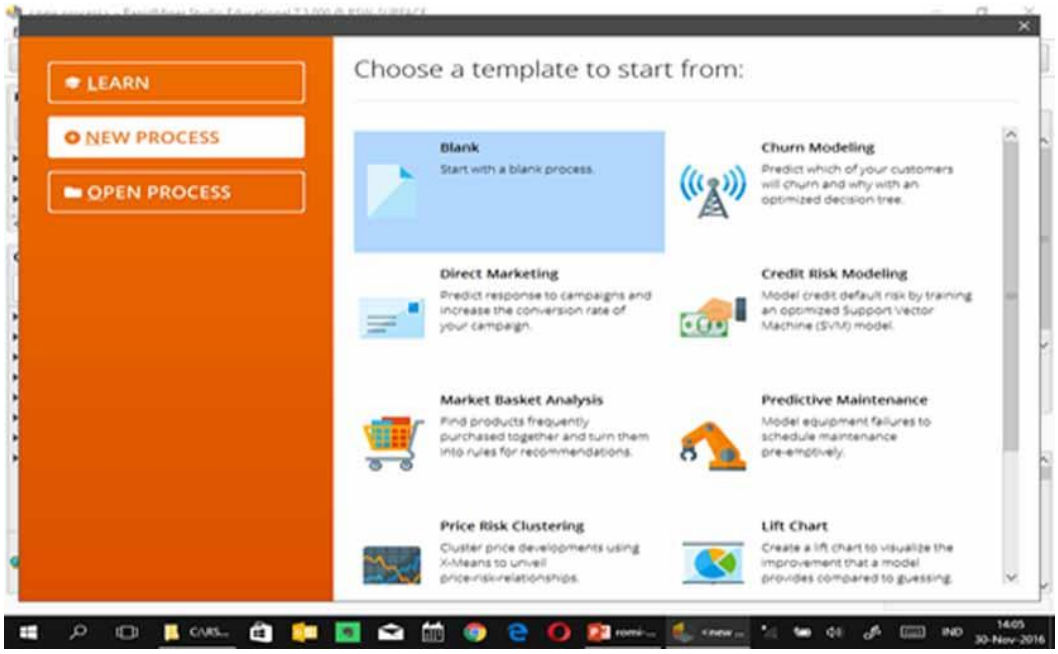
1. Latihan: Rekomendasi Main Golf

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

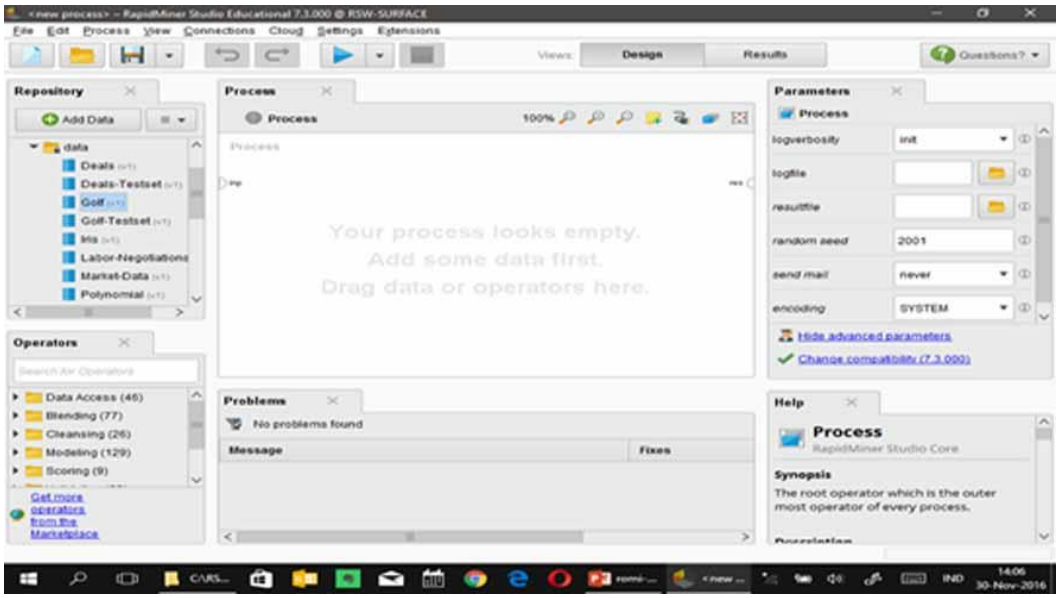
Buka RapidMiner



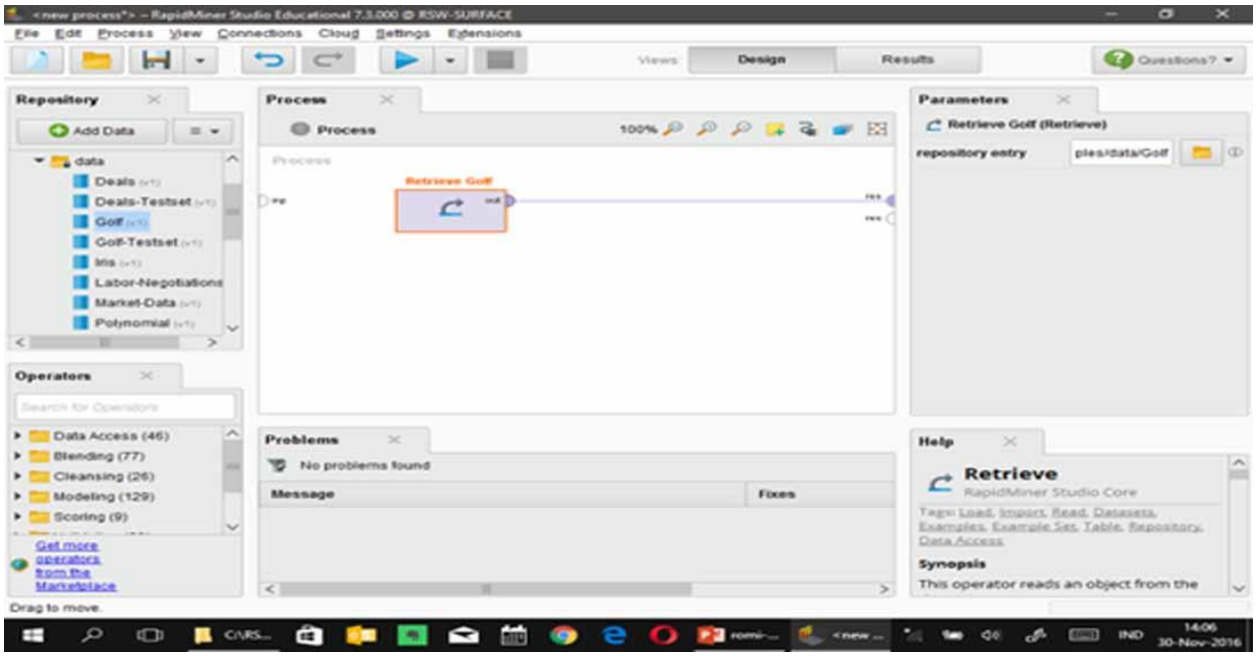
Klik Blank

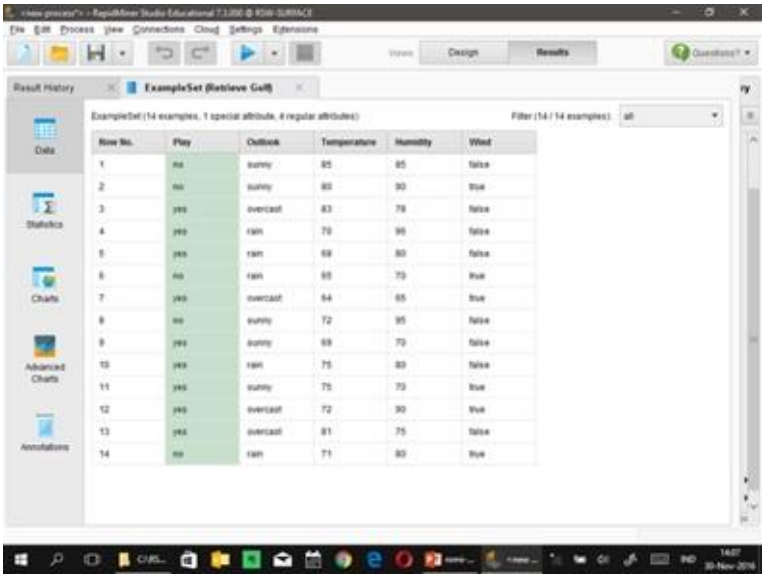


Muncul Lembar Kerja

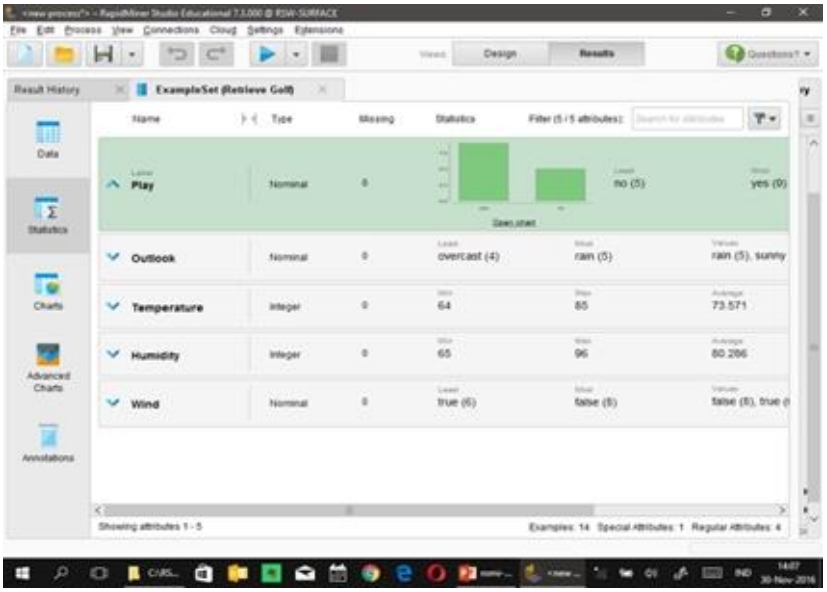


Membaca Data dan menampilkan data

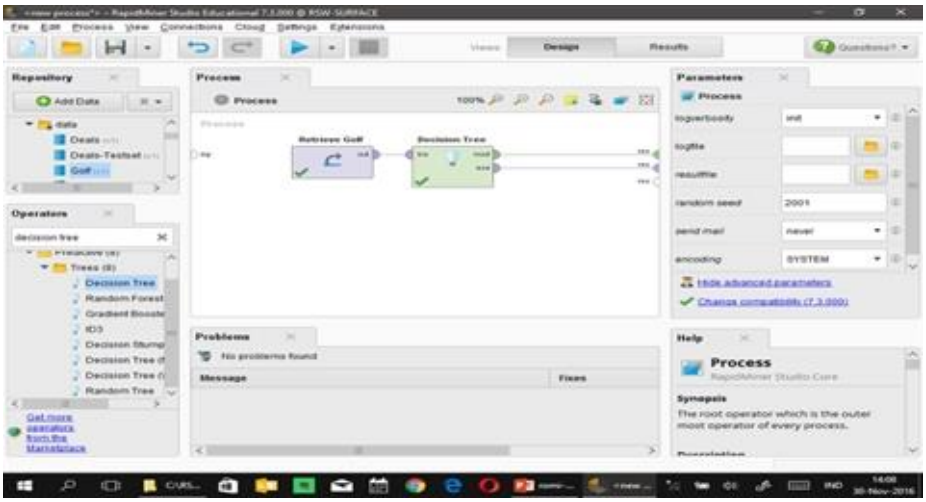




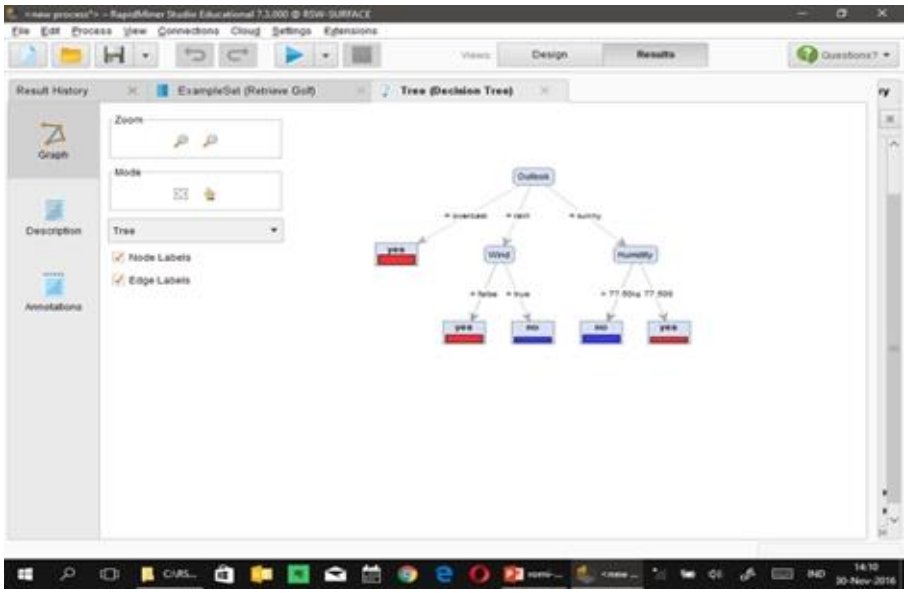
Menampilkan Statistik Data



Membuat Model



Menampilkan Hasil

