

Data Mining



Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion



Session 01

PROGRAM STUDI TEKNIK INFORMATIKA

Sumber: <https://ft.uhamka.ac.id/prodi/informatika/>

- **VISI.**

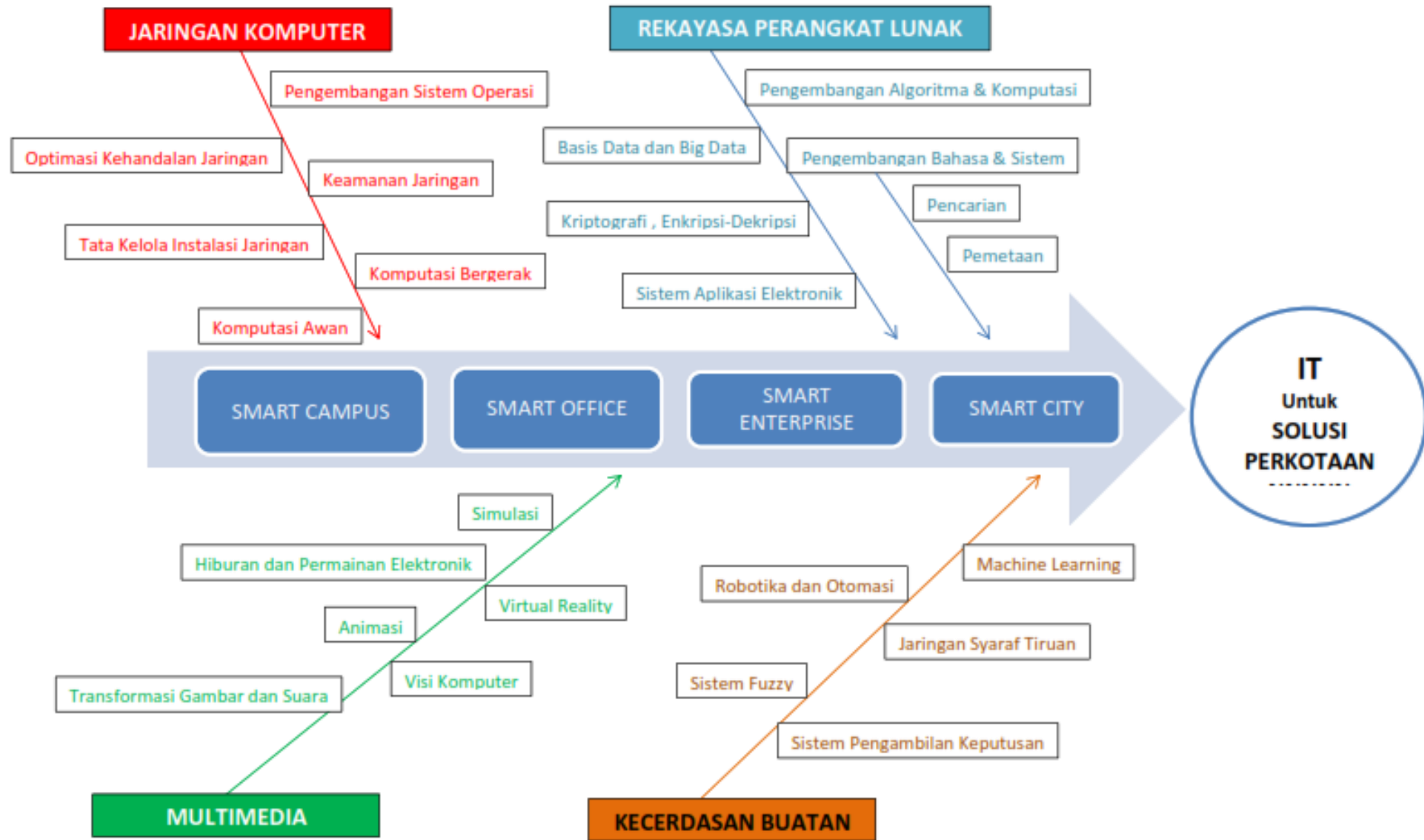
- √ Menjadi program studi informatika yang menghasilkan karya unggul di bidang teknologi informasi melalui pembelajaran berlandaskan prophetic teaching untuk membentuk lulusan yang cerdas secara spiritual, intelektual, emosional, dan sosial.

- **MISI.**

1. Melaksanakan pendidikan di bidang informatika yang terintegrasi dengan nilai-nilai Al Islam Kemuhammadiyahan.
2. Melaksanakan penelitian untuk menghasilkan karya teknologi informasi yang unggul untuk masyarakat berkembang
3. Melaksanakan pengabdian kepada masyarakat untuk kesejahteraan umat.
4. Menghasilkan mahasiswa yang berwawasan global dalam teknologi informasi.
5. Mendorong jiwa kewirausahaan mahasiswa untuk berkiprah di tengah masyarakat dalam meningkatkan kegiatan ekonomi.
6. Menerapkan standar mutu pendidikan informatika melalui layanan mutu yang berkesinambungan.
7. Membangun kerja sama terpadu dengan mitra yang berkontribusi untuk kemajuan pendidikan di bidang informatika.



Road Map TI UHAMKA





Sistem Penilaian

- SKS : 3
 - Teori 3 SKS – (14 Kali Pertemuan)
- Keaktifan (bobot **10%**)
 - Terdiri dari: Aktivitas Perkuliahan (*Keaktifan, Etika, dan Sikap*).
- Project (bobot **25%**)
 - Terdiri dari: Tugas Besar / Project (*Penguasaan Keterampilan*) dengan cara observasi, mengimplementasikan dan menghasilkan Proyek Akhir, serta mempresentasikannya.
- Tes (UTS = bobot **25%** , UAS = bobot **40%**)
 - Terdiri dari: UTS dan UAS (*Penguasaan Pengetahuan*) dengan cara tes tertulis maupun tes praktikum.



Tujuan/Objectives (Session.01)

- Objectives (Tujuan Instruksional Umum).
 - RPS/SAP.
 - Pengantar & Silabus.
- Deskripsi Singkat
 - Matakuliah ini membekali mahasiswa kemampuan untuk memahami teknik data mining, aplikasinya dalam kehidupan sehari-hari, serta mampu merancang gagasan riset tentang data mining.
- Indikator Kompetensi.
 - Mahasiswa dapat menjelaskan teknik-teknik merepresentasikan data.
 - Mahasiswa memahami tentang jenis dan kualitas data lalu memahami konsep dan fungsi preposisi data dalam data mining.
 - Mahasiswa memahami teknik klasifikasi, teknik asosiasi, dan teknik klustering dalam data mining.
 - Mahasiswa memahami konsep Anomali Data.
 - Mahasiswa memahami konsep spasial dan multimedia data mining.
 - Mahasiswa memahami konsep web dan text mining.



اقْرَأْ بِاسْمِ رَبِّكَ الَّذِي خَلَقَ
خَلَقَ الْإِنْسَانَ مِنْ عَلَقٍ اقْرَأْ وَرَبُّكَ الْأَكْرَمُ
الَّذِي عَلَّمَ بِالْقَلَمِ عَلَّمَ الْإِنْسَانَ مَا لَمْ يَعْلَمْ

- Qs. Al-Alaq, ayat 1 s/d 5
 - Bacalah dengan (menyebut) nama Tuhanmu yang menciptakan,
 - Dia telah menciptakan manusia dari segumpal darah.
 - Bacalah, dan Tuhanmulah Yang Mahamulia,
 - Yang mengajar (manusia) dengan pena.
 - Dia mengajarkan manusia apa yang tidak diketahuinya.

Data Mining



Uhamka

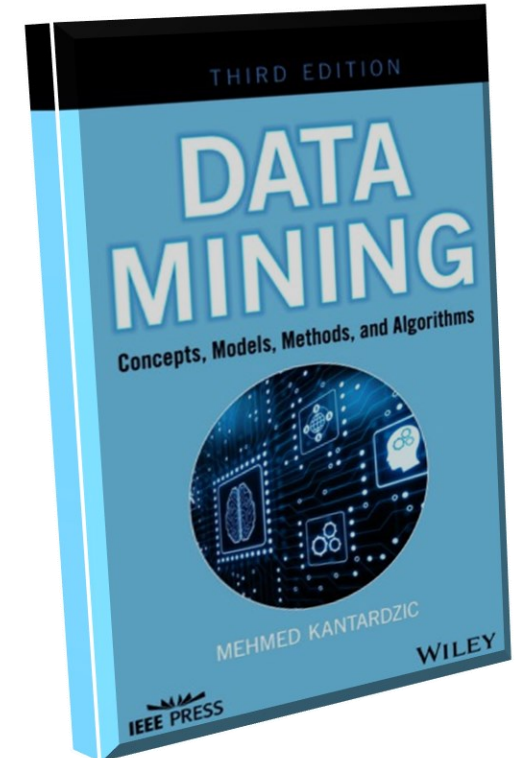
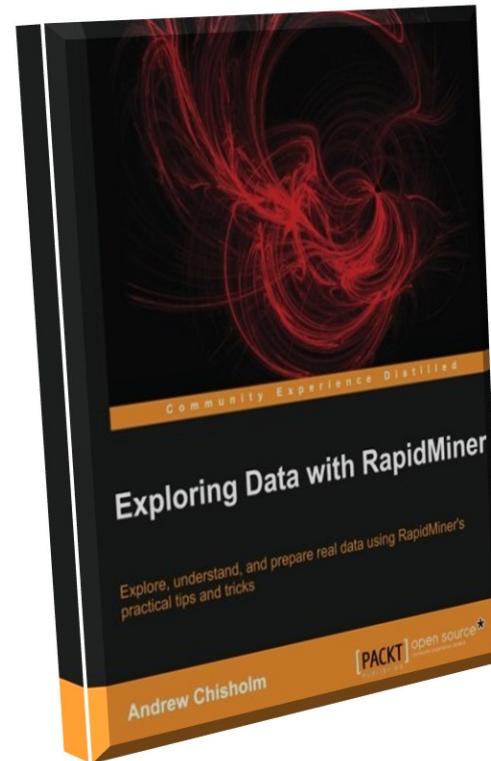
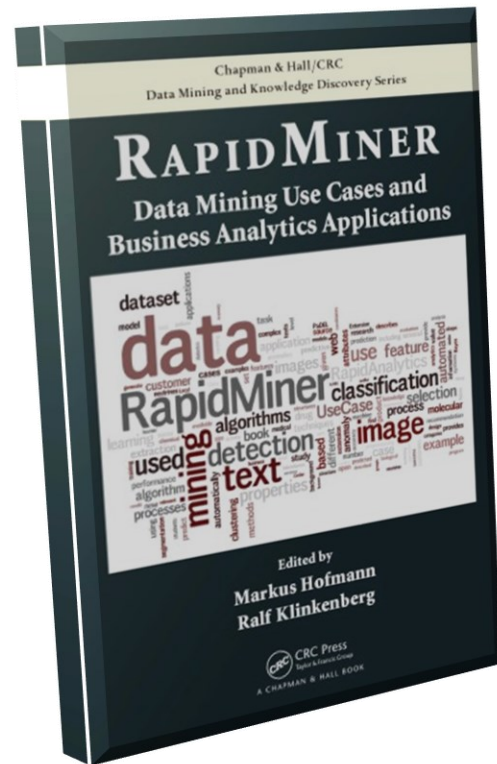
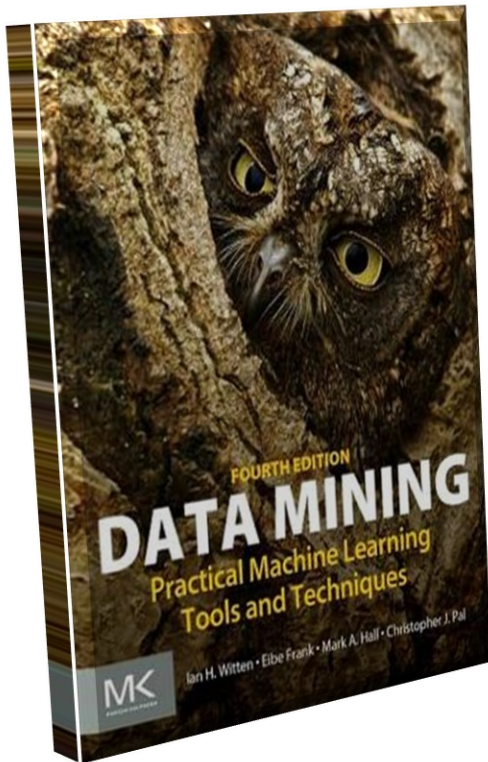
UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion



Session 01

Pengantar & Silabus



RPS/SAP;

Rencana Pembelajaran Semester / Satuan Acara Perkuliahan.



Session 01

- Perkenalan dan Silabus.
- Buku dan Referensi.

Session 03

- Eksplorasi Data

Session 05

- Klasifikasi Data 1

Session 07

- Asosiasi Data 1

Session 02

- Pengantar Mata Kuliah
- Pengenalan Data Mining

Session 04

- Kualitas Data
- Teknik Preposisi

Session 06

- Klasifikasi Data 2

Session 08

- U.T.S.



Session 09

- Asosiasi Data 2

Session 11

- Klustering 2

Session 13

- Spasial
- Multimedia Data Mining

Session 15

- Presentasi

Session 10

- Klustering 1

Session 12

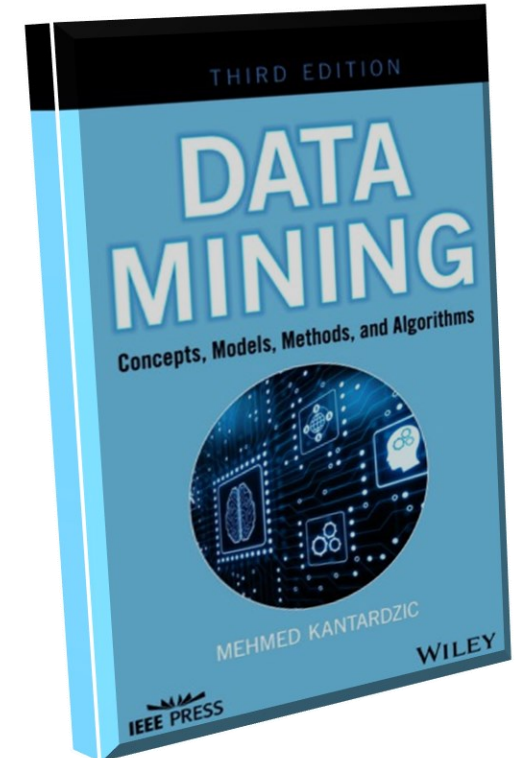
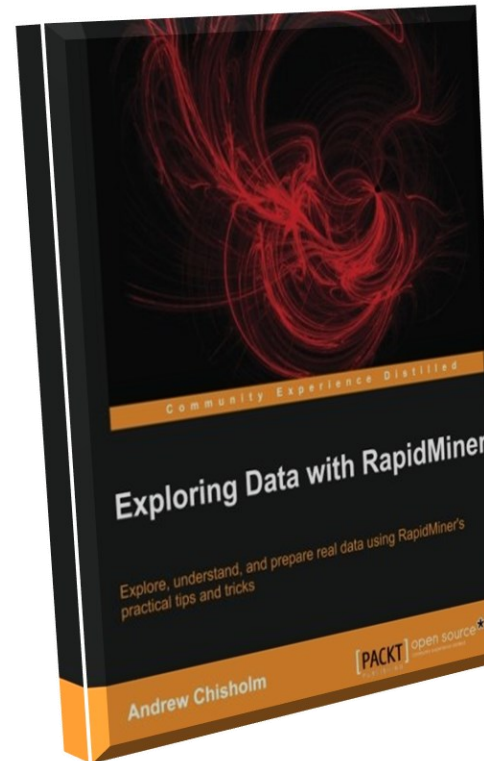
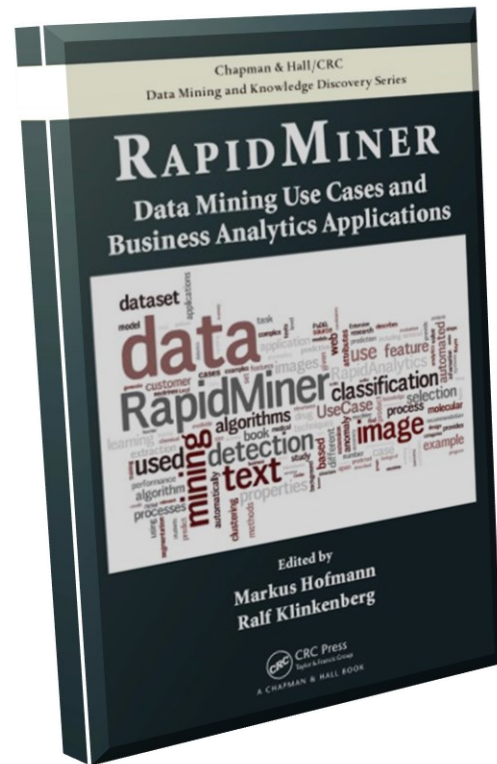
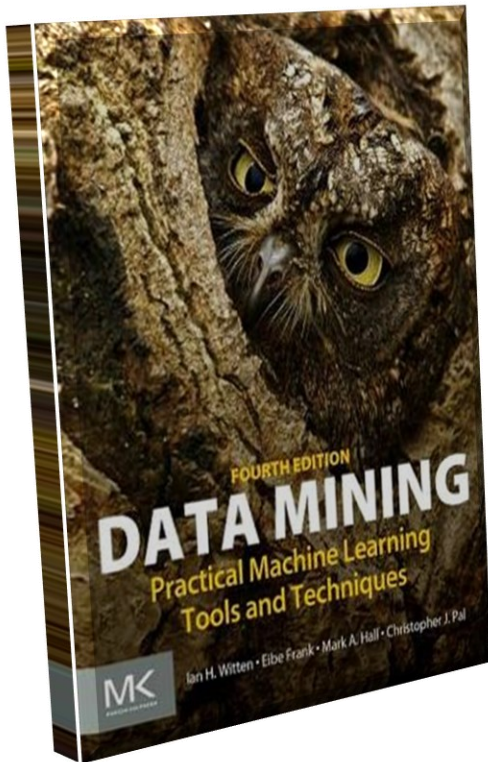
- Anomali Data

Session 14

- Analisis Sentimen dan Text Mining

Session 16

- U.A.S.



Pendahuluan; Referensi dan Buku-Buku



Uhamka
UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA
integrity, trust, compassion



YouTube



zoom



Uhamka

TEKNIK
INFORMATIKA
FAKULTAS TEKNIK

Online Learning UHAMKA

Integrity, Trust, Compassion

**ONLINE
LEARNING**



Uhamka
UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

Tools



Google Classroom



Mentimeter



rapidminer

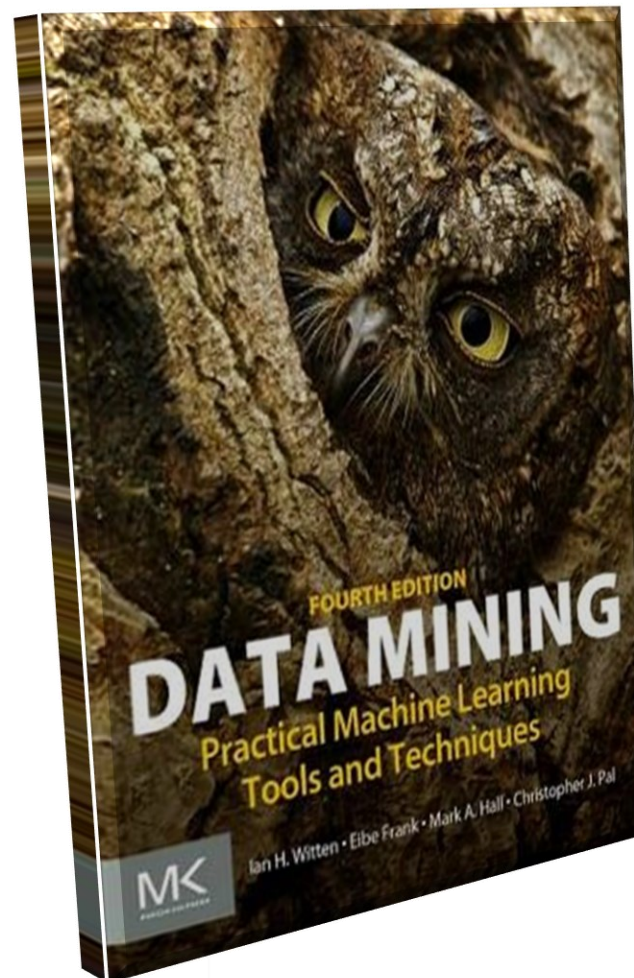
Google 
FORMS

Kahoot!

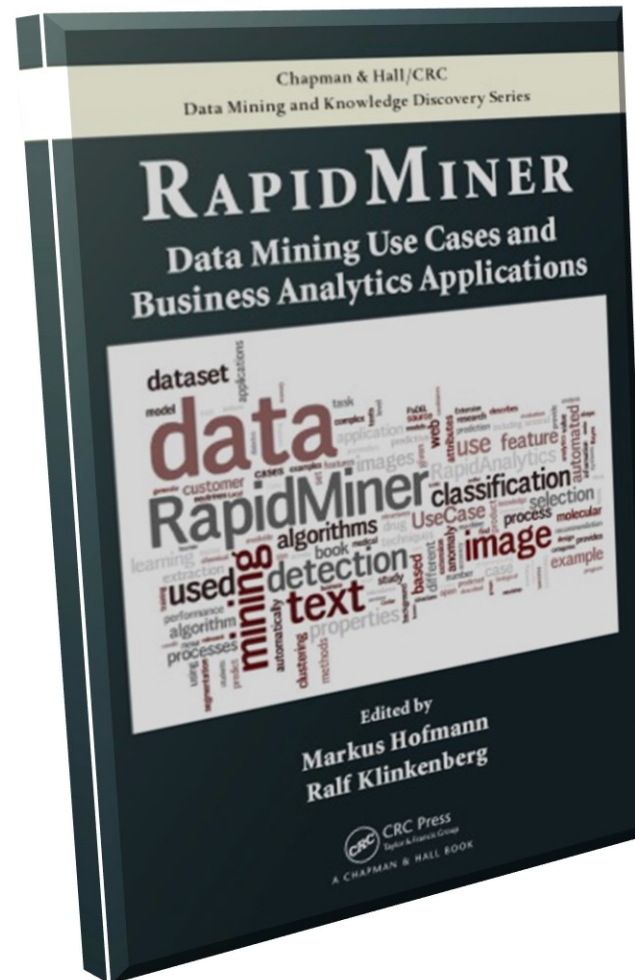


Buku-Buku

- Ian H Witten, Eibe Frank, Mark A Hall. (2016). *Data Mining - Practical Machine Learning Tools and Techniques*, 4th Edition. Morgan-Kaufman.



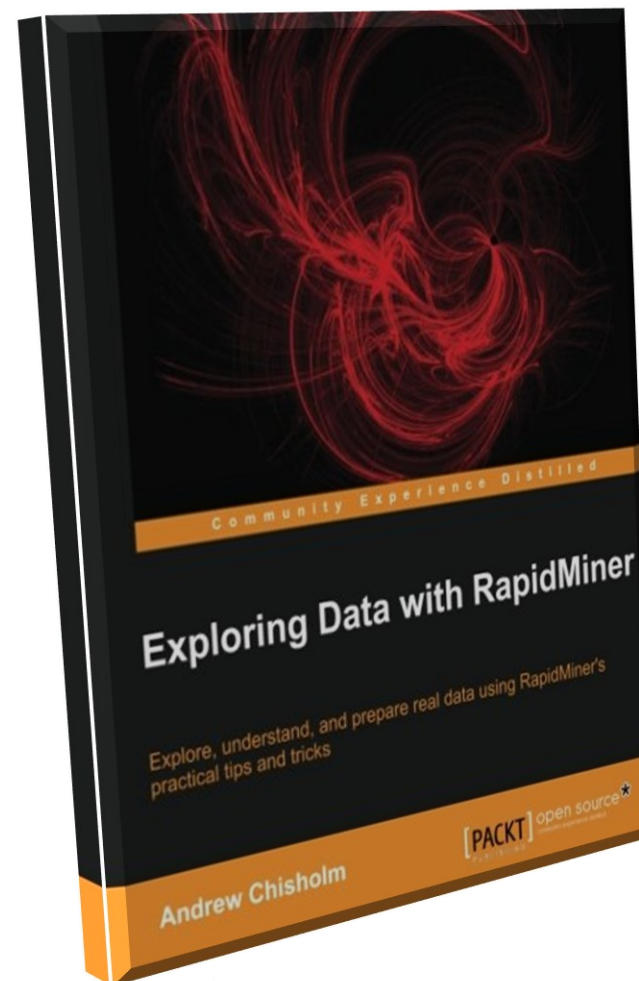
- Hofmann, Markus., Klinkenberg, Ralf. (2016). *Data Mining Use Cases and Business Analytics Application*. CRC Press.





Buku-Buku

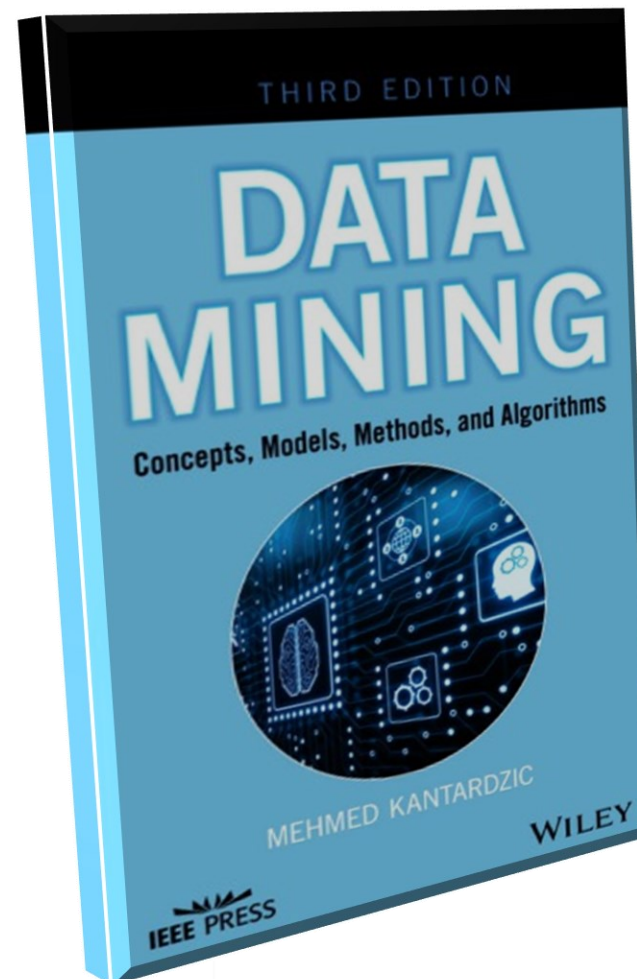
- Crisholm, Andrew. (2013). *Exploring Data With Rapidminer*. Packt Publishing.





Buku-Buku

- Mehmed Kantardzic. (2020). Data Mining - Concepts, Models, Methods, and Algorithms, 3rd edition. Wiley - IEEE Press.





Referensi

- Ian H Witten, Eibe Frank, Mark A Hall. (2016). *Data Mining - Practical Machine Learning Tools and Techniques*, 4th Edition. Morgan-Kaufman.
- Hofmann, Markus., Klinkenberg, Ralf. (2016). *Data Mining Use Cases and Business Analytics Application*. CRC Press.
- Crisholm, Andrew. (2013). *Exploring Data With Rapidminer*. Packt Publishing.
- Mehmed Kantardzic. (2020). *Data Mining - Concepts, Models, Methods, and Algorithms*, 3rd edition. Wiley - IEEE Press.
- rapidminer.com, 2021,. Rapidminer Manual and Documentation, <https://docs.rapidminer.com/latest/studio/operators/>.



Data Mining



Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

● ● ●

Session 02



PROGRAM STUDI TEKNIK INFORMATIKA

Sumber: <https://ft.uhamka.ac.id/prodi/informatika/>

- VISI.
 - ✓ Menjadi program studi informatika yang menghasilkan karya unggul di bidang teknologi informasi melalui pembelajaran berlandaskan prophetic teaching untuk membentuk lulusan yang cerdas secara spiritual, intelektual, emosional, dan sosial.
- MISI.
 1. Melaksanakan pendidikan di bidang informatika yang terintegrasi dengan nilai-nilai Al Islam Kemuhammadiyah.
 2. Melaksanakan penelitian untuk menghasilkan karya teknologi informasi yang unggul untuk masyarakat berkembang.
 3. Melaksanakan pengabdian kepada masyarakat untuk kesejahteraan umat.
 4. Menghasilkan mahasiswa yang berwawasan global dalam teknologi informasi.
 5. Mendorong jiwa kewirausahaan mahasiswa untuk berkiprah di tengah masyarakat dalam meningkatkan kegiatan ekonomi.
 6. Menerapkan standar mutu pendidikan informatika melalui layanan mutu yang berkesinambungan.
 7. Membangun kerja sama terpadu dengan mitra yang berkontribusi untuk kemajuan pendidikan di bidang informatika.

Semester Genap - TA.2022/2023 2



Sistem Penilaian

- SKS : 3
 - Teori 3 SKS – (14 Kali Pertemuan)
- Keaktifan (bobot 10%)
 - Terdiri dari: Aktivitas Perkuliahan (*Kehadiran, Etika, dan Sikap*).
- Project (bobot 25%)
 - Terdiri dari: Tugas Besar / Project (*Penguasaan Keterampilan*) dengan cara observasi, mengimplementasikan dan menghasilkan Proyek Akhir, serta mempresentasikannya.
- Tes (UTS = bobot 25% , UAS = bobot 40%)
 - Terdiri dari: UTS dan UAS (*Penguasaan Pengetahuan*) dengan cara tes tertulis maupun tes praktikum.

Semester Genap - TA.2022/2023 3



Tujuan/Objectives (Session.02)

- Objectives (Tujuan Instruksional Umum).
 - RPS/SAP.
 - Pengantar & Pengenalan Data Mining
- Indikator Kompetensi.
 - Mahasiswa mampu mengetahui dan memahami apa itu *data mining*, dan mengapa harus menggunakan *data mining*.
 - Mahasiswa mampu mengetahui dan memahami tentang peran utama dan metode *data mining*.
 - Mahasiswa mampu memahami sejarah dan penerapan *data mining*.

Semester Genap - TA.2022/2023 4

Muhasabah

يَا أَيُّهَا الَّذِينَ آمَنُوا اتَّقُوا اللَّهَ
وَكُونُوا مَعَ الصَّادِقِينَ

“Wahai orang-orang yang
beriman, bertakwalah kepada
Allah dan hendaklah kamu
bersama orang-orang yang
jujur.”

[Al-Taubah: 119]

Al-Imam Ibnu Katsir rahimahullah berkata, “Maknanya:
Jujurlah dan teruslah berlaku jujur maka kamu akan bersama
orang-orang yang jujur dan selamat dari kebinasaan, dan
Allah akan menjadikan bagimu kelapangan serta jalan keluar
dalam urusanmu.” [Tafsir Ibnu Katsir, 4/230]

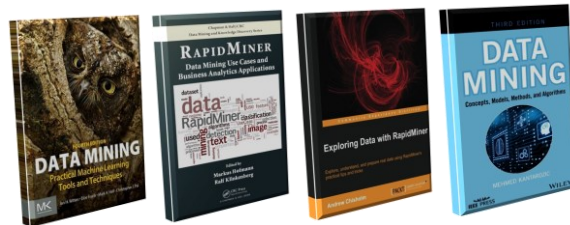
Data Mining



Uhamka
UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA
integrity, trust, compassion

Session 02

Pengantar & Pengenalan Data Mining



Pengantar & Pengenalan Data Mining

Mengapa Data Mining ?

Semester Genap - TA.2022/2023

7

Mengapa Data Mining ?

- Manusia memproduksi beragam data yang jumlah dan ukurannya sangat besar

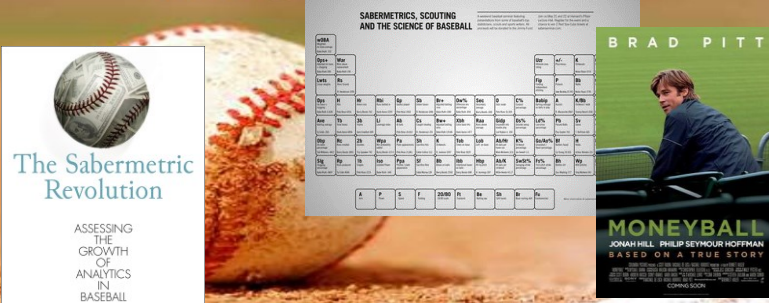
- Astronomi
- Bisnis
- Kedokteran
- Ekonomi
- Olahraga
- Cuaca
- Financial
- ...



Semester Genap - TA.2022/2023

8

Winning An Unfair Game



The Sabermetric Revolution
ASSESSING THE GROWTH OF ANALYTICS IN BASEBALL

How Do The Oakland A's Win With a Poor Team?

Mengapa Data Mining ? (lanjutan)

Astronomi

- **Sloan Digital Sky Survey**
 - New Mexico, 2000
 - **140TB** over 10 years
- **Large Synoptic Survey Telescope**
 - Chile, 2016
 - Will acquire **140TB** every five days

Biologi dan Kedokteran

- **European Bioinformatics Institute (EBI)**
 - **20PB** of data (genomic data doubles in size each year)
 - A single sequenced human genome can be around **140GB** in size

Unit	Value
kilobyte (kB)	10^3
megabyte (MB)	10^6
gigabyte (GB)	10^9
terabyte (TB)	10^{12}
petabyte (PB)	10^{15}
exabyte (EB)	10^{18}
zettabyte (ZB)	10^{21}
yottabyte (YB)	10^{24}

Semester Genap - TA.2022/2023 10

Mengapa Data Mining ? (lanjutan)

Unit	Value
kilobyte (kB)	10^3
megabyte (MB)	10^6
gigabyte (GB)	10^9
terabyte (TB)	10^{12}
petabyte (PB)	10^{15}
exabyte (EB)	10^{18}
zettabyte (ZB)	10^{21}
yottabyte (YB)	10^{24}

- **Mobile Electronics** market
 - 4.43B mobile phone users in 2015
 - 7B mobile phone subscriptions in 2015
- **Web and 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
 - **235 TBs** data collected by the US Library of Congress
 - 15 out of 17 sectors in the US have **more data stored per company than the US Library of Congress** 10^{18}

Semester Genap - TA.2022/2023 11



Mengapa Data Mining ?
(lanjutan)

We are drowning in data, but starving for knowledge!

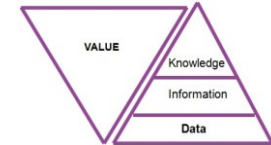
(John Naisbitt, Megatrends, 1988)



Semester Genap - TA.2022/2023 13

Mengapa Data Mining ?
(lanjutan)

- Data harus kita 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**



Semester Genap - TA.2022/2023 14

“Data is the new Oil”

— Clive Humby



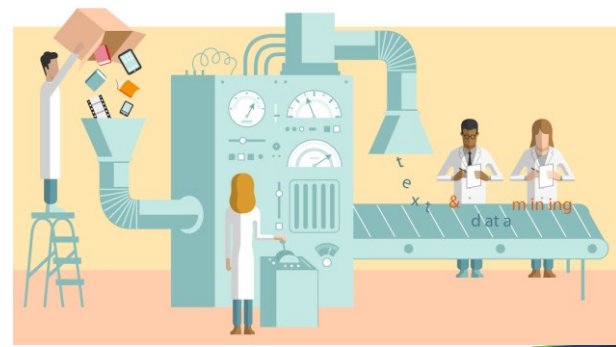
8 YOUTUBER TERKAYA INDONESIA

Youtube menjadi mesin uang. Miliaran rupiah bisa diraup setiap bulannya. Beberapa youtuber Indonesia telah menikmatinya.