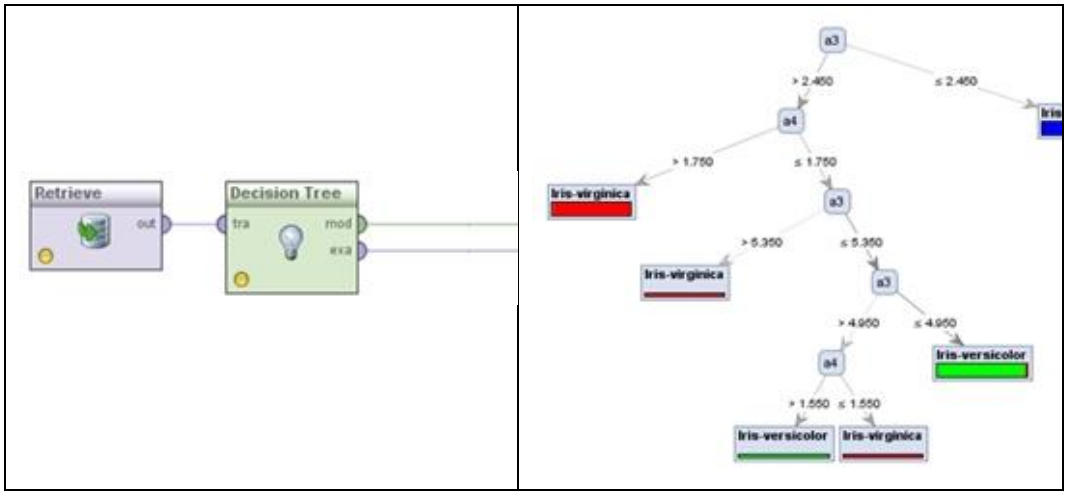


2. Latihan: Rekomendasi Main Tenis

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

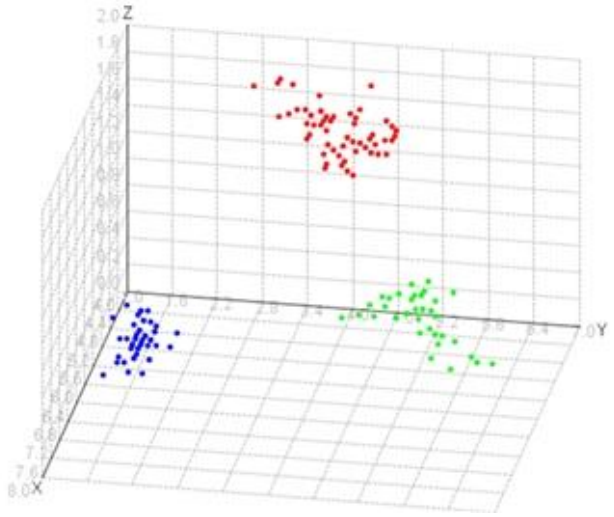
3. Latihan: Penentuan Jenis Bunga Iris

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

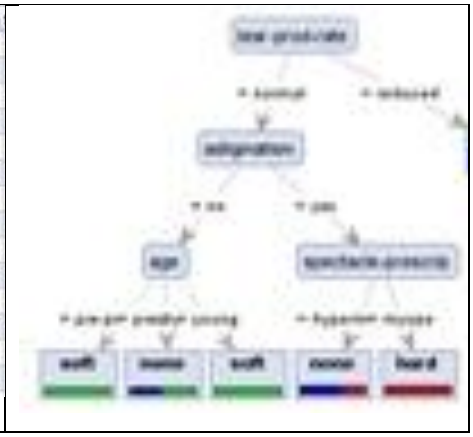
1. Lakukan training pada data Bunga Iris (ambil dari repositories rapidminer) dengan menggunakan algoritma k-Means
2. Tampilkan himpunan data (dataset) dan pengetahuan (model tree) yang terbentuk
3. Tampilkan grafik dari cluster yang terbentuk seperti di bawah ini.



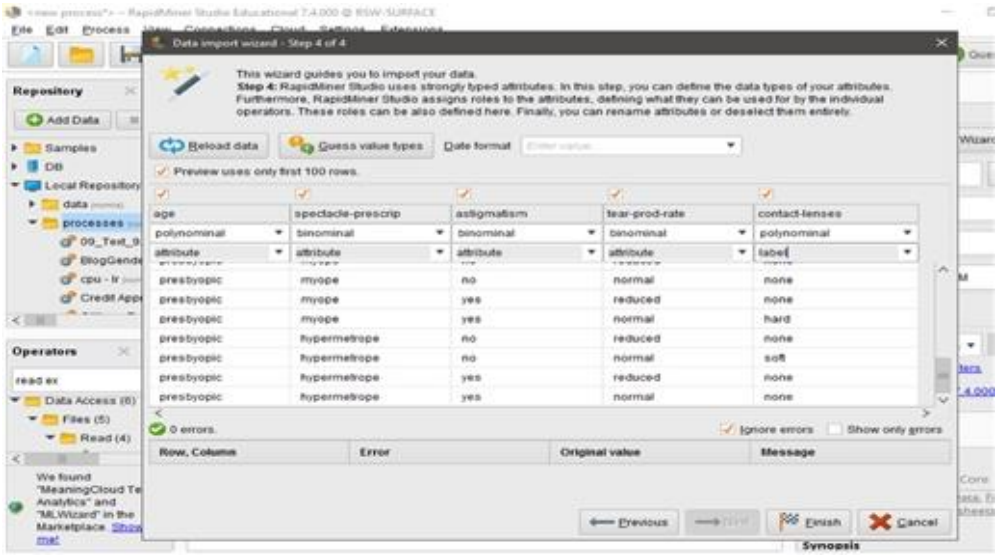
5. Latihan: Rekomendasi Contact Lenses

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

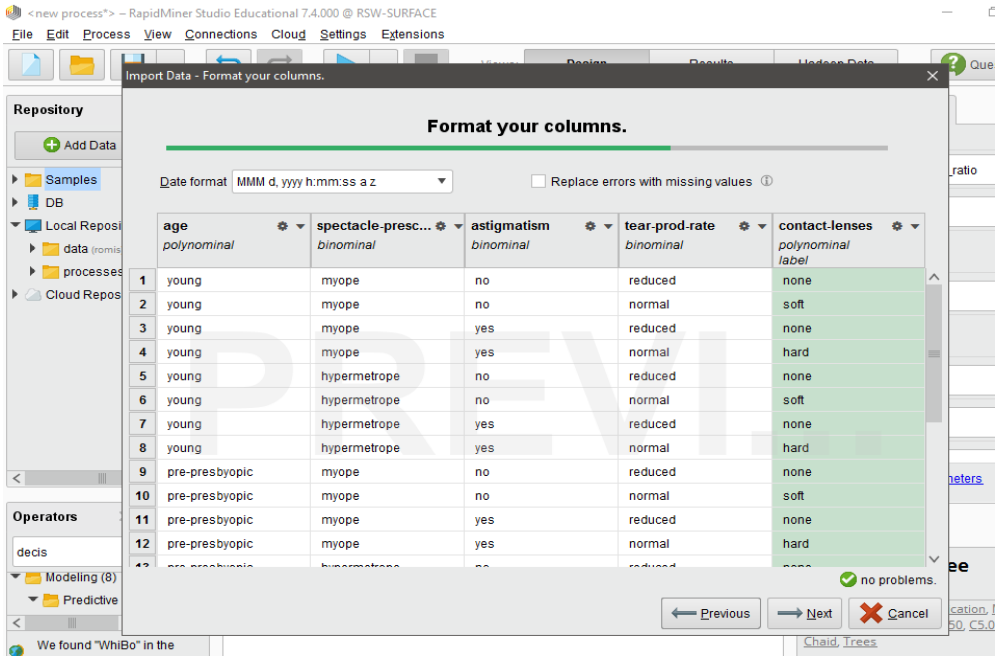
Row No.	contact-len...	age	spectacle-p...	astigmatism
1	none	young	myope	no
2	soft	young	myope	no
3	none	young	myope	yes
4	hard	young	myope	yes
5	none	young	hypermetrop	no
6	soft	young	hypermetrop	no
7	none	young	hypermetrop	yes
8	hard	young	hypermetrop	yes
9	none	pre-presbyo	myope	no
10	soft	pre-presbyo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



Read Excel Operator

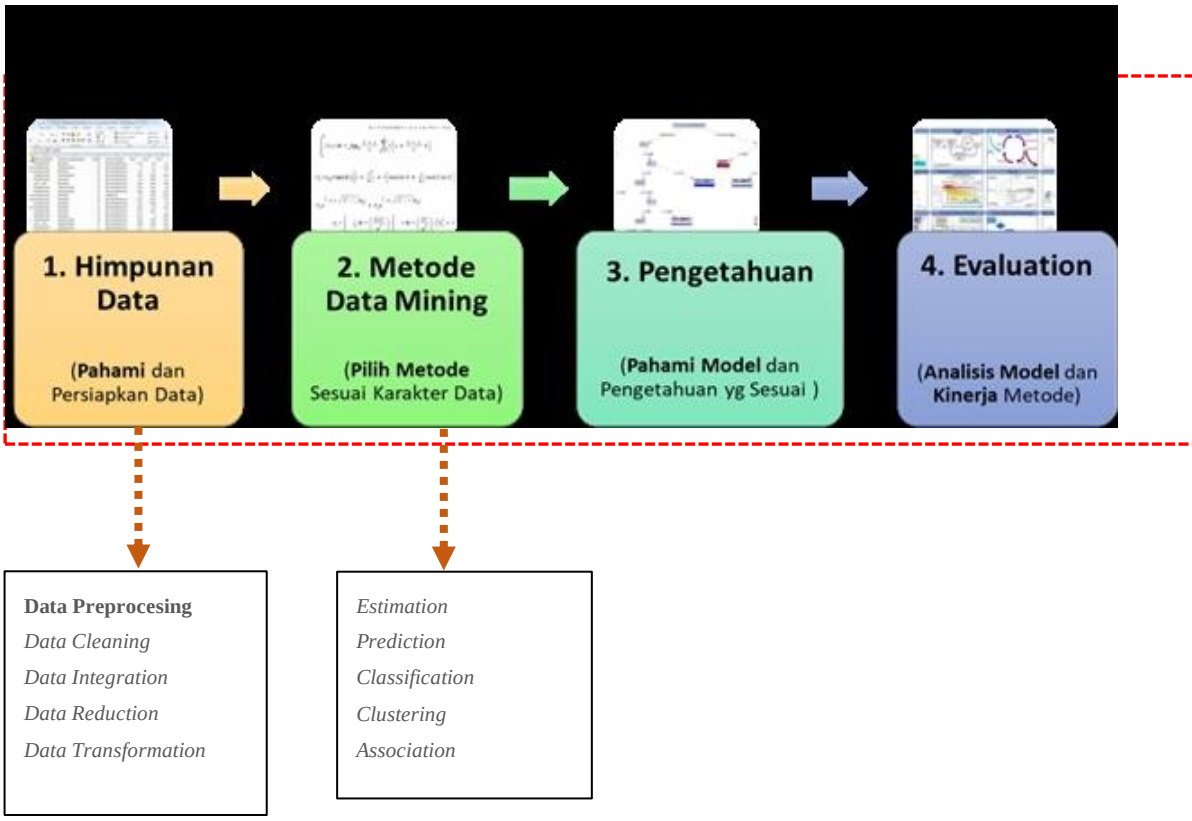


Import Data Function


<new process*> – RapidMiner Studio Educational 7.4.000 @ RSW-SURFACE
File Edit Process View Connections Cloud Settings Extensions
Repository
Add Data
Samples
DB
Local Reposi
data
processes
Cloud Repos
Operators
decis
Modeling (8)
Predictive
We found "WhiBo" in the
Import Data - Format your columns.
Format your columns.
Date format MMM d, yyyy h:mm:ss a z
Replace errors with missing values
age polynominal
spectacle-presc... binominal
astigmatism binominal
tear-prod-rate binominal
contact-lenses polynominal label
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
no problems.
Previous
Next
Cancel
Chaid, Trees

Modul 7. Evaluasi Model Data Mining

7.1 Proses Data Mining



7.2 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:

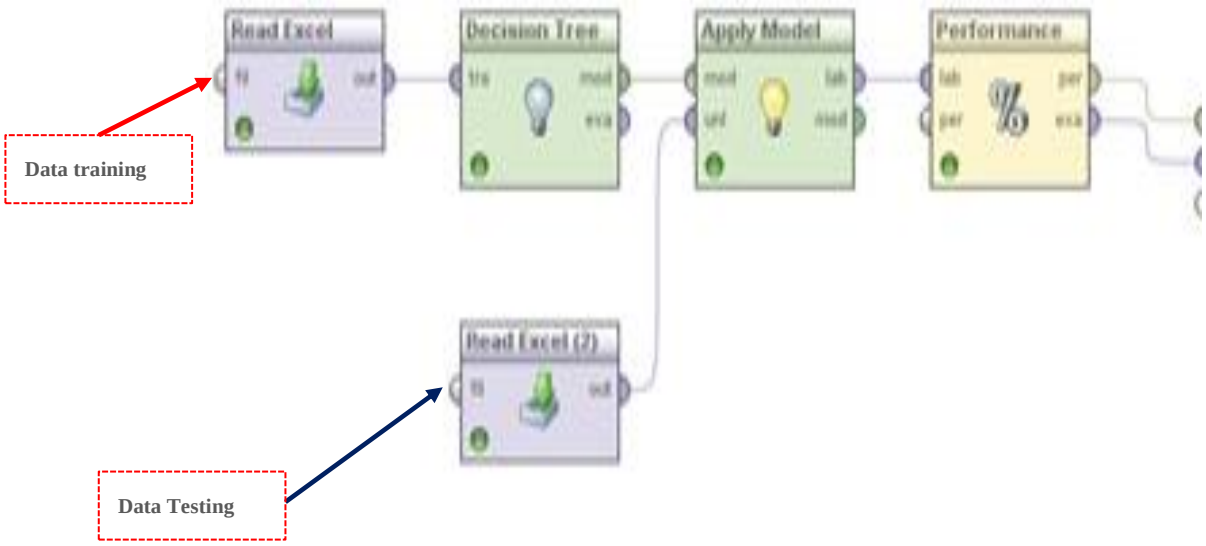
1. Data dipisahkan secara manual
2. Data dipisahkan otomatis dengan operator Split Data
3. Data dipisahkan otomatis dengan X Validation