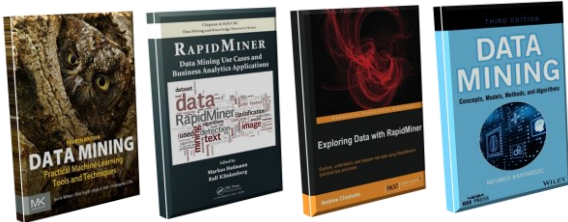
Youtuber	Video	Jumlah	Estimasi Pendapatan
Atta Halilintar	821 video	22,6 Juta	Rp2,33 Miliar
Riesli Official	891 video	20,4 Juta	Rp2,89 Miliar
Jessy No Limit	963 video	15,8 Juta	Rp1,24 Miliar
Isabel Pradie	584 video	11,5 Juta	Rp1,48 Miliar
Gen Halilintar	738 video	11,5 Juta	Rp1,48 Miliar
Rizki Dendang	1.242 video	11,5 Juta	Rp1,48 Miliar
Reza Entertainmen	1.242 video	11,5 Juta	Rp1,48 Miliar
Frank Diamond	1.048 video	10,8 Juta	Rp1,48 Miliar
Saah Halilintar	1.242 video	11,5 Juta	Rp1,48 Miliar
Pratik Dendang	1.242 video	11,5 Juta	Rp1,48 Miliar
Pratik Dendang	1.242 video	11,5 Juta	Rp1,48 Miliar
Pratik Dendang	1.242 video	11,5 Juta	Rp1,48 Miliar

Semester Genap - TA.2022/2023

Mengapa Data Mining ?
(lanjutan)



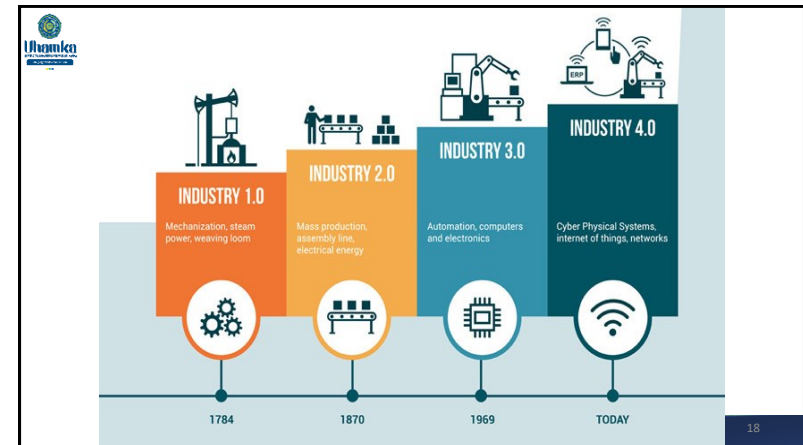
Semester Genap - TA.2022/2023 16



Pengantar & Pengenalan Data Mining

Data Mining

Semester Genap - TA.2022/2023 17



Perusahaan Pengolah Pengetahuan

- **Uber** - the world's largest taxi company, *owns no vehicles*.
- **Google** - world's largest media/advertising company, *creates no content*.
- **Alibaba** - the most valuable retailer, *has no inventory*.
- **Airbnb** - the world's largest accommodation provider, *owns no real estate*.
- **Gojek** - perusahaan serba ada, *tanpa memiliki apapun*

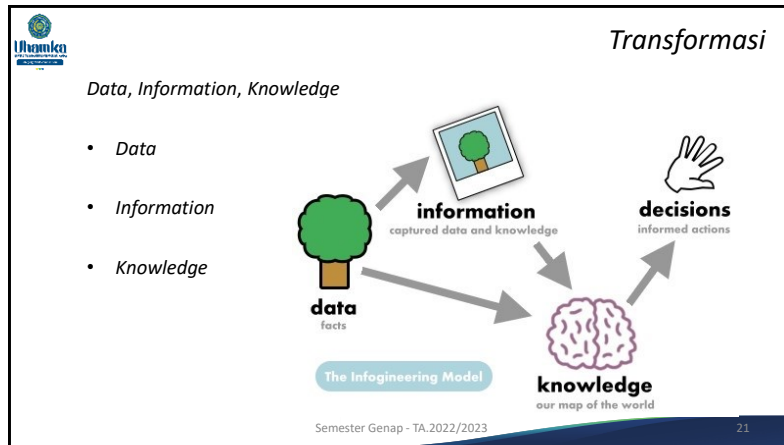


Semester Genap - TA.2022/2023 19

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)

Semester Genap - TA.2022/2023 20



Data - Informasi – Pengetahuan

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

Data Kehadiran Pegawai

Semester Genap - TA.2022/2023 22

Data - Informasi – Pengetahuan

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

Informasi Akumulasi Bulanan Kehadiran Pegawai

Semester Genap - TA.2022/2023 23

Data - Informasi – Pengetahuan

Keterangan	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

Pola Kebiasaan Kehadiran Mingguan Pegawai

Semester Genap - TA.2022/2023 24

Data - Informasi – Pengetahuan - Kebijakan

- 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**

Semester Genap - TA.2022/2023 25

Mengubah Pengetahuan Menjadi Kebijakan

Semester Genap - TA.2022/2023 26

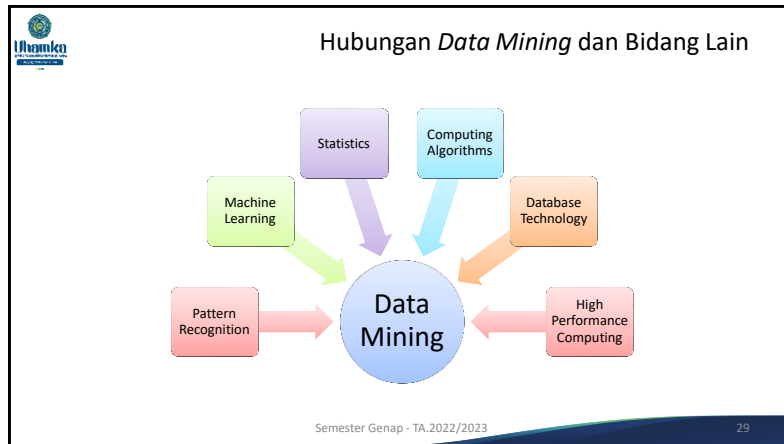
Data Mining Tasks and Roles

Semester Genap - TA.2022/2023 27

Pengantar & Pengenalan Data Mining

Hubungan Data Mining Dengan Bidang Lain

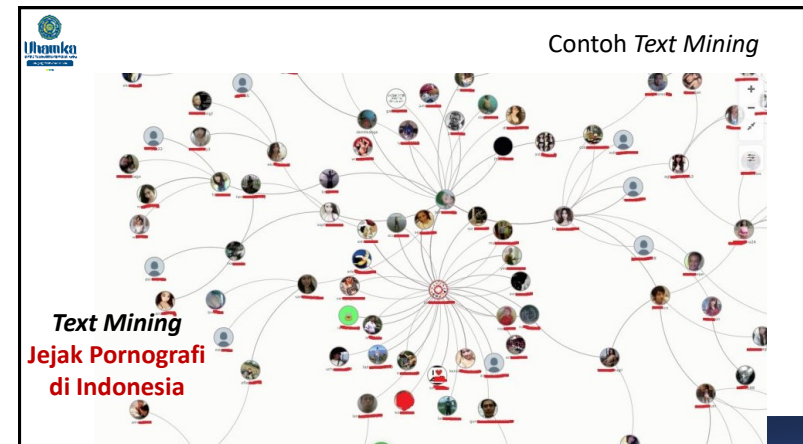
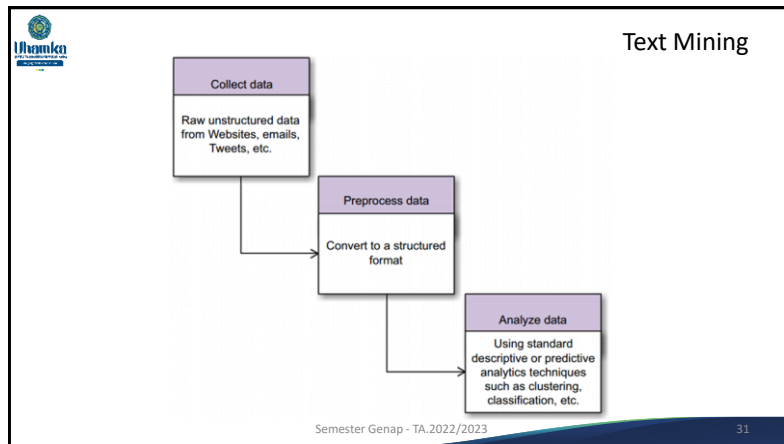
Semester Genap - TA.2022/2023 28

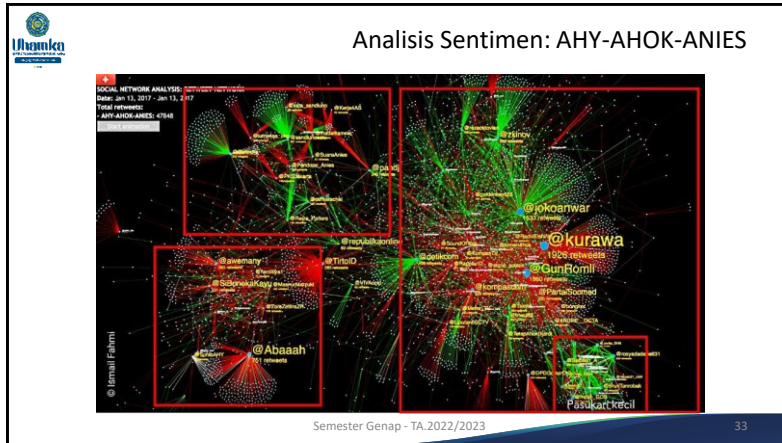


Data Mining vs Text Mining

- 1. Text Mining:**
 - Mengolah **data tidak terstruktur** dalam bentuk *text, web, social media*, dsb
 - Menggunakan **metode text processing** untuk mengkonversi data tidak terstruktur menjadi terstruktur
 - Kemudian **diolah dengan data mining**.
- 2. Data Mining:**
 - Mengolah **data terstruktur** dalam bentuk tabel yang memiliki atribut dan kelas.
 - Menggunakan **metode data mining**, yang terbagi menjadi metode estimasi, *forecasting*, klasifikasi, klastering atau asosiasi.
 - Yang dasar berpikirnya menggunakan konsep **statistika** atau heuristik ala *machine learning*.

Semester Genap - TA.2022/2023 30





UHAMKA

Permasalahan yang ada pada Data Mining

- Tremendous amount of data**
 - Algorithms must be **highly scalable** to handle such as tera-bytes of data
- High-dimensionality of data**
 - Micro-array may have tens of **thousands of dimensions**
- High complexity of data**
 - Data streams** and sensor data
 - Time-series data**, temporal data, sequence data
 - Structure data, graphs, **social networks** and multi-linked data
 - Heterogeneous **databases** and legacy databases
 - Spatial, spatiotemporal, **multimedia**, text and **Web data**
 - Software programs**, scientific simulations
- New and sophisticated applications**

Semester Genap - TA.2022/2023 34

UHAMKA

Pengantar & Pengenalan Data Mining

Pengertian

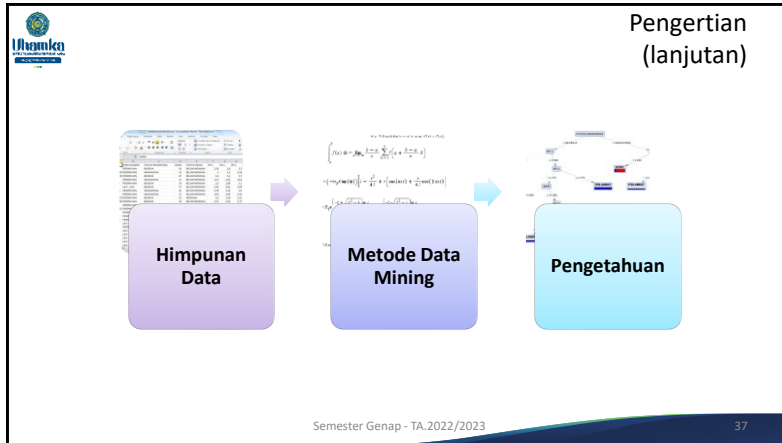
Semester Genap - TA.2022/2023 35

UHAMKA

Pengertian

- Disiplin ilmu yang mempelajari **metode** untuk **mengeksrak pengetahuan** atau **menemukan pola** dari suatu data yang besar
- Ekstraksi dari **data** ke **pengetahuan**:
 - Data**: **fakta yang terekam** dan tidak membawa arti.
 - Pengetahuan**: **pola, rumus**, aturan atau model yang muncul dari data.
- Nama lain **data mining**:
 - Knowledge Discovery in Database (KDD)**
 - Knowledge extraction**
 - Pattern analysis**
 - Information harvesting**
 - Business intelligence**
 - Big data**

Semester Genap - TA.2022/2023 36

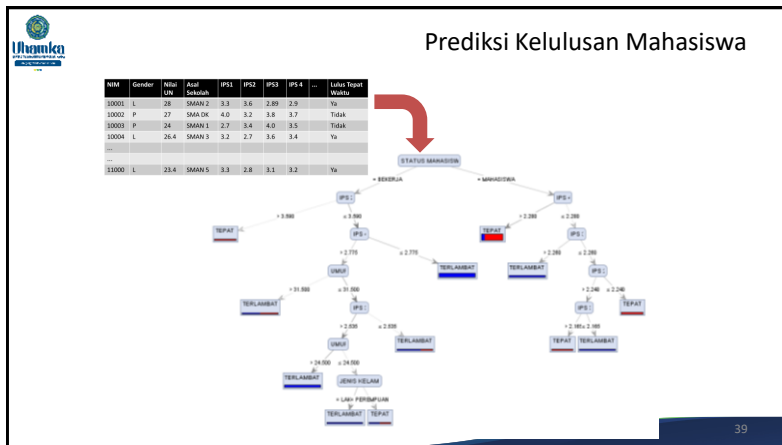


Contoh Data di Kampus

- 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

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

Semester Genap - TA.2022/2023 38

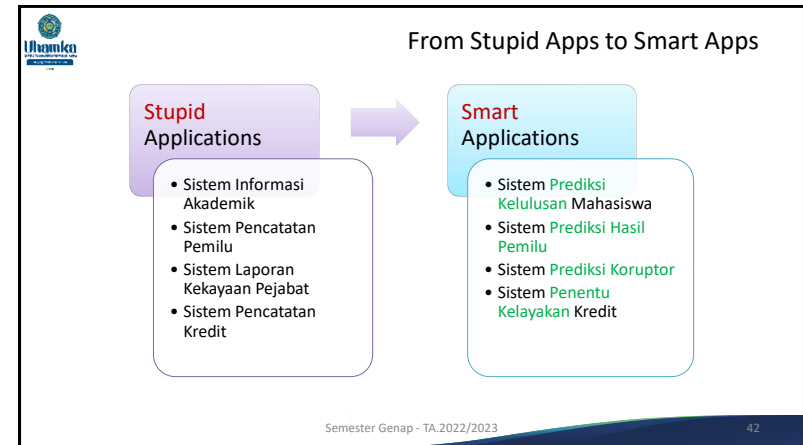
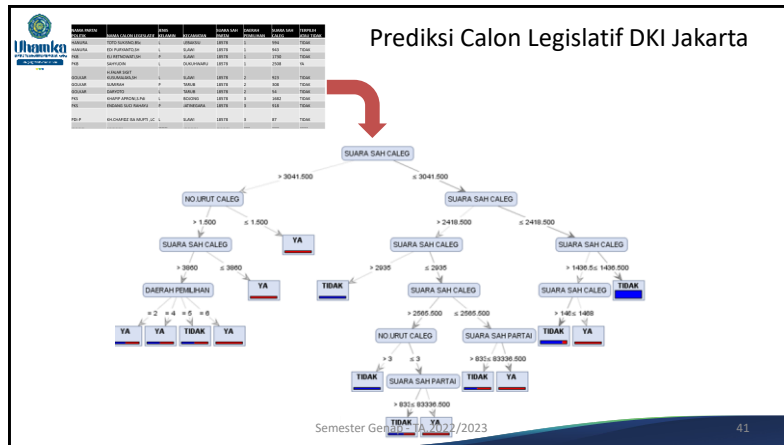


Contoh Data di KPU

- Puluhan ribu data calon anggota legislatif di KPU
- Apakah pernah kita ubah menjadi pengetahuan yang lebih bermanfaat? TIDAK!

NAMA PARTAI POLITIK	NAMA CALON LEGISLATIF	JENIS KELAMIN	KECAMATAN	SUARA SAH PARTAI	DAERAH PEMILIHAN	SUARA SAH KALEG	TERPILIH ATAU TIDAK
HANURA	TOTO SUKISNO.BSc	L	LEBAKSIU	18578	1	594	TIDAK
HANURA	EDI PURYANTO.SH	L	SLAWI	18578	1	943	TIDAK
PKB	ELI RETNOWATLSH	P	SLAWI	18578	1	1730	TIDAK
PKB	SAHYUDIN	L	DUKUHWARU	18578	1	2508	YA
GOLKAR	H.FAJAR SIGIT KUSUMAJAYA.SH	L	SLAWI	18578	2	923	TIDAK
GOLKAR	SUMIRAH	P	TARUB	18578	2	308	TIDAK
GOLKAR	DARYOTO	L	TARUB	18578	2	54	TIDAK
PKS	KHAPIP APRONI.S.Pdi	L	BOJONG	18578	3	1682	TIDAK
PKS	ENDANG SUCI RAHAYU	P	JATINEGARA	18578	3	918	TIDAK
PDI-P	KH.CHAFIDZ ISA MUFTI .LC	L	SLAWI	18578	3	87	TIDAK

Semester Genap - TA.2022/2023 40



Referensi

- Ian H Witten, Eibe Frank, Mark A Hall. (2016). *Data Mining - Practical Machine Learning Tools and Techniques*, 4th Edition. Morgan-Kaufman.
- Hofmann, Markus., Klinkenberg, Ralf. (2016). *Data Mining Use Cases and Business Analytics Application*. CRC Press.
- Crisholm, Andrew. (2013). *Exploring Data With Rapidminer*. Packt Publishing.
- Mehmed Kantardzic. (2020). *Data Mining - Concepts, Models, Methods, and Algorithms*, 3rd edition. Wiley - IEEE Press.
- rapidminer.com, 2021,. Rapidminer Manual and Documentation, <https://docs.rapidminer.com/latest/studio/operators/>.

Semester Genap - TA.2022/2023

Terima Kasih

HADITS KEJUJURAN

عَلَيْكُمْ بِالصِّدْقِ فَإِنَّ الصِّدْقَ يَهْدِي إِلَى الْبِرِّ
وَإِنَّ الْبِرَّ يَهْدِي إِلَى الْجَنَّةِ

"Hendaklah kalian senantiasa berlaku jujur, karena sesungguhnya kejujuran akan mengantarkan pada kebaikan dan sesungguhnya kebaikan akan mengantarkan pada surga. (HR. Muslim)"



Data Mining



Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

● ● ●

Session 03



PROGRAM STUDI TEKNIK INFORMATIKA

Sumber: <https://ft.uhamka.ac.id/prodi/informatika/>

- VISI.
 - ✓ Menjadi program studi informatika yang menghasilkan karya unggul di bidang teknologi informasi melalui pembelajaran berlandaskan prophetic teaching untuk membentuk lulusan yang cerdas secara spiritual, intelektual, emosional, dan sosial.
- MISI.
 1. Melaksanakan pendidikan di bidang informatika yang terintegrasi dengan nilai-nilai Al Islam Kemuhammadiyah.
 2. Melaksanakan penelitian untuk menghasilkan karya teknologi informasi yang unggul untuk masyarakat berkembang.
 3. Melaksanakan pengabdian kepada masyarakat untuk kesejahteraan umat.
 4. Menghasilkan mahasiswa yang berwawasan global dalam teknologi informasi.
 5. Mendorong jiwa kewirausahaan mahasiswa untuk berkiprah di tengah masyarakat dalam meningkatkan kegiatan ekonomi.
 6. Menerapkan standar mutu pendidikan informatika melalui layanan mutu yang berkesinambungan.
 7. Membangun kerja sama terpadu dengan mitra yang berkontribusi untuk kemajuan pendidikan di bidang informatika.

Semester Genap - TA.2022/2023 2



Sistem Penilaian

- SKS : 3
 - Teori 3 SKS – (14 Kali Pertemuan)
- Keaktifan (bobot 10%)
 - Terdiri dari: Aktivitas Perkuliahan (*Kehadiran, Etika, dan Sikap*).
- Project (bobot 25%)
 - Terdiri dari: Tugas Besar / Project (*Penguasaan Keterampilan*) dengan cara observasi, mengimplementasikan dan menghasilkan Proyek Akhir, serta mempresentasikannya.
- Tes (UTS = bobot 25% , UAS = bobot 40%)
 - Terdiri dari: UTS dan UAS (*Penguasaan Pengetahuan*) dengan cara tes tertulis maupun tes praktikum.

Semester Genap - TA.2022/2023 3



Tujuan/Objectives (Session.03)

- Objectives (Tujuan Instruksional Umum).
 - RPS/SAP.
 - Eksplorasi Data
- Indikator Kompetensi.
 - Mahasiswa mampu mengetahui dan memahami tentang teknik-teknik merepresentasikan data.
 - Mahasiswa mampu mengetahui dan memahami tentang teknik Statistik Data dan Visualisasi Data.
 - Mahasiswa mampu mengetahui dan memahami tentang *data preparation*.
 - Mahasiswa mampu mengetahui dan memahami tentang *understanding data*.

Semester Genap - TA.2022/2023 4



Muhasabah

شَهْرُ رَمَضَانَ الَّذِي أُنْزِلَ فِيهِ الْقُرْآنُ هُدًى لِّلنَّاسِ وَبَيِّنَاتٍ مِّنَ الْهُدَىٰ وَالْفُرْقَانِ فَمَن شَهِدَ مِنْكُمُ
 الشَّهْرَ فَلْيَصُمْهُ وَمَن كَانَ مَرِيضًا أَوْ عَلَىٰ سَفَرٍ فَعِدَّةٌ مِّنْ أَيَّامٍ أُخَرَ يُرِيدُ اللَّهُ بِكُمُ الْيُسْرَ وَلَا يُرِيدُ
 بِكُمُ الْعُسْرَ وَلِتُكْمِلُوا الْعِدَّةَ وَلِتُكَبِّرُوا اللَّهَ عَلَىٰ مَا هَدَاكُمْ وَلَعَلَّكُمْ تَشْكُرُونَ

- Qs. Al-Baqarah, ayat 185
 - Bulan Ramadan adalah (bulan) yang di dalamnya diturunkan Al-Qur'an sebagai petunjuk bagi manusia dan penjelasan-penjelasan mengenai petunjuk itu serta pembeda (antara yang hak dan yang batil).
 - Oleh karena itu, siapa di antara kamu hadir (di tempat tinggalnya atau bukan musafir) pada bulan itu, berpuasalah.
 - Siapa yang sakit atau dalam perjalanan (lalu tidak berpuasa), maka (wajib menggantinya) sebanyak hari (yang ditinggalkannya) pada hari-hari yang lain. Allah menghendaki kemudahan bagimu dan tidak menghendaki kesukaran.
 - Hendaklah kamu mencukupkan bilangannya dan mengagungkan Allah atas petunjuk-Nya yang diberikan kepadamu agar kamu bersyukur.

Semester Genap - TA.2022/2023



Data Mining



Uhamka

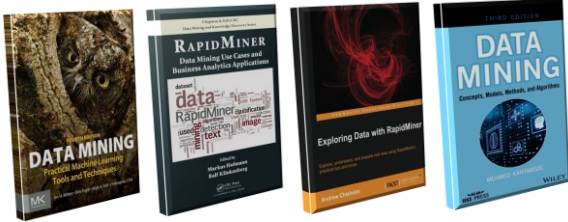
UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

Session 03
 Eksplorasi Data



Eksplorasi Data



PendahuTuan

Semester Genap - TA.2022/2023



Pendahuluan

- Sebelum melakukan penambangan emas, tembaga, minyak, atau apapun, maka tentunya Anda perlu tahu:
 - Seberapa banyak kandungan tambang yang ada
 - Seberapa luas areanya
 - Seberapa rumit topologi wilayahnya
 - dll .
- Sehingga Anda dapat menentukan teknik, metode, dan *tools* yang harus digunakan.

Semester Genap - TA.2022/2023

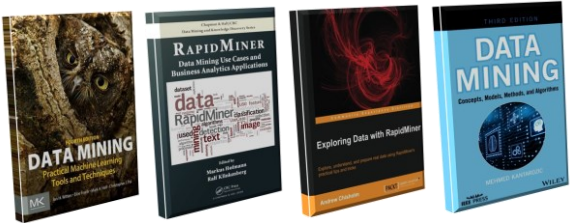


Pendahuluan (lanjutan)

- Begitu juga dengan penambangan data (*data mining*).
- Sebelum Anda melakukan *data mining*, maka Anda perlu memahami data dengan baik.
- Melakukan *pre-processing* (pra-pemrosesan) untuk meningkatkan kualitasnya, setelah itu baru kemudian dapat melakukan penambangan.

Semester Genap - TA.2022/2023
9





Eksplorasi Data

Himpunan Data

Semester Genap - TA.2022/2023
10



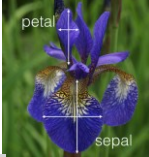
Himpunan Data

- Himpunan data dibangun dari objek-objek data, dimana objek data menyatakan sebuah entitas.
 - Misalnya pada himpunan data Universitas.
 - Objek dapat berupa mahasiswa, dosen, mata kuliah.
 - Misalnya pada himpunan data operator seluler.
 - Objek dapat berupa pelanggan, layanan, fitur, dll.
- Objek biasanya digambarkan dengan menggunakan atribut. Objek data disebut juga sebagai sampel, contoh, atau titik data.
- Objek-objek data yang disimpan dalam suatu basisdata (*database*) disebut *tuple*, dimana baris menyatakan objek-objek data dan kolom adalah atribut.

Semester Genap - TA.2022/2023
11



Dataset (Himpunan Data)



Attribute/Feature/Dimension
Class/Label/Target

	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

Record/
Object/
Sample/
Tuple/
Data
Nominal
Numerik

Semester Genap - TA.2022/2023
12



Himpunan Data (lanjutan)

- Dalam kehidupan sehari-hari, Anda pasti sering mendengar kata atribut.
- Atribut adalah simbol yang menggambarkan identitas/karakteristik suatu kelompok.
 - Misalnya, atribut sebagai siswa SMA:
 - Celana abu-abu
 - Baju putih dengan tulisan nama SMA dibagian lengan
 - Topi abu-abu
 - dll .

Semester Genap - TA.2022/2023

13



Himpunan Data (lanjutan)

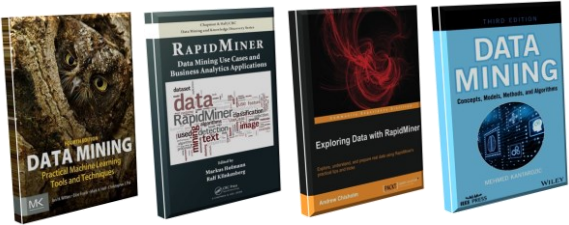
- Dibiidang *data mining*, atribut adalah *data field*, yang merepresentasikan karakteristik atau fitur dari objek data.
- Pada bidang yang lain, atribut biasanya disebut juga sebagai **dimensi** (pada bidang *data warehouse*), **fitur** (pada bidang *machine learning*), atau **variabel**.
- Sebagai contoh lain, atribut yang menggambarkan objek siswa SMA:
 - Nomor induk siswa
 - Nama kelas
 - Alamat
 - dll .

Semester Genap - TA.2022/2023

14



Eksplorasi Data Jenis-jenis Atribut

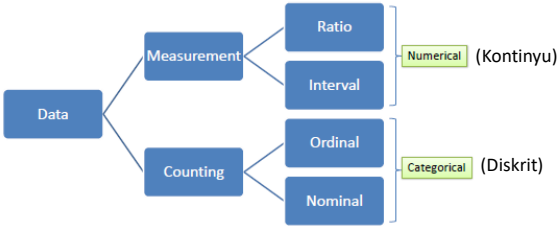


Semester Genap - TA.2022/2023

15



Tipe Data



Semester Genap - TA.2022/2023

16

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 ($=$, $\neq$)	- Kode pos - Jenis kelamin - Nomer id karyawan - Nama kota	mode, entropy, contingency correlation, χ^2 test

17

Jenis-jenis Atribut

- Dibidang *data mining*, atribut adalah *data field*, yang merepresentasikan karakteristik atau fitur dari objek data.
- Pada bidang yang lain, atribut biasanya disebut juga sebagai **dimensi** (pada bidang *data warehouse*), **fitur** (pada bidang *machine learning*), atau **variabel**.
- Atribut terdiri dari berbagai jenis, seperti:
 - Atribut Nominal
 - Atribut Biner
 - Atribut Numerik
 - Atribut Ordinal
 - Atribut Diskrit dan Kontinu

Semester Genap - TA.2022/2023

18

Jenis-jenis Atribut (lanjutan)

a. Atribut Nominal

- Atribut nominal disebut juga kategorial, karena nilainya yang menggambarkan kategori, kode, atau status, yang tidak memiliki urutan.
- Misalnya:
 - Atribut Warna, yang mempunyai 2 kemungkinan nilai: yaitu Merah, atau Putih.
 - Atribut Kategori Pelanggan, yang dapat bernilai: Silver, Gold, atau Platinum.
 - Atribut Kode Wilayah, yang dapat bernilai: A, B, C, D, atau E.

Semester Genap - TA.2022/2023

19

Jenis-jenis Atribut (lanjutan)

- Apakah atribut nominal harus berupa teks? Bisakah bernilai Numerik? Tentu saja bisa.
- Misalnya:
 - Atribut Warna bisa bernilai 0 atau 1. Dimana 0 menyatakan Merah, sedangkan 1 menyatakan Putih.
 - Atribut Kategori Pelanggan, bisa bernilai 1, 2, atau 3. Dimana 1 menyatakan Silver, 2 adalah Gold, dan 3 menyatakan Platinum.

Semester Genap - TA.2022/2023

20



Jenis-jenis Atribut (lanjutan)

b. Atribut Biner

- Atribut biner, atau disebut juga dengan atribut *Boolean*, merupakan atribut nominal yang hanya memiliki 2 kategori nilai: yaitu 0 atau 1.
- Dimana 0 biasanya menyatakan Tidak (sesuatu yang negatif, atau yang berdampak kecil), sedangkan 1 menyatakan Ya (sesuatu yang positif, atau yang berdampak besar).
- Walaupun memiliki nilai kategorial seperti atribut nominal, dalam *data mining* atribut biner dipandang secara khusus, karena memiliki nilai yang unik.

Semester Genap - TA.2022/2023 21



Jenis-jenis Atribut (lanjutan)

– Atribut biner dibedakan menjadi 2, yaitu:

1. Atribut biner simetris.
 - Jika nilainya dianggap memberikan dampak yang setara.
 - Misalnya atribut Jenis Kelamin yang bernilai Pria atau Wanita (tidak penting mana yang harus bernilai 0, dan mana yang bernilai 1, karena keduanya dianggap setara).
2. Atribut biner asimetris.
 - Jika nilainya memberikan dampak berbeda, yang secara konvensi bernilai 1 untuk sesuatu yang jarang terjadi, dan bernilai 0 untuk sesuatu yang umum terjadi.
 - Misalnya atribut Hasil Tes Buta Warna, yang bernilai 1 (Buta Warna) dan 0 (Tidak Buta Warna).

Semester Genap - TA.2022/2023 22



Jenis-jenis Atribut (lanjutan)

c. Atribut Numerik

- Atribut numerik adalah kuantitatif, yang memiliki nilai berupa kuantitas yang terukur dan dinyatakan dalam nilai-nilai bulat (*integer*) atau riil (*real*).
- Atribut numerik dapat diskalakan secara interval atau secara rasio.
- Sebagai contoh, temperature udara dalam satuan Celcius adalah atribut numerik yang dapat diskalakan secara interval, namun tidak dapat diskalakan secara rasio.
 - Temperatur udara di ruangan A adalah 30°C, sedangkan di ruangan B adalah 15°C.
 - Kedua ruangan dapat dikatakan memiliki perbedaan temperatur lima belas derajat Celcius, namun tidak dapat dikatakan bahwa ruangan A dua kali lebih panas dibanding ruangan B. Mengapa? . . . Karena satuan Celcius tidak pernah memiliki titik 0.
 - Dengan kata lain, 0°C bukan menyatakan “Tidak ada temperatur”.

Semester Genap - TA.2022/2023 23



Jenis-jenis Atribut (lanjutan)

- Agar lebih jelas, coba bedakan dengan atribut Gaji Pegawai, yang dapat bernilai 0.
- Gaji seorang pegawai yang mencapai 30 juta rupiah per bulan, dapat dikatakan tiga kali lebih besar dibanding pegawai lain yang bergaji 10 juta rupiah per bulan.
- Dengan demikian, atribut Gaji Pegawai adalah atribut numerik yang dapat diskalakan secara rasio.

Semester Genap - TA.2022/2023 24



Jenis-jenis Atribut (lanjutan)

d. Atribut Ordinal

- Atribut ordinal memiliki nilai yang menggambarkan urutan atau peringkat (*ranking*).
- Namun, ukuran perbedaan antara dua nilai yang berurutan tidak diketahui.
 - Misalnya, atribut Kategori Pelanggan yang dapat bernilai: Silver, Gold, atau Platinum. Ketiga nilai tersebut memiliki urutan atau tingkatan, namun tidak menjelaskan seberapa besar perbedaan antara Silver, Gold, dan Platinum.
 - Contoh lain, adalah atribut Cuaca yang memiliki empat nilai: Beku, Dingin, Hangat, dan Panas. Keempat nilai tersebut menyatakan urutan atau tingkatan, namun nilai-nilai tersebut tidak menjelaskan seberapa besar perbedaan diantara keempat nilai tersebut.

Semester Genap - TA.2022/2023 25



Jenis-jenis Atribut (lanjutan)

- Atribut ordinal sangat berguna didalam survei, yaitu untuk penilaian subjektif (kualitatif) yang tidak dapat diukur secara objektif.
 - Misalnya survei tentang kepuasan pelanggan yang menghasilkan atribut bernilai ordinal, yaitu: 1 (Sangat Tidak Puas), 2 (Tidak Puas), 3 (Puas), 4 (Sangat Puas).
- Sama dengan atribut nominal dan biner, atribut ordinal adalah kualitatif.
 - Artinya, ketiga atribut ini menggambarkan fitur dari suatu objek namun tanpa memberikan ukuran atau kuantitas yang sebenarnya.
 - Ketiga atribut ini memiliki nilai yang berupa teks, yang menyatakan kategori atau bisa juga berupa angka yang menggambarkan kode (yang hanya digunakan dalam komputer) untuk kategori tersebut (bukan angka yang menyatakan ukuran atau kuantitas), seperti angka 1 s/d 4 yang menyatakan Kepuasan Pelanggan pada contoh diatas.

Semester Genap - TA.2022/2023 26



Jenis-jenis Atribut (lanjutan)

e. Atribut Diskrit dan Kontinu

- Semua atribut diatas (atribut nominal, atribut biner, atribut numerik, atribut ordinal), pada dasarnya dapat dikelompokkan kedalam jenis atribut diskrit atau atribut kontinu.
- **Atribut diskrit** memiliki nilai terbatas, atau nilai terbatas tapi masih dapat dihitung, yang dapat bernilai bulat. Misalnya :
 - Atribut Temperatur Udara dalam satuan Celcius, yang dapat bernilai bulat antara 0 sampai 100.
 - Atribut Nomor Pelanggan (walaupun jumlah pelanggan terus bertambah, namun jumlah pelanggan masih dapat dihitung), sehingga jumlah pelanggan sebenarnya saat ini dapat bernilai bulat.

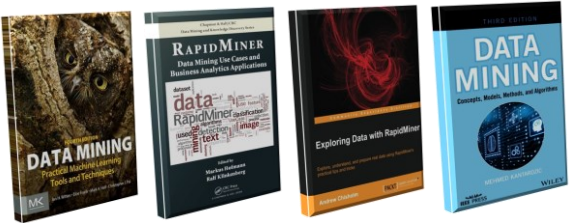
Semester Genap - TA.2022/2023 27



Jenis-jenis Atribut (lanjutan)

- **Atribut Kontinu**, memiliki nilai riil atau pecahan, yang dalam komputer biasanya direpresentasikan sebagai bilangan *real* atau *floating point*.
- Misalnya, atribut Gaji Pegawai yang dapat bernilai Rp.15.550.000,50,-

Semester Genap - TA.2022/2023 28



Eksplorasi Data

Data Understanding

Semester Genap - TA.2022/2023 29

Memahami Data

- Setelah mengetahui himpunan data dan jenis atribut, maka Anda harus dapat memahami data terlebih dahulu sebelum melakukan pra-pemrosesan.
- Tujuan dari memahami data, adalah agar Anda dapat memilih teknik dan metode pemrosesan yang sesuai menghasilkan data baru yang lebih mudah untuk dianalisis.
- Terdapat 4 cara dalam memahami data (J Han et al, 2012), yaitu:
 1. Tendensi Sentral
 2. Sebaran Data
 3. Grafik Statistik
 4. Visualisasi Data

Semester Genap - TA.2022/2023 30

Memahami Data (lanjutan)

1. Tendensi Sentral

- Tendensi sentral, atau biasa disebut sebagai kecenderungan pusat data.
- Dapat diukur dengan menggunakan empat cara, yaitu:
 - *Rata-rata*
 - Merupakan cara paling umum dan efektif untuk mengukur pusat dari himpunan N nilai dalam suatu atribut X.
 - *Median*
 - Merupakan nilai tengah.
 - *Modus*
 - Adalah nilai yang paling sering muncul.
 - *Midrange*
 - Adalah rata-rata dari nilai minimum dan maksimum.
- Setiap ukuran memiliki kelebihan dan kekurangan masing-masing, baik berdasarkan informasi yang dapat diberikan maupun berdasarkan kompleksitas komputasinya.