Pemisahan Data Manual

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

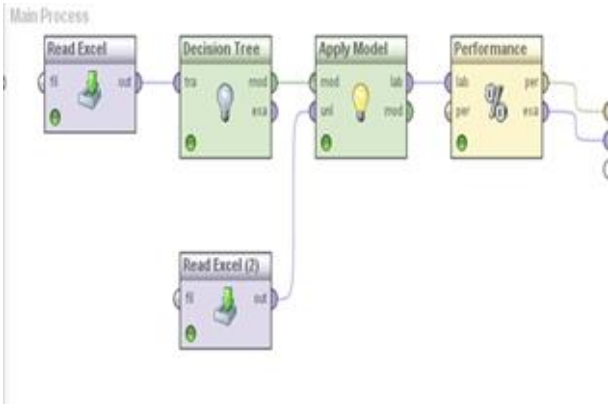
Latihan: Penentuan Kelayakan Kredit

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)



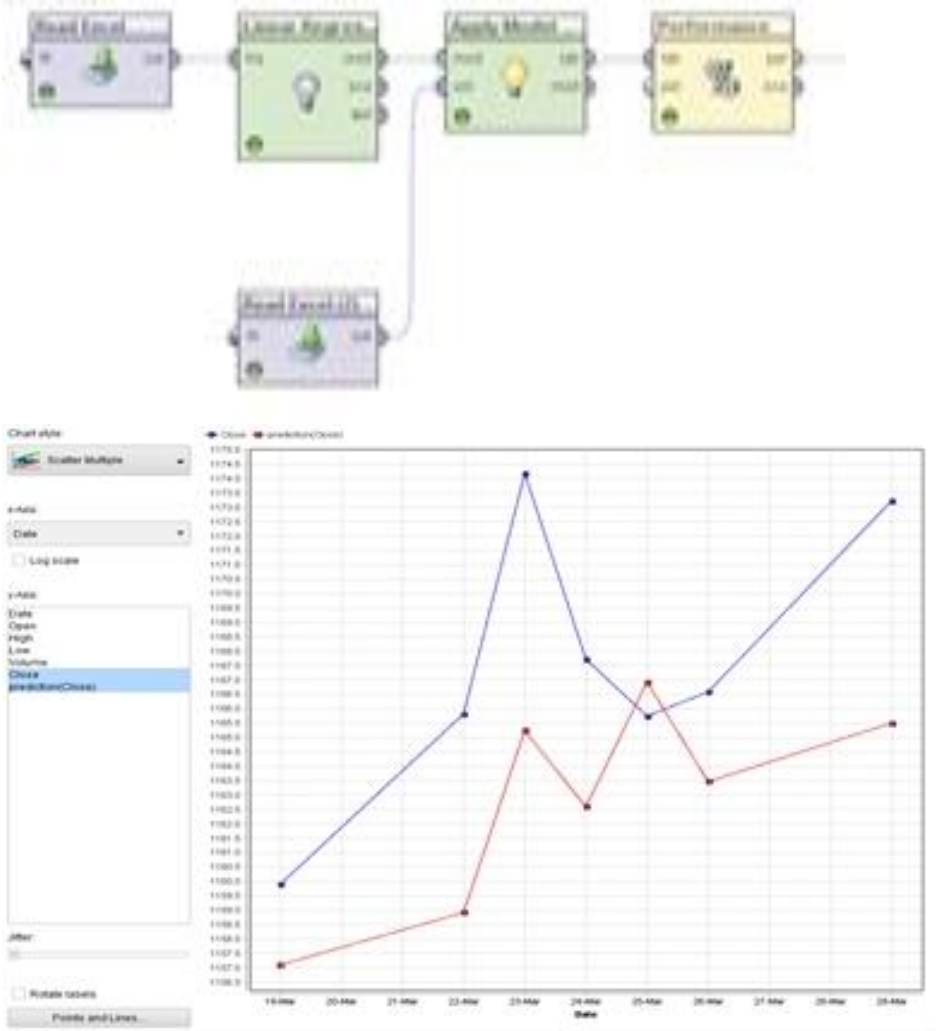
	C4.5
Accuracy	58%
AUC	0.86

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



7.3 Pemisahan Data otomatis dengan operator Split Data

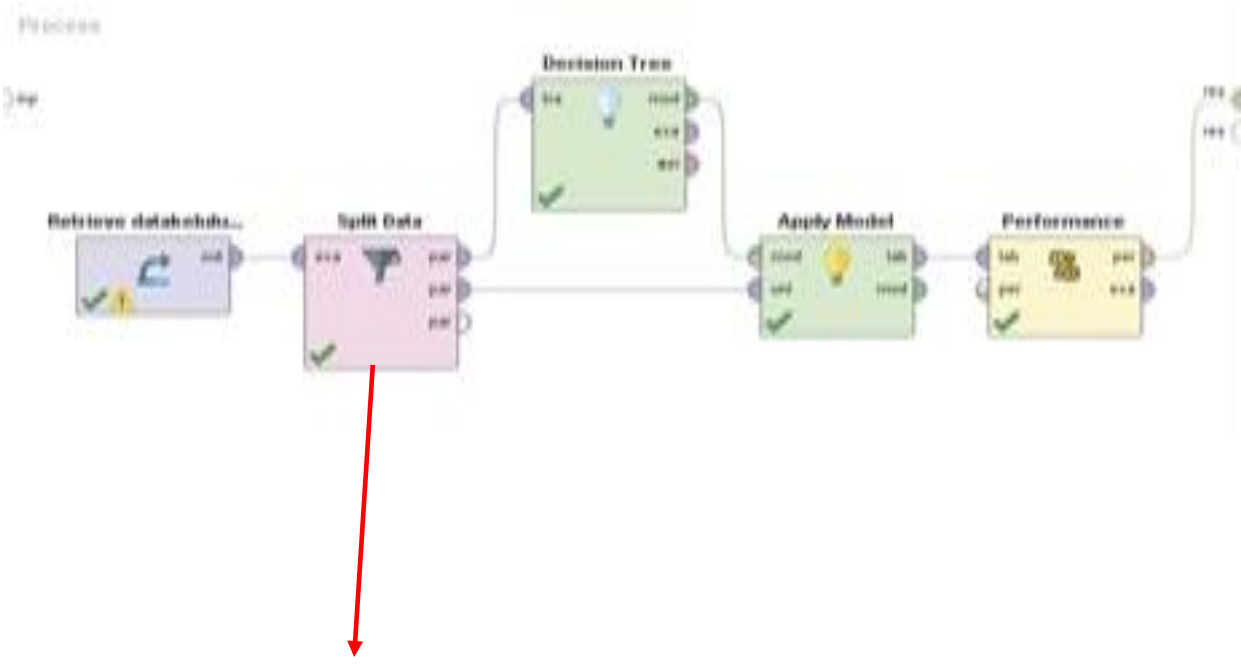
The Split Data operator takes a dataset as its input and delivers the subsets of that dataset through its output ports. The sampling type parameter decides how the examples should be shuffled in the resultant partitions:

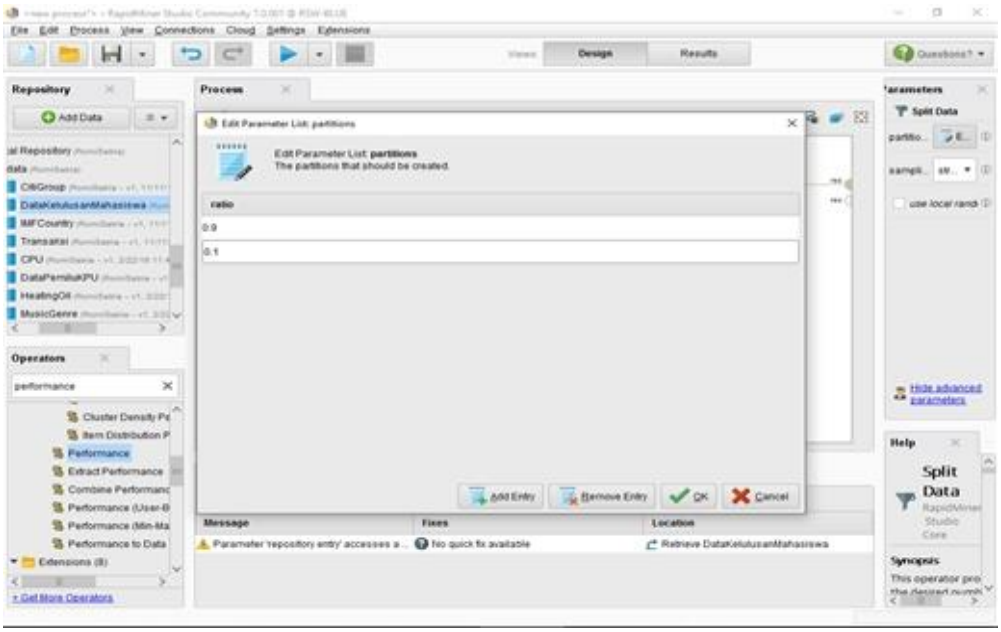
a. Linear sampling: Divides the dataset into partitions without changing the order of the examples
b. Shuffled sampling: Builds random subsets of the dataset
c. Stratified sampling: Builds random subsets and ensures that the class distribution in the subsets is the same as in the whole dataset

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





7.3 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

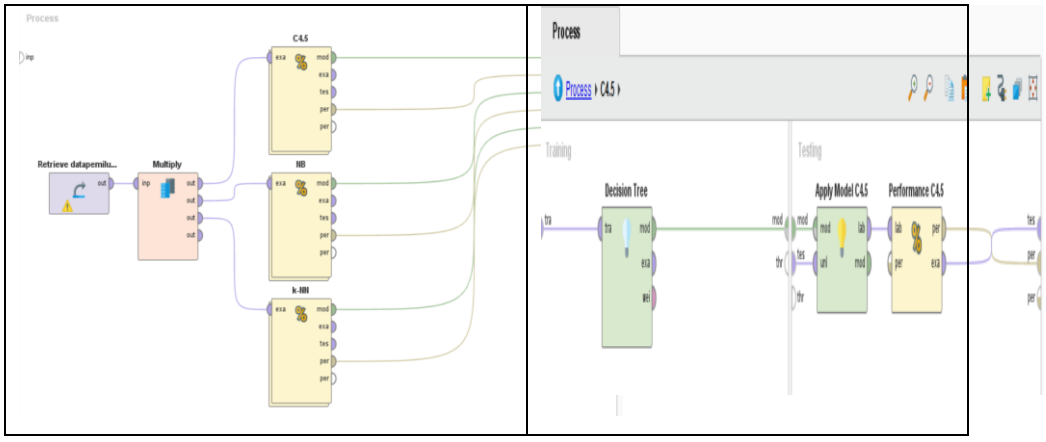
Metode evaluasi standard: stratified 10-fold cross-validation. Mengapa 10? Hasil dari berbagai percobaan yang ekstensif dan pembuktian teoritis, menunjukkan bahwa 10-fold cross-validation adalah pilihan terbaik untuk mendapatkan hasil validasi yang akurat. 10-fold cross-validation akan mengulang pengujian 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%
	Akurasi Rata-Rata	92%

Gambar fold cross-validation

Latihan: Prediksi Elektabilitas Caleg

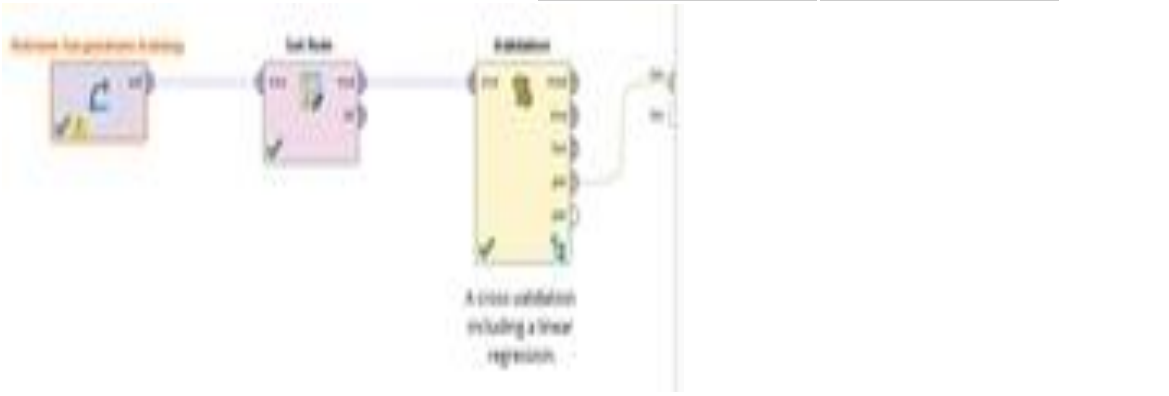
1. Lakukan training pada data pemilu (datapemilukpu.xls)
2. Lakukan pengujian dengan menggunakan 10-fold X Validation
3. Ukur performance-nya dengan confusion matrix dan ROC Curve
4. Lakukan ujicoba, ubah algoritma menjadi C4.5, Naive Bayes, dan k-NN, analisis mana algoritma yang menghasilkan model yang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



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

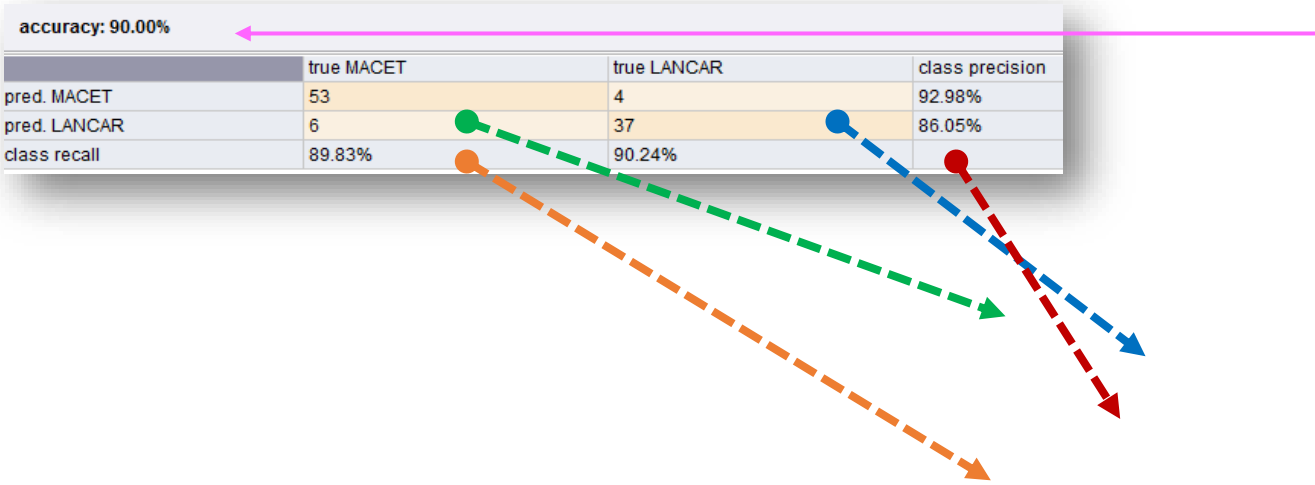
Latihan: Prediksi Harga Saham

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



	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 lancer (**FP**)

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

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

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

Precision and Recall, and F-measures

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

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

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 ß 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

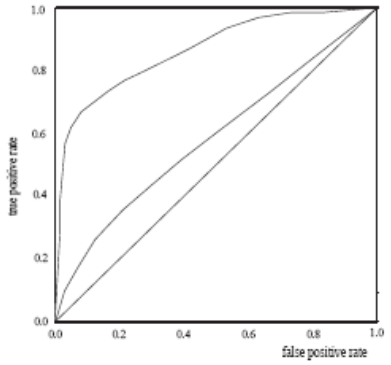
$$PPV = \frac{Number\ of\ 'True\ Positives'}{Number\ of\ 'True\ Positives' + Number\ of\ 'False\ Positives''}$$

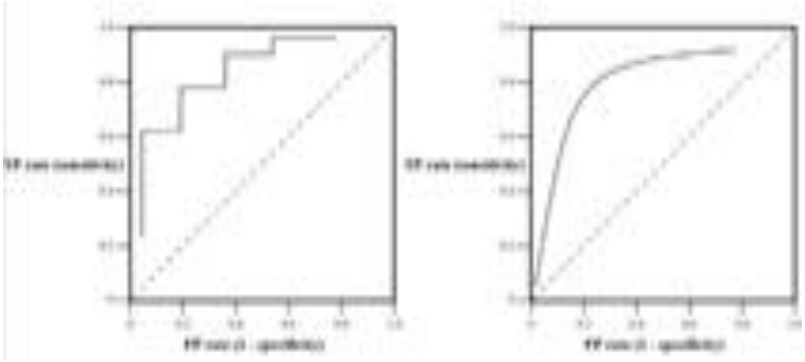
- Negative Predictive Value (NPV) is the proportion of cases with 'negative' test results that are correctly diagnosed

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 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)

Modul 8 Algoritma Klasifikasi

8.1 Pengertian

Klasifikasi adalah Teknik yang digunakan untuk membuat skema yang menunjukkan pengkategorian data dimulai dengan precursor variable (Clarisse, 2018).

Cara kerja dari metode Classification adalah sebuah proses 2 langkah. Langkah pertama dari classification juga disebut sebagai learning of mapping atau function, suatu fungsi pemetaan yang bisa memprediksi class label ypada suatu tuple X. Pemetaan ini direpresentasikan dalam bentuk classification rules, decision tree atau formula matematika. Dari rules atau tree tersebut dapat digunakan untuk mengklasifikasi tuple baru (Han, 2006). Langkah kedua adalah classifier yang sudah dibangun akan digunakan untuk mengklasifikasi data. Pertama, akurasi dari prediksi classifier tersebut diperkirakan. Jika menggunakan training set untuk mengukur akurasi dari classifier, maka estimasi akan optimis karena data yang digunakan untuk membentuk classifier adalah training set juga. Oleh karena itu, digunakan test set, yaitu sekumpulan tuple beserta class label-nya yang dipilih secara acak dari dataset. Test set bersifat independen dari training set dikarenakan test set tidak digunakan untuk membangun classifier (Han, 2006).

Metode Classification termasuk dari "supervise learning" karena class label dari setiap tuple sudah disediakan. Berbeda dengan "unsupervised learning" dimana class label dari setiap tuple tidak diketahui.Metode yang menggunakan unsupervised learning adalah metode Clustering (Han, 2006). Terdapat beberapa algoritma machine learning yang menggunakan metode Classification ini, seperti K Nearest Neighbours, Decision tree (C4.5) dan Naïve Bayes, Neural Network.