Semester Genap - TA.2022/2023 31

Memahami Data (lanjutan)

2. Sebaran Data

- Dengan menggunakan tendensi sentral, Anda bisa mendapatkan pusat dari himpunan data.
- Namun untuk memahami data dengan lebih baik, maka Anda perlu mengetahui *disperse* atau sebaran data tersebut.
- Setelah kita mendapatkan sebaran data, maka kita akan dapat menentukan data mana yang termasuk pencilan (*outlier*), dan data mana yang bukan pencilan.
- Terdapat dua cara untuk mengukur sebaran data, yaitu:
 - *Interquartile Range (IQR)*
 - *Standard Deviation (STD)*

Semester Genap - TA.2022/2023 32

Memahami Data
(lanjutan)

3. Grafik Statistik

- Jika Anda suka statistik, barangkali mudah saja memahami data hanya dengan melihat angka-angka statistik.
- Namun jika Anda seorang yang kurang suka atau bahkan sangat awam terhadap statistik, rasanya sulit mengimajinasikan angka-angka statistik untuk memahami data.
- Anda dapat menggunakan grafik statistik sebagai alat bantu untuk memahami data, diantaranya adalah:
 - *Quantile Plot*
 - *Q-Q Plot*
 - *Histogram*
 - *Scatter Plot*

Semester Genap - TA.2022/2023 33

Memahami Data
(lanjutan)

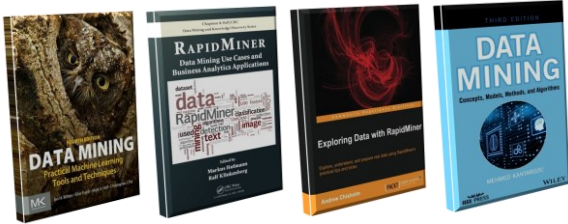
4. Visualisasi Data

- Pada banyak kasus, analisis data relatif sulit dilakukan hanya dengan mengandalkan nilai-nilai tendensi sentral maupun sebaran data. Bahkan grafik statistik juga tidak cukup untuk menjelaskan karakteristik data. Oleh karena itu, diperlukan teknik khusus untuk memvisualisasikan data.
- Visualisasi data memiliki dua kegunaan, yaitu:
 - Memudahkan analisis bagi perancang.
 - Memudahkan pembacaan laporan bagi pengguna (*user*).
- Terdapat banyak teknik visualisasi data yang dapat digunakan (J Han et al, 2012), diantaranya:
 - Visualisasi berorientasi pixel
 - Visualisasi proyeksi geometris
 - Visualisasi berbasis icon

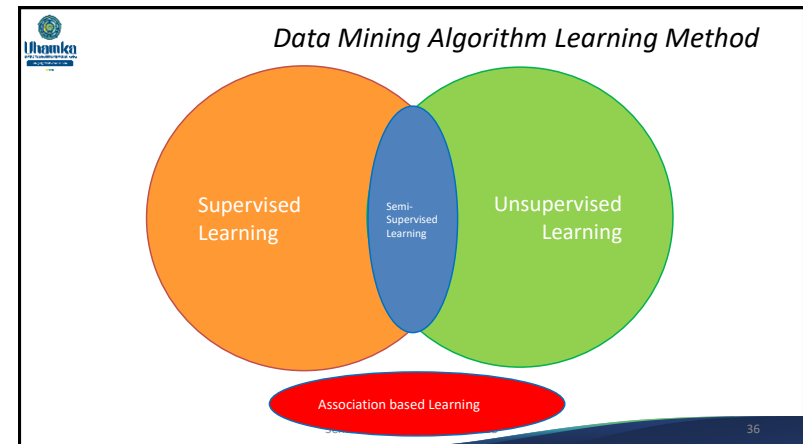
Semester Genap - TA.2022/2023 34

Eksplorasi Data

Metode Pembelajaran
didalam Algoritma *Data Mining*



Semester Genap - TA.2022/2023 35





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 variabel prediktor.

Semester Genap - TA.2022/2023

37



Dataset dengan Class

Attribute/Feature/Dimension Class/Label/Target

	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

Nominal

Numerik

Semester Genap - TA.2022/2023

38



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*.

Semester Genap - TA.2022/2023

39



Dataset tanpa Class

Attribute/Feature/Dimension

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

Semester Genap - TA.2022/2023

40

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), sedangkan data tanpa label digunakan untuk membuat batasan antara kelas.

⊕ Positive example --- Decision boundary without unlabeled examples
⊖ Negative example — Decision boundary with unlabeled examples
○ Unlabeled example

Semester Genap - TA.2022/2023 41

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

Semester Genap - TA.2022/2023 42

Output/Pola/Model/Knowledge

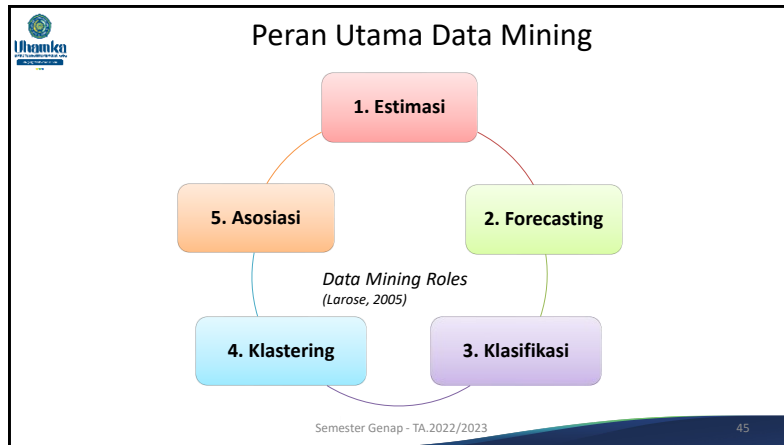
- 1. Formula/Function (Rumus atau Fungsi Regresi)**
 - $WAKTU\ TEMPUH = 0.48 + 0.6\ JARAK + 0.34\ LAMPU + 0.2\ PESANAN$
- 2. Decision Tree (Pohon Keputusan)**
- 3. Tingkat Korelasi**
- 4. Rule (Aturan)**
 - IF $ips3=2.8$ THEN luluscepatwaktu
- 5. Cluster (Klaster)**

43

Eksplorasi Data

Data Mining Roles

Semester Genap - TA.2022/2023 44



1. Estimasi Waktu Pengiriman Pizza

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

Semester Genap - TA.2022/2023 46

Contoh: Estimasi Performansi CPU

• **Example:** 209 different computer configurations

	Cycle time (ns)	Main memory (Kb)	Cache (Kb)	Channels	Performance		
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

• **Linear regression function**

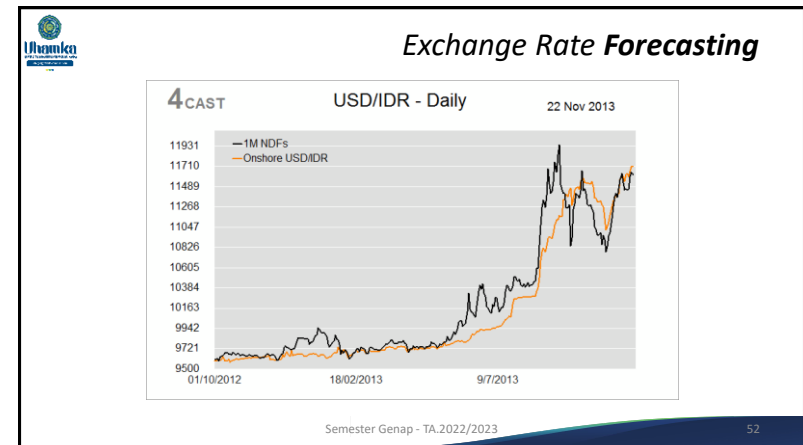
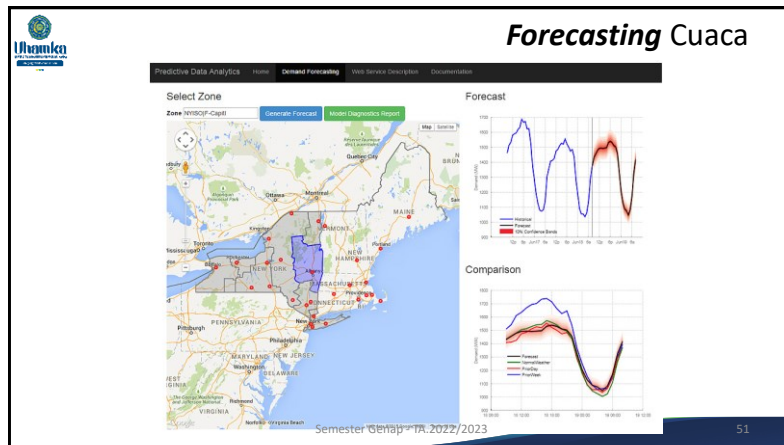
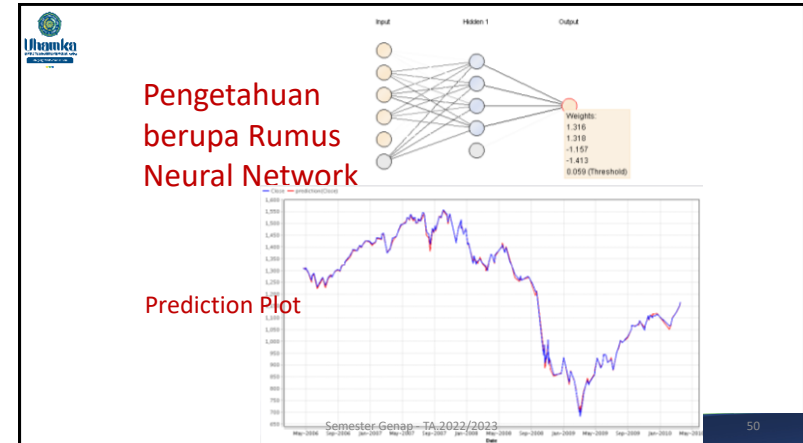
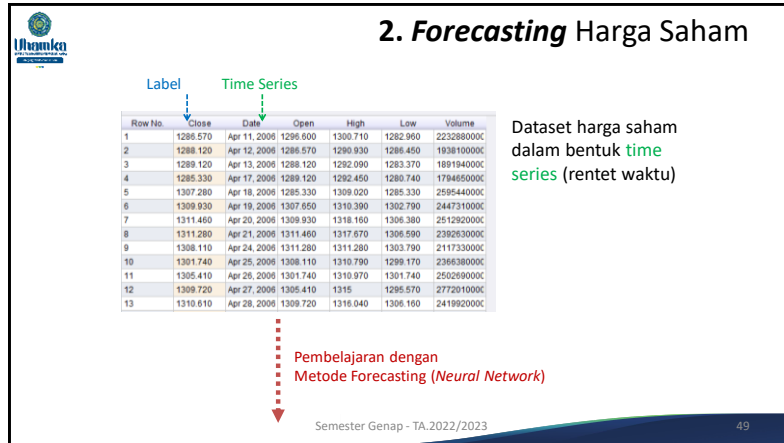
PRP = -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX

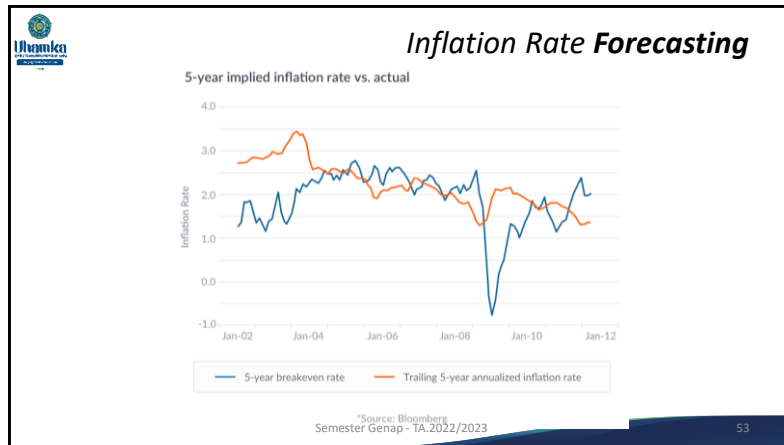
Semester Genap - TA.2022/2023 47

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 lulustepatwaktu
5. **Cluster** (Klaster)

Semester Genap - TA.2022/2023 48





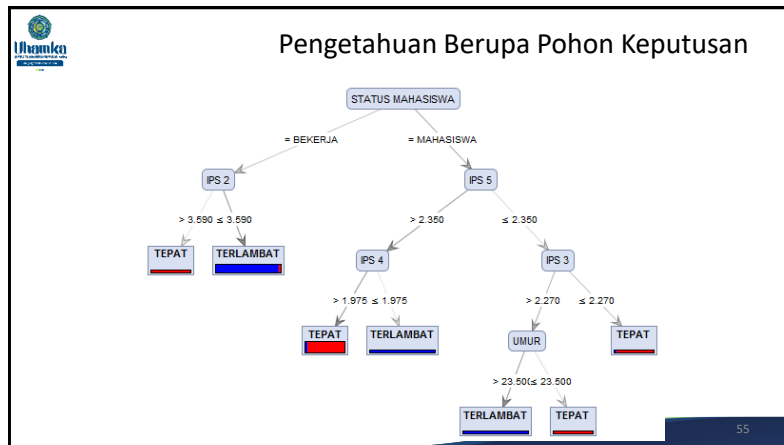
3. Klasifikasi Kelulusan Mahasiswa

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)

Semester Genap - TA.2022/2023



Contoh: Rekomendasi Main Golf

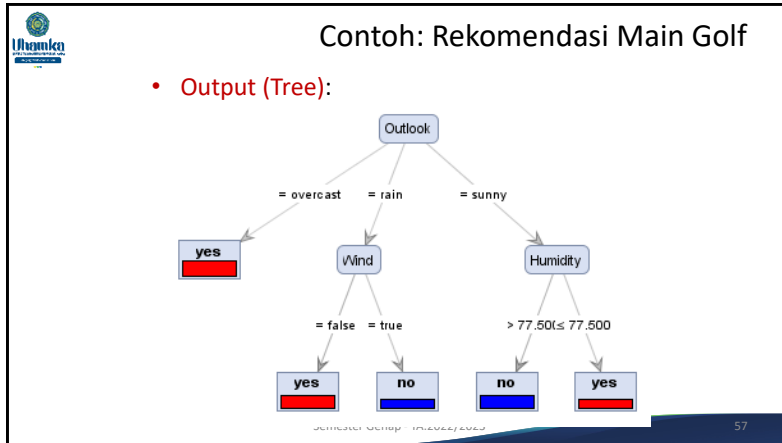
Input

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

Semester Genap - TA.2022/2023



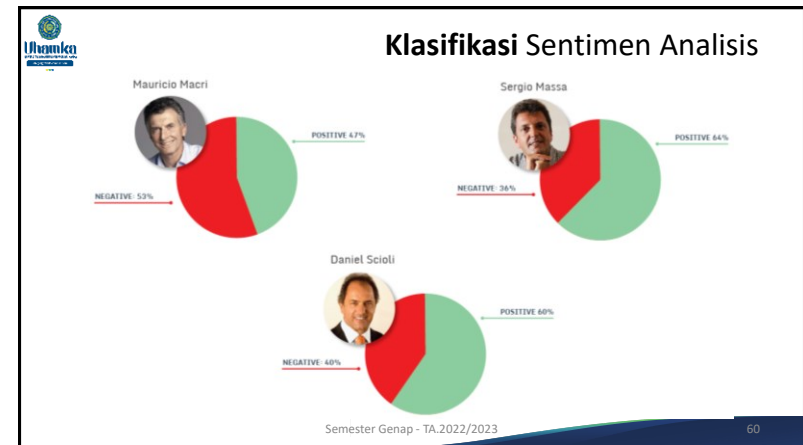
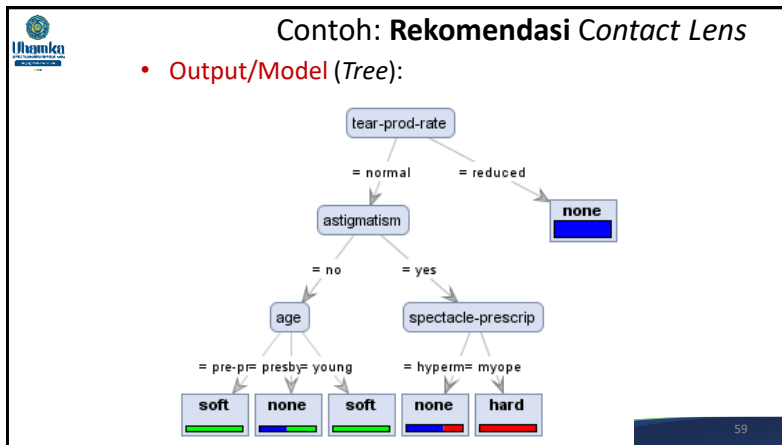
Contoh: Rekomendasi Contact Lens

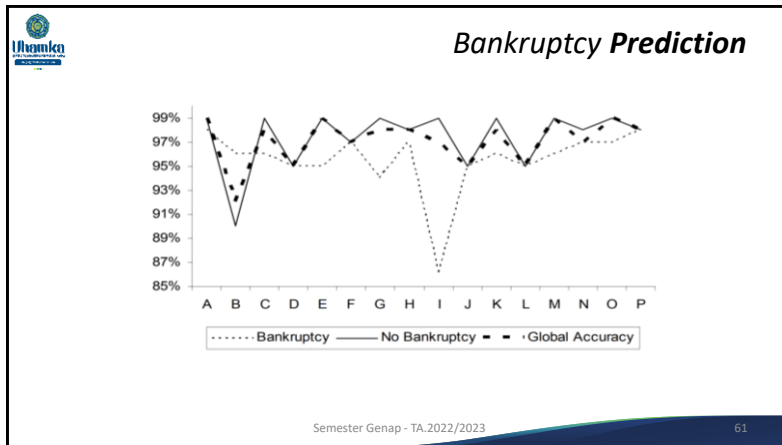
- **Input:**

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

Semester Genap - TA.2022/2023

58





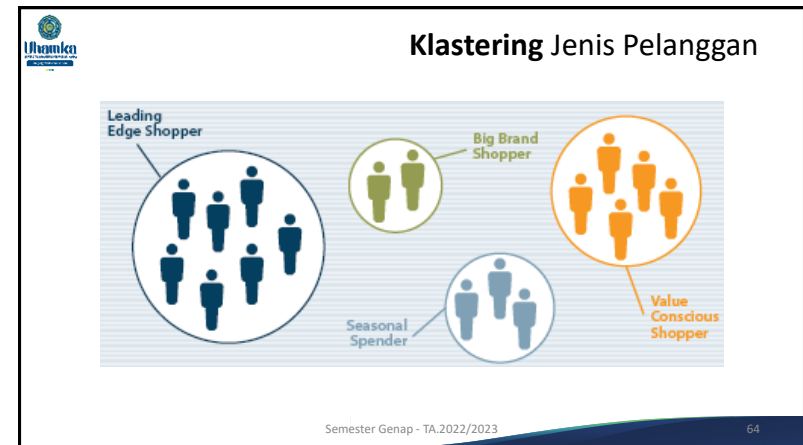
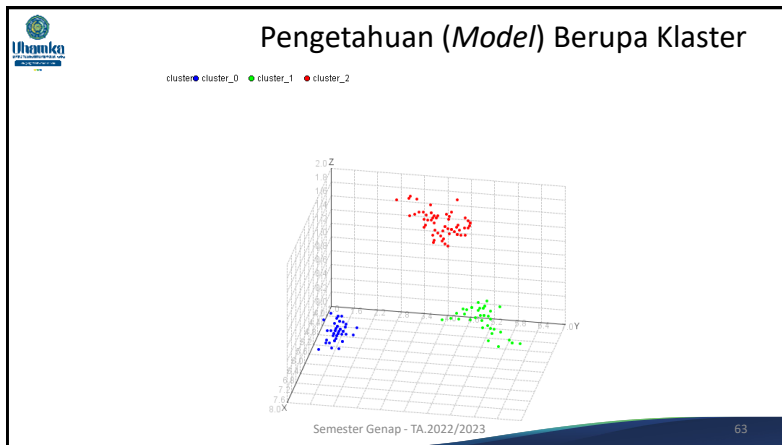
4. Klastering Bunga Iris

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)

Semester Genap - TA.2022/2023 62





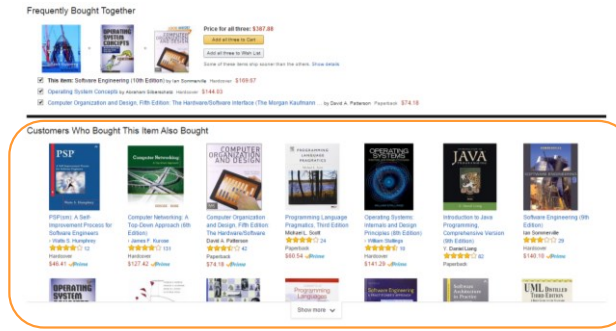
Contoh Aturan Asosiasi

- Algoritma *association rule* (aturan asosiasi) adalah algoritma yang menemukan atribut yang “**muncul bersamaan**”
- Contoh, pada hari Kamis malam, 1000 pelanggan telah melakukan belanja di supermarket 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** = $50/200 = 25\%$
- Algoritma association rule diantaranya adalah: **A priori algorithm**, **FP-Growth algorithm**, **GRI algorithm**

Semester Genap - TA.2022/2023
69



Aturan Asosiasi di Amazon.com



The screenshot shows a 'Frequently Bought Together' section on Amazon.com. It displays a grid of book covers with their titles and prices. Below this, there is a section titled 'Customers Who Bought This Item Also Bought' which shows a grid of related book covers. The books are primarily computer science and programming related, such as 'Introduction to Java Programming', 'Computer Organization and Design', and 'Software Engineering'.

Semester Genap - TA.2022/2023
70



Heating Oil Consumption

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

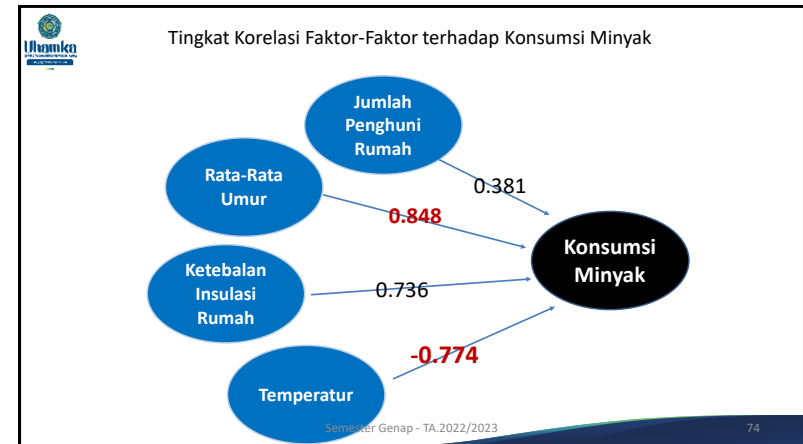
- Insulation**: Ketebalan insulasi rumah
- Temperatur**: Suhu udara sekitar rumah
- Heating Oil**: Jumlah konsumsi minyak pertahun perumah
- Number of Occupant**: Jumlah penghuni rumah
- Average Age**: Rata-rata umur penghuni rumah
- Home Size**: Ukuran rumah

Semester Genap - TA.2022/2023
71



Row No.	Insulation	Temperature	Heating_Oil	Num_Occup...	Avg_Age	Home_Size
1	6	74	132	4	23.800	4
2	10	43	263	4	56.700	4
3	3	81	145	2	28	6
4	9	50	196	4	45.100	3
5	2	80	131	5	20.800	2
6	5	76	129	3	21.500	3
7	5	72	131	4	23.500	3
8	6	88	161	2	38.200	6
9	5	77	184	3	42.500	3
10	10	42	225	3	51.100	1
11	6	90	178	2	42.100	2
12	3	83	121	1	19.800	2
13	10	43	186	5	45.100	6
14	8	59	206	2	50.100	8
15	4	86	179	5	41.400	6

Semester Genap - TA.2022/2023
72



Insight Law (Data Mining Law 6)

Data mining amplifies perception in the business domain

- 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
- 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
 - Thus data mining is, or assists as part of, a perceptual process
 - Data mining algorithms reveal patterns that are not normally visible to human perception
- Within the data mining process, the human problem solver interprets the results of data mining algorithms and integrates them into their business understanding

Semester Genap - TA.2022/2023

Referensi

- Ian H Witten, Eibe Frank, Mark A Hall. (2016). *Data Mining - Practical Machine Learning Tools and Techniques*, 4th Edition. Morgan-Kaufman.
- Hofmann, Markus., Klinkenberg, Ralf. (2016). *Data Mining Use Cases and Business Analytics Application*. CRC Press.
- Crisholm, Andrew. (2013). *Exploring Data With Rapidminer*. Packt Publishing.
- Mehmed Kantardzic. (2020). *Data Mining - Concepts, Models, Methods, and Algorithms*, 3rd edition. Wiley - IEEE Press.
- rapidminer.com, 2021., Rapidminer Manual and Documentation, <https://docs.rapidminer.com/latest/studio/operators/>.

Semester Genap - TA.2022/2023

Terima Kasih

KEDATANGAN BULAN RAMADHAN

عَنْ أَبِي هُرَيْرَةَ رَضِيَ اللَّهُ عَنْهُ يَقُولُ قَالَ رَسُولُ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ
إِذَا دَخَلَ شَهْرُ رَمَضَانَ فَتُخْتَلَفُ أَبْوَابُ السَّمَاءِ وَتُغْلَقُ أَبْوَابُ جَهَنَّمَ وَتُسَلِّسُ
الشَّيَاطِينُ

Maksudnya :

Dari Abu Hurairah RA berkata, Rasulullah SAW bersabda: "Apabila bulan Ramadan datang, maka pintu - pintu langit dibuka sedangkan pintu-pintu Jahannam ditutup dan syaitan-syaitan dibelenggu".

(al-Bukhārī, Kitāb al-Ṣawm, Bāb Hal yuqāl ramadān aw shahr ramadān wa man ra'a kullahu wāḥid'an, No. 1800)



Data Mining



Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

● ● ●

Session 04

Jumat & Sabtu , 01 & 02 April 2022



PROGRAM STUDI TEKNIK INFORMATIKA

Sumber: <https://ft.uhamka.ac.id/prodi/informatika/>

- VISI.
 - ✓ Menjadi program studi informatika yang menghasilkan karya unggul di bidang teknologi informasi melalui pembelajaran berlandaskan prophetic teaching untuk membentuk lulusan yang cerdas secara spiritual, intelektual, emosional, dan sosial.
- MISI.
 1. Melaksanakan pendidikan di bidang informatika yang terintegrasi dengan nilai-nilai Al Islam Kemuhammadiyah.
 2. Melaksanakan penelitian untuk menghasilkan karya teknologi informasi yang unggul untuk masyarakat berkembang.
 3. Melaksanakan pengabdian kepada masyarakat untuk kesejahteraan umat.
 4. Menghasilkan mahasiswa yang berwawasan global dalam teknologi informasi.
 5. Mendorong jiwa kewirausahaan mahasiswa untuk berkiprah di tengah masyarakat dalam meningkatkan kegiatan ekonomi.
 6. Menerapkan standar mutu pendidikan informatika melalui layanan mutu yang berkesinambungan.
 7. Membangun kerja sama terpadu dengan mitra yang berkontribusi untuk kemajuan pendidikan di bidang informatika.

Semester Genap - TA.2022/2023 2



Sistem Penilaian

- SKS : 3
 - Teori 3 SKS – (14 Kali Pertemuan)
- Keaktifan (bobot 10%)
 - Terdiri dari: Aktivitas Perkuliahan (*Kehadiran, Etika, dan Sikap*).
- Project (bobot 25%)
 - Terdiri dari: Tugas Besar / Project (*Penguasaan Keterampilan*) dengan cara observasi, mengimplementasikan dan menghasilkan Proyek Akhir, serta mempresentasikannya.
- Tes (UTS = bobot 25% , UAS = bobot 40%)
 - Terdiri dari: UTS dan UAS (*Penguasaan Pengetahuan*) dengan cara tes tertulis maupun tes praktikum.

Semester Genap - TA.2022/2023 3



Tujuan/Objectives (Session.04)

- Objectives (Tujuan Instruksional Umum).
 - RPS/SAP.
 - *Data Quality & Data Mining Tools*
- Indikator Kompetensi.
 - Mahasiswa mampu mengetahui dan memahami tentang kualitas data dan *tools* yang digunakan pada *data mining*.
 - Mahasiswa mampu mengetahui, memahami, serta dapat menjelaskan tentang perbedaan *supervised learning* dengan *unsupervised learning*, dan *data mining tools* yang digunakan pada mata kuliah ini yaitu *rapidminer studio*.

Semester Genap - TA.2022/2023 4

Muhasabah

هَلْ جَزَاءُ الْإِحْسَانِ إِلَّا الْإِحْسَانُ
فَبِأَيِّ آلَاءِ رَبِّكُمَا تُكَذِّبِينَ

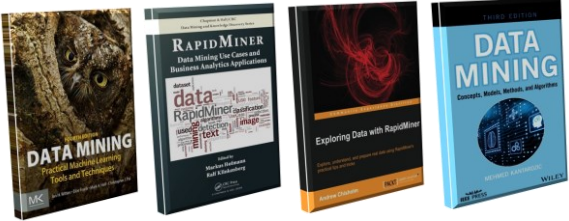
- Qs. Ar-Rahman, ayat 60 - 61
 - Sungguh, Tidak ada balasan untuk kebaikan selain kebaikan (pula).
 - Maka nikmat Tuhanmu yang manakah yang kamu dustakan ?

Semester Genap - TA.2022/2023 5

Data Mining


Uhamka
 UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA
integrity, trust, compassion

Session 04
Data Quality & Data Mining Tools

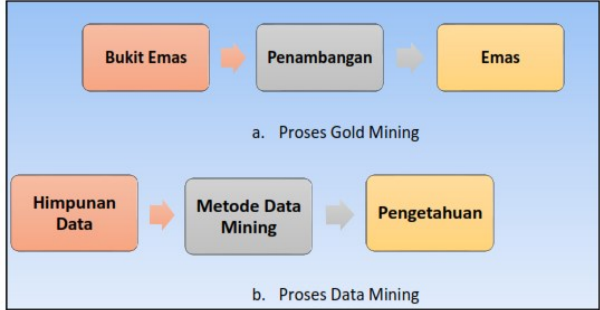


Data Quality & Data Mining Tools

Pendahuluan

Semester Genap - TA.2022/2023 7

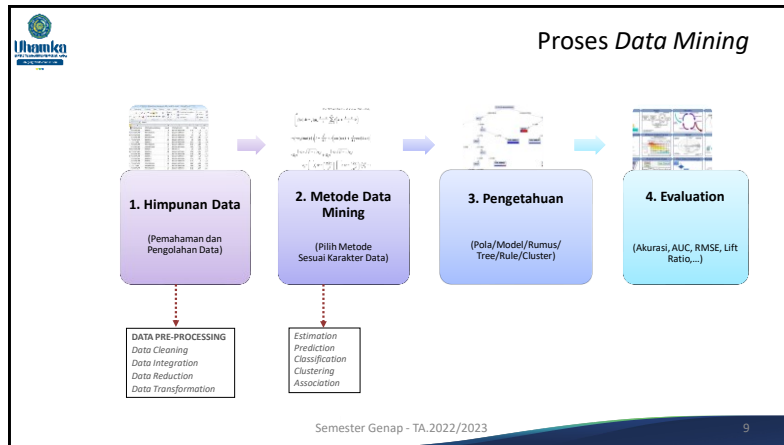
Pendahuluan



```

graph LR
    subgraph "a. Proses Gold Mining"
        A[Bukit Emas] --> B[Penambangan] --> C[Emas]
    end
    subgraph "b. Proses Data Mining"
        D[Himpunan Data] --> E[Metode Data Mining] --> F[Pengetahuan]
    end
      
```

Semester Genap - TA.2022/2023 8



Pendahuluan
(lanjutan)

- Sebelumnya kita mengetahui bahwa yang akan ditambah atau digali dalam tanda kutip, adalah Himpunan Data/Basis Data (*database*).
- Database* (himpunan data) itulah, yang kemudian akan diekstraksi menjadi sebuah pengetahuan, baik Pola, Klaster, *Decision Tree*, dan lain-lain.
- Sebelum kita melakukan proses *data mining* tentunya kita terlebih dahulu mengetahui beberapa elemen dalam sebuah himpunan data seperti pada gambar di bawah ini:

Semester Genap - TA.2022/2023 10

Pendahuluan
(lanjutan)

Tabel 1.1: Contoh Himpunan Data

Attribut / Feature/Selection					Label/Class/ Target
No	NAMA	NILAI IPK	NILAI ABSENSI	NILAI ETIKA	Kel
1	Dini	0.25	73.6	79.3	Gagal
2	Dino	3.75	98.9	87	Lulus
3	Dina	3.85	99	85	Lulus
4	Dani	0.56	60.3	65	Gagal
5	Dana	3.15	95.7	84.3	Lulus
6	Danu	0.35	52.6	56	Gagal
7	Doni	1.72	68.3	73	Gagal

Semester Genap - TA.2022/2023 11

Dataset (Himpunan Data)

	Attribute/Feature/Dimension				Class/Label/Target	
	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type	
1	5.1	3.5	1.4	0.2	Iris setosa	Record/ Object/ Sample/ Tuple/ Data
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	Nominal
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	Numerik
102	5.8	2.7	5.1	1.9	Iris virginica	
103	7.1	3.0	5.9	2.1	Iris virginica	

Semester Genap - TA.2022/2023 12



Pendahuluan (lanjutan)

- **Attribut**, adalah deskripsi data yang dapat mengidentifikasi entitas. *Field*, adalah lokasi penyimpanan. *Record*, adalah kumpulan dari berbagai *field* yang saling berhubungan.
- **Class/Label/Target**, dapat disebut sebagai atribut keputusan.

Semester Genap - TA.2022/2023

13



Pendahuluan (lanjutan)

- Pada *Data Mining*, secara garis besar terdapat 2 (dua) tipe data yang harus dipahami, yaitu:
 1. **Numeric**
 - Merupakan tipe data yang dapat di kalkulasi, seperti operasi tambah, kurang, kali, maupun bagi.
 2. **Nominal**
 - Merupakan tipe data yang tidak dapat di kalkulasi.

Semester Genap - TA.2022/2023

14



Pendahuluan (lanjutan)

Tabel 1.2 : Tipe Data Dalam Data Mining

No	NAMA	V1	V2	V3	Ket
1	Dini	0.25	73.6	79.3	Gagal
2	Dino	3.75	98.9	87	Lulus
3	Dina	3.85	99	85	Lulus
4	Dani	0.56	60.3	65	Gagal
5	Dana	3.15	95.7	84.3	Lulus
6	Danu	0.35	52.6	56	Gagal
7	Doni	1.72	68.3	73	Gagal
8	Doño	0.75	79.4	80	Gagal

Numeric

Nominal

Semester Genap - TA.2022/2023

15



2. Metode Data Mining

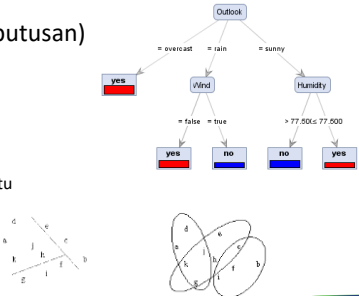
1. **Estimation (Estimasi):**
 - Linear Regression, Neural Network, Support Vector Machine, Deep Learning, etc
2. **Prediction/Forecasting (Prediksi/Peramalan):**
 - Linear Regression, Neural Network, Support Vector Machine, Deep Learning, etc
3. **Classification (Klasifikasi):**
 - Decision Tree (CART,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

Semester Genap - TA.2022/2023

16

3. Pengetahuan (Pola/Model)

1. Formula/Function (Rumus atau Fungsi Regresi)
 - $\text{WAKTU TEMPUH} = 0.48 + 0.6 \text{ JARAK} + 0.34 \text{ LAMPU} + 0.2 \text{ PESANAN}$
2. Decision Tree (Pohon Keputusan)
3. Tingkat Korelasi
4. Rule (Aturan)
 - IF $\text{ips3} \geq 2.8$ THEN lulustepatwaktu
5. Cluster (Klaster)



Semester Genap - TA.2022/2023 17

4. Evaluasi (Akurasi, Error, etc)

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)

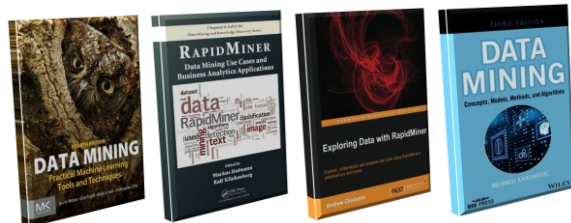
Semester Genap - TA.2022/2023 18

Kriteria Evaluasi dan Validasi Model

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

Semester Genap - TA.2022/2023 19



Data Quality & Data Mining Tools

Supervised Learning vs Unsupervised Learning

Semester Genap - TA.2022/2023 20



Supervised Learning

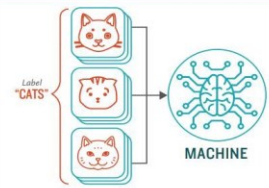
- *Supervised Learning* dalam bahasa Indonesia adalah pembelajaran yang ada *supervisor*-nya.
- Maksud disini ada *supervisor*-nya adalah *label* di tiap data nya. *Label* maksudnya adalah *tag* dari data yang ditambahkan.
- Contohnya gambar kucing di *tag* “kucing” di tiap masing masing *image* kucing, dan gambar anjing di *tag* “anjing” di tiap masing gambar anjing.
- *Data mining* kategori dapat berupa *clasification* (“anjing”, “kucing”, “beruang”, dsb), dan *regression* (berat badan, tinggi badan dsb).

Semester Genap - TA.2022/2023
21

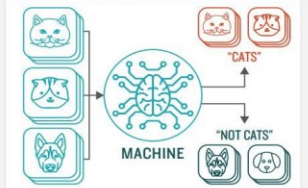


How Supervised Machine Learning Works

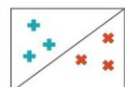
STEP 1
Provide the machine learning algorithm categorized or “labeled” input and output data from to learn



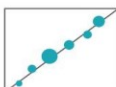
STEP 2
Feed the machine new, unlabeled information to see if it tags new data appropriately. If not, continue refining the algorithm



TYPES OF PROBLEMS TO WHICH IT'S SUITED



CLASSIFICATION
Sorting items into categories



REGRESSION
Identifying real values (dollars, weight, etc.)