8.2 Algoritma Decision Tree

Decision tree merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon (tree) di mana setiap node merepresentasikan atribut, cabangnya merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas. Node yang paling atas dari decision tree disebut sebagai root (Gorunescu, 2011).

Algoritma C4.5 adalah algoritma klasifikasi data dengan teknik pohon keputusan yang memiliki kelebihan-kelebihan. Kelebihan ini misalnya dapat mengolah data numerik (kontinyu) dan diskret, dapat menangani nilai atribut yang hilang, menghasilkan aturan-aturan yang mudah diintrepetasikan dan tercepat diantara algoritma-algoritma yang lain (Luthfi. 2009). Tahapan Algoritma C45 adalah:

1. Pilih atribut sebagai akar.
2. Buat cabang untuk tiap-tiap nilai.
3. Bagi kasus dalam cabang.
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

8.3 Latihan Algoritma C45 Mengunakan Rapid miner

- Lakukan eksperimen mengikuti buku Matthew North, Data Mining for the Masses 2nd Edition, 2016, Chapter 10 (Decision Tree), p 195-217
- Datasets:
    - eReaderAdoption-Training.csv
    - eReaderAdoption-Scoring.csv
- Analisis peran metode prunning pada decision tree dan hubungannya dengan nilai confidence
- Analisis jenis decision tree apa saja yang digunakan dan mengapa perlu dilakukan pada dataset tersebut

8.4 Bayesian Classification

Proses klasifikasi menggunakan metode probabilitas dan statistika sebagaimana yang telah dikenalkan oleh Thomas Bayes. Menurut Olson dan Delen (2008, p102) menjelaskan Naïve Bayes untuk setiap kelas keputusan, menghitung probabilitas dengan syarat bahwa kelas keputusan adalah benar, mengingat vektor informasi obyek. Algoritma ini mengasumsikan bahwa atribut obyek adalah independen. Probabilitas yang terlibat dalam memproduksi perkiraan akhir dihitung sebagai jumlah frekuensi dari "master" tabel keputusan.

Naïve Bayes Classifier atau bisa disebut sebagai multinomial naïve bayes merupakan model penyederhanaan dari algoritma bayes yang cocok dalam pengklasifikasian text atau dokumen. Tahapan Bayesian classifier adalah:

1. Step 1 : Hitung P(vj)Π P(wk | vj) untuk setiap kategori.
2. Step 2 : Tentukan kategori dengan nilai P(vj)Π P(wk | vj) maksimal

Dimana untuk:

- P(vj) : Probabilitas setiap dokumen terhadap sekumpulan dokumen.
- P(wk|vj) : Probabilitas kemunculan kata wk pada suatu dokumen dengan kategori class vj.

8.5 Latihan Bayesian Classification

Latihan: Prediksi Kelulusan Mahasiswa menggunakan naïve Bayesian classifier, dengan tahapan:

- Lakukan training pada data mahasiswa (datakelulusanmahasiswa.xls).
- Lakukan pengujian dengan menggunakan 10-fold X Validation
- Uji beda dengan t-Test untuk mendapatkan model terbaik

8.6 Neural Network

Neural network (NN) adalah suatu model matematis yang meniru sistem kerja dari jaringan neural otak manusia. NN telah dikenal sukses untuk solusi dari suatu problem dengan data yang bersifat kompleks (uncertainty), memiliki noise atau incomplete

Neural Netwok mampu menyelesaikan:

- Permasalahan yang tidak terstruktur dan sulit didefinisikan, dapat belajar dari pengalaman
- Mampu memilih suatu input data ke dalam kategori tertentu yang sudah ditetapkan (klasifikasi)
- Mampu menggambarkan suatu obyek secara keseluruhan walau dengan data yang tidak lengkap
- Mempunyai kemampuan mengolah data-data input tanpa harus mempunyai target (Self organizing),
- Mampu menemukan suatu jawaban terbaik sehingga mampu meminimalisasi fungsi biaya (optimasi).

langkah-langkah penting yang dilakukan sebagai berikut :

- Menentukan fitur-fitur atau ciri yang dapat memudahkan klasifikasi kelaskelas, yang akan dijadikan input. Pengetahuan akan fitur kelas-kelas, akan membantu dalam klasifikasi.
- Mengumpulkan data pelatihan yang representatif dengan jumlah yang memadai bila memungkinkan
- Pemilihan model jaringan, dan algoritme pelatihan. Setiap model arsitektur dan algoritme memiliki kelebihan dan kekurangan dalam hal waktu dan akurasi, sehingga pemilihan model neural network bergantung pada faktor mana yang lebih diprioritaskan.

8.7 Latihan Neural Network

Latihan Neural Network seperti tahapan dibawah ini:

- Lakukan training dengan neural network untuk dataset TeamValue Training.csv
- Gunakan 10-fold cross validation
- Lakukan adjusment terhadap hidden layer dan neuron size, misal: hidden layer saya buat 3, neuron size masing-masing 5
- Apa yang terjadi, apakah ada peningkatan akurasi?

	NN	NN (HL 2, NS 3)	NN (HL 2, NS 5)	NN (HL 3, NS 3)	NN (HL 3, NS 5)	NN (HL 4, NS 3)	NN (HL 4, NS 5)
Accuracy							

Tentukan Hidden Layer

Hidden Layer	Capabilities
0	Only capable of representing linear separable functions or decisions
1	Can approximate any function that contains a continuous mapping from one finite space to another
2	Can represent an arbitrary decision boundary to arbitrary accuracy with rational activation functions and can approximate any smooth mapping to any accuracy

Penentuan Neuron Size

1. Trial and Error

2. Rule of Thumb:
    - Between the size of the input layer and the size of the output layer
    - 2/3 the size of the input layer, plus the size of the output layer
    - Less than twice the size of the input layer
3. Search Algorithm:
    - Greedy
    - Genetic Algorithm
    - Particle Swarm Optimization
    - etc

Modul 9. Algoritma Klastering

9.1 Pengertian Clustering

Suatu cluster merupakan sekelompok entitas yang memiliki kesamaan dan memiliki perbedaan dengan entitas dari kelompok lain(Everitt,1980).

Algoritma Clustering bekerja dengan mengelompokkan obyek-obyek data (pola, entitas, kejadian, unit,hasil observasi) ke dalam sejumlah cluster tertentu (Xu and Wunsch,2009). Dengan kata lain algoritma Clustering melakukan pemisahan/ pemecahan/ segmentasi data ke dalam sejumlah kelompok (cluster) menurut karakteristik tertentu

Tujuan clustering (pengelompokan) data dapat dibedakan menjadi dua, yaitu pengelompokan untuk pemahaman dan clustering untuk penggunaan (Prasetyo,2012).

9.2 K-Mean Klustering

K-Means yaitu algoritma yang dimulai dengan menetapkan nilai pusat terlebih dahulu untuk menjadi pusat sementara dari centroid atau cluster, kemudian dengan menggunakan rumus menghitung jarak setiap data ke pusat sehingga data yang lebih dekat ke pusat menjadi satu kelompok dan data yang jauh menjadi kelompok lainnya.

Tahapan Algoritma K-mean clustering sebagai berikut:

1. Tentukan jumlah cluster
2. Alokasikan data ke dalam kelompok secara acak
3. Hitung pusat cluster (centroid) menggunakan mean utk masing-masing kelompok
4. Alokasikan masing-masing data ke centroid terdekat

5. Kembali ke langkah 3, jika masih ada data yang berpindah cluster atau jika nilai centroid diatas nilai ambang, atau jika nilai pada fungsi obyektif yang digunakan masih diatas ambang

Latihan

- Lakukan eksperimen mengikuti buku Matthew North, Data Mining for the Masses, 2012, Chapter 6 k-Means Clustering, pp. 91-103 (CoronaryHeartDisease.csv)
- Gambarkan grafik (chart) dan pilih Scatter 3D Color untuk menggambarkan data hasil klastering yang telah dilakukan
- Lakukan pengukuran performance dengan menggunakan Cluster Distance Performance, untuk mendapatkan nilai Davies Bouldin Index (DBI)
- Nilai DBI semakin rendah berarti cluster yang kita bentuk semakin baik

Modul 10 Algoritma Asosiasi

10.1 Pengertian

Analisis asosiasi atau association rule mining adalah teknik data mining untuk menemukan aturan assosiatif antara suatu kombinasi item. Contoh aturan assosiatif dari analisa pembelian di suatu pasar swalayan adalah dapat diketahuinya berapa besar kemungkinan seorang pelanggan membeli roti bersamaan dengan susu. Dengan pengetahuan tersebut pemilik pasar swalayan dapat mengatur penempatan barangnya atau merancang kampanye pemasaran dengan memakai kupon diskon untuk kombinasi barang tertentu.

Analisis asosiasi menjadi terkenal karena aplikasinya untuk menganalisa isi keranjang belanja di pasar swalayan. Analisis asosiasi juga sering disebut dengan istilah market basket analysis.

Analisis asosiasi dikenal juga sebagai salah satu teknik data mining yang menjadi dasar dari berbagai teknik data mining lainnya. Khususnya salah satu tahap dari analisis asosiasi yang disebut analisis pola frequensi tinggi (frequent pattern mining) menarik perhatian banyak peneliti untuk menghasilkan algoritma yang efisien.