Semester Genap - TA.2022/2023
22



Supervised Learning (lanjutan)

- *Supervised learning* banyak digunakan dalam memprediksi pola, dimana pola tersebut sudah ada contoh data yang lengkap, jadi pola yang terbentuk adalah hasil pembelajaran data lengkap tersebut.
- Tentunya jika kita memasukan data baru, setelah kita melakukan ETL (*Extract Transform Load*) maka kita mendapat info *feature-feature* dari *sample* baru tersebut.
- Kemudian dari *feature-feature* tersebut, di *compare* dengan *pattern clasification* dari model yang didapat dari *labeled data*.
- Setiap *label* akan di-*compare* sampai selesai, dan yang memiliki *percentage* lebih banyak akan diambil sebagai prediksi akhir.

Semester Genap - TA.2022/2023
23



Supervised Learning (lanjutan)

- Contoh algoritma yang digunakan pada *supervised learning* meliputi :

- 1) *Clasification (Categorical), and Regression (Numerical)*
- 2) *Logistic Regression*
- 3) *Model Ensemble*
- 4) *Time series*

Semester Genap - TA.2022/2023
24

Supervised Learning
(lanjutan)

SQFT	BEDS	BATHS		PRICE
3,125	5	3		\$530,000
2,100	4	2		\$460,000
1,200	3	1.5		\$250,000
3,950	6	4		???

Example of Data in ERP
Source : <https://medium.com/@gowthamy>

Semester Genap - TA.2022/2023 25

Supervised Learning
(lanjutan)

- Pada contoh tabel diatas (pada slide sebelumnya), kita akan memprediksi harga rumah dengan luas tanah 3950 *sqft* dengan 6 kamar tidur, dan 4 kamar mandi.
- Harga disini adalah target.
- Kita akan men-*train* model *machine learning* kita, dengan menggunakan 3 data yang ada dengan harga propertinya.
- Kita memasukan 3 data yang lengkap ini ke dalam *machine learning model* kemudian setelah berhasil di *train*, kita harus menghitung besar *accuracy* dari model yang di-*develop*.
- Kemudian jika kurang dari 80%, maka diperlukan *remodel*, sampai didapatkan *accuracy* sebesar-besarnya dan *loss* sekecil-kecilnya.

Semester Genap - TA.2022/2023 26

Unsupervised Learning

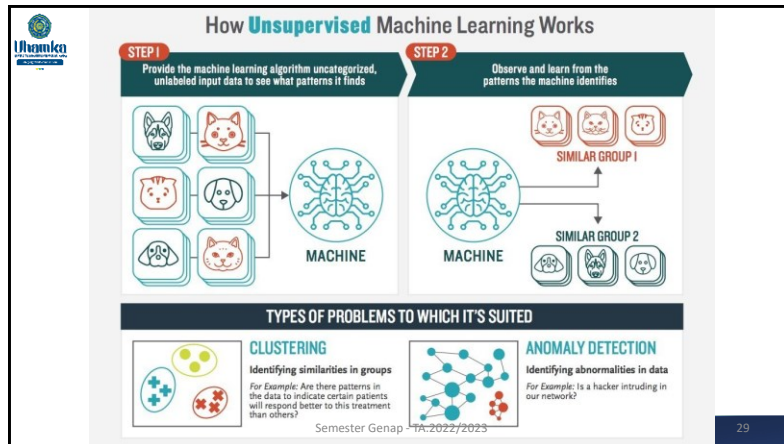
- *Unsupervised learning* memiliki keunggulan dari *supervised learning*.
- Jika *supervised learning* memiliki label sebagai dasar prediksi, baik untuk membuat *clasification* dan *regression algorithm*, maka *unsupervised learning* tidak.
- Tetapi dalam realitanya, data real itu banyak yang tidak memiliki label.
- Label kebanyakan jika data sudah masuk ke ERP, apapun bentuk ERP-nya dan bagaimana kalo datanya berupa natural input seperti suara, gambar, dan video.

Semester Genap - TA.2022/2023 27

Unsupervised Learning
(lanjutan)

- *Unsupervised learning* tidak menggunakan label dalam memprediksi target *features/variable*. Melainkan menggunakan kesamaan dari attribute-attribute yang dimiliki.
- Jika attribute dan sifat-sifat dari data-data *feature* yang diekstrak memiliki kemirip-miripan, maka akan dikelompok-kelompokan (*clustering*).
- Sehingga hal ini akan menimbulkan kelompok-kelompok (*cluster*).
- Jumlah *cluster* bisa *unlimited*. Dari kelompok kelompok itu model melabelkan, dan jika data baru mau di prediksi, maka akan dicocokkan dengan kelompok yang mirip-mirip *feature*-nya.

Semester Genap - TA.2022/2023 28



29

Unsupervised Learning (lanjutan)

- Tetapi *unsupervised learning* tidak memiliki *outcome* yang spesifik layaknya pada *supervise learning*.
- Hal ini dikarenakan tidak adanya *ground truth*/label dasar.
- Walaupun begitu, *unsupervised learning* masih dapat memprediksi dari ketidakadaan label dari kemiripan *attribute* yang dimiliki data.

Semester Genap - TA.2022/2023

30

Unsupervised Learning (lanjutan)

- Algoritma yang digunakan di *unsupervised learning* :

- 1) *Clustering*
- 2) *Anomaly Detection*
- 3) *Training Model*
- 4) *Association Discovery*

Semester Genap - TA.2022/2023

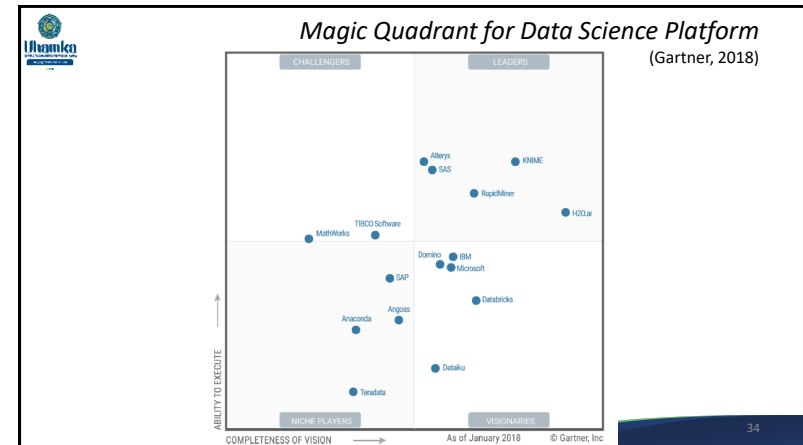
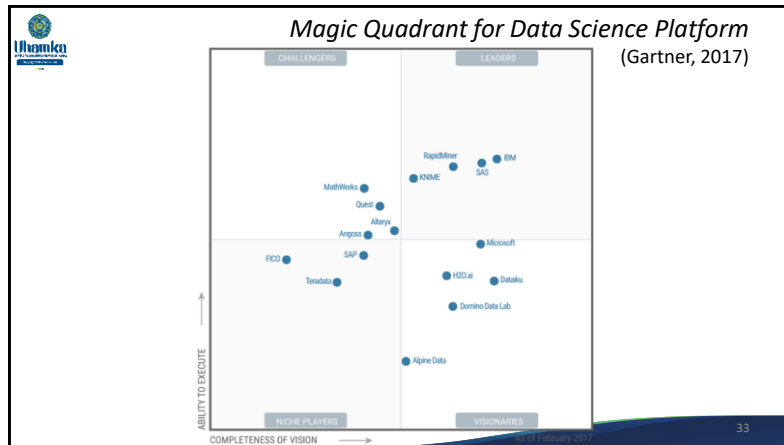
31

Data Quality & Data Mining Tools

Data Mining Tools (Rapidminer Studio)

Semester Genap - TA.2022/2023

32



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.

Semester Genap - TA.2022/2023

35

Fitur Rapidminer Studio



- 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.

Semester Genap - TA.2022/2023

36



Atribut Pada Rapidminer Studio

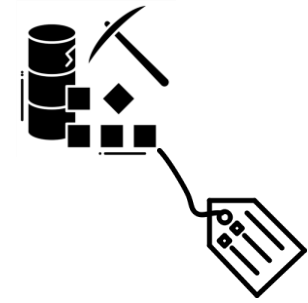
1. **Atribut:** karakteristik atau fitur dari data yang menggambarkan sebuah proses atau situasi.
 - ID, atribut biasa
2. **Atribut target:** atribut yang menjadi tujuan untuk diisi oleh proses *data mining*.
 - *Label, cluster, weight*

Semester Genap - TA.2022/2023
37



Tipe Nilai Atribut pada Rapidminer Studio

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



Semester Genap - TA.2022/2023
38



Data dan Format Data

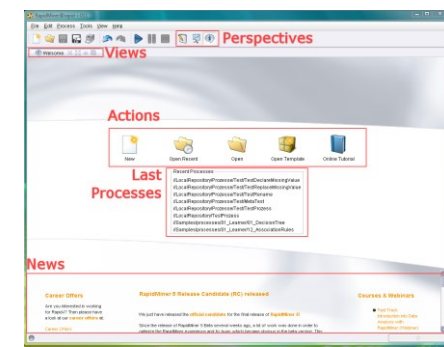
- Data menyebutkan obyek-obyek dari sebuah konsep.
 - Ditunjukkan sebagai baris dari tabel
- Metadata menggambarkan karakteristik dari konsep tersebut.
 - Ditunjukkan sebagai kolom dari tabel
- Dukungan Format data.
 - Oracle, IBM DB2, Microsoft SQL Server, MySQL, PostgreSQL, Ingres, Excel, Access, SPSS,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

Semester Genap - TA.2022/2023
39



Perspectives dan View

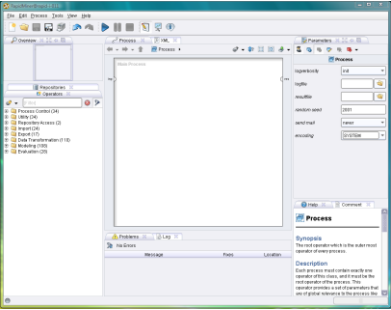
1. Perspektif Selamat Datang (*Welcome perspective*)
2. Perspektif Desain (*Design perspective*)
3. Perspektif Hasil (*Result perspective*)



Semester Genap - TA.2022/2023
40

Perspektif Desain

- Perspektif pusat di mana semua proses analisa dibuat dan dimanage
- Pindah ke Perspektif Desain dengan:
 - Klik tombol paling kiri
 - Atau gunakan menu View → Perspectives → Design



Semester Genap - TA.2022/2023 41

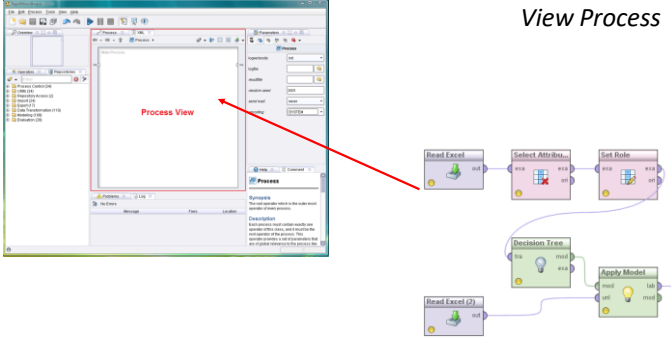
View Operator

- Process Control**
Untuk mengontrol aliran proses, seperti *loop* atau *conditional branch*
- Utility**
Untuk mengelompokkan *subprocess*, juga *macro* dan *logger*
- Repository Access**
Untuk membaca dan menulis repositori
- Import**
Untuk membaca data dari berbagai format eksternal
- Export**
Untuk menulis data ke berbagai format eksternal
- Data Transformation**
Untuk transformasi data dan metadata
- Modelling**
Untuk proses data mining yang sesungguhnya seperti klasifikasi, regresi, *clustering*, aturan asosiasi dll
- Evaluation**
Untuk menghitung kualitas dan perfomansi dari model



Semester Genap - TA.2022/2023 42

View Process



Semester Genap - TA.2022/2023 43

View Parameter

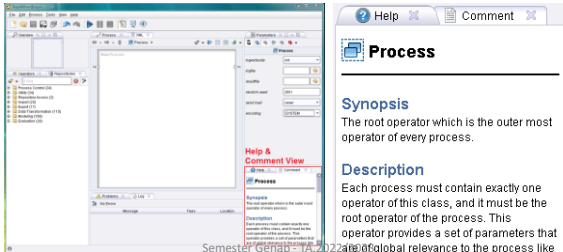
- Operator kadang memerlukan parameter untuk dapat berfungsi
- Setelah operator dipilih di view Proses, maka parameternya akan ditampilkan pada view ini



Semester Genap - TA.2022/2023 44

View Help dan View Comment

- View Help menampilkan deskripsi dari operator
- View Comment menampilkan komentar yang dapat diedit terhadap operator



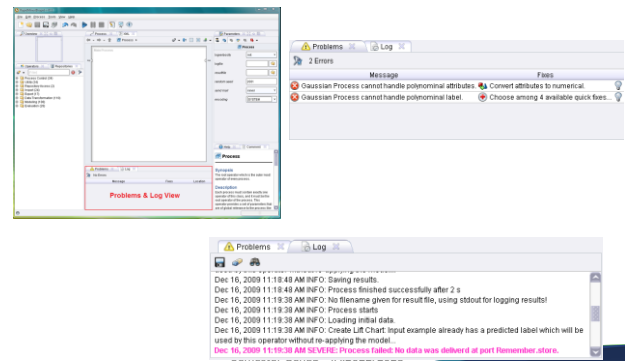
Process

Synopsis
The root operator which is the outer most operator of every process.

Description
Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like

Semester Genap - TA.2022/2023 45

View Problems dan View Log



Problems & Log View

2 Errors

Message	Files
Gaussian Process cannot handle polynomial attributes. Convert attributes to numerical.	
Gaussian Process cannot handle polynomial label. Choose among 4 available quick fixes.	

Log

Dec 16, 2009 11:18:48 AM INFO: Saving results.
 Dec 16, 2009 11:18:48 AM INFO: Process finished successfully after 2 s
 Dec 16, 2009 11:19:38 AM INFO: No filename given for result file, using default for logging results!
 Dec 16, 2009 11:19:38 AM INFO: Process starts
 Dec 16, 2009 11:19:38 AM INFO: Loading initial data
 Dec 16, 2009 11:19:38 AM INFO: Create Lift Chart. Input example already has a predicted label which will be used by this operator without re-applying the model.
 Dec 16, 2009 11:19:38 AM SEVERE: Process failed: No data was delivered at port Remember store.

Semester Genap - TA.2022/2023 46

Operator dan Process

- Proses *data mining* pada dasarnya adalah proses analisa yang berisi alur kerja dari komponen *data mining*.
- Komponen dari proses ini disebut operator, yang didefinisikan dengan:
 - Deskripsi input
 - Deskripsi output
 - Aksi yang dilakukan
 - Parameter yang diperlukan

Semester Genap - TA.2022/2023 47

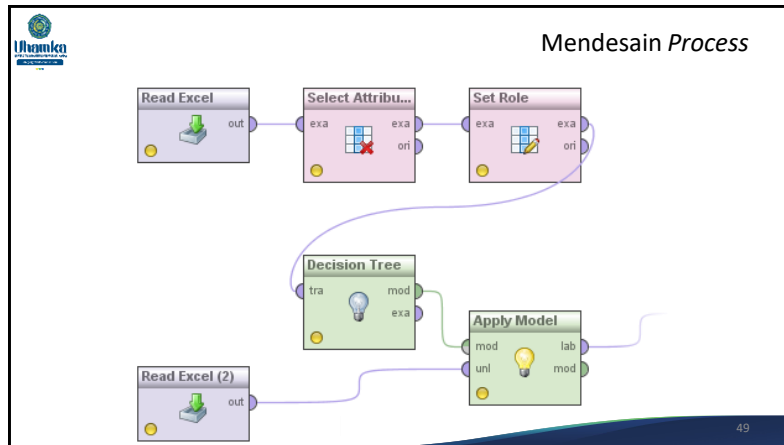
Operator dan Process (lanjutan)

- Sebuah operator bisa disambungkan melalui port masukan (kiri) dan port keluaran (kanan)



- Indikator status dari operator:
 - Lampu** status: merah (tak tersambung), kuning (lengkap tetapi belum dijalankan), hijau (sudah berhasil dijalankan)
 - Segitiga warning**: bila ada pesan status
 - Breakpoint**: bila ada breakpoint sebelum/sesudahnya
 - Comment**: bila ada komentar
 - Subprocess**: bila mempunyai subprocess

Semester Genap - TA.2022/2023 48



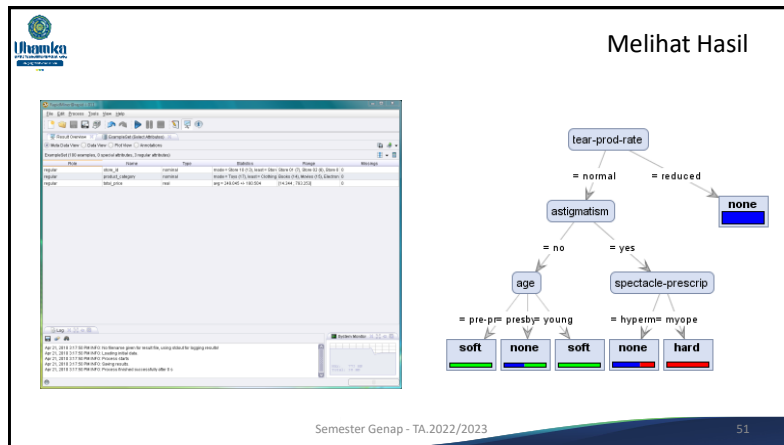
Menjalankan Process

Proses dapat dijalankan dengan:

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

Semester Genap - TA.2022/2023

50

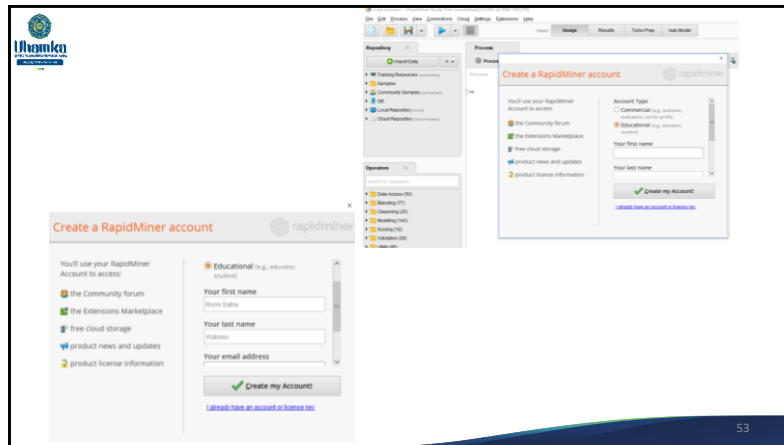


Instalasi dan Registrasi Lisensi Rapidminer

Instal Rapidminer Studio trial version (berlaku hanya 30 hari)

- Registrasi account di rapidminer.com, dan lakukan dapatkan lisensi **Educational Program** (masa berlaku 1 tahun, dan dapat diperpanjang) untuk mengolah data tanpa batasan *record*

52



53

Referensi

- Ian H Witten, Eibe Frank, Mark A Hall. (2016). *Data Mining - Practical Machine Learning Tools and Techniques*, 4th Edition. Morgan-Kaufman.
- Hofmann, Markus., Klinkenberg, Ralf. (2016). *Data Mining Use Cases and Business Analytics Application*. CRC Press.
- Crisholm, Andrew. (2013). *Exploring Data With Rapidminer*. Packt Publishing.
- Mehmed Kantardzic. (2020). *Data Mining - Concepts, Models, Methods, and Algorithms*, 3rd edition. Wiley - IEEE Press.
- rapidminer.com, 2021,. Rapidminer Manual and Documentation, <https://docs.rapidminer.com/latest/studio/operators/>.

Semester Genap - TA.2022/2023

54





Data Mining



Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

● ● ●

Session 05



PROGRAM STUDI TEKNIK INFORMATIKA

Sumber: <https://ft.uhamka.ac.id/prodi/informatika/>

- VISI.
 - ✓ Menjadi program studi informatika yang menghasilkan karya unggul di bidang teknologi informasi melalui pembelajaran berlandaskan prophetic teaching untuk membentuk lulusan yang cerdas secara spiritual, intelektual, emosional, dan sosial.
- MISI.
 1. Melaksanakan pendidikan di bidang informatika yang terintegrasi dengan nilai-nilai Al Islam Kemuhammadiyah.
 2. Melaksanakan penelitian untuk menghasilkan karya teknologi informasi yang unggul untuk masyarakat berkembang.
 3. Melaksanakan pengabdian kepada masyarakat untuk kesejahteraan umat.
 4. Menghasilkan mahasiswa yang berwawasan global dalam teknologi informasi.
 5. Mendorong jiwa kewirausahaan mahasiswa untuk berkiprah di tengah masyarakat dalam meningkatkan kegiatan ekonomi.
 6. Menerapkan standar mutu pendidikan informatika melalui layanan mutu yang berkesinambungan.
 7. Membangun kerja sama terpadu dengan mitra yang berkontribusi untuk kemajuan pendidikan di bidang informatika.

Semester Genap - TA.2022/2023 2



Sistem Penilaian

- SKS : 3
 - Teori 3 SKS – (14 Kali Pertemuan)
- Keaktifan (bobot 10%)
 - Terdiri dari: Aktivitas Perkuliahan (*Kehadiran, Etika, dan Sikap*).
- Project (bobot 25%)
 - Terdiri dari: Tugas Besar / Project (*Penguasaan Keterampilan*) dengan cara observasi, mengimplementasikan dan menghasilkan Proyek Akhir, serta mempresentasikannya.
- Tes (UTS = bobot 25% , UAS = bobot 40%)
 - Terdiri dari: UTS dan UAS (*Penguasaan Pengetahuan*) dengan cara tes tertulis maupun tes praktikum.

Semester Genap - TA.2022/2023 3



Tujuan/Objectives (Session.05)

- Objectives (Tujuan Instruksional Umum).
 - RPS/SAP.
 - Klasifikasi (*part.1*)
- Indikator Kompetensi.
 - Mahasiswa mampu mengetahui dan memahami tentang teknik klasifikasi pada *data mining*.
 - Mahasiswa mampu mengetahui, memahami, serta dapat menjelaskan tentang penggunaan *Decision Tree* dan *Overfitting Model*, evaluasi kinerja pengklasifikasi, serta metode untuk membandingkan pengklasifikasi didalam *data mining*.

Semester Genap - TA.2022/2023 4



Muhasabah

إِنَّ الْمُصَّدِّقِينَ وَالْمُصَّدِّقَاتِ وَأَقْرَضُوا اللَّهَ قَرْضًا حَسَنًا يُضَعَّفُ لَهُمْ وَلَهُمْ أَجْرٌ كَرِيمٌ

- Qs. Al Hadid, ayat 18
 - Sesungguhnya orang-orang yang bersedekah, baik laki-laki maupun perempuan, dan meminjamkan (kepada) Allah pinjaman yang baik, akan dilipatgandakan (balasannya) kepada mereka dan baginya (diberikan) ganjaran yang sangat mulia (surga).

Semester Genap - TA.2022/2023

5



Data Mining



Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

Session 05

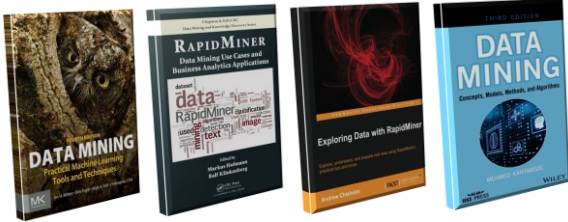
Klasifikasi (*part.1*)

Semester Genap - TA.2022/2023

6



Klasifikasi (*part.1*)



Pendahuluan

Semester Genap - TA.2022/2023

7



Pendahuluan

- Bagian sangat penting dalam *data mining*, adalah teknik klasifikasi. Yaitu bagaimana mempelajari sekumpulan data sehingga dihasilkan aturan yang dapat mengklasifikasi atau mengenali data-data baru yang belum pernah dipelajari.
- Klasifikasi dapat didefinisikan sebagai proses untuk menyatakan suatu objek data sebagai salah satu kategori (kelas) yang telah didefinisikan sebelumnya (Zaki et al, 2013).
- Klasifikasi banyak digunakan dalam berbagai aplikasi, diantaranya adalah klasifikasi profil pelanggan, deteksi kecurangan (*fraud detection*), diagnosis medis, prediksi penjualan, dsb.

Semester Genap - TA.2022/2023

8



Pendahuluan (lanjutan)

- Teknik Klasifikasi ini cocok digunakan dalam mendeskripsikan *dataset* dengan tipe data dari suatu himpunan data, yaitu biner atau nominal.
- Adapun kekurangan dari teknik ini yaitu tidak tepat untuk himpunan data ordinal, karena pendekatan-pendekatan yang digunakan secara implisit dalam kategori data.

Semester Genap - TA.2022/2023

9



Pendahuluan (lanjutan)

- Ada beberapa teknik klasifikasi yang digunakan sebagai solusi pemecahan kasus, diantaranya yaitu:
 - Algoritma C4.5
 - Algoritma K-Nearest Neighbor
 - ID3
 - Naive Bayesian *Classification*
 - CART (*Classification And Regression Tree*)
 - Dan lain-lain
- Output atau keluaran dari metode klasifikasi ini biasanya dalam bentuk "*Decision Tree* (pohon keputusan)".

Semester Genap - TA.2022/2023

10



Pendahuluan (lanjutan)

- Model klasifikasi dapat dibangun berdasarkan pengetahuan seorang pakar (ahli).
- Namun mengingat himpunan data yang relative besar, model klasifikasi lebih sering dibangun menggunakan teknik pembelajaran dalam bidang *machine learning*.
- Terdapat banyak teknik klasifikasi yang telah diusulkan para ahli, yang dapat dikelompokkan ke dalam 2 kategori, yang masing-masing teknik mempunyai kelebihan dan kekurangan masing-masing:
 - Teknik Klasifikasi Global
 - Memperhitungkan semua data latih.
 - Teknik Klasifikasi Lokal
 - Hanya memperhitungkan sebagian data latih.

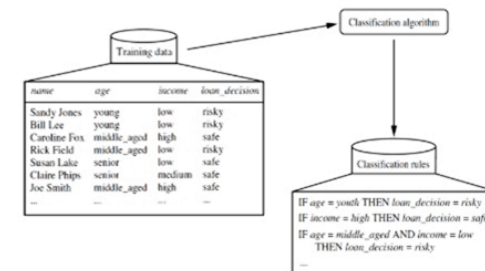
Semester Genap - TA.2022/2023

11



Pendahuluan (lanjutan)

Gambaran Umum Data Mining Klasifikasi



Semester Genap - TA.2022/2023

12

Pendahuluan
(lanjutan)

Gambaran Umum Data Mining Klasifikasi (2)

name	age	income	loan_decision
Juan Bello	senior	low	safe
Sylvia Crest	middle_aged	low	risky
Anne Yee	middle_aged	high	safe
...	...	...	...

(John Henry, middle_aged, low)
Loan decision?
↓
risky

Semester Genap - TA.2022/2023 13

Klasifikasi (part.1)

Tahapan & Proses

Semester Genap - TA.2022/2023 14

Tahapan & Proses Klasifikasi

- Terdapat sejumlah record (*training set*)
 - Tiap *record* terdiri dari sejumlah atribut dan salah satunya merupakan atribut kelas /label.
- Cari model untuk atribut kelas sebagai fungsi nilai dari atribut lain.
- Tujuannya: *record* yang belum diketahui kelasnya akan diberi label seakurat mungkin.
 - Untuk mengetahui dan menentukan tingkat akurasi model maka digunakan *test set*.
 - Biasanya *dataset* akan dibagi menjadi dua bagian yang pertama untuk pembangunan model (*training*), dan yang kedua sebagai data untuk memvalidasi model tersebut (*test set*)

Semester Genap - TA.2022/2023 15

Tahapan & Proses Klasifikasi
(lanjutan)

- Terdapat 2 (dua) langkah didalam klasifikasi, yaitu:
 - 1. Pembangunan Model**
 - Dengan mendeskripsikan kelas yang sudah ditentukan sebelumnya.
 - 2. Penggunaan Model**
 - Untuk mengklasifikasikan objek yang selanjutnya, atau yang belum diketahui.

Semester Genap - TA.2022/2023 16

Tahapan & Proses Klasifikasi (lanjutan)

1. Pembangunan Model

- Dengan mendeskripsikan kelas yang sudah ditentukan sebelumnya.
- Tiap *tuple*/sample diasumsikan termasuk dalam salah satu kelas yang ditunjukkan oleh atribut kelas.
- *Tuple* atau *record* yang digunakan untuk pembangunan model, disebut sebagai *training set*.
- Model yang dihasilkan dapat berupa:
 - *rule* klasifikasi,
 - *decision trees*,
 - atau rumus matematika

Semester Genap - TA.2022/2023 17

Tahapan & Proses Klasifikasi (lanjutan)

Training Data

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Classifier (Model)

Algoritma klasifikasi

IF rank = 'professor'
OR years > 6
THEN tenured = 'yes'

Gambar -> Pembangunan Model
Semester Genap - TA.2022/2023 18

Tahapan & Proses Klasifikasi (lanjutan)

2. Penggunaan Model :

- Untuk mengklasifikasikan objek yang selanjutnya, atau yang belum diketahui.
- Perhitungan akurasi model:
 - Label *test set* yang sebenarnya akan dibandingkan dengan hasil pelabelan *test set* dari penggunaan model.
 - Tingkat akurasi adalah persentase *test set* yang terklasifikasi dengan benar oleh model.
 - *Test set* independen terhadap *training set*, jika tidak maka kondisi *over-fitting* akan terjadi.
- Jika akurasi dapat diterima, maka model dapat digunakan untuk mengklasifikasikan data yang labelnya belum diketahui.

Semester Genap - TA.2022/2023 19

Tahapan & Proses Klasifikasi (lanjutan)

Testing Data

NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Classifier

Unseen Data
(Jeff, Professor, 4)

Tenured? **Yes**

Gambar -> Penggunaan Model untuk Prediksi
Semester Genap - TA.2022/2023 20

Teknik Klasifikasi

- Beberapa teknik klasifikasi, yaitu:
 - Metode Berdasarkan *Decision Tree*
 - Metode Berdasarkan *Rule*
 - Memory based reasoning*
 - Neural Networks*
 - Naïve Bayes* dan *Bayesian Belief Networks*
 - Support Vector Machines*

Semester Genap - TA.2022/2023 21

Klasifikasi (part.1)

Pendahuluan

Semester Genap - TA.2022/2023 22

Contoh Decision Tree

age	income	student	credit rating	buys computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

Training Data

Model: Decision Tree

Class

Splitting Attributes

```

graph TD
    A[age?] -->|<=30| B[student?]
    A -->|30...40| C[yes]
    A -->|>40| D[credit rating?]
    B -->|no| E[no]
    B -->|yes| F[yes]
    D -->|excellent| G[no]
    D -->|fair| H[yes]
  
```

Semester Genap - TA.2022/2023 23

Contoh Lain Decision Tree

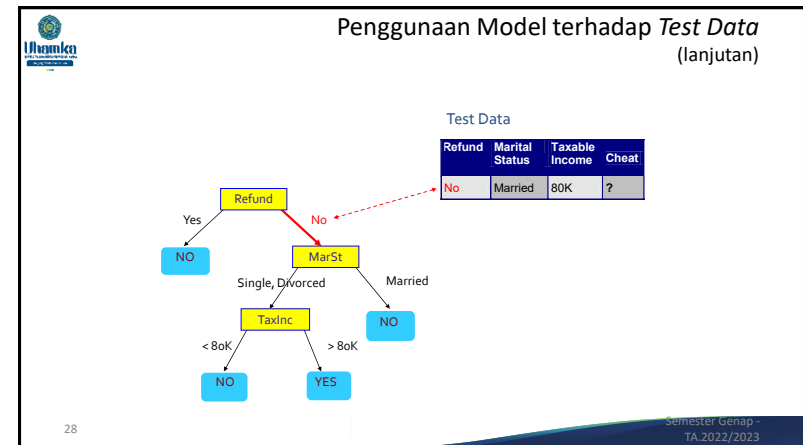
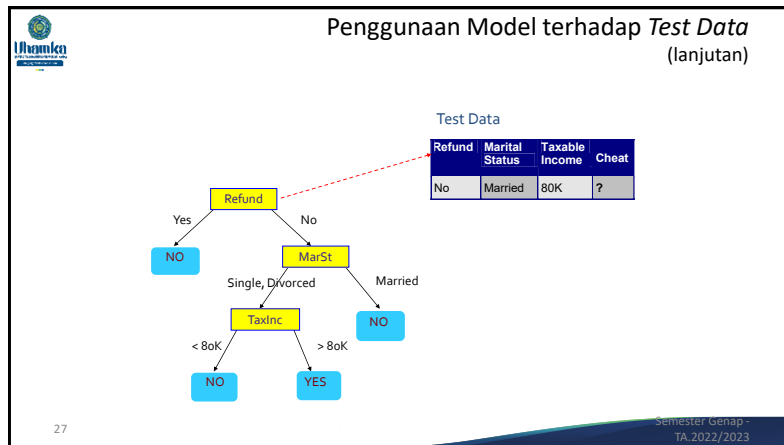
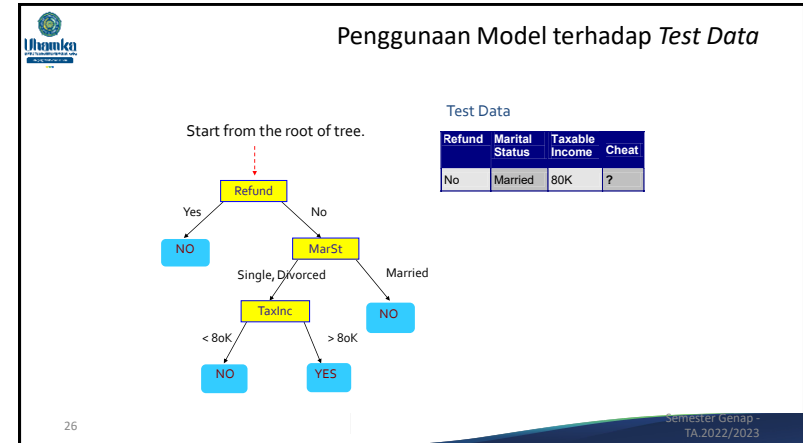
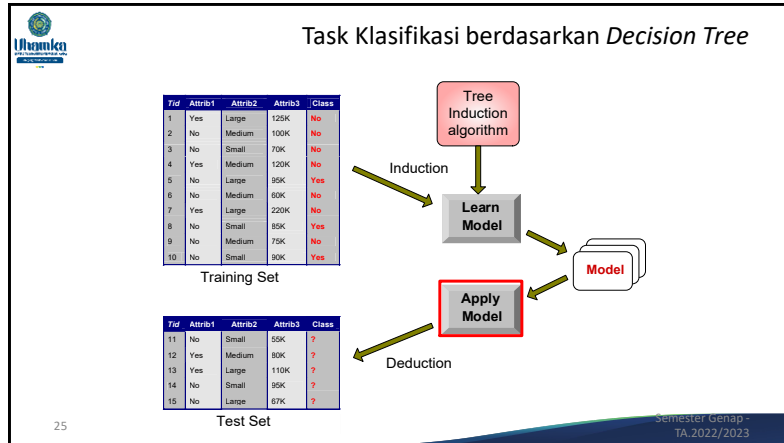
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund: categorical, Marital Status: categorical, Taxable Income: continuous, Cheat: class

```

graph TD
    A[MarSt] -- Married --> B[NO]
    A -- Single, Divorced --> C[Refund]
    C -- Yes --> D[NO]
    C -- No --> E[TaxInc]
    E -- < 80K --> F[NO]
    E -- > 80K --> G[YES]
  
```

Semester Genap - TA.2022/2023 24



Penggunaan Model terhadap *Test Data* (lanjutan)

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Semester Genap - TA. 2022/2023

Penggunaan Model terhadap *Test Data* (lanjutan)

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Semester Genap - TA. 2022/2023

Penggunaan Model terhadap *Test Data* (lanjutan)

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Semester Genap - TA. 2022/2023

Klasifikasi (part.1)

Induksi Decision Tree

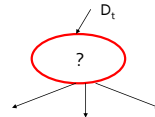
Semester Genap - TA. 2022/2023

Induksi Decision Tree

- Algoritma:
 - Algoritma Hunt's (salah satu yang paling awal)
 - CART
 - ID3, C4.5
 - SLIQ, SPRINT

Semester Genap - TA.2022/2023

Struktur Umum Algorithm Hunts

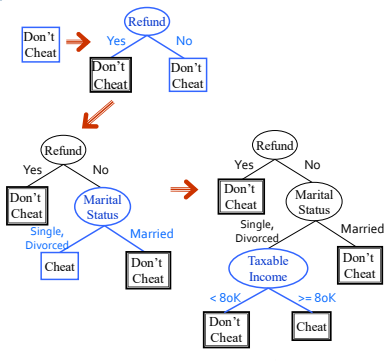


- Let D_t be the set of training records that reach a node t
- General Procedure:
 - If D_t contains records that belong the same class y_t , then t is a leaf node labeled as y_t
 - If D_t is an empty set, then t is a leaf node labeled by the default class, y_d
 - If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets. Recursively apply the procedure to each subset.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Semester Genap - TA.2022/2023

Algoritma Hunt's



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Semester Genap - TA.2022/2023

Induksi Tree

- Algoritma Dasar (algoritma *greedy*)
 - Tree dibangun secara *top-down recursive divide-and-conquer*
 - Awalnya, semua *training examples* sebagai *root*
 - Atribut kategoris (jika kontinu, maka harus didiskritkan lebih dulu)
 - Examples* dipartisi secara rekursif berdasarkan atribut tertentu
 - Atribut Test dipilih berdasarkan perhitungan heuristik atau statistik (contoh: *information gain*)
- Kondisi untuk berhenti mempartisi:
 - Semua *samples* di *node* tersebut merupakan kelas yang sama
 - Tidak ada atribut lagi untuk dipartisi *majority voting* digunakan untuk mengklasifikasikan *leaf*
 - Samples* sudah habis

Semester Genap - TA.2022/2023

Bagaimana menentukan kondisi *test* ?