Penting tidaknya suatu aturan assosiatif dapat diketahui dengan dua parameter, support (nilai penunjang) yaitu persentase kombinasi item tersebut dalam database dan confidence (nilai kepastian) yaitu kuatnya hubungan antar item dalam aturan assosiatif. Aturan assosiatif biasanya dinyatakan dalam bentuk : {roti, mentega} → {susu} (support = 40%, confidence = 50%)

Yang artinya : "50% dari transaksi di database yang memuat item roti dan mentega juga memuat item susu. Sedangkan 40% dari seluruh transaksi yang ada di database memuat ketiga item itu." Dapat juga diartikan : "Seorang konsumen yang membeli roti dan mentega punya kemungkinan 50% untuk juga membeli susu. Aturan ini cukup signifikan karena mewakili 40% dari catatan transaksi selama ini." Analisis asosiasi didefinisikan suatu proses untuk menemukan semua aturan assosiatif yang memenuhi syarat minimum untuk support (minimum support) dan syarat minimum untuk confidence (minimum confidence).

Metodologi dasar analisis asosiasi terbagi menjadi dua tahap :

- Analisa pola frekuensi tinggi

  Tahap ini mencari kombinasi item yang memenuhi syarat minimum dari nilai support dalam database. Nilai support sebuah item diperoleh dengan rumus berikut:

$$\text{Support (A)} = \frac{\text{Jumlah Transaksi mengandung A}}{\text{Total Transaksi}}$$

sedangkan nilai support dari 2 item diperoleh dari rumus berikut:

$$\text{Support } (A \cap B) = \frac{\text{Jumlah Transaksi mengandung A dan B}}{\text{Total Transaksi}}$$

- Pembentukan aturan assosiatif

  Setelah semua pola frekuensi tinggi ditemukan, barulah dicari aturan assosiatif yang memenuhi syarat minimum untuk confidence dengan menghitung confidence aturan assosiatif A →B Nilai confidence dari aturan A →B diperoleh dari rumus berikut:

$$\text{Confidence} = P(B \mid A) = \frac{\text{Jumlah Transaksi mengandung A dan B}}{\text{Jumlah Transaksi mengandung A}}$$

10.2 Latihan

- Lakukan eksperimen mengikuti buku Matthew North, Data Mining for the Masses 2nd Edition, 2016, Chapter 5 (Association Rules), p 85-97

Modul 11. Algoritma Estimasi

11.1 Pengertian

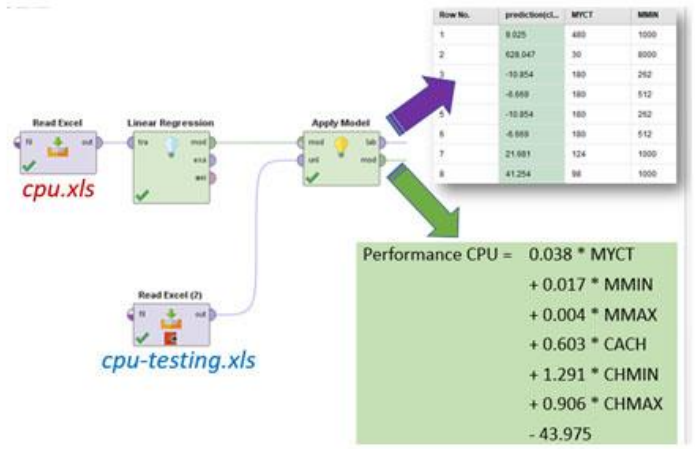
Algoritma estimasi dalam Data Mining – Algoritma Estimasi merupakan metode paling tepat untuk menyelesaikan yang berkaitan dengan memperkirakan seberapa banyak hasil produksi.

Estimasi adalah suatu metode dimana kita dapat memperkirakan nilai Populasi dengan memakai nilai sampel. Estimasi biasanya diperlukan untuk mendukung keputusan yang baik,menjadwalkan pekerjaan, menentukan berapa lama proyek perlu dilakukan dan berapa biayanya, menentukan apakah proyek layak dikerjakan, mengembangkan kebutuhan arus kas, menentukan seberapa baik kemajuan proyek, menyusun anggaran time phased dan menetapkan baseline proyek (Prasetyo, 2014). Salah satu algoritma yang dapat memodelkan persamaan untuk menghitung estimasi yakni Algoritma Linear Regression.

11.2 Latihan: Estimasi Performance CPU

1. Lakukan training pada data CPU (cpu.xls) dengan menggunakan algoritma linear regression
2. Lakukan pengujian terhadap data baru (cpu-testing.xls), untuk model yang dihasilkan dari tahapan 1. Data baru berisi 10 setting konfigurasi, yang belum diketahui berapa performancenya
3. Amati hasil estimasi performance dari 10 setting konfigurasi di atas

Berikut Hasil Estimasi Performace cpu-testing.xls dapat dilihat pada gambar 13.1.

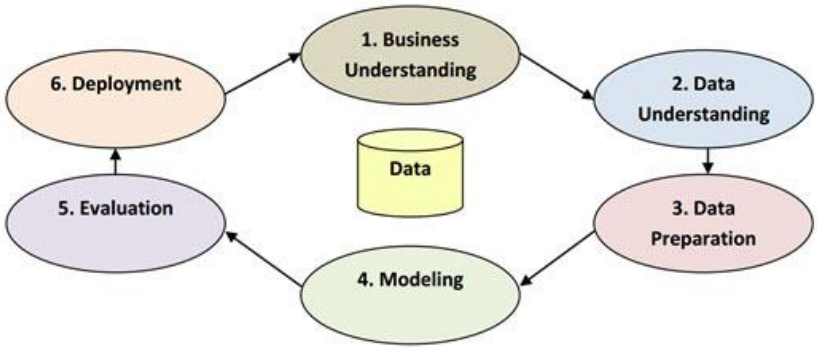


Gambar Hasil Estimasi

11.3 CRISP-DM

Awalnya dibangun oleh 3 perusahaan yaitu SPSS, Daimler Chrysler, dan NCR yang kemudian dikembangan ratusan organisasi dan perusahaan secara bersama-sama. CRISP-DM merupakan singkatan dari Cross-Industry Standard Process for Data Mining dan memiliki 6 tahapan yaitu Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment.

Gambar 13.2 adalah tahapan CRISP_DM.



Gambar 13. Tahapan CRISP_DM.

11.4 Penerapan Crisp-DM

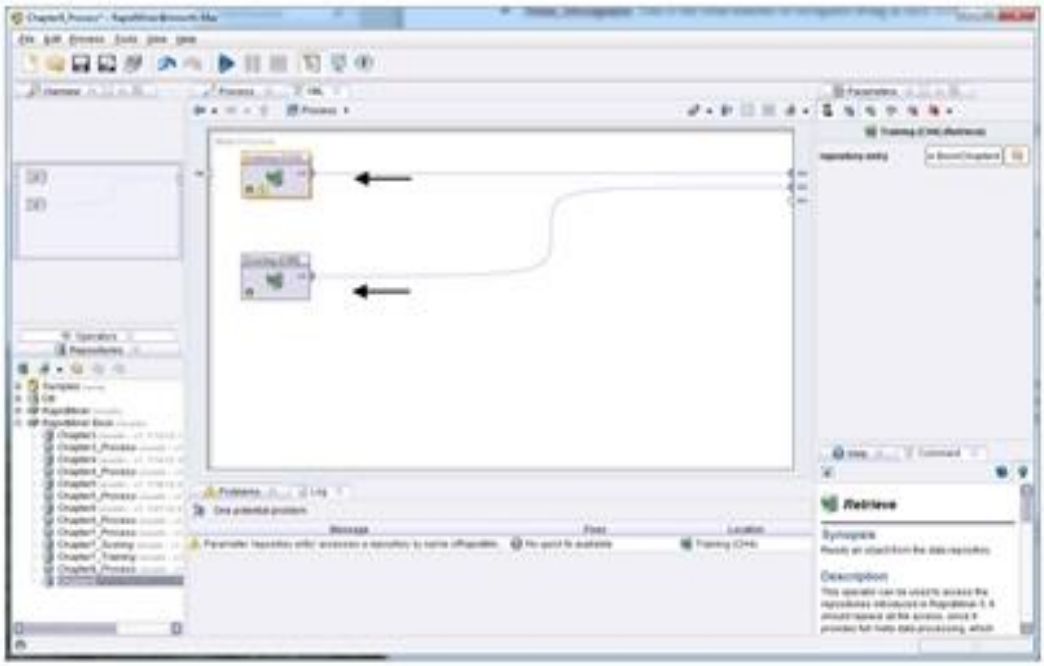
1. **Business Understanding**
   - Sarah's new data mining objective is pretty clear: she wants to anticipate demand for a consumable product
   - We will use a linear regression model to help her with her desired predictions
   - She has data, 1,218 observations that give an attribute profile for each home, along with those homes' annual heating oil consumption
   - She wants to use this data set as training data to predict the usage that 42,650 new clients will bring to her company
   - She knows that these new clients' homes are similar in nature to her existing client base, so the existing customers' usage behavior should serve as a solid gauge for predicting future usage by new customers.
2. **Data Understanding**

   We create a data set comprised of the following attributes:
   - Insulation: This is a density rating, ranging from one to ten, indicating the thickness of each home's insulation. A home with a density rating of one is poorly insulated, while a home with a density of ten has excellent insulation
   - Temperature: This is the average outdoor ambient temperature at each home for the most recent year, measure in degree Fahrenheit

- Heating_Oil: This is the total number of units of heating oil purchased by the owner of each home in the most recent year
- Num_Occupants: This is the total number of occupants living in each home
- Avg_Age: This is the average age of those occupants
- Home_Size: This is a rating, on a scale of one to eight, of the home's overall size. The higher the number, the larger the home

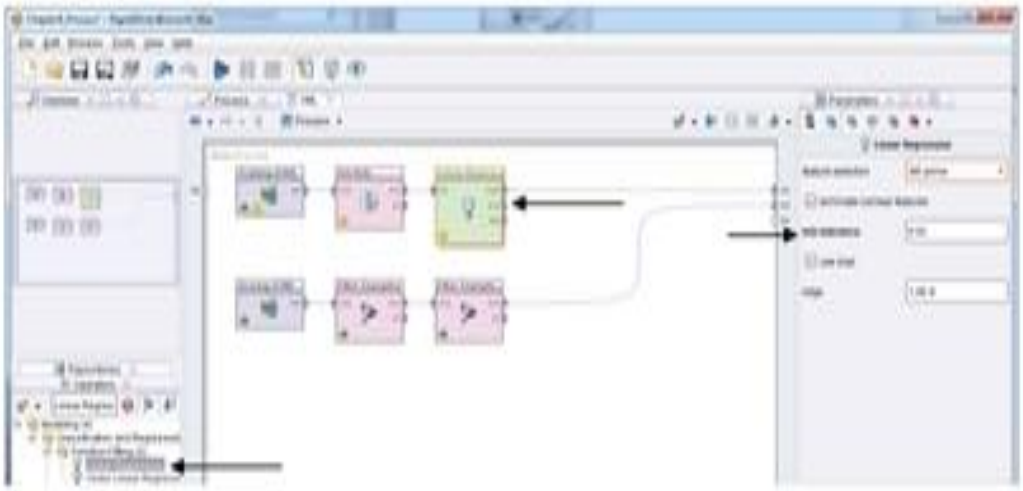
3. Data Preparation

A CSV data set for this chapter's example is available for download at the book's companion web site (https://sites.google.com/site/dataminingforthemasses/)

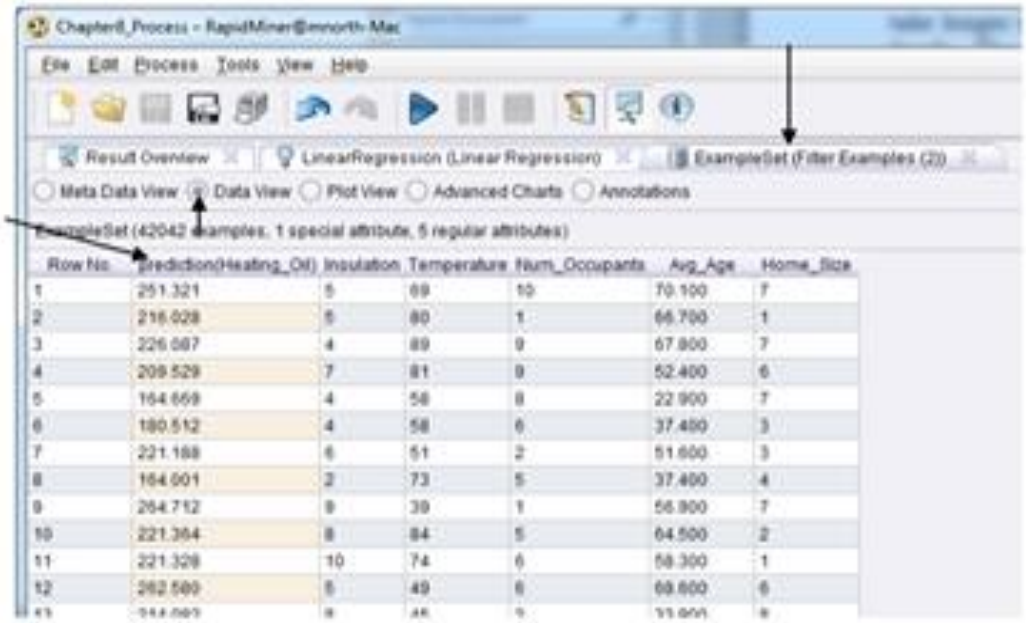


Gambar. Tahap data preparation

4. Modeling

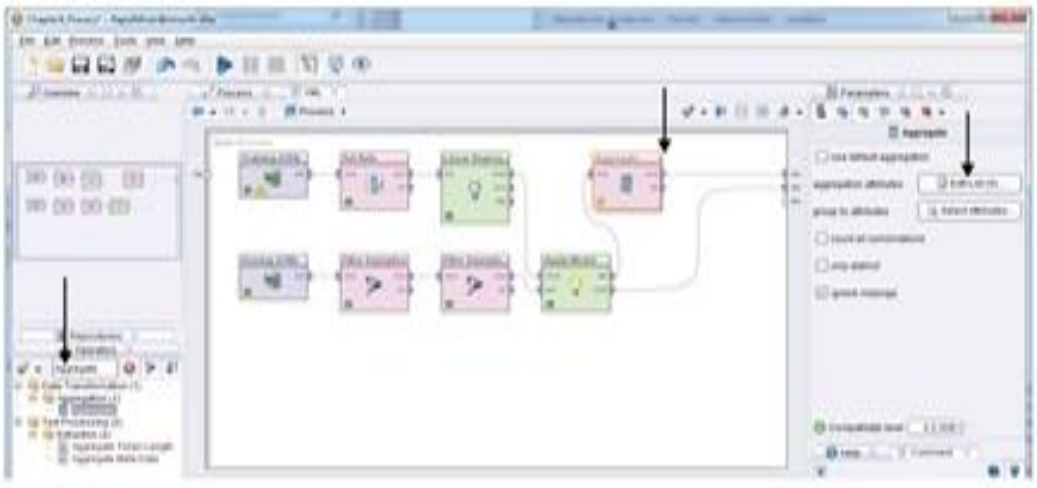


5. Evaluation



Row No.	prediction(Heating_Oil)	Insulation	Temperature	Num_Occupants	Avg_Age	Home_Size
1	251.321	5	69	10	70.100	7
2	216.028	5	80	1	66.700	1
3	226.087	4	69	9	67.900	7
4	209.529	7	81	9	52.400	6
5	164.659	4	58	8	22.900	7
6	180.512	4	58	6	37.480	3
7	221.188	6	51	2	51.600	3
8	164.001	2	73	5	37.400	4
9	264.712	9	39	1	56.900	7
10	221.364	8	84	5	64.500	2
11	221.328	10	74	6	58.300	1
12	262.580	5	49	6	69.600	6

6. Deployment



Modul 12 Algoritma Forecasting

12.1 Pengertian

Time Series Forecasting adalah salah satu teknik analisis prediktif paling tua yang diketahui. Teknik ini telah ada dan telah digunakan secara luas bahkan sebelum istilah "analytics prediktif

Algoritma forecasting (prediksi) sama dengan algoritma estimasi di mana label/target/class bertipe numerik, bedanya adalah data yang digunakan merupakan data rentet waktu (data time series)

Variabel independen atau prediktor tidak sepenuhnya diperlukan untuk Time Series Forecasting univariat, tetapi sangat disarankan untuk deret waktu multivarian. Metode Time Series Forecasting yakni Metode Data Driven dan Metode Model Driven:

- Metode Data Driven: Tidak ada perbedaan antara prediktor dan target. Teknik seperti rata-rata deret waktu atau perataan dianggap pendekatan berbasis data untuk peramalan deret waktu
- Model Driven Method: Mirip dengan model prediksi "konvensional", yang memiliki variabel independen dan dependen, tetapi dengan twist: variabel independen sekarang waktunya

12.2 Implementasi Algoritma Forecasting

- Pendekatan RapidMiner terhadap deret waktu didasarkan pada dua proses transformasi data utama.
- Yang pertama adalah windowing untuk mengubah data deret waktu menjadi kumpulan data generik:
- Langkah ini akan mengubah baris terakhir dari suatu jendela dalam rangkaian waktu menjadi label atau variabel target. Kita menerapkan salah satu dari "pelajar" atau algoritma untuk memprediksi variabel target dan dengan demikian memprediksi langkah waktu berikutnya dalam seri.

12.3 Latihan 1

- Lakukan training dengan menggunakan linear regression pada dataset hargasaham-training.xls
- Gunakan Split Data untuk memisahkan dataset di atas, 90% training dan 10% untuk testing

12.4 Latihan 2

- Lakukan training dengan menggunakan NN pada dataset hargasaham-training.xls
- Terapkan model yang dihasilkan untuk data hargasaham-testing.xls

DAFTAR PUSTAKA

1. Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques Third Edition, Elsevier, 2012
2. Ian H. Witten, Frank Eibe, Mark A. Hall, Data mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Elsevier, 2011
3. Daniel T. Larose, Discovering Knowledge in Data: an Introduction to Data Mining, John Wiley & Sons, 2005
4. Matthew North, Data Mining for the Masses, 2012
5. Daniel T. Larose, Discovering Knowledge in Data An Introduction to Data Mining 2nd ed., John Wiley & Sons , 2014
6. Charu C. Aggarwal , Data Mining: The Textbook, Springer, 2015
7. Nong Ye, Data Mining Theories, Algorithms, and Examples, CRC Press, 2014
8. Vijay Kotu, Bala Deshpande , Predictive Analytics and Data Mining Concepts and Practice with RapidMiner, 1st ed. Elsevier,2015