- Tergantung pada tipe atribut
 - Nominal
 - Ordinal
 - Kontinu
- Tergantung jumlah cara untuk split
 - 2-way split
 - Multi-way split

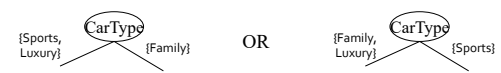
37 Semester Genap - TA 2022/2023

Splitting berdasar Atribut Nominal

❖ **Multi-way split:** jumlah partisi sesuai dengan jumlah distinct values.



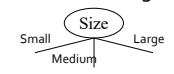
❖ **Binary split:** membagi nilai menjadi 2 subset dan mencari yang paling optimal



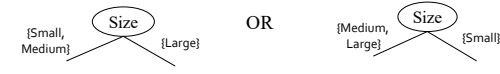
38 Semester Genap - TA 2022/2023

Splitting berdasar Atribut Ordinal

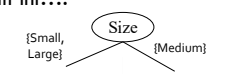
■ **Multi-way split:** jumlah partisi sesuai dengan jumlah *distinct values*.



■ **Binary split:** membagi nilai menjadi 2 subset dan mencari yang paling optimal.



■ Bagaimana split ini....



39 Semester Genap - TA 2022/2023

Splitting berdasar Atribut Kontinu

■ **Diskretisasi** menjadi bentuk atribut kategoris ordinal

- *Static – discretize once at the beginning*
- *Dynamic – ranges* bisa ditentukan dengan menggunakan cara *equal interval bucketing*, *equal frequency bucketing (percentiles)*, atau *clustering*.

■ **Binary Decision:** $(A < v)$ or $(A \geq v)$

- Memperhitungkan semua kemungkinan split dan mencari yang paling baik.
- Secara komputasi harus intensif.

40 Semester Genap - TA 2022/2023

Splitting berdasar Atribut Kontinu

(i) Binary split

(ii) Multi-way split

Semester Genap - TA 2022/2023

Tree Induction

- *Greedy strategy.*
 - Split the records based on an attribute test that optimizes certain criterion.
- *Issues*
 - Determine how to split the records
 - How to specify the attribute test condition?
 - How to determine the best split?
 - Determine when to stop splitting.

Semester Genap - TA 2022/2023

Menentukan Best Split

Sebelum Splitting: 10 records of class 0, 10 records of class 1

Mana yang terbaik?

Semester Genap - TA 2022/2023

Menentukan *Best Split*

- Pendekatan *Greedy* :
 - Node dengan distribusi kelas yang **homogen** yang dipilih
- Perlu pengukuran *node impurity*:

C0: 5
C1: 5

Non-homogeneous,
High degree of impurity

C0: 9
C1: 1

Homogeneous,
Low degree of impurity

Semester Genap - TA 2022/2023



Tree Induction

- *Greedy strategy.*
 - *Split the records based on an attribute test that optimizes certain criterion.*
- *Issues*
 - *Determine how to split the records*
 - *How to specify the attribute test condition?*
 - *How to determine the best split?*
 - *Determine when to stop splitting*

45 Semester Genap - TA 2022/2023



Kriteria Berhenti untuk *Induction Tree*

- *Stop* membagi *node* seandainya semua *record* yang ada merupakan kelas yang sama.
- *Stop* membagi *node* saat semua *record* yang ada memiliki nilai atribut yang sama.
- Pemberhentian secara awal (*Early termination*)

46 Semester Genap - TA 2022/2023



Klasifikasi berdasarkan *Decision Tree*

- Keuntungan:
 - Murah untuk dibangun
 - Dapat mengklasifikasikan *unknown record* dengan sangat cepat
 - Mudah untuk menginterpretasikan tree yang berukuran kecil
 - Akurasi bisa dibandingkan dengan teknik klasifikasi lain untuk dataset yang sederhana

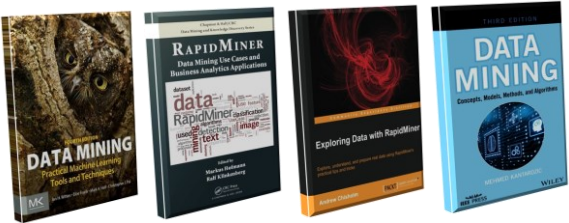
47 Semester Genap - TA 2022/2023



Contoh: C4.5

- Pembangunan Simple depth-first
- Menggunakan Information Gain
- Mengurutkan atribut kontinu di setiap node.
- Data harus cukup dalam memori.
- Tidak cocok untuk Dataset yang besar.
 - Needs out-of-core sorting.
 - You can download the software from:
<http://www.cse.unsw.edu.au/~quinlan/c4.5r8.tar.gz>

48 Semester Genap - TA 2022/2023



Klasifikasi (part.1)

Pengukuran

Semester Genap - TA.2022/2023 49

Pengukuran Node Impurity

- Gini Index
- Entropy
- Misclassification error

Semester Genap - TA.2022/2023 50

Pengukuran Impurity: GINI

- Gini Index untuk node t :

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

(NOTE: $p(j|t)$ adalah frekuensi relatif class j padanode t).

 - Maximum ($1 - 1/n_c$) jika record terdistribusi merata untuk semua kelas,
 - Minimum (0.0) jika semua records adalah satu kelas

C1	0
C2	6
Gini=0.000	

C1	1
C2	5
Gini=0.278	

C1	2
C2	4
Gini=0.444	

C1	3
C2	3
Gini=0.500	

Semester Genap - TA.2022/2023 51

Contoh Perhitungan GINI

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

C1	0
C2	6

$P(C_1) = 0/6 = 0$ $P(C_2) = 6/6 = 1$
 $Gini = 1 - P(C_1)^2 - P(C_2)^2 = 1 - 0 - 1 = 0$

C1	1
C2	5

$P(C_1) = 1/6$ $P(C_2) = 5/6$
 $Gini = 1 - (1/6)^2 - (5/6)^2 = 0.278$

C1	2
C2	4

$P(C_1) = 2/6$ $P(C_2) = 4/6$
 $Gini = 1 - (2/6)^2 - (4/6)^2 = 0.444$

Semester Genap - TA.2022/2023 52

Splitting berdasarkan GINI

- Digunakan pada CART, SLIQ, SPRINT.
- Saat *node* p dibagi jadi k partisi (*children*), kualitas *split* dihitung sebagai ,

$$GINI_{split} = \sum_{i=1}^k \frac{n_i}{n} GINI(i)$$

dimana, n_i = jumlah *record* pada *child* i,
 n = jumlah *record* pada *node* p.

Semester Genap -
TA. 2022/2023

Atribut Binary : Menghitung GINI Index

- Split jadi 2 partisi
- cek yang akan menghasilkan purity yang lebih baik

GINI(N1)
 $= 1 - (5/6)^2 - (2/6)^2$
 $= 0.194$

GINI(N2)
 $= 1 - (1/6)^2 - (4/6)^2$
 $= 0.528$

GINI(Children)
 $= 7/12 * 0.194 +$
 $5/12 * 0.528$
 $= 0.333$

Parent

	Parent
C1	6
C2	6
GINI = 0.500	

Node N1

	N1	N2
C1	5	1
C2	2	4
GINI=0.333		

Semester Genap -
TA. 2022/2023

Atribut kategoris : Menghitung GINI Index

- Untuk setiap *distinct value*, hitung distribusi kelasnya
- Gunakan *count matrix* untuk menentukan pilihan

Multi-way split

	CarType		
	Family	Sports	Luxury
C1	1	2	1
C2	4	1	1
Gini	0.393		

Two-way split
(find best partition of values)

	CarType	
	{Sports, Luxury}	{Family}
C1	3	1
C2	2	4
Gini	0.400	

	CarType	
	{Sports}	{Family, Luxury}
C1	2	2
C2	1	5
Gini	0.419	

Semester Genap -
TA. 2022/2023

Atribut kontinu : Menghitung GINI Index

- Menggunakan *Binary Decisions* berdasarkan satu nilai
- Several Choices for the splitting value
 - Number of possible splitting values = Number of distinct values
- Each splitting value has a count matrix associated with it
 - Class counts in each of the partitions, $A < v$ and $A \geq v$
- Simple method to choose best v
 - For each v_i , scan the database to gather count matrix and compute its Gini index
 - Computationally Inefficient! Repetition of work.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Semester Genap -
TA. 2022/2023

Atribut kontinu : Menghitung GINI Index...

- For efficient computation: for each attribute,
 - Sort the attribute on values
 - Linearly scan these values, each time updating the count matrix and computing gini index
 - Choose the split position that has the least gini index

Cheat	No	No	No	Yes	Yes	Yes	No	No	No	No
	60	70	75	85	90	95	100	120	125	220
Sorted Values	55	65	72	80	87	92	97	110	122	172
Split Positions	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3	0	3
No	0	7	1	6	2	5	3	4	3	0
Gini	0.420	0.400	0.375	0.343	0.417	0.400	0.300	0.343	0.375	0.420

Pilih GINI index terkecil

Semester Genap - TA.2022/2023

Mencari Best Split

Sebelum Splitting:

C0	N00
C1	N01

 → Mo

Decision Tree Structure:

- Root Node: A?
 - Yes: Node N1

C0	N10
C1	N11

 ↓ M1
 - No: Node N2

C0	N20
C1	N21

 ↓ M2
- Root Node: B?
 - Yes: Node N3

C0	N30
C1	N31

 ↓ M3
 - No: Node N4

C0	N40
C1	N41

 ↓ M4

Gain = Mo - M12 vs Mo - M34

Semester Genap - TA.2022/2023

Kriteria Splitting Alternatif berdasarkan INFO

- Entropy at a given node t:

$$Entropy(t) = -\sum_j p(j|t) \log p(j|t)$$

(NOTE: $p(j|t)$ is the relative frequency of class j at node t).
- Measures homogeneity of a node.
 - Maximum ($\log n_c$) when records are equally distributed among all classes implying least information
 - Minimum (0.0) when all records belong to one class, implying most information
- Entropy based computations are similar to the GINI index computations

Semester Genap - TA.2022/2023

Examples for computing Entropy

$$Entropy(t) = -\sum_j p(j|t) \log_2 p(j|t)$$

C1	0	P(C1) = 0/6 = 0	P(C2) = 6/6 = 1
C2	6	Entropy = -0 log 0 - 1 log 1 = -0 - 0 = 0	

C1	1	P(C1) = 1/6	P(C2) = 5/6
C2	5	Entropy = -(1/6) log ₂ (1/6) - (5/6) log ₂ (5/6) = 0.65	

C1	2	P(C1) = 2/6	P(C2) = 4/6
C2	4	Entropy = -(2/6) log ₂ (2/6) - (4/6) log ₂ (4/6) = 0.92	

Semester Genap - TA.2022/2023

Splitting Berdasarkan INFO...

■ Information Gain:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^k \frac{n_i}{n} Entropy(i) \right)$$

Parent Node, p displit menjadi k partisi
 n_i jumlah record dalam partisi i

- Measures Reduction in Entropy achieved because of the split. Pilih split yang Choose the split that achieves most reduction (maximizes GAIN)
- Digunakan dalam ID3 dan C4.5
- Kekurangan: kecenderungan membagi partisi dalam jumlah banyak, tiap partisi kecil namun 'murni'

61 Semester Genap - TA.2022/2023

Splitting Berdasarkan INFO...

■ Gain Ratio:

$$GainRatio_{split} = \frac{GAIN_{split}}{SplitINFO} \quad SplitINFO = - \sum_{i=1}^k \frac{n_i}{n} \log \frac{n_i}{n}$$

Parent Node, p di-split menjadi k partisi
 n_i jumlah record dalam partisi i

- Adjusts Information Gain by the entropy of the partitioning (SplitINFO). Higher entropy partitioning (large number of small partitions) is penalized!
- Digunakan dalam C4.5
- Didesain untuk mengatasi kekurangan dari Information Gain

62 Semester Genap - TA.2022/2023

Kriteria Splitting berdasarkan Classification Error

■ Error klasifikasi pada node-t :

$$Error(t) = 1 - \max_i P(i | t)$$

■ Menghitung error misklasifikasi oleh satu node.

- Maximum ($1 - 1/n_i$) when records are equally distributed among all classes, implying least interesting information
- Minimum (0.0) when all records belong to one class, implying most interesting information

63 Semester Genap - TA.2022/2023

Contoh untuk Computing Error

$$Error(t) = 1 - \max_i P(i | t)$$

C1	0
C2	6

$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$
 Error = $1 - \max(0, 1) = 1 - 1 = 0$

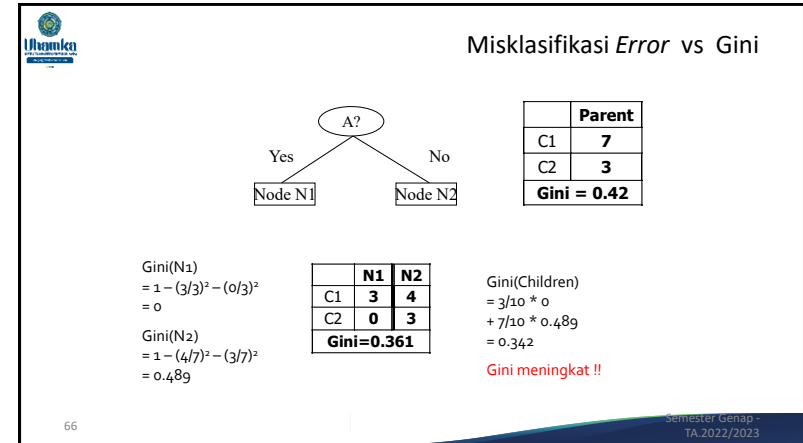
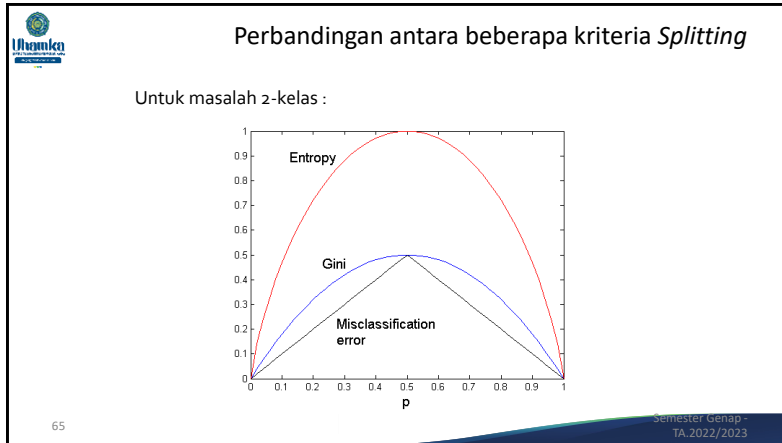
C1	1
C2	5

$P(C1) = 1/6 \quad P(C2) = 5/6$
 Error = $1 - \max(1/6, 5/6) = 1 - 5/6 = 1/6$

C1	2
C2	4

$P(C1) = 2/6 \quad P(C2) = 4/6$
 Error = $1 - \max(2/6, 4/6) = 1 - 4/6 = 1/3$

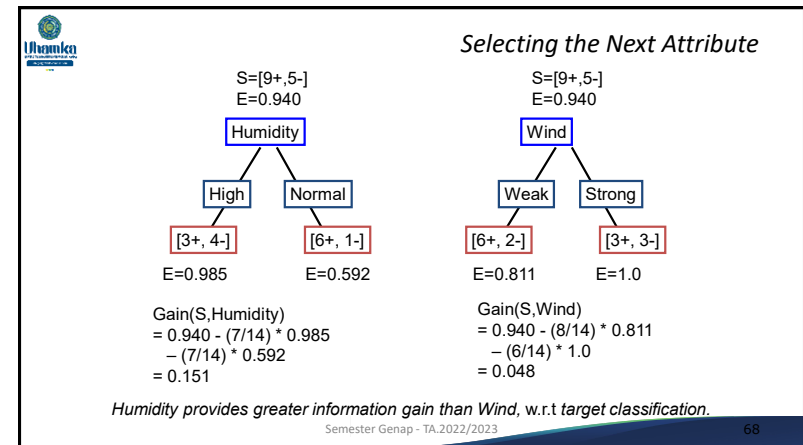
64 Semester Genap - TA.2022/2023



Training Examples

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Semester Genap - TA.2022/2023



Selecting the Next Attribute

$S = [9+, 5-]$
 $E = 0.940$

Outlook

- Sunny: $[2+, 3-]$, $E = 0.971$
- Overcast: $[4+, 0]$, $E = 0.0$
- Rain: $[3+, 2-]$, $E = 0.971$

Gain(S, Outlook)
 $= 0.940 - (5/14) * 0.971$
 $- (4/14) * 0.0 - (5/14) * 0.971$
 $= 0.247$

Semester Genap - TA.2022/2023 69

Selecting the Next Attribute

The information gain values for the 4 attributes are:

- Gain(S, Outlook) = 0.247
- Gain(S, Humidity) = 0.151
- Gain(S, Wind) = 0.048
- Gain(S, Temperature) = 0.029

where S denotes the collection of training examples

Semester Genap - TA.2022/2023 70

Selecting the Next Attribute

$[D1, D2, \dots, D14]$
 $[9+, 5-]$

Outlook

- Sunny: $S_{\text{sunny}} = [D1, D2, D8, D9, D11]$, $[2+, 3-]$, ?
- Overcast: $[D3, D7, D12, D13]$, $[4+, 0-]$, Yes
- Rain: $[D4, D5, D6, D10, D14]$, $[3+, 2-]$, ?

Gain(S_{sunny} , Humidity) = $0.970 - (3/5) 0.0 - 2/5 (0.0) = 0.970$
 Gain(S_{sunny} , Temp.) = $0.970 - (2/5) 0.0 - 2/5 (1.0) - (1/5) 0.0 = 0.570$
 Gain(S_{sunny} , Wind) = $0.970 - (2/5) 1.0 - 3/5 (0.918) = 0.019$

Semester Genap - TA.2022/2023 71

ID3 Algorithm

Outlook

- Sunny: Humidity
 - High: No ($[D1, D2]$)
 - Normal: Yes ($[D8, D9, D11]$)
- Overcast: Yes ($[D3, D7, D12, D13]$)
- Rain: Wind
 - Strong: No ($[D6, D14]$)
 - Weak: Yes ($[D4, D5, D10]$)

Semester Genap - TA.2022/2023 72



Referensi

- Ian H Witten, Eibe Frank, Mark A Hall. (2016). *Data Mining - Practical Machine Learning Tools and Techniques*, 4th Edition. Morgan-Kaufman.
- Hofmann, Markus., Klinkenberg, Ralf. (2016). *Data Mining Use Cases and Business Analytics Application*. CRC Press.
- Crisholm, Andrew. (2013). *Exploring Data With Rapidminer*. Packt Publishing.
- Mehmed Kantardzic. (2020). *Data Mining - Concepts, Models, Methods, and Algorithms*, 3rd edition. Wiley - IEEE Press.
- rapidminer.com, 2021,. Rapidminer Manual and Documentation, <https://docs.rapidminer.com/latest/studio/operators/>.

Semester Genap - TA.2022/2023

73

AMALAN HARIAN SEPANJANG RAMADHAN



MENDIRIKAN QIAMULLAIL	SAHUR & BERBUKA	SOLAT BERJEMAAH	MEMBACA AL-QURAN	MENGHULURKAN SEDEKAH
				
MENGHIDUPKAN 2/3 MALAM DENGAN IBADAH DAN BERMUNAJAT KEPADA ALLAH SWT	MELEWATKAN SAHUR DAN MENYEGERAKAN BERBUKA SERTA MEMPERBANYAKKAN DOA	SOLAT FARDHU DAN SOLAT TERAWIH BERJEMAAH DI MASJID, BESERTA SOLAT SUNAT	MEMPERBANYAK BACAAN AL-QURAN	MENGHULURKAN SEDEKAH DAN MEMBERI MAKAN KEPADA ORANG BERPUASA

Terima Kasih

Sabda Rasulullah S.A.W.

"Barang siapa merasa gembira dengan sebab kedatangan bulan Ramadhan nescaya Allah Taala mengharumkan jasadnya daripada api neraka".
(Rasulawati Ahmad)



Data Mining



Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

● ● ●

Session 06



PROGRAM STUDI TEKNIK INFORMATIKA

Sumber: <https://ft.uhamka.ac.id/prodi/informatika/>

- VISI.
 - ✓ Menjadi program studi informatika yang menghasilkan karya unggul di bidang teknologi informasi melalui pembelajaran berlandaskan prophetic teaching untuk membentuk lulusan yang cerdas secara spiritual, intelektual, emosional, dan sosial.
- MISI.
 1. Melaksanakan pendidikan di bidang informatika yang terintegrasi dengan nilai-nilai Al Islam Kemuhammadiyah.
 2. Melaksanakan penelitian untuk menghasilkan karya teknologi informasi yang unggul untuk masyarakat berkembang.
 3. Melaksanakan pengabdian kepada masyarakat untuk kesejahteraan umat.
 4. Menghasilkan mahasiswa yang berwawasan global dalam teknologi informasi.
 5. Mendorong jiwa kewirausahaan mahasiswa untuk berkiprah di tengah masyarakat dalam meningkatkan kegiatan ekonomi.
 6. Menerapkan standar mutu pendidikan informatika melalui layanan mutu yang berkesinambungan.
 7. Membangun kerja sama terpadu dengan mitra yang berkontribusi untuk kemajuan pendidikan di bidang informatika.

Semester Genap - TA.2022/2023 2



Sistem Penilaian

- SKS : 3
 - Teori 3 SKS – (14 Kali Pertemuan)
- Keaktifan (bobot 10%)
 - Terdiri dari: Aktivitas Perkuliahan (*Kehadiran, Etika, dan Sikap*).
- Project (bobot 25%)
 - Terdiri dari: Tugas Besar / Project (*Penguasaan Keterampilan*) dengan cara observasi, mengimplementasikan dan menghasilkan Proyek Akhir, serta mempresentasikannya.
- Tes (UTS = bobot 25% , UAS = bobot 40%)
 - Terdiri dari: UTS dan UAS (*Penguasaan Pengetahuan*) dengan cara tes tertulis maupun tes praktikum.

Semester Genap - TA.2022/2023 3



Tujuan/Objectives (Session.06)

- Objectives (Tujuan Instruksional Umum).
 - RPS/SAP.
 - Klasifikasi (*part.2*)
- Indikator Kompetensi.
 - Mahasiswa mampu mengetahui dan memahami tentang teknik klasifikasi pada *data mining*.
 - Mahasiswa mampu mengetahui, memahami, serta dapat menjelaskan tentang metode dan teknik klasifikasi, *decision tree*, algoritma C.45, dan *naïve bayes* didalam *data mining*.

Semester Genap - TA.2022/2023 4

Muhasabah

يَا أَيُّهَا الَّذِينَ آمَنُوا كُتِبَ عَلَيْكُمُ الصِّيَامُ
كَمَا كُتِبَ عَلَى الَّذِينَ مِنْ قَبْلِكُمْ لَعَلَّكُمْ تَتَّقُونَ

- Qs. Al Baqarah, ayat 183
 - Wahai orang-orang yang beriman! Diwajibkan atas kamu berpuasa sebagaimana diwajibkan atas orang sebelum kamu agar kamu bertakwa,

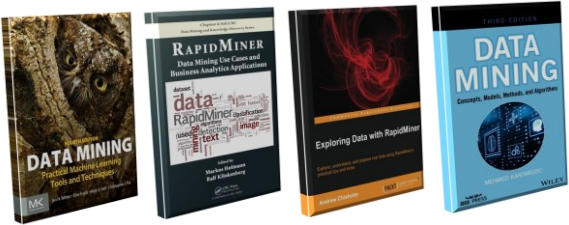
Semester Genap - TA.2022/2023 5

Data Mining



Uhamka
UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA
integrity, trust, compassion

Session 06
Klasifikasi (*part.2*)



Klasifikasi (*part.2*)

Pendahuluan

Semester Genap - TA.2022/2023 7

Pengantar

- **Classification**
 - Memprediksi kelas suatu item
 - Membuat model berdasarkan data pelatihan dan digunakan untuk mengklasifikasi data.
- **Prediction**
 - Memprediksi nilai yang belum diketahui
- **Aplikasi**
 - Persetujuan kredit
 - Diagnosis penyakit
 - Target marketing
 - Fraud detection

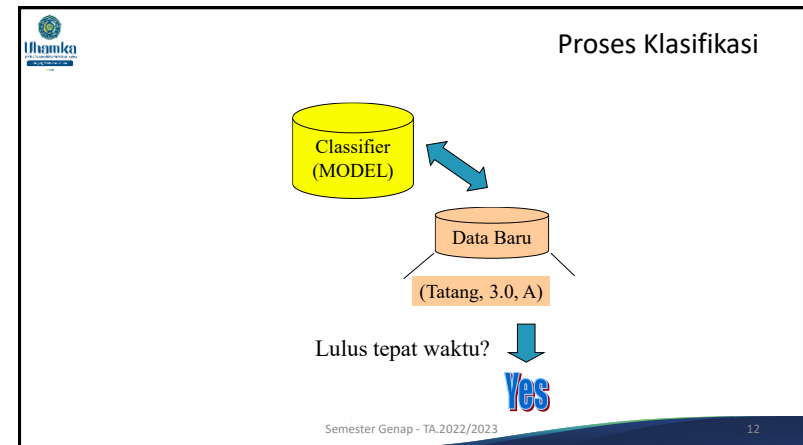
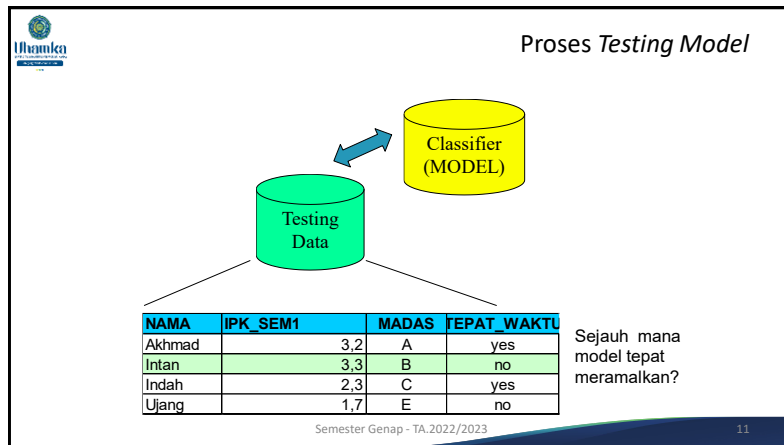
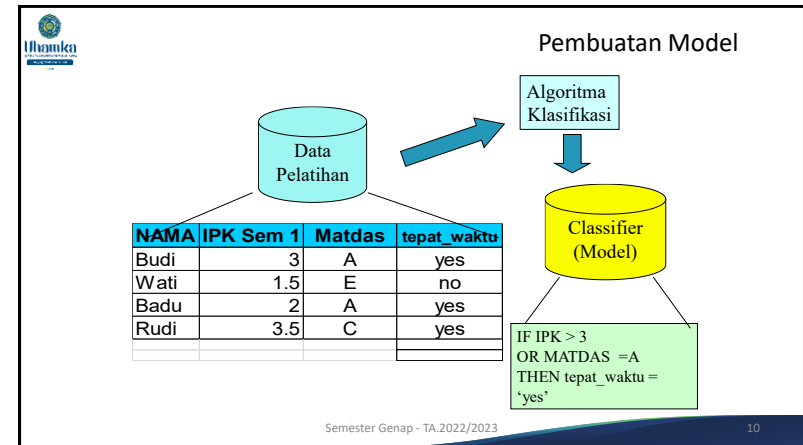
Semester Genap - TA.2022/2023 8

Contoh Kasus

- Input: data mahasiswa
- Output: dua kelas (lulus_tepat_waktu dan lulus_terlambat)

Bagaimana kalau diberikan data input mahasiswa, sistem secara otomatis menentukan mhs tersebut akan lulus tepat waktu atau terlambat?

Semester Genap - TA.2022/2023 9





- Proses pembuatan model
 - *Data training* → Model Klasifikasi
- Proses pengujian model
 - *Data testing* → Apakah model sudah benar?
- Proses klasifikasi
 - Data yang tidak diketahui kelasnya → kelas data

Semester Genap - TA.2022/2023 13



Sebelum Klasifikasi

- *Data cleaning*
 - *Pre-process* data untuk mengurangi *noise* dan *missing value*
- *Relevance analysis (feature selection)*
 - Memilih atribut yang penting
 - Membuang atribut yang tidak terkait atau duplikasi.
- Data transformation
 - *Generalize and/or normalize data*

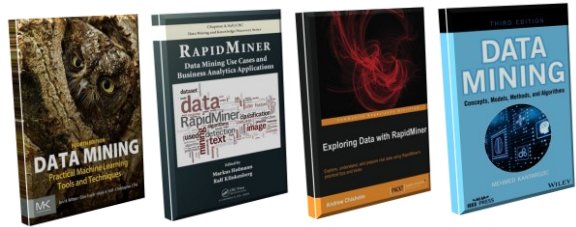
Semester Genap - TA.2022/2023 14



Evaluasi Metode Klasifikasi

- Akurasi
 - *classifier accuracy*: memprediksi label kelas
 - *predictor accuracy*: memprediksi nilai atribut
- Kecepatan
 - Waktu untuk membuat model (*training time*)
 - Waktu untuk menggunakan model (*classification/prediction time*)
- *Robustness*: menangani *noise* dan *missing value*.
- *Scalability*: efisien untuk proses dengan DBMS
- *Interpretability*
 - Model mudah dimengerti

Semester Genap - TA.2022/2023 15

Klasifikasi (part.2)

Decision Tree (Pohon Keputusan)

Semester Genap - TA.2022/2023 16



Decision Tree

- Diciptakan oleh Ross Quinlan
- ID3, C4.5, C5.0
- Model direpresentasikan dalam bentuk *tree*

Semester Genap - TA.2022/2023

17



Decision Tree: Contoh Input (Data Latih/Data Training)

age	income	student	credit rating	buys computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

Semester Genap - TA.2022/2023

18



Masalah

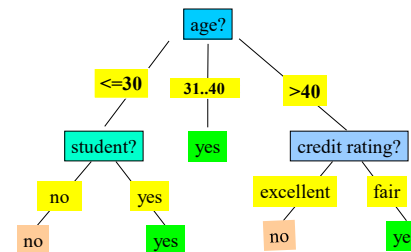
- Bagaimana dari data latih tersebut dapat diperoleh model yang dapat mengklasifikasikan secara otomatis?

Semester Genap - TA.2022/2023

19



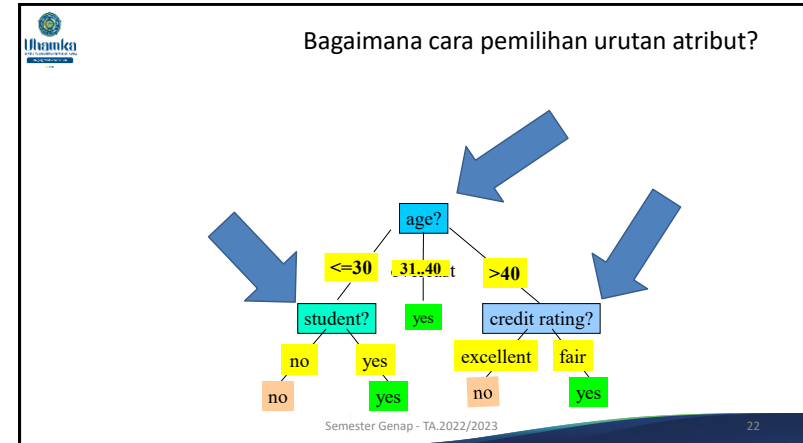
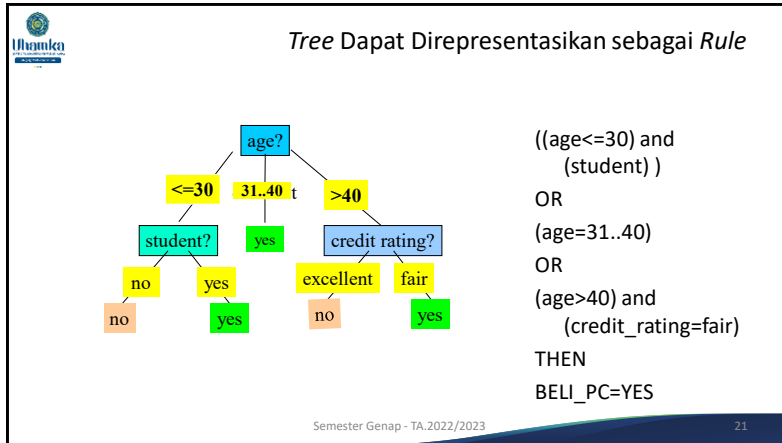
Model: Decision Tree



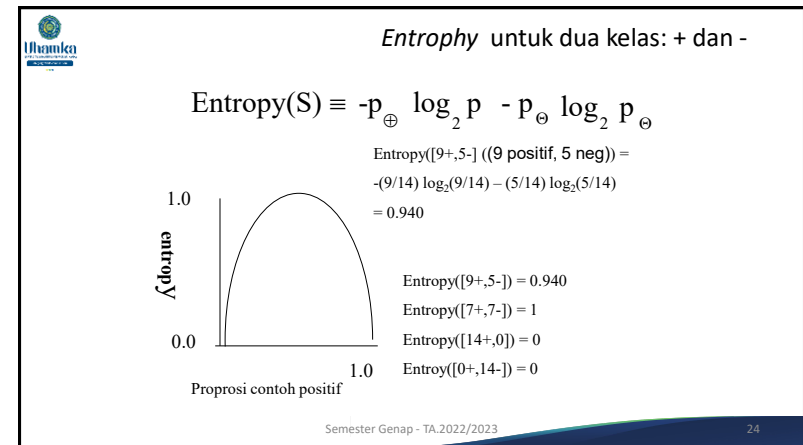
Dari data latih, model ini dibangun secara otomatis...

Semester Genap - TA.2022/2023

20



- Cara Pemilihan Atribut**
- **Entropy:** Ukuran kemurnian, semakin murni, semakin homogen, maka semakin rendah nilainya.
 - **Information Gain:** pengurangan *entropy* disebabkan oleh partisi berdasarkan suatu atribut.
 - Semakin besar info *gain* = atribut itu semakin membuat homogen = semakin bagus
 - Idenya → pilih atribut dengan info *gain* yg paling besar
- Semester Genap - TA.2022/2023 23



UHAMKA

Entropy untuk kelas > 2

$$Info(D) = - \sum_{i=1}^m p_i \log_2(p_i)$$

Info (D) = Entropy (D) (istilah dibuku J. HAN)

Semester Genap - TA.2022/2023 25

UHAMKA

Information Gain

$$Info_A(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times I(D_j)$$

$$Gain(A) = Info(D) - Info_A(D)$$

Gain(A) seberapa besar entropy berkurang akibat atribut A. Makin besar makin bagus.

Semester Genap - TA.2022/2023 26

UHAMKA

Contoh Pemilihan Atribut

- Class P: buys_computer = "yes"
- Class N: buys_computer = "no"

$$Info(D) = I(9,5) = -\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right) = 0.940$$

age	p _i	n _i	I(p _i , n _i)
<=30	2	3	0.971
31...40	4	0	0
>40	3	2	0.971

$$Info_{age}(D) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

$\frac{5}{14} I(2,3)$ berarti ada 5 dari 14
"age <=30" dgn 2 yes dan 3 no.

Gain (Age) = Info(D) - Info age (D)
= 0.940 - 0.694 = 0.246

Gain(income) = 0.029
Gain(student) = 0.151
Gain(credit_rating) = 0.048

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

A.2022/2023 27

UHAMKA

Pemilihan Atribut (lanj)

Gain (Age) = 0.246 ← yang terbesar, dipilih
Gain (income) = 0.029
Gain (student) = 0.151
Gain (credit_rating) = 0.048
Setelah AGE, atribut apa selanjutnya?

Diproses untuk setiap cabang selama masih ada > 1 kelas

Selanjutnya... proses data yang <=30

```

graph TD
    A[age?] --> B[<=30]
    A --> C[31...40]
    A --> D[>40]
    B --> E[yes]
    E --> F[Tidak perlu diproses lagi]
  
```

Semester Genap - TA.2022/2023 28

Pemilihan Atribut (lanj)

Selanjutnya... proses data $age \leq 30$

age	income	student	credit_rating	buys_computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
≤ 30	medium	no	fair	no
≤ 30	low	yes	fair	yes
≤ 30	medium	yes	excellent	yes

$$Inf(D) = I(2,3) = -\frac{2}{5} \log_2\left(\frac{2}{5}\right) - \frac{3}{5} \log_2\left(\frac{3}{5}\right) = 0.97$$

Gain(age) tidak perlu dihitung lagi, hitung
gain(student), gain(credit_rating)

$$Info_{student}(D) = \frac{3}{5} I(0,3) + \frac{2}{5} I(2,0) = 0$$

Gain(student) = $Inf(D) - Info_{student}(D)$
 $= 0.97 - 0$
 $= 0.97$

29

Pilihan Atribut (lanj)

Bandungkan semua gain, ambil yang paling besar

Gain(student) = 0.97
 Gain(credit_rating) = 0.02
 Gain(income) = 0.4

➔ Paling besar student

Semester Genap - TA.2022/2023 30

Pemilihan Atribut (lanj)

```

graph TD
    age[age?] --> age30["<=30"]
    age --> age31["31..40"]
    age --> age40[">40"]
    age30 --> student[student?]
    student --> no1[no]
    student --> yes1[yes]
    no1 --> no2[no]
    yes1 --> yes2[yes]
    age31 --> yes3[yes]
    age40 --> yes4[yes]
  
```

Semester Genap - TA.2022/2023 31

Latihan

No	Kelas	Kulit Buah	Warna	Ukuran	Bau
1	Aman	Kasar	Coklat	Besar	keras
2	Aman	Kasar	Hijau	Besar	Keras
3	Berbahaya	Halus	Merah	Besar	Lunak
4	Aman	Kasar	Hijau	Besar	Lunak
5	Aman	Kasar	Merah	Kecil	Keras
6	Aman	Halus	Merah	Kecil	Keras
7	Aman	Halus	Coklat	Kecil	Keras
8	Berbahaya	Kasar	Hijau	Kecil	Lunak
9	Berbahaya	Halus	Hijau	Kecil	Keras
10	Aman	Kasar	Merah	Besar	Keras
11	Aman	Halus	Coklat	Besar	Lunak
12	Berbahaya	Halus	Hijau	Kecil	Keras
13	Aman	Kasar	Merah	Kecil	Lunak
14	Berbahaya	Halus	Merah	Besar	Keras
15	Aman	Halus	Merah	Kecil	Keras
16	Berbahaya	Kasar	Hijau	Kecil	Keras

Semester Genap - TA.2022/2023 32

 Mengapa *Decision Tree* ?

- Mudah diimplementasikan
- Hipotesis yang dihasilkan mudah dipahami
- Efisien

Semester Genap - TA.2022/2023 33

 *Decision Tree* Cocok untuk Masalah:

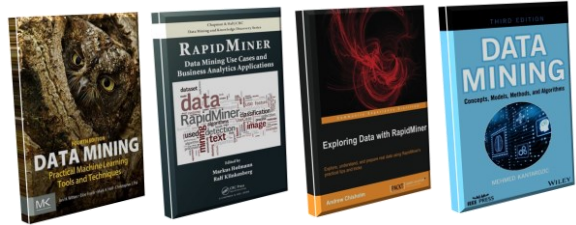
- Data dalam bentuk atribut-nilai.
- Kondisi ideal adalah jika isi nilai jumlahnya sedikit.
 - Misalnya: “panas”, “sedang”, “dingin”.
- Output diskrit.
- *Training data* dapat tidak lengkap

Semester Genap - TA.2022/2023 34

 Masalah didalam *Decission Tree*

- *Overfitting*: terlalu mengikuti *training data*
 - Terlalu banyak cabang, merefleksikan anomali akibat *noise* atau *outlier*.
 - Akurasi rendah untuk data baru
- Dua pendekatan untuk menghindari *overfitting*
 - *Pre-pruning*:
 - Hentikan pembuatan *tree* di awal. Tidak men-*split node* jika *goodness measure* dibawah *threshold*.
 - Sulit untuk menentukan *threshold*
 - *Post-pruning*: Buang cabang setelah *tree* jadi
 - Menggunakan data yang berbeda dengan *training* untuk menentukan *pruned tree* yang terbaik.

Semester Genap - TA.2022/2023 35

Klasifikasi (part.2)

Bayesian & Naïve Bayesian

Semester Genap - TA.2022/2023 36



Bayesian Classification: Why ?

- A statistical classifier:
 - membangun *probabilistic prediction*, yaitu memprediksi probabilitas keanggotaan kelas
- Didasarkan pada Bayes' Theorem.
- Performance:
 - sederhana ---- *naïve Bayesian classifier*

Semester Genap - TA.2022/2023 37



Bayesian Theorem: Basics

- **X** adalah data sample ("*evidence*"): label kelas tidak diketahui
- **H** adalah dugaan (*hypothesis*) bahwa **X** adalah anggota **C**
- Klasifikasi ditentukan $P(H|X)$, (*posteriori probability*), probabilitas bahwa dugaan terhadap data sample **X**
- $P(H)$ (*prior probability*), initial probability
 - Misal **X** akan membeli computer, tidak memperhatikan age, income, ...
- $P(X)$: probabilitas dari sample data yang diamati
- $P(X|H)$ (*likelihood*), probabilitas dari sample **X**, dengan the memperhatikan dugaan
 - Misal , **X** akan membeli computer, probabilitas bahwa **X**. Adalah 31..40, penghasilan sedang

Semester Genap - TA.2022/2023 38



Bayesian Theorem

- Dari training data **X**, *posteriori probabilitas dari hypothesis H* atau class, $P(H|X)$, *teorema Bayes*

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

- Ini dapat ditulis dengan
posterior = likelihood x prior/evidence
- Prediksi **X** anggota C_2 jika dan hanya jika probabilitas $P(C_2|X)$ paling tinggi diantara semua $P(C_k|X)$ dari semua *kelas k*

Semester Genap - TA.2022/2023 39



Naïve Bayesian Classifier: Training Dataset

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

Semester Genap - TA.2022/2023 40

Klasifikasi Naïve Bayesian

- Perhatikan D adalah record training dan ditetapkan label-label kelasnya dan masing-masing record dinyatakan n atribut (n field) $X = (x_1, x_2, \dots, x_n)$
- Misalkan terdapat m kelas $C_1, C_2, \dots, C_m$.
- Klasifikasi adalah diperoleh maximum posteriori yaitu maximum $P(C_i|X)$
- Ini dapat diperoleh dari teorema Bayes
$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$
- Karena $P(X)$ adalah konstan untuk semua kelas, hanya perlu dimaksimumkan

$$P(C_i|X) = P(X|C_i)P(C_i)$$

Semester Genap - TA.2022/2023 41

Derivation of Naïve Bayes Classifier

- Diasumsikan:** atribut dalam kondisi saling bebas (independent) yaitu tidak ada ketergantungan antara atribut-atribut :

$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \dots \times P(x_n|C_i)$$

- A_k adalah categorical, $P(x_k|C_i)$ adalah jumlah record dalam kelas C_i yang memiliki nilai x_k sama dengan A_k dibagi dengan $|C_i|$ jumlah record dalam C_i dalam D
- Jika A_k bernilai kontinu , $P(x_k|C_i)$ biasanya dihitung berdasarkan pada distribusi Gaussian dengan [mean](#) μ and standar deviasi σ

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Dan $P(x_k|C_i)$ adalah

$$P(X|C_i) = g(x_1, \mu_{C_i}, \sigma_{C_i})$$

Semester Genap - TA.2022/2023 42

Naïve Bayesian Classifier: Training Dataset

Class:
 $C_1: \text{buys_computer} = \text{'yes'}$
 $C_2: \text{buys_computer} = \text{'no'}$

Data sample
 $X = (\text{age} \leq 30,$
 $\text{Income} = \text{medium},$
 $\text{Student} = \text{yes}$
 $\text{Credit_rating} = \text{Fair})$

D = 14

age	income	student	credit_rating	com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

Semester Genap - TA.2022/2023 43

Naïve Bayesian Classifier: An Example

- $P(C_1): P(\text{buys_computer} = \text{'yes'}) = 9/14 = 0.643$
 $P(\text{buys_computer} = \text{'no'}) = 5/14 = 0.357$
- Compute $P(X|C_i)$ for each class
 $P(\text{age} = \text{'<=30'} | \text{buys_computer} = \text{'yes'}) = 2/9 = 0.222$
 $P(\text{age} = \text{'<=30'} | \text{buys_computer} = \text{'no'}) = 3/5 = 0.6$
 $P(\text{income} = \text{'medium'} | \text{buys_computer} = \text{'yes'}) = 4/9 = 0.444$
 $P(\text{income} = \text{'medium'} | \text{buys_computer} = \text{'no'}) = 2/5 = 0.4$
 $P(\text{student} = \text{'yes'} | \text{buys_computer} = \text{'yes'}) = 6/9 = 0.667$
 $P(\text{student} = \text{'yes'} | \text{buys_computer} = \text{'no'}) = 1/5 = 0.2$
 $P(\text{credit_rating} = \text{'fair'} | \text{buys_computer} = \text{'yes'}) = 6/9 = 0.667$
 $P(\text{credit_rating} = \text{'fair'} | \text{buys_computer} = \text{'no'}) = 2/5 = 0.4$
- $X = (\text{age} \leq 30, \text{income} = \text{medium}, \text{student} = \text{yes}, \text{credit_rating} = \text{fair})$
 $P(X|C_1) : P(X|\text{buys_computer} = \text{'yes'}) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$
 $P(X|C_2) : P(X|\text{buys_computer} = \text{'no'}) = 0.6 \times 0.4 \times 0.2 \times 0.4 = 0.019$
 $P(X|C_1) * P(C_1) = P(X|\text{buys_computer} = \text{'yes'}) * P(\text{buys_computer} = \text{'yes'})$
 $= 0.044 * 0.643 = 0.028$
 $P(X|C_2) * P(C_2) = P(X|\text{buys_computer} = \text{'no'}) * P(\text{buys_computer} = \text{'no'}) = 0.007$

Sehingga , X belongs to class ("buys_computer = yes")

Semester Genap - TA.2022/2023 44

Menghindari masalah Probabilitas 0

- Prediksi *Naïve Bayesian* membutuhkan masing-masing probabilitas tidak nol, Dengan kata lain. Probabilitas yang dihitung tidak menjadi nol

$$P(X | C_i) = \prod_{k=1}^n P(x_k | C_i)$$

- Misalkan data dengan 1000 record, *income=low* (0), *income=medium* (990), and *income=high* (10),
- Menggunakan *Laplacian correction* (atau *Laplacian estimator*)
 - Tambahkan 1 untuk masing-masing case
 - $Prob(income = low) = 1/1003$
 - $Prob(income = medium) = 991/1003$
 - $Prob(income = high) = 11/1003$

Semester Genap - TA.2022/2023 45

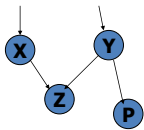
Penjelasan *Naïve Bayesian Classifier*:

- Keuntungan
 - Mudah diimplementasikan
 - Hasil baik dalam banyak kasus
- Kerugian
 - Asumsi : kondisi kelas saling bebas, sehingga kurang akurat
 - Pada prakteknya, kebergantungan anda diantara variabel
 - Misal hospitals: patients: Profile: age, family history, etc.
 - Gejala : demam (fever), batuk (cough) etc., Disease: lung cancer, diabetes, etc.
 - Kebergantunagn diantara variabel ini tidak dapat dimodelkan dengan menggunakan *Naïve Bayesian Classifier*
- How to deal with these dependencies?
 - Bayesian Belief Networks*

Semester Genap - TA.2022/2023 46

Bayesian Belief Networks

- Bayesian belief network*, memungkinkan untuk memodelkan sebagian variabel dalam kondisi saling bergantung.
- Model grafik menyatakan keterhubungan sebab akibat
 - Menyatakan kebergantungan (*dependency*) diantara variabel-variabel
 - Menggambarkan distribusi probabilitas gabungan



- Node (simpul) : variabel-variabel bebas
- Links: kebergantungan
- X dan Y adalah parents dari Z, dan Y adalah parent dari P
- Tidak ada kebergantungan diantara Z dan P
- Tidak memiliki loop atau siklus

Semester Genap - TA.2022/2023 47

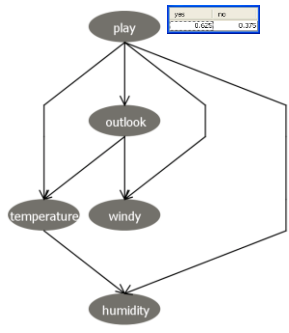
Play probability table

Based on the data...

$P(\text{play=yes}) = 9/14$
 $P(\text{play=no}) = 5/14$

Let's correct with Laplace ...

$P(\text{play=yes}) = (9+1)/(14+2) = .625$
 $P(\text{play=yes}) = (5+1)/(14+2) = .375$



Outlook	Temp.	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

Semester Genap - TA.2022/2023 48

Outlook probability table

Based on the data...

$P(\text{outlook}=\text{sunny}|\text{play}=\text{yes}) = (2+1)/(9+3) = .25$
 $P(\text{outlook}=\text{overcast}|\text{play}=\text{yes}) = (4+1)/(9+3) = .417$
 $P(\text{outlook}=\text{rainy}|\text{play}=\text{yes}) = (3+1)/(9+3) = .333$
 $P(\text{outlook}=\text{sunny}|\text{play}=\text{no}) = (3+1)/(5+3) = .5$
 $P(\text{outlook}=\text{overcast}|\text{play}=\text{no}) = (0+1)/(5+3) = .125$
 $P(\text{outlook}=\text{rainy}|\text{play}=\text{no}) = (2+1)/(5+3) = .375$

Semester Genap - TA.2022/2023 49

Windy probability table

Based on the data...let's find the conditional probabilities for "windy"

$P(\text{windy}=\text{true}|\text{play}=\text{yes}, \text{outlook}=\text{sunny}) = (1+1)/(2+2) = .5$

Outlook	Temp	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

Semester Genap - TA.2022/2023 50

Windy probability table

Based on the data...

$P(\text{windy}=\text{true}|\text{play}=\text{yes}, \text{outlook}=\text{sunny}) = (1+1)/(2+2) = .5$
 $P(\text{windy}=\text{true}|\text{play}=\text{yes}, \text{outlook}=\text{overcast}) = 0.5$
 $P(\text{windy}=\text{true}|\text{play}=\text{yes}, \text{outlook}=\text{rainy}) = 0.2$
 $P(\text{windy}=\text{true}|\text{play}=\text{no}, \text{outlook}=\text{sunny}) = 0.4$
 $P(\text{windy}=\text{true}|\text{play}=\text{no}, \text{outlook}=\text{overcast}) = 0.5$
 $P(\text{windy}=\text{true}|\text{play}=\text{no}, \text{outlook}=\text{rainy}) = 0.75$

Semester Genap - TA.2022/2023 51

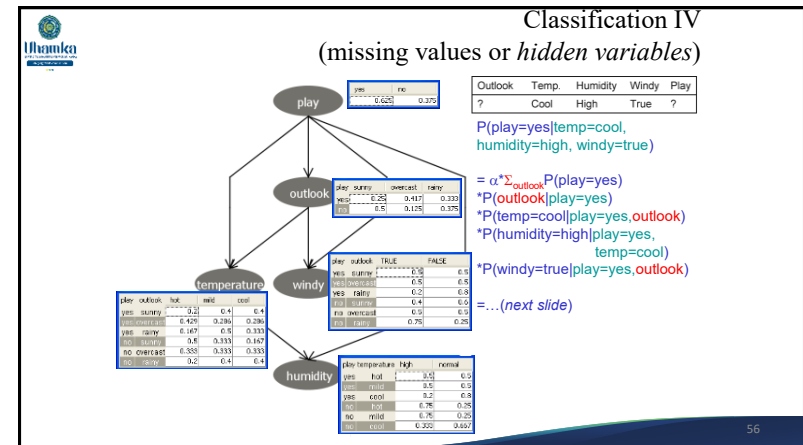
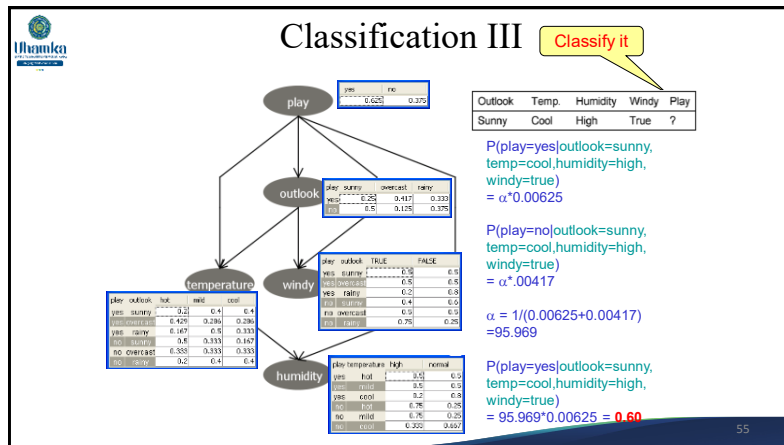
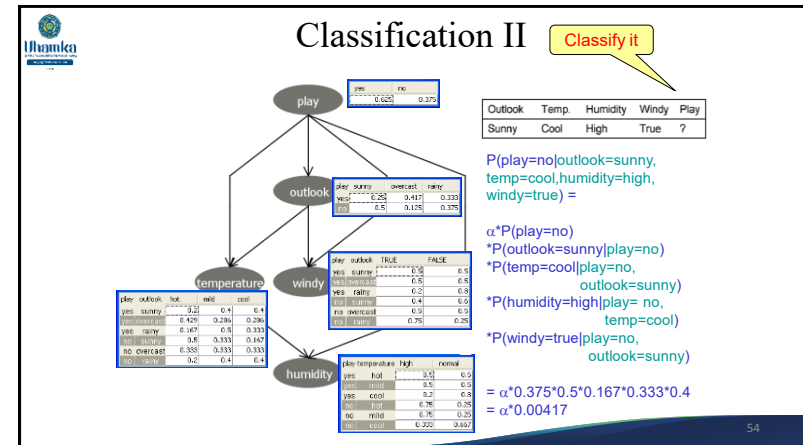
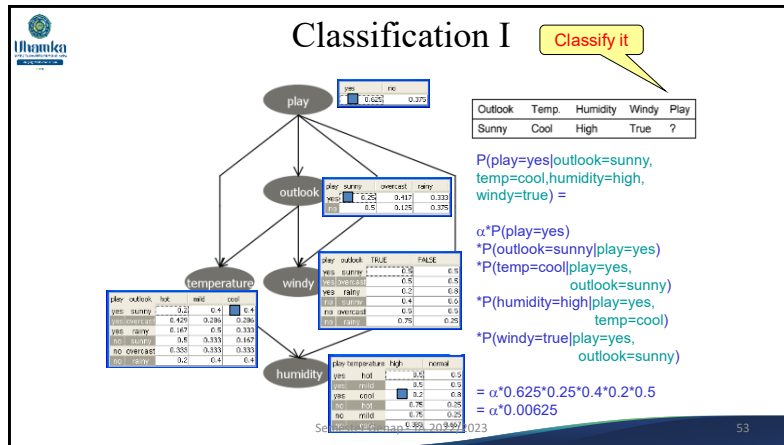
Final figure

Classify it

Classify it

Outlook Temp Humidity Windy Play
 Sunny Cool High True ?
 ? Cool High True ?

Semester Genap - TA.2022/2023 52



Classification V
(missing values or *hidden variables*)

$P(\text{play}=\text{yes}|\text{temp}=\text{cool}, \text{humidity}=\text{high}, \text{windy}=\text{true})$

$$= \alpha \sum_{\text{outlook}} P(\text{play}=\text{yes}) \cdot P(\text{outlook}|\text{play}=\text{yes}) \cdot P(\text{temp}=\text{cool}|\text{play}=\text{yes}, \text{outlook})$$

$$\cdot P(\text{humidity}=\text{high}|\text{play}=\text{yes}, \text{temp}=\text{cool}) \cdot P(\text{windy}=\text{true}|\text{play}=\text{yes}, \text{outlook})$$

$$= \alpha [$$

$$P(\text{play}=\text{yes}) \cdot P(\text{outlook}=\text{sunny}|\text{play}=\text{yes}) \cdot P(\text{temp}=\text{cool}|\text{play}=\text{yes}, \text{outlook}=\text{sunny})$$

$$\cdot P(\text{humidity}=\text{high}|\text{play}=\text{yes}, \text{temp}=\text{cool}) \cdot P(\text{windy}=\text{true}|\text{play}=\text{yes}, \text{outlook}=\text{sunny})$$

$$+ P(\text{play}=\text{yes}) \cdot P(\text{outlook}=\text{overcast}|\text{play}=\text{yes}) \cdot P(\text{temp}=\text{cool}|\text{play}=\text{yes}, \text{outlook}=\text{overcast})$$

$$\cdot P(\text{humidity}=\text{high}|\text{play}=\text{yes}, \text{temp}=\text{cool}) \cdot P(\text{windy}=\text{true}|\text{play}=\text{yes}, \text{outlook}=\text{overcast})$$

$$+ P(\text{play}=\text{yes}) \cdot P(\text{outlook}=\text{rainy}|\text{play}=\text{yes}) \cdot P(\text{temp}=\text{cool}|\text{play}=\text{yes}, \text{outlook}=\text{rainy})$$

$$\cdot P(\text{humidity}=\text{high}|\text{play}=\text{yes}, \text{temp}=\text{cool}) \cdot P(\text{windy}=\text{true}|\text{play}=\text{yes}, \text{outlook}=\text{rainy})$$

$$]$$

$$= \alpha [$$

$$0.625 \cdot 0.25 \cdot 0.4 \cdot 0.2 \cdot 0.5 + 0.625 \cdot 0.417 \cdot 0.286 \cdot 0.2 \cdot 0.5 + 0.625 \cdot 0.33 \cdot 0.333 \cdot 0.2 \cdot 0.2]$$

$$= \alpha \cdot 0.01645$$

Classification VI
(missing values or *hidden variables*)

$P(\text{play}=\text{no}|\text{temp}=\text{cool}, \text{humidity}=\text{high}, \text{windy}=\text{true})$

$$= \alpha \sum_{\text{outlook}} P(\text{play}=\text{no}) \cdot P(\text{outlook}|\text{play}=\text{no}) \cdot P(\text{temp}=\text{cool}|\text{play}=\text{no}, \text{outlook})$$

$$\cdot P(\text{humidity}=\text{high}|\text{play}=\text{no}, \text{temp}=\text{cool}) \cdot P(\text{windy}=\text{true}|\text{play}=\text{no}, \text{outlook})$$

$$= \alpha [$$

$$P(\text{play}=\text{no}) \cdot P(\text{outlook}=\text{sunny}|\text{play}=\text{no}) \cdot P(\text{temp}=\text{cool}|\text{play}=\text{no}, \text{outlook}=\text{sunny})$$

$$\cdot P(\text{humidity}=\text{high}|\text{play}=\text{no}, \text{temp}=\text{cool}) \cdot P(\text{windy}=\text{true}|\text{play}=\text{no}, \text{outlook}=\text{sunny})$$

$$+ P(\text{play}=\text{no}) \cdot P(\text{outlook}=\text{overcast}|\text{play}=\text{no}) \cdot P(\text{temp}=\text{cool}|\text{play}=\text{no}, \text{outlook}=\text{overcast})$$

$$\cdot P(\text{humidity}=\text{high}|\text{play}=\text{no}, \text{temp}=\text{cool}) \cdot P(\text{windy}=\text{true}|\text{play}=\text{no}, \text{outlook}=\text{overcast})$$

$$+ P(\text{play}=\text{no}) \cdot P(\text{outlook}=\text{rainy}|\text{play}=\text{no}) \cdot P(\text{temp}=\text{cool}|\text{play}=\text{no}, \text{outlook}=\text{rainy})$$

$$\cdot P(\text{humidity}=\text{high}|\text{play}=\text{no}, \text{temp}=\text{cool}) \cdot P(\text{windy}=\text{true}|\text{play}=\text{no}, \text{outlook}=\text{rainy})$$

$$]$$

$$= \alpha [$$

$$0.375 \cdot 0.5 \cdot 0.167 \cdot 0.333 \cdot 0.4 + 0.375 \cdot 0.125 \cdot 0.333 \cdot 0.333 \cdot 0.5 +$$

$$0.375 \cdot 0.375 \cdot 0.4 \cdot 0.333 \cdot 0.75]$$

$$= \alpha \cdot 0.0208$$

Classification VII
(missing values or *hidden variables*)

$P(\text{play}=\text{yes}|\text{temp}=\text{cool}, \text{humidity}=\text{high}, \text{windy}=\text{true}) = \alpha \cdot 0.01645$

$P(\text{play}=\text{no}|\text{temp}=\text{cool}, \text{humidity}=\text{high}, \text{windy}=\text{true}) = \alpha \cdot 0.0208$

$$\alpha = 1 / (0.01645 + 0.0208) = 26.846$$

$$P(\text{play}=\text{yes}|\text{temp}=\text{cool}, \text{humidity}=\text{high}, \text{windy}=\text{true}) = 26.846 \cdot 0.01645 = 0.44$$

$$P(\text{play}=\text{no}|\text{temp}=\text{cool}, \text{humidity}=\text{high}, \text{windy}=\text{true}) = 26.846 \cdot 0.0208 = 0.56$$

I.e. $P(\text{play}=\text{yes}|\text{temp}=\text{cool}, \text{humidity}=\text{high}, \text{windy}=\text{true})$ is 44% and
 $P(\text{play}=\text{no}|\text{temp}=\text{cool}, \text{humidity}=\text{high}, \text{windy}=\text{true})$ is 56%

So, we predict 'play=no.'

Referensi

- Ian H Witten, Eibe Frank, Mark A Hall. (2016). *Data Mining - Practical Machine Learning Tools and Techniques*, 4th Edition. Morgan-Kaufman.
- Hofmann, Markus., Klinkenberg, Ralf. (2016). *Data Mining Use Cases and Business Analytics Application*. CRC Press.
- Crisholm, Andrew. (2013). *Exploring Data With Rapidminer*. Packt Publishing.
- Mehmed Kantardzic. (2020). *Data Mining - Concepts, Models, Methods, and Algorithms*, 3rd edition. Wiley - IEEE Press.
- rapidminer.com, 2021., Rapidminer Manual and Documentation, <https://docs.rapidminer.com/latest/studio/operators/>.

Dari Abu Umamah RA, beliau berkata,
saya mendengar Rasulullah SAW bersabda:

Terima Kasih

“Bacalah al-Qur’an itu, karena sesungguhnya
ia akan datang pada Hari Kiamat sebagai
pemberi syafa’at bagi para pembaca lainnya.”

(Hadits Riwayat Imam Muslim)





Data Mining



Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

● ● ●

Session 08



PROGRAM STUDI TEKNIK INFORMATIKA

Sumber: <https://ft.uhamka.ac.id/prodi/informatika/>

- VISI.
 - ✓ Menjadi program studi informatika yang menghasilkan karya unggul di bidang teknologi informasi melalui pembelajaran berlandaskan prophetic teaching untuk membentuk lulusan yang cerdas secara spiritual, intelektual, emosional, dan sosial.
- MISI.
 1. Melaksanakan pendidikan di bidang informatika yang terintegrasi dengan nilai-nilai Al Islam Kemuhammadiyah.
 2. Melaksanakan penelitian untuk menghasilkan karya teknologi informasi yang unggul untuk masyarakat berkembang.
 3. Melaksanakan pengabdian kepada masyarakat untuk kesejahteraan umat.
 4. Menghasilkan mahasiswa yang berwawasan global dalam teknologi informasi.
 5. Mendorong jiwa kewirausahaan mahasiswa untuk berkiprah di tengah masyarakat dalam meningkatkan kegiatan ekonomi.
 6. Menerapkan standar mutu pendidikan informatika melalui layanan mutu yang berkesinambungan.
 7. Membangun kerja sama terpadu dengan mitra yang berkontribusi untuk kemajuan pendidikan di bidang informatika.

Semester Genap - TA.2022/2023 2



Sistem Penilaian

- SKS : 3
 - Teori 3 SKS – (14 Kali Pertemuan)
- Keaktifan (bobot 10%)
 - Terdiri dari: Aktivitas Perkuliahan (*Kehadiran, Etika, dan Sikap*).
- Project (bobot 25%)
 - Terdiri dari: Tugas Besar / Project (*Penguasaan Keterampilan*) dengan cara observasi, mengimplementasikan dan menghasilkan Proyek Akhir, serta mempresentasikannya.
- Tes (UTS = bobot 25% , UAS = bobot 40%)
 - Terdiri dari: UTS dan UAS (*Penguasaan Pengetahuan*) dengan cara tes tertulis maupun tes praktikum.

Semester Genap - TA.2022/2023 3



Tujuan/Objectives (Session.08)

- Objectives (Tujuan Instruksional Umum).
 - RPS/SAP.
 - Text Mining
- Indikator Kompetensi.
 - Mahasiswa mampu mengetahui dan memahami tentang teknik klasifikasi pada *data mining*, khususnya untuk topik *Text Mining*.
 - Mahasiswa mampu mengetahui, memahami, serta dapat menjelaskan tentang metode dan teknik klasifikasi untuk *text mining*, tahapan *pre-processing* dan proses pada *text mining*, dan *web mining*.

Semester Genap - TA.2022/2023 4

Muhasabah

﴿وَاعْبُدُوا اللَّهَ وَلَا تُفْرِكُوا بِهِ شَيْئًا ۚ وَالْأُولَ الَّذِينَ أَحْسَنًا وَيَذِي الْقُرْبَىٰ وَالْيَتَامَىٰ وَالْمَسْكِينِ وَالْجَارِ ذِي الْقُرْبَىٰ وَالْجَارِ الْجُنُبِ وَالصَّاحِبِ بِالْجَنُبِ وَإِنِ السَّبِيلُ ۚ وَمَا مَلَكَتْ أَيْمَانُكُمْ ۚ إِنَّ اللَّهَ لَا يُحِبُّ مَنْ كَانَ مُخْتَالًا فَخُورًا﴾

- Qs. An Nisa, ayat 36
 - Dan sembahlah Allah dan janganlah kamu mempersekutukan-Nya dengan sesuatu apa pun.
 - Dan berbuat-baiklah kepada kedua orang tua, karib-kerabat, anak-anak yatim, orang-orang miskin, tetangga dekat dan tetangga jauh, teman sejawat, ibnu sabil dan hamba sahaya yang kamu miliki.
 - Sungguh, Allah tidak menyukai orang yang sombong dan membanggakan diri,

Semester Genap - TA.2022/2023 5

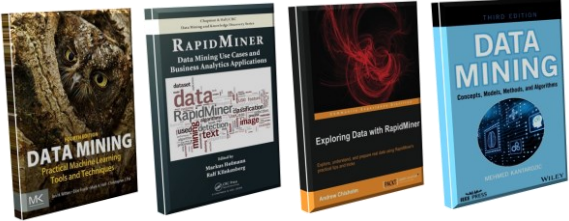
Data Mining



Uhamka
UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

Session 08
Text Mining



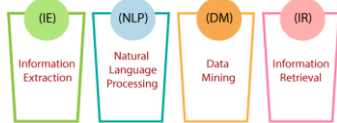
Text Mining

Pendahuluan

Semester Genap - TA.2022/2023 7

Pendahuluan

Area's of Text Mining



- Text mining*, merupakan suatu proses untuk mengambil informasi dari teks yang ada.
- Text mining* mencari pola-pola yang ada di teks-teks dalam bahasa natural yang tidak terstruktur seperti buku, email, artikel, halaman *web*, *social media*, dll.
- Kegiatan yang biasa dilakukan oleh *text mining* adalah *text categorization*, *text clustering*, *conception/entity extraction*, dll.

Semester Genap - TA.2022/2023 8

Pendahuluan
(lanjutan)

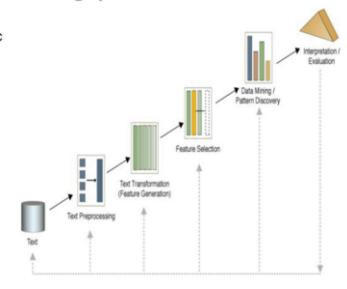
- Proses kegiatan pada text mining, terbagi menjadi 3 bagian:
 - Characterization of data**
 - Teks yang ada distrukturkan dengan proses seperti *parsing*, dan dimasukkan ke dalam sebuah *database*
 - Data mining**
 - Dari data yang ada, dilakukan sebuah pencarian dengan algoritma tertentu untuk mendapatkan pola dari data tersebut
 - Data visualization**
 - Hasil pencarian yang ada akan di-interpretasi dan dikeluarkan dalam bentuk output yang dapat dimengerti dengan mudah.

Semester Genap - TA.2022/2023 9

Pendahuluan
(lanjutan)

Text mining process

- Text preprocessing
 - Syntactic/Semantic text analysis
- Features Generation
 - Bag of words
- Features Selection
 - Simple counting
 - Statistics
- Text/Data Mining
 - Classification-Supervised learning
 - Clustering-Unsupervised learning
- Analyzing results



Semester Genap - TA.2022/2023 10

Pendahuluan
(lanjutan)

- Aplikasi Marketing**
 - Text mining* dapat digunakan untuk *cross-selling* dan *up-selling* dengan menganalisis data yang tidak terstruktur yang dihasilkan oleh *call center*.
- Aplikasi Keamanan**
 - Pada tahun 2007, EUROPOL mengembangkan sistem terintegrasi yang mampu mengakses, menyimpan dan menganalisis sejumlah besar sumber data terstruktur dan tak terstruktur untuk melacak organisasi kriminal trans-nasional.
- Aplikasi Biomedis**
 - PubGene yang menggabungkan *text mining biomedis* dengan visualisasi jaringan sebagai sebuah layanan *Internet*. Contoh lainnya yaitu GoPubMed *Semantic similarity* yang juga telah digunakan oleh sistem *text mining*, yaitu, GOAnnotato.
- Aplikasi Akademik**
 - National Centre for Text Mining* yang merupakan hasil kolaborasi Universitas Manchester dan Liverpool, digunakan untuk menyediakan *customized tools*, fasilitas penelitian, dan saran pada *text mining* untuk komunitas akademik.

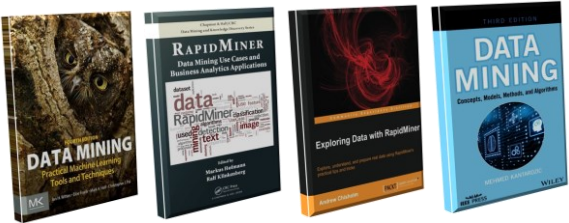
Semester Genap - TA.2022/2023 11

Pendahuluan
(lanjutan)

- Commercial Software Tools :**
 - RapidMiner, ClearForest, IBM Intelligent Miner Data Mining Suite, Megaputer Text Analyst, SAS Text Miner, SPSS Text Mining, The Statistica Text Mining, VantagePoint, The WordStat Analysis module, dll.
- Free Software Tools :**
 - Bersifat *open source* diantaranya, GATE, LingPipe, S-EM(Spy-EM), Vivisimo/Clusty.



Semester Genap - TA.2022/2023 12



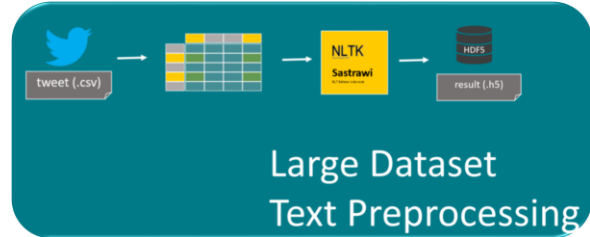
Text Mining

Pre-Processing

Semester Genap - TA.2022/2023

Pre-Processing

- yaitu proses pendahuluan yang diterapkan terhadap data teks, yang bertujuan untuk menghasilkan data numerik.



Large Dataset Text Preprocessing

Semester Genap - TA.2022/2023

Pre-Processing (lanjutan)

- Penghapusan format dan markup
 - Jika dokumen yang digunakan bukan berupa teks murni, maka tahap ini dilakukan. Karena dokumen *teks* yang biasanya kita lihat berupa format non-teks seperti *html*, *pdf*, atau dalam bentuk *word*.
 - Format-format ini mengharuskan sebuah teks dilengkapi unsur-unsur tambahan untuk dapat menghasilkan tampilan yang *friendly* dimata kita.
 - Informasi-informasi itu dihilangkan karena dianggap tidak perlu dan tidak mencerminkan isi sebuah dokumen teks.

Semester Genap - TA.2022/2023

Pre-Processing (lanjutan)

- Penghapusan tanda baca
 - Tanda baca juga dianggap tidak penting, karena biasanya didalam penelitian tidak memperhatikan keterkaitan kata, kalimat ataupun sejenisnya.
 - Sehingga semua kata dianggap berdiri sendiri.



Semester Genap - TA.2022/2023

TABEL II
PROSES CLEANSING

No	Proses Input	Proses Output
1	malam kak leny @leny_briyan, maaf banget udah bikin kakak kecewa	malam kak leny leny briyan maaf banget udah bikin kakak kecewa
2	jgn syedih dong kak, mantikan terus kejutan promo di wilayah lainnya yahl.	jgn syedih dong kak mantikan terus kejutan promo di wilayah lainnya yahl
3	apakah translation di grabchat harus punya toefl 550, biar pas pesen grabfood terus nge-chat abangnya "siap mas!" gak di-translate "ready gold!" ?	apakah translation di grabchat harus punya toefl 550 biar pas pesen grabfood terus nge-chat abangnya siap mas gak di-translate ready gold
4	maaf sudah buat kakak ga nyaman ya, sudah aku bantu berikan jawaban nya ya?	maaf sudah buat kakak ga nyaman ya sudah aku bantu berikan jawaban nya ya
5	buat ongkir dipengaruhi juga dari tingginya permintaan serta ketersediaan pengemudi di sekitar yahl.	buat ongkir dipengaruhi juga dari tingginya permintaan serta ketersediaan pengemudi di sekitar yahl

Pre-Processing (lanjutan)

- Pengubahan dari huruf besar ke huruf kecil semua.

Aa	Bb	Cc	Dd
Ee	Ff	Gg	Hh
Ii	Jj	Kk	Ll
Mm	Nn	Oo	Pp

TABEL I
PROSES CASE FOLDING

No	Proses Input	Proses Output
1	Malam Kak Leny @Leny_Brilyan, maaf banget udah bikin Kakak kecewa	malam kak leny @leny_brilyan, maaf banget udah bikin kakak kecewa
2	Ign syedih dong Kak, nantikan terus kejutan promo di wilayah lainnya yahl.	ign syedih dong kak, nantikan terus kejutan promo di wilayah lainnya yahl.
3	Apakah translation di GrabChat harus punya TOEFL 550, biar pas pesen Grabfood terus nge-chat abangnya "Siap mas!" gak di-translate "Ready gold!"?	apakah translation di grabchat harus punya toefl 550, biar pas pesen grabfood terus nge-chat abangnya "siap mas!" gak di-translate "ready gold!"?
4	Maaf sudah buat kakak ga nyaman ya, sudah aku bantu berikan jawaban nya ya	maaf sudah buat kakak ga nyaman ya, sudah aku bantu berikan jawaban nya ya
5	Buat ongkir dipengaruhi juga dari tingginya permintaan serta ketersediaan pengemudi di sekitar yahl.	buat ongkir dipengaruhi juga dari tingginya permintaan serta ketersediaan pengemudi di sekitar yahl.

Semester Genap - TA.2022/2023 17

Pre-Processing (lanjutan)

- Parsing dan Stemming**
 - Penguraian kata kedalam bentuk tunggal dan pembentukan kata kedalam bentuk dasarnya.
 - Sehingga kata-kata yang mempunyai bentuk kata dasar yang sama akan dikelompokkan.

TABEL III
PROSES STEMMING

No	Sebelum Proses	Setelah Proses
1	malam kak leny leny brilyan maaf banget udah bikin kakak kecewa	malam kak leny leny brilyan maaf sudah bikin kakak kecewa
2	ign syedih dong kak nantikan terus kejutan promo di wilayah lainnya yahl.	janan sedih kak nanti promo di wilayah lain ya
3	apakah translation di grabchat harus punya toefl 550 biar pas pesen grabfood terus nge-chat abangnya "Siap mas!" gak di-translate "Ready gold!"?	apa translation grabchat harus punya toefl 550 biar pesen grabfood chat abang siap mas translate ready gold
4	maaf sudah buat kakak ga nyaman ya, sudah aku bantu berikan jawaban nya ya	maaf sudah buat kakak ga nyaman ya sudah aku bantu berikan jawaban nya ya
5	buat ongkir dipengaruhi juga dari tingginya permintaan serta ketersediaan pengemudi di sekitar yahl.	buat ongkir pengaruh dari tinggi minta serta sedia pengemudi di sekitar ya

Semester Genap - TA.2022/2023 18

Pre-Processing (lanjutan)

- Pembobotan**
 - Dimulai dengan perhitungan jumlah kata dalam setiap dokumen, yang kemudian akan dihitung menggunakan skema pembobotan yang dikehendaki.

	A	B
1 word		weight
2 hai		-2
3 merestam		-2
4 ekstrensif		-2
5 persimpang		-4
6 detail		-4
7 pemak		-3
8 belat		-4
9 weta		-4
10 kabung		-2
11 rahayu		-4
12 maaf		-1
13 hualo		-4
14 pramo		-2
15 termakasih		-2

	A	B
1 word		weight
2 julus tali gantun		-2
3 gubahan		-2
4 gobat hali		-2
5 beresuluh (geras)		-1
6 isak		-4
7 larat hali		-3
8 melangsa		-3
9 remuk nadam		-4
10 tidak segan		-2
11 gemar		-1
12 tak segan		-1
13 keas		-4
14 jengon		-2
15 penghayatan		-2

malam maaf banget aku bantu cek ya

0	1	1	1	3	1	1	2
1	0	0	1	1	0	0	2
2	0	1	0	4	1	2	3
3	0	0	0	0	0	0	0
4	1	0	0	3	1	1	2
5	1	1	1	3	1	1	2
6	0	1	0	1	0	0	2
7	0	0	0	0	0	0	1
8	0	0	0	0	0	0	0
9	0	0	0	2	0	1	1

10 rows x 390 columns

Gambar. 7 Hasil pemberian bobot pada dataset crawling..

Semester Genap - TA.2022/2023 19

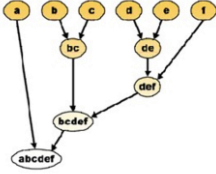
Text Mining

Proses pada Text Mining

Semester Genap - TA.2022/2023 20

Proses pada Text Mining

- Input dari proses *text mining* berupa kumpulan data terstruktur maupun tidak terstruktur.
- Sedangkan *output*-nya merupakan pengetahuan berkonteks khusus yang dapat digunakan untuk mengambil keputusan.
- Kontrol atau hambatan prosesnya mencakup keterbatasan *hardware*, masalah privasi, kesulitan pemrosesan teks yang ditampilkan dalam bentuk *natural language*.
- Mekanisme proses termasuk teknik yang tepat, peralatan *software*, dan domain keahlian.



Garis besar cara kerja text mining

Semester Genap - TA.2022/2023 21

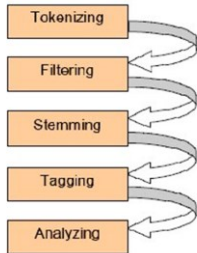
Proses pada Text Mining (lanjutan)

Pada level yang sangat tinggi, proses *text mining* dapat dipecah menjadi 3 *task* yaitu :

- 1) Membentuk Korpus :
 - Bertujuan mengumpulkan semua dokumen yang berhubungan dengan konteks yang sedang dipelajari. Setelah dikumpulkan, dokumen-dokumen teks di ubah dan diorganisir dalam suatu bentuk sehingga dokumen-dokumen tersebut berada dalam bentuk representasi yang sama.
- 2) Menciptakan *term-document matrix* :
 - Pada tahap ini, pendigitalan dan pengorganisasian dokumen (korpus) digunakan untuk menciptakan *term-document matrix* (TDM). Tujuan utama tahap ini adalah meng-konvert daftar dokumen yang terorganisir kedalam TDM dimana sel-sel nya berisi indeks yang paling tepat.
- 3) Menggali pengetahuan :
 - Text mining* menggunakan *data mining* algoritma seperti *classification*, *clustering*, *association*, dan *sequence discovery* untuk menggali pengetahuan dari algoritma tersebut.

Semester Genap - TA.2022/2023 22

Proses pada Text Mining (lanjutan)



Tahapan yang dilakukan secara umum



Semester Genap - TA.2022/2023 23

Proses pada Text Mining (lanjutan)

- Tokenizing**
 - Merupakan tahap pemotongan *string input* berdasarkan tiap kata yang menyusunnya.

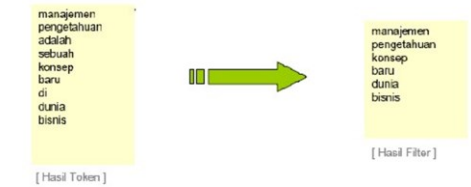


Contoh Tokenizing

Semester Genap - TA.2022/2023 24

Proses pada Text Mining (lanjutan)

- **Filtering**
 - Merupakan tahap mengambil kata-kata penting dari hasil token.
 - Biasanya menggunakan algoritma *stop-list* (membuang kata yang kurang penting), atau *word-list* (menyimpan kata penting).

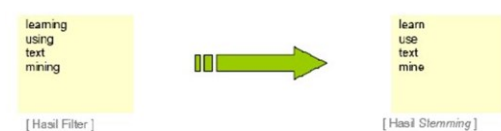


Contoh Filtering

Semester Genap - TA.2022/2023 25

Proses pada Text Mining (lanjutan)

- **Stemming**
 - Merupakan tahap mencari *root*/bentuk asal kata dari setiap kata hasil *filtering*.

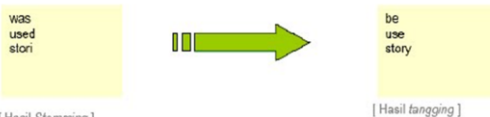


Contoh Stemming

Semester Genap - TA.2022/2023 26

Proses pada Text Mining (lanjutan)

- **Tagging**
 - Merupakan tahap mencari bentuk asal kata/*root* dari setiap kata lampau, atau kata hasil *stemming*.

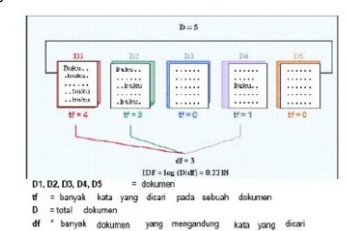


Contoh Tagging

Semester Genap - TA.2022/2023 27

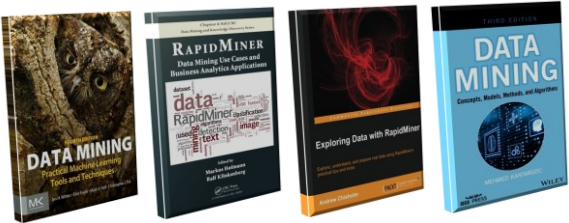
Proses pada Text Mining (lanjutan)

- **Analyzing**
 - Merupakan tahap penentuan seberapa jauh keterhubungan antar kata-kata dari dokumen yang ada.



Ilustrasi Algoritma Text Mining

Semester Genap - TA.2022/2023 28



Text Mining

Web Mining

Semester Genap - TA.2022/2023 29

Web Mining

- Merupakan proses menemukan hubungan intrinsik dari data *web* yang diekspresikan dalam bentuk tekstual, *linkage*, atau informasi yang berguna.
- *Web mining* dapat didefinisikan sebagai penemuan dan analisis informasi yang menarik dan berguna dari *web*, mengenai web, dan biasanya dilakukan dengan menggunakan peralatan berbasis *web*.
- Berdasarkan analisis target, *web mining* dapat dibagi menjadi 3 jenis yaitu :
 - *Web Usage Mining*
 - *Web Content Mining*
 - *Web Structure Mining*

Semester Genap - TA.2022/2023 30

Web Mining (lanjutan)

- **Web Usage Mining**
 - Merupakan penggalian informasi yang berguna dari data yang dihasilkan melalui kunjungan dan transaksi halaman *web*.
 - *Web usage mining* mengacu pada pengembangan informasi yang berguna melalui analisis *web server log*, profil pengguna, dan informasi transaksi.



Semester Genap - TA.2022/2023 31

Web Mining (lanjutan)

- **Web Content Mining**
 - Mengacu pada penggalian data yang berguna dari halaman *web*.
 - Dokumen-dokumen mungkin digali dalam beberapa mesin dengan format yang dapat dibaca, sehingga teknik otomatis dapat menghasilkan beberapa informasi mengenai halaman *web*.
 - *Web content mining* kadang disebut juga *web text mining*, karena isi teks adalah daerah yang paling banyak diteliti.
 - Teknologi yang biasa digunakan dalam *web content mining*, adalah NLP (*Natural Language Processing*) dan IR (*Information Retrieval*).



Semester Genap - TA.2022/2023 32

Web Mining
(lanjutan)

- **Web Structure Mining**
 - Proses penggalian informasi yang berguna dari *link-link* yang tertanam pada dokumen *web*.
 - *Web structure mining* digunakan untuk mengidentifikasi kepemilikan *web* dan hubungannya, yang mana merupakan pilar dari algoritma *page-rank* (peringkat halaman).

33

Web Mining
(lanjutan)

- **Implementasi Web Mining**
 - *Web Mining* menganalisis web logs untuk informasi berguna yang berkaitan dengan pelanggan yang dapat membantu personalisasi situs web berdasarkan perilaku pengguna.
 - *Web mining* juga digunakan untuk mencari web untuk kata kunci, frase dan konten lainnya yang berhubungan.
 - Perangkat lunak open source untuk web mining termasuk RapidMiner, yang menyediakan modul untuk pengelompokan teks, kategorisasi teks, ekstraksi informasi, yang dinamakan pengakuan entitas, dan analisis sentimen.
 - RapidMiner digunakan misalnya pada aplikasi penyaringan berita otomatis untuk personalisasi survey berita.

Semester Genap - TA.2022/2023

34

Referensi

- Ian H Witten, Eibe Frank, Mark A Hall. (2016). *Data Mining - Practical Machine Learning Tools and Techniques*, 4th Edition. Morgan-Kaufman.
- Hofmann, Markus., Klinkenberg, Ralf. (2016). *Data Mining Use Cases and Business Analytics Application*. CRC Press.
- Crisholm, Andrew. (2013). *Exploring Data With Rapidminer*. Packt Publishing.
- Mehmed Kantardzic. (2020). *Data Mining - Concepts, Models, Methods, and Algorithms*, 3rd edition. Wiley - IEEE Press.
- rapidminer.com, 2021,. Rapidminer Manual and Documentation, <https://docs.rapidminer.com/latest/studio/operators/>.

Semester Genap - TA.2022/2023

35

Silaturahmi
adalah menjalin hubungan
dengan kerabat.

Terima Kasih

مَنْ سَرَّهُ أَنْ يُبْسَطَ لَهُ فِي رِزْقِهِ ، وَأَنْ يُنْسَأَ لَهُ فِي أَثَرِهِ ، فَلْيَصِلْ رَجُلَهُ
"Siapa yang suka dilapangkan rizqnya dan dipanjangkan umurnya hendaklah dia menyambung silaturahmi." (HR. Bukhari-Muslim)



Data Mining



Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

● ● ●

Session 09



PROGRAM STUDI TEKNIK INFORMATIKA

Sumber: <https://ft.uhamka.ac.id/prodi/informatika/>

- VISI.
 - ✓ Menjadi program studi informatika yang menghasilkan karya unggul di bidang teknologi informasi melalui pembelajaran berlandaskan prophetic teaching untuk membentuk lulusan yang cerdas secara spiritual, intelektual, emosional, dan sosial.
- MISI.
 1. Melaksanakan pendidikan di bidang informatika yang terintegrasi dengan nilai-nilai Al Islam Kemuhammadiyah.
 2. Melaksanakan penelitian untuk menghasilkan karya teknologi informasi yang unggul untuk masyarakat berkembang.
 3. Melaksanakan pengabdian kepada masyarakat untuk kesejahteraan umat.
 4. Menghasilkan mahasiswa yang berwawasan global dalam teknologi informasi.
 5. Mendorong jiwa kewirausahaan mahasiswa untuk berkiprah di tengah masyarakat dalam meningkatkan kegiatan ekonomi.
 6. Menerapkan standar mutu pendidikan informatika melalui layanan mutu yang berkesinambungan.
 7. Membangun kerja sama terpadu dengan mitra yang berkontribusi untuk kemajuan pendidikan di bidang informatika.

Semester Genap - TA.2022/2023 2



Sistem Penilaian

- SKS : 3
 - Teori 3 SKS – (14 Kali Pertemuan)
- Keaktifan (bobot 10%)
 - Terdiri dari: Aktivitas Perkuliahan (*Kehadiran, Etika, dan Sikap*).
- Project (bobot 25%)
 - Terdiri dari: Tugas Besar / Project (*Penguasaan Keterampilan*) dengan cara observasi, mengimplementasikan dan menghasilkan Proyek Akhir, serta mempresentasikannya.
- Tes (UTS = bobot 25% , UAS = bobot 40%)
 - Terdiri dari: UTS dan UAS (*Penguasaan Pengetahuan*) dengan cara tes tertulis maupun tes praktikum.

Semester Genap - TA.2022/2023 3



Tujuan/Objectives (Session.09)

- Objectives (Tujuan Instruksional Umum).
 - RPS/SAP.
 - Asosiasi (part.1)
- Indikator Kompetensi.
 - Mahasiswa mampu mengetahui dan memahami tentang teknik asosiasi pada *data mining*.
 - Mahasiswa mampu mengetahui, memahami, serta dapat menjelaskan tentang metode dan teknik asosiasi didalam *data mining*.

Semester Genap - TA.2022/2023 4

Muhasabah

وَلَنْ تَرْضَىٰ عَنْكَ الْيَهُودُ وَلَا النَّصَارَىٰ حَتَّىٰ تَتَّبِعَ مِلَّتَهُمْ
 قُلْ إِنَّ هُدَى اللَّهِ هُوَ الْهُدَىٰ وَلَئِنْ اتَّبَعْتَ
 أَهْوَاءَهُمْ بَعْدَ الَّذِي جَاءَكَ مِنَ الْعِلْمِ
 مَا لَكَ مِنَ اللَّهِ مِن وَّلِيٍّ وَلَا نَصِيرٍ

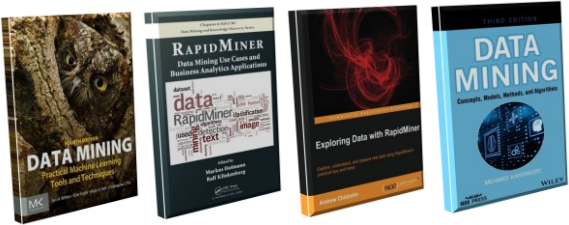
- Qs. Al Baqarah, ayat 120
 - Dan orang-orang Yahudi dan Nasrani tidak akan rela kepadamu (Muhammad) sebelum engkau mengikuti agama mereka. Katakanlah, "Sesungguhnya petunjuk Allah itulah petunjuk (yang sebenarnya)."
 - Dan jika engkau mengikuti keinginan mereka setelah ilmu (kebenaran) sampai kepadamu, tidak akan ada bagimu pelindung dan penolong dari Allah.

Semester Genap - TA.2022/2023

Data Mining


Uhamka
 UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA
integrity, trust, compassion

Session 09
 Asosiasi (part.1)



Asosiasi (part.1)

Pendahuluan

Semester Genap - TA.2022/2023

Pendahuluan

- Teknik Asosiasi (*association*) atau biasa disebut juga sebagai *association rule mining*.
- Merupakan teknik didalam *data mining* untuk menemukan aturan asosiatif antara suatu kombinasi *item*.
- Diusulkan oleh Agrawal, Imielinski, and Swami (1993).
- Teknik ini biasa digunakan untuk menemukan pola yang mendeteksi kumpulan atribut-atribut yang muncul bersamaan dalam frekuensi yang sering.

Semester Genap - TA.2022/2023



Pendahuluan (lanjutan)

- Biasa disebut dengan *affinity analysis* atau *market basket analysis* (menganalisa kebiasaan *customer* dengan mencari asosiasi dan korelasi dari data transaksi).
- Contoh penggunaan:
 - Dalam suatu supermarket kita ingin mengetahui seberapa banyak orang yang membeli celana juga membeli sabuk/gesper ?
 - Sebagai saran penempatan barang atau *item* dalam supermarket.
 - Sebagai saran produk apa yang akan digunakan dalam promosi / *bundling*.

Semester Genap - TA.2022/2023

9



Pendahuluan (lanjutan)

- Didalam teknik asosiasi, menggunakan algoritma sebagai berikut:
 - Algoritma:
 - A Priori
 - FP-Growth
 - GRI
 - dll



Customers who bought this item
also bought

Semester Genap - TA.2022/2023

10



Pendahuluan (lanjutan)

- Input dan Problema
 - Input
 - Adanya sejumlah transaksi.
 - Setiap transaksi memuat kumpulan *item*.
 - Problema
 - Bagaimana caranya menemukan *association rule* yang memenuhi minimum *support* dan minimum *confidence* yang kita berikan.

Semester Genap - TA.2022/2023

11



Pendahuluan (lanjutan)

- *Association rule* biasanya dinyatakan dalam bentuk :

$\{\text{TEH}\} \rightarrow \{\text{GULA}\} \text{ (support = 40\%, confidence = 50\%)}$
- Yang artinya : "50% dari transaksi di *database* yang memuat *item* **TEH** juga memuat *item* **GULA**.
- Sedangkan 40% dari seluruh transaksi yang ada di *database* memuat kedua *item* tersebut."

Semester Genap - TA.2022/2023

12

Pendahuluan
(lanjutan)

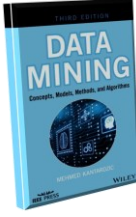
- Dapat juga diartikan bahwa :
 "Seorang konsumen yang membeli **TEH** punya kemungkinan 50% untuk juga membeli **GULA**. Aturan ini cukup signifikan karena mewakili 40% dari catatan transaksi selama ini."




$\{\text{TEH}\} \rightarrow \{\text{GULA}\}$ (support = 40%, confidence = 50%)

Semester Genap - TA.2022/2023 13



Asosiasi (part.1)
Definisi Umum

Semester Genap - TA.2022/2023 14

Definisi umum

- Itemset:**
 - Himpunan dari *item-item* yang muncul bersama-sama.
- Kaidah asosiasi:**
 - Peluang bahwa *item-item* tertentu hadir bersama-sama.
- Support** dari suatu *itemset* X ($\text{supp}(X)$) adalah rasio dari jumlah transaksi, dimana *itemset* muncul dengan total jumlah transaksi.

Semester Genap - TA.2022/2023 15

Definisi umum
(lanjutan)

- Konfidence (keyakinan) dari kaidah $X \rightarrow Y$, ditulis $\text{conf}(X \rightarrow Y)$ adalah:
 - $\text{conf}(X \rightarrow Y) = \text{supp}(X \cup Y) / \text{supp}(X)$
 - Konfidence bisa juga didefinisikan dalam terminologi peluang bersyarat sebagai berikut:

$$\text{conf}(X \rightarrow Y) = P(Y|X) = P(X \cap Y) / P(X)$$
- Database** transaksi menyimpan data transaksi.
- Data transaksi bisa juga disimpan dalam suatu bentuk lain dari suatu *database* $m \times n$.

Semester Genap - TA.2022/2023 16



Ukuran Support

- Misalkan $I = \{I_1, I_2, \dots, I_m\}$ merupakan suatu himpunan dari literal, yang disebut item-item.
- Misalkan $D = \{T_1, T_2, \dots, T_n\}$ merupakan suatu himpunan dari n transaksi, dimana untuk setiap transaksi $T \in D, T \subseteq I$.
- Suatu himpunan item $X \subseteq I$ disebut *itemset*.
- Suatu transaksi T memuat suatu *itemset* X jika $X \subseteq T$.
- Setiap *itemset* X diasosiasikan dengan suatu himpunan transaksi $T_x = \{T \in D \mid T \supseteq X\}$ yang merupakan himpunan transaksi yang memuat *itemset*.
- Support dari *itemset* $X \rightarrow \text{supp}(X)$:

$$|T_x| / |D|$$

Semester Genap - TA.2022/2023 17



Contoh

Transaksi	A	B	C	D
T1	1	0	1	14
T2	0	0	6	0
T3	1	0	2	4
T4	0	0	4	0
T5	0	0	3	1
T6	0	0	1	13
T7	0	0	8	0
T8	4	0	0	7
T9	0	1	1	10
T10	0	0	0	18

Jumlah transaksi $|D| = 10$

Kemunculan item A pada transaksi ($|T_a|$) sebanyak 3 kali yaitu pada T1, T3, T8.

$\text{Supp}(A) = |T_a| / |D|$
 $= 3/10 = 0.3$

$|T_{cd}|$ sebanyak 5 kali, yaitu pada T1, T3, T5, T6, T9.

$\text{Supp}(CD) = |T_{cd}| / |D|$
 $= 5/10 = 0.5$

Semester Genap - TA.2022/2023 18



Apriori

Itemset	Sp
A	0.3
B	0.1
C	0.8
D	0.7
AB	0
AC	0.2
AD	0.3
BC	0.1
BD	0.1
CD	0.5
ABC	0
ABD	0
ACD	0.2
BCD	0.1
ABCD	0

Jika *minsupport* diberikan oleh user sebagai *threshold* adalah 0.2, maka *frequent itemset* adalah semua *itemset* yang support-nya ≥ 0.2 , yakni:
 A, C, D, AC, AD, CD, ACD

Dari *frequent itemset* bisa dibangun kaidah asosiasi sbb:

$A \rightarrow C \quad C \rightarrow A \quad A \rightarrow D$
 $D \rightarrow A \quad C \rightarrow D \quad D \rightarrow C$
 $A, C \rightarrow DA, D \rightarrow CC, D \rightarrow A$

$\text{Conf}(A \rightarrow C) = \text{supp}(A, C) / \text{supp}(A)$

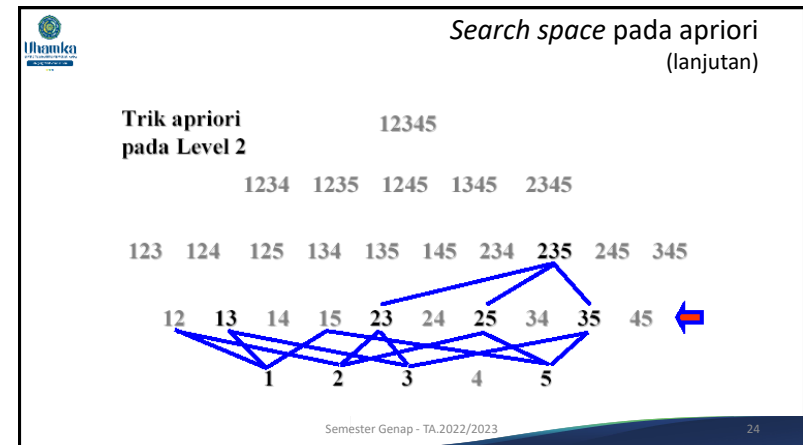
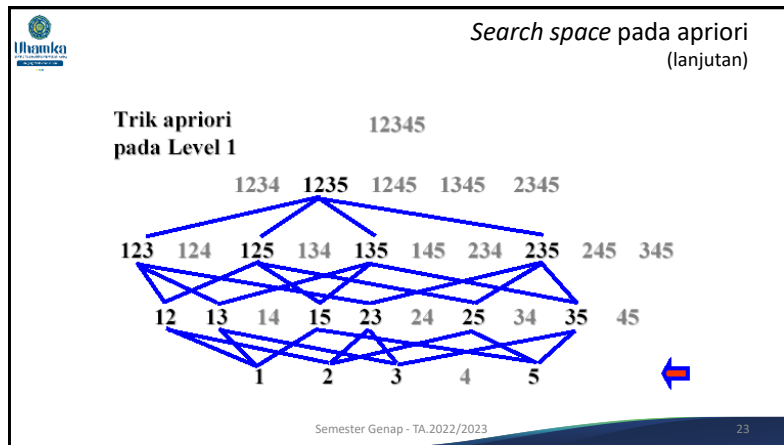
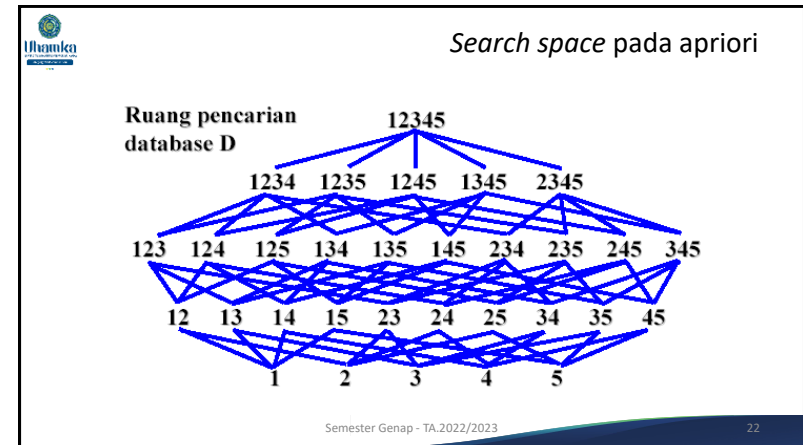
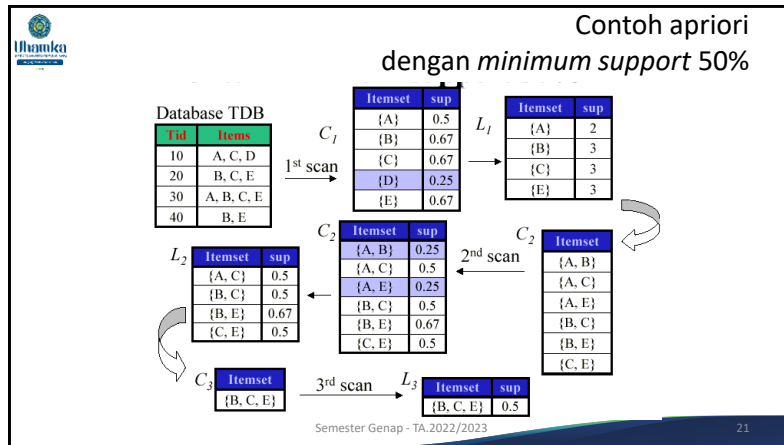
Semester Genap - TA.2022/2023 19



Prinsip apriori :

- Subset apapun dari suatu *frequent itemset* harus *frequent*
- $L_3 = \{abc, abd, acd, ace, bcd\}$
- Penggabungan sendiri : $L_3 * L * L_3$
 - $abcd$ dari abc dan abd
 - $acde$ dari acd dan ace
- Pemangkasan Pemangkasan:
 - $acde$ dibuang sebab ade tidak dalam L_3
- $C_4 = \{abcd\}$

Semester Genap - TA.2022/2023 20



Latihan

T1	{roti, selai, mentega}
T2	{roti, mentega}
T3	{roti, susu, mentega}
T4	{coklat, roti}
T5	{coklat, susu}

Itemset	Sp
{roti}	0.8
{selai}	0.2
{mentega}	0.6
{susu}	0.4
{coklat}	0.4

Itemset	Sp
{roti, mentega}	0.6
{roti, susu}	0.2
{roti, coklat}	0.2
{mentega, susu}	0.2
{mentega, coklat}	0
{susu, coklat}	0.2

$\text{Conf}(\text{roti} \rightarrow \text{mentega}) = \frac{\text{Supp}(\{\text{roti, mentega}\})}{\text{Supp}(\{\text{roti}\})} = \frac{0.6}{0.8} = 0.75 \rightarrow 75\%$
 $\text{Conf}(\text{mentega} \rightarrow \text{roti}) = \frac{\text{Supp}(\{\text{mentega, roti}\})}{\text{Supp}(\{\text{mentega}\})} = \frac{0.6}{0.6} = 1 \rightarrow 100\%$

Semester Genap - TA.2022/2023

Latihan

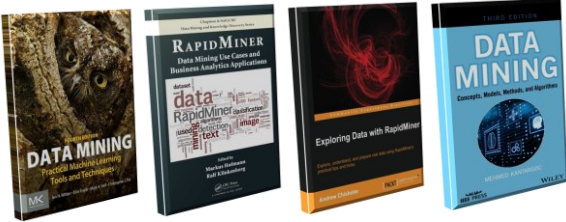
T1	{roti, selai, mentega}
T2	{roti, mentega}
T3	{roti, susu, mentega}
T4	{coklat, roti}
T5	{coklat, susu}

- Suatu supermarket mempunyai sejumlah transaksi seperti dalam tabel
- Buatlah *association rule* dari data tersebut dengan cara menghitung *support* dan *confidence*
- Pakailah metode apriori dengan minimum *support*=0.3 dan *confidence*=0.8

Semester Genap - TA.2022/2023

Asosiasi (part.1)

Metodologi Dasar



Semester Genap - TA.2022/2023

Metodologi Dasar

- Metodologi dasar analisis pada teknik asosiasi terbagi menjadi dua tahap, yaitu:
 - Analisis pola frekuensi tinggi.
 - Pembentukan aturan asosiasi.

Semester Genap - TA.2022/2023

Metodologi Dasar
(lanjutan)

1) Analisis 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:

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

- Sementara nilai *support* dari 2 *item* diperoleh dengan rumus berikut:

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

Semester Genap - TA.2022/2023 29

Metodologi Dasar
(lanjutan)

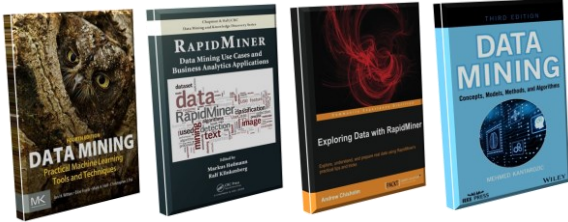
2) Pembentukan Aturan Asosiasi

- Setelah semua pola frekuensi tinggi ditemukan, maka setelah itu kita dapat mencari aturan asosiatif yang memenuhi syarat minimum untuk *confidence*, dengan cara menghitung *confidence* dari aturan asosiatif $A \rightarrow B$.
- Nilai *confidence* dari aturan $A \rightarrow B$ diperoleh dari rumus berikut:

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

Semester Genap - TA.2022/2023 30

Asosiasi (part.1)
Contoh Kasus



Semester Genap - TA.2022/2023 31

Contoh Data Transaksi

- Dibawah ini adalah contoh 10 data transaksi dari sebuah mini-market 212.

Data Transaksi

Transaksi	Item yang dibeli
1	Teh , Gula , Susu
2	Roti , Teh , Gula
3	Teh , Gula
4	Susu , Roti
5	Susu , Gula , Roti
6	Susu , Gula , Teh
7	Roti , Teh
8	Gula , Roti
9	Gula , Susu , Roti
10	Teh , Susu , Roti

Daftar item yang dibeli: Teh, Gula, Susu, Roti

Semester Genap - TA.2022/2023 32



Daftar Istilah

I adalah himpunan yang tengah dibicarakan.
Contoh: $I = \{\text{Teh, Gula, Susu, Roti}\}$

D adalah Himpunan seluruh transaksi yang tengah dibicarakan
Contoh: $D = \{\text{Transaksi 1, Transaksi 2, ..., Transaksi 10}\}$

Semester Genap - TA.2022/2023 33



Daftar Istilah (lanjutan)

Proper Subset adalah himpunan bagian murni

Contoh:
Ada suatu himpunan $I = \{a, b, c\}$
Himpunan Bagian dari I adalah
Himpunan Kosong = $\{\}$
Himpunan 1 Unsur = $\{a\}, \{b\}, \{c\}$
Himpunan 2 Unsur = $\{a, b\}, \{a, c\}, \{b, c\}$
Himpunan 3 Unsur = $\{a, b, c\}$
Proper subset nya adalah Himpunan 1 Unsur dan Himpunan 2 Unsur

Semester Genap - TA.2022/2023 34



Daftar Istilah (lanjutan)

Item Set adalah himpunan item atau item-item di I
Contoh:
Ada suatu himpunan $I = \{a, b, c\}$
Item set nya adalah
 $\{a\}; \{b\}; \{c\}; \{a, b\}; \{a, c\}; \{b, c\}$

K-item set adalah Item set yang terdiri dari K buah item yang ada pada I . Intinya K itu adalah jumlah unsur yang terdapat pada suatu himpunan.
Contoh:
2-item set adalah yang bersifat 2 unsur.

Semester Genap - TA.2022/2023 35



Daftar Istilah (lanjutan)

Data Transaksi

Transaksi	Item yang dibeli
1	Teh , Gula , Susu
2	Roti , Teh , Gula
3	Teh , Gula
4	Susu , Roti
5	Susu , Gula , Roti
6	Susu , Gula , Teh
7	Roti , Teh
8	Gula , Roti
9	Gula , Susu , Roti
10	Teh , Susu , Roti

Daftar item yang dibeli: Teh, Gula, Susu, Roti

Item Set Frekuensi adalah Jumlah transaksi di I yang mengandung jumlah item set tertentu.
Intinya jumlah transaksi yang membeli suatu item set.
Contoh:
Kita gunakan tabel transaksi di atas:
Frekuensi Item set yang sekaligus membeli Teh dan Gula adalah 4

Semester Genap - TA.2022/2023 36



Daftar Istilah (lanjutan)

Frekuen Item Set adalah item set yang muncul sekurang-kurangnya "sekian" kali di D. Kata "sekian" biasanya di simbolkan dengan Φ . Φ merupakan batas minimum dalam suatu transaksi.

Contoh:

Pertama kita tentukan $\Phi = 3$, karena jika tidak di tentukan maka frekuen item set tidak dapat di hitung. Jika $\Phi=3$ untuk {Teh, Gula} apakah frekuen Item set?

Jika kita hitung maka jumlah transaksi yang membeli Teh sekaligus membeli Gula adalah 4. Karena $4 \geq 3$ maka {Teh, Gula} merupakan Frekuen Item set.

Fk adalah Himpunan semua frekuen Item Set yang terdiri dari K item.

Semester Genap - TA.2022/2023

37



Tahapan Algoritma

1. Tentukan Φ
2. Tentukan semua Frekuen Item Set
3. Untuk setiap Frekuen Item Set lakukan hal sbb:
 - i. Ambil sebuah unsur, namakanlah s
 - ii. Untuk sisanya namakanlah ss-s
 - iii. Masukkan unsur-unsur yang telah di umpamakan ke dalam *rule If (ss-s) then s* (untuk langkah ke 3 lakukan untuk semua unsur)

Semester Genap - TA.2022/2023

38



Contoh Data Transaksi

Transaksi	Item yang dibeli
1	A, B, C
2	D, A, B
3	A, B
4	C, D
5	C, B, D
6	C, B, A
7	D, A
8	B, D
9	B, C, D
10	A, C, D

Nilai minimum frekuensi kemunculan $\Phi = 4$ dan nilai minimal confidence = 65%

Semester Genap - TA.2022/2023

39



Format Tabular Data Transaksi

1-Item Set / 1 Unsur / F1

Transaksi	A	B	C	D
1	1	1	1	0
2	1	1	0	1
3	1	1	0	0
4	0	0	1	1
5	0	1	1	1
6	1	1	1	0
7	1	0	0	1
8	0	1	0	1
9	0	1	1	1
10	1	0	1	1
	6	7	6	7

Semester Genap - TA.2022/2023

40

Format Tabular Data Transaksi
(lanjutan)

2-Item Set / 2 Unsur / F2

Transaksi	A	B	f
1	1	1	P
2	1	1	P
3	1	1	P
4	0	0	S
5	0	1	S
6	1	1	P
7	1	0	S
8	0	1	S
9	0	1	S
10	1	0	S
	6	7	4

Transaksi	A	C	f
1	1	1	P
2	1	0	S
3	1	0	S
4	0	1	S
5	0	1	S
6	1	1	P
7	1	0	S
8	0	0	S
9	0	1	S
10	1	1	P
	6	6	3

Transaksi	A	D	f
1	1	0	S
2	1	1	P
3	1	0	S
4	0	1	S
5	0	1	S
6	1	0	S
7	1	1	P
8	0	1	S
9	0	1	S
10	1	1	P
	6	7	3

Transaksi	B	C	f
1	1	1	P
2	1	0	S
3	1	0	S
4	0	1	S
5	1	1	P
6	1	1	P
7	0	0	S
8	1	0	S
9	1	1	P
10	0	1	S
	7	6	4

Transaksi	B	D	f
1	1	0	S
2	1	1	P
3	1	0	S
4	0	1	S
5	1	1	P
6	1	0	S
7	0	1	S
8	1	1	P
9	1	1	P
10	0	1	S
	7	7	4

Transaksi	C	D	f
1	1	0	S
2	0	1	S
3	0	0	S
4	1	1	P
5	1	1	P
6	1	0	S
7	0	1	S
8	0	1	S
9	1	1	P
10	1	1	P
	6	7	4

41

Format Tabular Data Transaksi
(lanjutan)

3-Item Set / 3 Unsur / F3

Transaksi	A	B	C	f
1	1	1	1	P
2	1	1	0	S
3	1	1	0	S
4	0	0	1	S
5	0	1	1	S
6	1	1	1	P
7	1	0	0	S
8	0	1	0	S
9	0	1	1	S
10	1	0	1	S
	6	7	6	2

Transaksi	B	C	D	f
1	1	1	0	S
2	1	0	1	S
3	1	0	0	S
4	0	1	1	S
5	1	1	1	P
6	1	1	0	S
7	0	0	1	S
8	1	0	1	S
9	1	1	1	P
10	0	1	1	S
	7	6	7	2

Transaksi	A	B	D	f
1	1	1	0	S
2	1	1	1	P
3	1	1	0	S
4	0	0	1	S
5	0	1	1	S
6	1	1	0	S
7	1	0	1	S
8	0	1	1	S
9	0	1	1	S
10	1	0	1	S
	6	7	7	1

42

Rule IF X THEN Y

- *Rule* yang digunakan adalah **if x then y**, dimana x adalah *antecedent* dan y adalah *consequent*.
- Berdasarkan *rule* tersebut, maka dibutuhkan 2 buah *item* yang mana salah satunya sebagai *antecedent* dan sisanya sebagai *consequent*.
- **F1** tidak disertakan, karena hanya terdiri dari 1 *item* saja.
- Untuk *antecedent* boleh lebih dari 1 unsur, sedangkan untuk *consequent* terdiri dari 1 unsur.

Semester Genap - TA.2022/2023

43

Rule IF X THEN Y
(lanjutan)

If x then y	Support	Confidence
A - B	40	66.67
B - A	40	57.14
B - C	40	57.14
C - B	40	66.67
B - D	40	57.14
D - B	40	57.14
C - D	40	66.67
D - C	40	57.14

Semester Genap - TA.2022/2023

44



Referensi

- Ian H Witten, Eibe Frank, Mark A Hall. (2016). *Data Mining - Practical Machine Learning Tools and Techniques*, 4th Edition. Morgan-Kaufman.
- Hofmann, Markus., Klinkenberg, Ralf. (2016). *Data Mining Use Cases and Business Analytics Application*. CRC Press.
- Crisholm, Andrew. (2013). *Exploring Data With Rapidminer*. Packt Publishing.
- Mehmed Kantardzic. (2020). *Data Mining - Concepts, Models, Methods, and Algorithms*, 3rd edition. Wiley - IEEE Press.
- rapidminer.com, 2021,. Rapidminer Manual and Documentation, <https://docs.rapidminer.com/latest/studio/operators/>.

Semester Genap - TA.2022/2023

45





Data Mining



Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

Session 10



PROGRAM STUDI TEKNIK INFORMATIKA

Sumber: <https://ft.uhamka.ac.id/prodi/informatika/>

- VISI.
 - ✓ Menjadi program studi informatika yang menghasilkan karya unggul di bidang teknologi informasi melalui pembelajaran berlandaskan prophetic teaching untuk membentuk lulusan yang cerdas secara spiritual, intelektual, emosional, dan sosial.
- MISI.
 1. Melaksanakan pendidikan di bidang informatika yang terintegrasi dengan nilai-nilai Al Islam Kemuhammadiyah.
 2. Melaksanakan penelitian untuk menghasilkan karya teknologi informasi yang unggul untuk masyarakat berkembang.
 3. Melaksanakan pengabdian kepada masyarakat untuk kesejahteraan umat.
 4. Menghasilkan mahasiswa yang berwawasan global dalam teknologi informasi.
 5. Mendorong jiwa kewirausahaan mahasiswa untuk berkiprah di tengah masyarakat dalam meningkatkan kegiatan ekonomi.
 6. Menerapkan standar mutu pendidikan informatika melalui layanan mutu yang berkesinambungan.
 7. Membangun kerja sama terpadu dengan mitra yang berkontribusi untuk kemajuan pendidikan di bidang informatika.

Semester Genap - TA.2022/2023 2



Sistem Penilaian

- SKS : 3
 - Teori 3 SKS – (14 Kali Pertemuan)
- Keaktifan (bobot 10%)
 - Terdiri dari: Aktivitas Perkuliahan (*Kehadiran, Etika, dan Sikap*).
- Project (bobot 25%)
 - Terdiri dari: Tugas Besar / Project (*Penguasaan Keterampilan*) dengan cara observasi, mengimplementasikan dan menghasilkan Proyek Akhir, serta mempresentasikannya.
- Tes (UTS = bobot 25% , UAS = bobot 40%)
 - Terdiri dari: UTS dan UAS (*Penguasaan Pengetahuan*) dengan cara tes tertulis maupun tes praktikum.

Semester Genap - TA.2022/2023 3



Tujuan/Objectives (Session.10)

- Objectives (Tujuan Instruksional Umum).
 - RPS/SAP.
 - Asosiasi (part.2) / Market Basket Analysis
- Indikator Kompetensi.
 - Mahasiswa mampu mengetahui dan memahami tentang teknik asosiasi dan market basket analysis pada *data mining*.
 - Mahasiswa mampu mengetahui, memahami, serta dapat menjelaskan tentang metode dan teknik asosiasi, terkait *market basket analysis* didalam *data mining*.

Semester Genap - TA.2022/2023 4



Muhasabah

وَلْتَبْلُوْكُمْ بِشَيْءٍ مِّنَ الْخَوْفِ وَالْجُوعِ وَنَقْصٍ مِّنَ الْأَمْوَالِ وَالْأَنْفُسِ وَالثَّمَرَاتِ وَبَشِّرِ الصَّابِرِينَ الَّذِينَ إِذَا أَصَابَتْهُمُ مُّصِيبَةٌ قَالُوا إِنَّا لِلَّهِ وَإِنَّا إِلَيْهِ رَاجِعُونَ أُولَٰئِكَ عَلَيْهِمْ صَلَوَاتٌ مِّن رَّبِّهِمْ وَرَحْمَةٌ وَأُولَٰئِكَ هُمُ الْمُهْتَدُونَ

- Qs. Al-Baqarah, ayat 155-157
 - Dan Kami pasti akan menguji kamu dengan sedikit ketakutan, kelaparan, kekurangan harta, jiwa, dan buah-buahan. Dan sampaikanlah kabar gembira kepada orang-orang yang sabar,
 - (yaitu) orang-orang yang apabila ditimpa musibah, mereka berkata “Inna lillahi wa inna ilaihi raji’un” (sesungguhnya kami milik Allah dan kepada-Nya lah kami kembali).
 - Mereka itulah yang memperoleh ampunan dan rahmat dari Tuhannya, dan mereka itulah orang-orang yang mendapat petunjuk.

Semester Genap - TA.2022/2023
5



Data Mining



Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

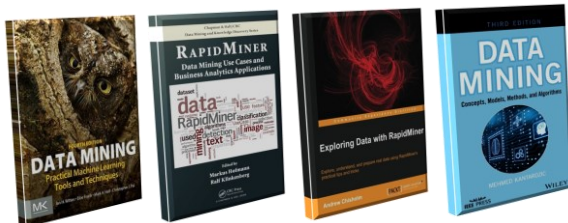
Session 10

Asosiasi (part.2)

Semester Genap - TA.2022/2023
6



Asosiasi (part.2)



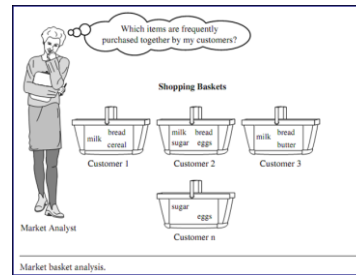
PendahuLuan

Semester Genap - TA.2022/2023
7

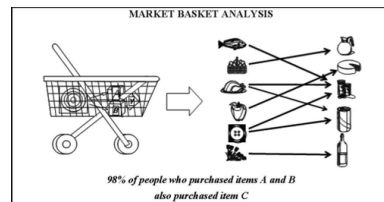


Pendahuluan

“Market Basket Analysis is, how to Understanding Your Customer Purchasing Patterns”
Anastasia Sherin



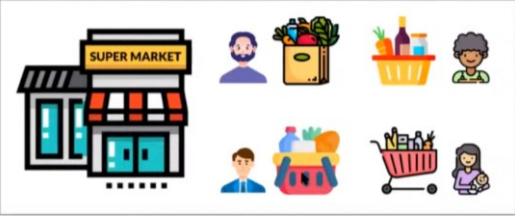
Market basket analysis.



Semester Genap - TA.2022/2023
8

Pendahuluan
(lanjutan)

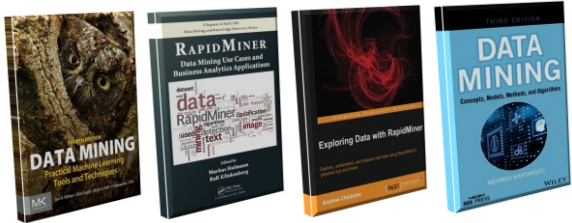
- What is Market Basket Analysis ?



- Uncover Associations between items :
 - Mencari pola, terkait barang apa yang sering dibeli bersamaan ?
 - Meng-identifikasi kebiasaan belanja pelanggan.

Semester Genap - TA.2022/2023 9

Asosiasi (part.2)
Story from Walmart



Semester Genap - TA.2022/2023 10

Story from Walmart



- Walmart merupakan salah satu *store chain* terbesar dari USA.
- Produk-produk yang dijual lengkap dan bermacam², mengusung konsep *one stop shopping*.
- Mempunyai lebih dari 245 juta pelanggan dan 10.500 toko di 24 negara.
- Annual revenue* sebesar \$572,8 milyar, atau setara dengan Rp.8.374 triliun (Rp.22,9 triliun per hari).

Semester Genap - TA.2022/2023 11

Story from Walmart
(lanjutan)

- Walmart's Market Basket Analysis Story.

Awal mula	Dampak	Why?	Lesson learned
2009 Walmart mengurangi beberapa jenis barang yang dijual untuk mengurangi biaya	↓ Berdampak negatif terhadap penjualan mereka di tahun itu. Penjualan turun 1,6% di kuartal keempat saja	? Pelanggan beralih ke pesaing untuk pilihan barang yang lebih luas Walmart mengabaikan Market Basket Effect untuk membantu memahami hubungan produk dan pola pembelian pelanggan	Walmart membuang barang yang salah Contoh: Barang berfrekuensi rendah sering dibeli dengan barang berfrekuensi rendah lainnya. Jika Walmart tidak lagi menjual salah satu barang ini, mereka akan kehilangan pelanggan.

Semester Genap - TA.2022/2023 12

Story from Walmart
(lanjutan)

- Market Basket Implementation.

Membuat bundling Membuat promosi Mengatur susunan rak

Cross and up selling Menentukan storefront Menyusun strategi

Semester Genap - TA.2022/2023 13

Asosiasi (part.2)
Association Rule

Semester Genap - TA.2022/2023 14

Association Rule

- Tujuan: Mencari asosiasi atau hubungan pada data.

A → B Hubungan **jika - maka**, misalkan:
→ Jika seseorang membeli roti, maka ia akan membeli mentega

- Beberapa metrik yang sering digunakan:

Support	Confidence	Lift
Peluang pembelian roti dan mentega secara bersamaan.	Peluang pembeli membeli mentega, ketika ia membeli roti. (dalam istilah Matematika, dikenal dengan istilah peluang bersyarat)	Melihat dependensi atau keterkaitan antara pembelian roti dan mentega (apakah ada kaitan atau ternyata saling bebas).

Semester Genap - TA.2022/2023 15

Association Rule
(lanjutan)

- Contoh Encoding

Transaction ID	List of Item
1	
2	
3	
4	
5	

→

Transaction ID	Susu	Roti	Apel	Mentega	Selai
1	1	1	1	0	0
2	0	0	1	1	1
3	1	0	0	0	0
4	1	1	0	1	1
5	0	1	0	1	1

Semester Genap - TA.2022/2023 16

Association Rule
(lanjutan)

- Contoh Perhitungan

Jika seseorang membeli roti, maka ia akan membeli mentega

- Peluang membeli roti = $\frac{4}{5} \rightarrow \text{support roti}$
- Peluang membeli mentega = $\frac{3}{5} \rightarrow \text{support mentega}$
- Peluang membeli roti dan mentega = $\frac{2}{5}$

Transaction ID	List of Item	Support	peluang pembelian roti dan mentega	$\frac{2}{5}$
1		Confidence (roti $\rightarrow$ mentega)	$\frac{\text{support roti dan mentega}}{\text{support roti}}$	$\frac{2/5}{4/5} = 1/2$
2		Confidence (mentega $\rightarrow$ roti)	$\frac{\text{support roti dan mentega}}{\text{support mentega}}$	$\frac{2/5}{3/5} = 2/3$
3		Lift	$\frac{\text{support roti dan mentega}}{(\text{support roti} \times \text{support mentega})}$	$\frac{2/5}{(4/5 \cdot 3/5)} = 10/9$
4				
5				

Semester Genap - TA.2022/2023 17

Association Rule
(lanjutan)

- Perhitungan

Confidence (roti $\rightarrow$ mentega)	Confidence (mentega $\rightarrow$ roti)
Explanation: Your Input: $\frac{2}{5} \div \frac{4}{5}$ Step # 1 = $\frac{2 \times 5}{4 \times 5}$ Step # 2 = $\frac{10}{20}$ Step # 3 = $\frac{10 \div 10}{20 \div 10}$ Ans = $\frac{1}{2}$	Explanation: Your Input: $\frac{2}{5} \div \frac{3}{5}$ Step # 1 = $\frac{2 \times 5}{3 \times 5}$ Step # 2 = $\frac{10}{15}$ Step # 3 = $\frac{10 \div 5}{15 \div 5}$ Ans = $\frac{2}{3}$

Semester Genap - TA.2022/2023 18

Association Rule
(lanjutan)

Hasil perhitungan *lift*

- Kejadian membeli roti dan mentega, dikatakan saling bebas apabila:

$$P(\text{roti dan mentega}) = P(\text{roti}) \cdot P(\text{mentega})$$

- Jika sebuah *rule* memiliki nilai *lift*:
 - ≥ 1 , artinya kejadian pembelian roti dan mentega independen (tidak ber-asosiasi) antara satu sama lain.
 - > 1 , artinya kejadian pembelian roti dan mentega memiliki tingkat dependensi antara satu sama lain.
 - < 1 , artinya kejadian pembelian roti dan mentega saling berlawanan. Yaitu jika roti dibeli, maka mentega akan cenderung tidak dibeli, atau sebaliknya.

Semester Genap - TA.2022/2023 19

Association Rule
(lanjutan)

Asosiasi (part.2)

Titanic Data set

Semester Genap - TA.2022/2023 20

Titanic Data Set

Kisah Titanic

- Titanic adalah sebuah kapal pesiar Britania Raya yang menabrak gunung es dan tenggelam pada tanggal 15 April 1912 dalam pelayaran perdananya ke New York City.

(sumber: wikipedia)



Semester Genap - TA.2022/2023 21

Titanic Data Set
(lanjutan)



No	Kolom	Keterangan
1	Jenis Kelamin	pria, wanita, anak-anak
2	Kelas Tiket	kelas 1, kelas 2, kelas 3
3	Asal Keberangkatan	Southampton, Cherbourg, Queenstown
4	Jumlah Saudara	sendiri, berdua, bertiga, grup
5	Selamat	ya, tidak

- Tujuan → Mencari kondisi yang berhubungan / ber-asosiasi erat dengan selamatnya penumpang Titanic.

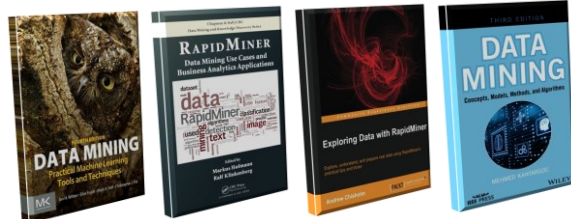
Semester Genap - TA.2022/2023 22

Titanic Data Set
(lanjutan)

- Praktikum Market Basket Analysis dengan Rapidminer Studio.
- Selamat mencoba dan berlatih.

Semester Genap - TA.2022/2023 23

Asosiasi (part.2)
Studi Kasus



Semester Genap - TA.2022/2023 24



Studi Kasus

- Penjelasan ada pada file *suplemen materia* :
 - Kasus-1 (*Market Basket Analysis*)
 - Kasus-2 (*Market Basket Analysis* dengan Rapidminer)

Semester Genap - TA.2022/2023

25



Referensi

- Ian H Witten, Eibe Frank, Mark A Hall. (2016). *Data Mining - Practical Machine Learning Tools and Techniques*, 4th Edition. Morgan-Kaufman.
- Hofmann, Markus., Klinkenberg, Ralf. (2016). *Data Mining Use Cases and Business Analytics Application*. CRC Press.
- Crisholm, Andrew. (2013). *Exploring Data With Rapidminer*. Packt Publishing.
- Mehmed Kantardzic. (2020). *Data Mining - Concepts, Models, Methods, and Algorithms*, 3rd edition. Wiley - IEEE Press.
- rapidminer.com, 2021,. Rapidminer Manual and Documentation, <https://docs.rapidminer.com/latest/studio/operators/>.

Semester Genap - TA.2022/2023

26





Data Mining



Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

● ● ●

Session 11



PROGRAM STUDI TEKNIK INFORMATIKA

Sumber: <https://ft.uhamka.ac.id/prodi/informatika/>

- VISI.
 - ✓ Menjadi program studi informatika yang menghasilkan karya unggul di bidang teknologi informasi melalui pembelajaran berlandaskan prophetic teaching untuk membentuk lulusan yang cerdas secara spiritual, intelektual, emosional, dan sosial.
- MISI.
 1. Melaksanakan pendidikan di bidang informatika yang terintegrasi dengan nilai-nilai Al Islam Kemuhammadiyah.
 2. Melaksanakan penelitian untuk menghasilkan karya teknologi informasi yang unggul untuk masyarakat berkembang.
 3. Melaksanakan pengabdian kepada masyarakat untuk kesejahteraan umat.
 4. Menghasilkan mahasiswa yang berwawasan global dalam teknologi informasi.
 5. Mendorong jiwa kewirausahaan mahasiswa untuk berkiprah di tengah masyarakat dalam meningkatkan kegiatan ekonomi.
 6. Menerapkan standar mutu pendidikan informatika melalui layanan mutu yang berkesinambungan.
 7. Membangun kerja sama terpadu dengan mitra yang berkontribusi untuk kemajuan pendidikan di bidang informatika.

Semester Genap - TA.2022/2023 2



Sistem Penilaian

- SKS : 3
 - Teori 3 SKS – (14 Kali Pertemuan)
- Keaktifan (bobot 10%)
 - Terdiri dari: Aktivitas Perkuliahan (*Kehadiran, Etika, dan Sikap*).
- Project (bobot 25%)
 - Terdiri dari: Tugas Besar / Project (*Penguasaan Keterampilan*) dengan cara observasi, mengimplementasikan dan menghasilkan Proyek Akhir, serta mempresentasikannya.
- Tes (UTS = bobot 25% , UAS = bobot 40%)
 - Terdiri dari: UTS dan UAS (*Penguasaan Pengetahuan*) dengan cara tes tertulis maupun tes praktikum.

Semester Genap - TA.2022/2023 3



Tujuan/Objectives (Session.11)

- Objectives (Tujuan Instruksional Umum).
 - RPS/SAP.
 - Klastering (part.1)
- Indikator Kompetensi.
 - Mahasiswa mampu mengetahui dan memahami tentang teknik klastering pada *data mining*.
 - Mahasiswa mampu mengetahui, memahami, serta dapat menjelaskan tentang metode dan teknik klastering serta algoritma K-means, studi kasus, dan latihan teknik klastering didalam *data mining*.

Semester Genap - TA.2022/2023 4



Muhasabah

فِيهِ آيَةٌ بَيِّنَةٌ مِّمَّا يُبْرَاهِيمَ ۚ وَمَنْ دَخَلَهُ كَانَ آمِنًا ۚ وَلِلَّهِ عَلَى النَّاسِ حِجُّ الْبَيْتِ
 مَنِ اسْتَطَاعَ إِلَيْهِ سَبِيلًا ۚ وَمَنْ كَفَرَ فَإِنَّ اللَّهَ غَفُورٌ عَلِيمٌ

- Qs. Ali Imran, ayat 97
 - Di sana terdapat tanda-tanda yang jelas, (di antaranya) maqam Ibrahim.
 - Barangsiapa memasukinya (Baitullah) amanlah dia. Dan (di antara) kewajiban manusia terhadap Allah adalah melaksanakan ibadah haji ke Baitullah, yaitu bagi orang-orang yang mampu mengadakan perjalanan ke sana.
 - Barangsiapa mengingkari (kewajiban) haji, maka ketahuilah bahwa Allah Mahakaya (tidak memerlukan sesuatu) dari seluruh alam.

Semester Genap - TA.2022/2023
5



Data Mining



Uhamka

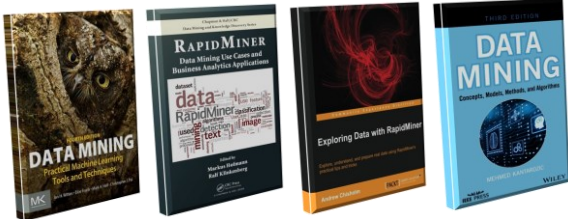
UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

Session 11
 Klastering (*part.1*)



Klastering (part.1)



Pendahuluan

Semester Genap - TA.2022/2023
7



Pendahuluan

- Teknik klustering/klasterisasi berbeda dengan teknik klasifikasi.
- Teknik klusterisasi mengelompokkan data secara otomatis, tanpa perlu diberitahu label kelasnya.
- Tetapi, teknik klusterisasi tentu saja dapat juga digunakan untuk himpunan data dengan label kelas yang diketahui.

Semester Genap - TA.2022/2023
8

Pendahuluan
(lanjutan)

- Clustering, adalah data yang tidak mempunyai label/kelas, sehingga sering disebut dengan teknik *unsupervised learning* (He & Tan, 2012).
- Tujuan dari clustering adalah pengelompokan data ke dalam beberapa cluster berdasarkan tingkat kemiripan (Xiang, Zhu, Ma, Meng, & An, 2015).
- Tingkat kemiripan dihitung berdasarkan jarak antara data (Zhang, Ouyang, & Ning, 2010).

Semester Genap - TA.2022/2023 9

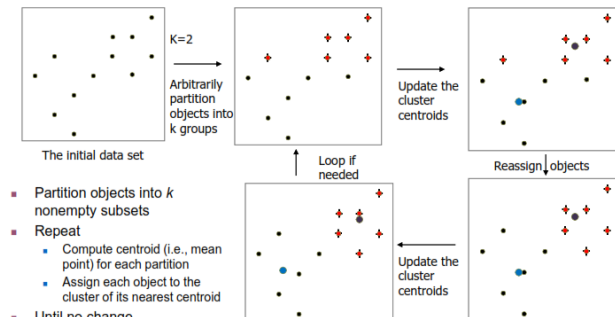
Pendahuluan
(lanjutan)

Penklusteran (clustering) digunakan untuk melakukan pengelompokan data-data ke dalam sejumlah kelompok (cluster) berdasarkan karakteristik masing-masing data pada kelompok-kelompok yang ada.



Semester Genap - TA.2022/2023 10

Pendahuluan
(lanjutan)

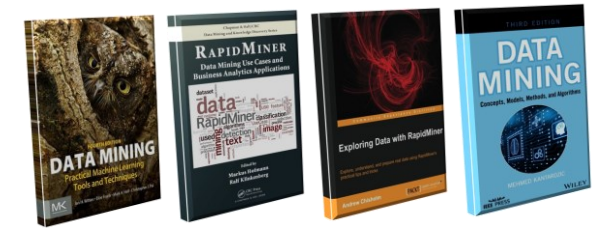


- Partition objects into k nonempty subsets
- Repeat
 - Compute centroid (i.e., mean point) for each partition
 - Assign each object to the cluster of its nearest centroid
- Until no change

Semester Genap - TA.2022/2023 11

Klastering (part.1)

Algoritma K-Means



Semester Genap - TA.2022/2023 12



Algoritma K-MEANS

- Algoritma K-MEANS masuk ke dalam penerapan *data mining clustering*.
- Clustering, adalah data yang tidak mempunyai label/kelas, sehingga sering disebut dengan teknik *unsupervised learning* (He & Tan, 2012).
- Tujuan dari clustering adalah pengelompokan data ke dalam beberapa cluster berdasarkan tingkat kemiripan (Xiang, Zhu, Ma, Meng, & An, 2015).
- Tingkat kemiripan dihitung berdasarkan jarak antara data (Zhang, Ouyang, & Ning, 2010).

Semester Genap - TA.2022/2023

13



Algoritma K-MEANS

(lanjutan)

- Algoritma K-MEANS merupakan algoritma klusterisasi yang paling tua dan paling banyak digunakan dalam berbagai aplikasi kecil hingga menengah, dikarenakan kemudahan didalam implementasinya.
- Ide dasar algoritma K-MEANS sangatlah sederhana, yaitu meminimalkan *Sum of Squared Error* (SSE) antara objek-objek data dengan sejumlah *k* centroid.
- Metode pengelompokan data *partitioning* (non hierarki).
- Data berkarakteristik sama dimasukkan ke dalam satu kelompok.
- Meminimalkan variasi dalam satu kelompok dan memaksimalkan variasi antar kelompok.

Semester Genap - TA.2022/2023

14



Algoritma K-MEANS

(lanjutan)

Tahapan pengelompokan didalam algoritma K-MEANS, yaitu:

1. Tentukan jumlah kelompok (K) dan nilai ambang batas atau *threshold* (T).
2. Alokasikan data ke dalam kelompok secara acak.
3. Hitung pusat kelompok atau *centroid* (C) untuk setiap kelompok.
4. Alokasikan semua data ke centroid terdekat
5. Kembali ke langkah 3 (iterasi), apabila masih ada:
 - data yang berpindah kelompok,
 - atau ada perubahan nilai centroid di atas nilai ambang yang ditentukan,
 - atau ada perubahan nilai pada fungsi objektif yang digunakan (di atas nilai ambang yang ditentukan).

Semester Genap - TA.2022/2023

15



Algoritma K-MEANS

(lanjutan)

- Formula yang digunakan didalam algoritma K-MEANS.

Formula untuk menghitung centroid:

$$C_i = \frac{1}{M} \sum_{j=1}^M x_j$$

Formula untuk menghitung jarak data dengan centroid:

$$D(x,y) = \sqrt{\sum_{j=1}^n |x - y|^2} \quad \dots \text{Jarak Euclidean}$$

$$D(x,y) = \sum_{j=1}^n |x - y| \quad \dots \text{Jarak Manhattan/City Block}$$

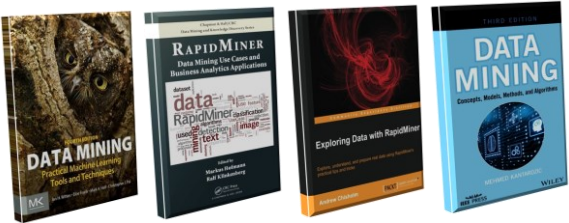
$$D(x,y) = \max(|x - y|) \quad \dots \text{Jarak Chebyshev}$$

Formula untuk menghitung fungsi objektif:

$$J = \sum_{i=1}^N \sum_{l=1}^K a_{il} D(x_i, C_l)^2$$

Semester Genap - TA.2022/2023

16



Klastering (part.1)

Studi Kasus

Semester Genap - TA.2022/2023

17

Studi Kasus

Data training (data latih): Pengelompokan data dua dimensi

Data ke-i	Fitur x	Fitur y
1	1	1
2	4	1
3	6	1
4	1	2
5	2	3
6	5	3
7	2	5
8	3	5
9	2	6
10	3	8

Terapkan teknik clustering untuk data latih seperti pada tabel diatas, dengan menggunakan algoritma K-MEANS untuk ketentuan sbb:

- Jumlah kelompok (K) = 3
- Threshold (T) = 0,1
- Fungsi Objektif (J) = 0

Semester Genap - TA.2022/2023

18

Studi Kasus (lanjutan)

Tahapan pengelompokan didalam algoritma K-MEANS, yaitu:

- Tentukan jumlah kelompok (K) dan nilai ambang batas atau *threshold* (T).
- Alokasikan data ke dalam kelompok secara acak.
- Hitung pusat kelompok atau *centroid* (C) untuk setiap kelompok.
- Alokasikan semua data ke centroid terdekat
- Kembali ke langkah 3 (iterasi), apabila masih ada:
 - data yang berpindah kelompok,
 - atau ada perubahan nilai centroid di atas nilai ambang yang ditentukan,
 - atau ada perubahan nilai pada fungsi objektif yang digunakan (di atas nilai ambang yang ditentukan).

Semester Genap - TA.2022/2023

19

Studi Kasus (lanjutan)

- Formula yang digunakan didalam algoritma K-MEANS.

Formula untuk menghitung centroid:

$$C_i = \frac{1}{M} \sum_{j=1}^M x_j$$

Formula untuk menghitung jarak data dengan centroid:

$D(x,y) = \sqrt{\sum_{j=1}^n |x - y|^2}$... Jarak Euclidean
 $D(x,y) = \sum_{j=1}^n |x - y|$... Jarak Manhattan/City Block
 $D(x,y) = \max(|x - y|)$... Jarak Chebyshev

Formula untuk menghitung fungsi objektif:

$$J = \sum_{i=1}^N \sum_{l=1}^K a_{il} D(x_i, C_l)^2$$

Semester Genap - TA.2022/2023

20



Studi Kasus (lanjutan)

Data training (data latih): Pengelompokan data dua dimensi

Data ke-i	Fitur x	Fitur y
1	1	1
2	4	1
3	6	1
4	1	2
5	2	3
6	5	3
7	2	5
8	3	5
9	2	6
10	3	8

Langkah 1: menentukan jumlah kelompok (K), $K = 3$

Ambang batas atau *threshold* (T) yang akan digunakan untuk perubahan fungsi objektif adalah 0.1

Fungsi Objektif (J) awal = 0

Semester Genap - TA.2022/2023

21



Studi Kasus (lanjutan)

Langkah 2: Alokasikan data ke dalam kelompok secara acak

Data ke-i	Fitur x	Fitur y	Kelompok 1	Kelompok 2	Kelompok 3
1	1	1	*		
2	4	1			*
3	6	1		*	
4	1	2		*	
5	2	3			*
6	5	3		*	
7	2	5		*	
8	3	5			*
9	2	6			*
10	3	8		*	

Semester Genap - TA.2022/2023

22



Studi Kasus (lanjutan)

Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 1:

Data ke-i	Fitur x	Fitur y
1	1	1
M (Jumlah Data)		
1	Jumlah x	Jumlah y
Rata-rata		
	1	1

Centroid untuk kelompok 2:

Data ke-i	Fitur x	Fitur y
3	6	1
4	1	2
6	5	3
7	2	5
10	3	8
M (Jumlah Data)		
5	Jumlah x	Jumlah y
Rata-rata		
	3.400	3.800

Semester Genap - TA.2022/2023

23



Studi Kasus (lanjutan)

Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 3:

Data ke-i	Fitur x	Fitur y
2	4	1
5	2	3
8	3	5
9	2	6
M (Jumlah Data)		
4	Jumlah x	Jumlah y
Rata-rata		
	2.750	3.750

Centroid pertama untuk setiap kelompok:

Kelompok	Fitur x	Fitur y
1	1	1
2	3.400	3.800
3	2.750	3.750

Semester Genap - TA.2022/2023

24

Studi Kasus
(lanjutan)

Menghitung fungsi objektif baru

Jarak data ke centroid pertama (menggunakan jarak euclidean):

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	K
1	1	1	0			1
2	4	1			3.021	3
3	6	1		3.821		2
4	1	2		3.000		2
5	2	3			1.061	3
6	5	3		1.789		2
7	2	5		1.844		2
8	3	5			1.275	3
9	2	6			2.372	3
10	3	8		4.219		2
			0	14.673	7.728	22.401
			Perubahan Fungsi Objektif			22.401

Semester Genap - TA.2022/2023 25

Studi Kasus
(lanjutan)

Menghitung perubahan fungsi objektif

Fungsi Objektif (J) lama/sebelumnya = 0

Fungsi Objektif (J) baru/sekarang = $0 + 14.673 + 7.728 = 22.401$

Perubahan Fungsi Objektif = $|J \text{ baru} - J \text{ lama}|$
 $= |22.401 - 0| = 22.401$

Perubahan masih di atas ambang batas atau threshold (T); >0.1 , artinya pencarian centroid masih terus dilakukan.

Semester Genap - TA.2022/2023 26

Studi Kasus
(lanjutan)

Langkah 4: Alokasikan semua data ke centroid terdekat (menggunakan jarak euclidean)

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	Terdekat	K Baru	K Lama
1	1	1	0	3.688	3.260	0	1	1
2	4	1	3	2.864	3.021	2.864	2	3
3	6	1	5	3.821	4.257	3.821	2	2
4	1	2	1	3	2.475	1	1	2
5	2	3	2.236	1.612	1.061	1.061	3	3
6	5	3	4.472	1.789	2.372	1.789	2	2
7	2	5	4.123	1.844	1.458	1.458	3	2
8	3	5	4.472	1.265	1.275	1.265	2	3
9	2	6	5.099	2.608	2.372	2.372	3	3
10	3	8	7.280	4.219	4.257	4.219	2	2

Masih terdapat data yang berpindah kelompok, artinya masih perlu dilakukan pencarian centroid baru (kembali ke langkah 3).

Semester Genap - TA.2022/2023 27

Studi Kasus
(lanjutan)

Langkah 4: Alokasikan semua data ke centroid terdekat

Data ke-i	Fitur x	Fitur y	Kelompok 1	Kelompok 2	Kelompok 3
1	1	1	*		
2	4	1		*	
3	6	1		*	
4	1	2	*		
5	2	3			*
6	5	3		*	
7	2	5			*
8	3	5		*	
9	2	6			*
10	3	8		*	

Masih terdapat data yang berpindah kelompok, artinya masih perlu dilakukan pencarian centroid baru (kembali ke langkah 3).

Semester Genap - TA.2022/2023 28



Studi Kasus (lanjutan)

Iterasi - 1

Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 1:

Data ke-i	Fitur x	Fitur y
1	1	1
4	1	2
M (Jumlah Data)	Jumlah x	Jumlah y
2	2	3
Rata-rata	1	1.500

Centroid untuk kelompok 2:

Data ke-i	Fitur x	Fitur y
2	4	1
3	6	1
6	5	3
8	3	5
10	3	8
M (Jumlah Data)	Jumlah x	Jumlah y
5	21	18
Rata-rata	4.200	3.600

Semester Genap - TA.2022/2023

29



Studi Kasus (lanjutan)

Iterasi - 1

Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 3:

Data ke-i	Fitur x	Fitur y
5	2	3
7	2	5
9	2	6
M (Jumlah Data)	Jumlah x	Jumlah y
3	6	14
Rata-rata	2	4.667

Centroid kedua untuk setiap kelompok (hasil iterasi - 1):

Kelompok	Fitur x	Fitur y
1	1	1.500
2	4.200	3.600
3	2	4.667

Semester Genap - TA.2022/2023

30



Studi Kasus (lanjutan)

Iterasi - 1

Menghitung fungsi objektif baru

Jarak data ke centroid kedua (menggunakan jarak euclidean):

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	K
1	1	1	0.500			1
2	4	1		2.608		2
3	6	1		3.162		2
4	1	2	0.500			1
5	2	3			1.667	3
6	5	3		1.000		2
7	2	5			0.333	3
8	3	5		1.844		2
9	2	6			1.333	3
10	3	8		4.561		2
			1.000	13.175	3.333	17.508
			Perubahan Fungsi Objektif			4.893

Semester Genap - TA.2022/2023

31



Studi Kasus (lanjutan)

Iterasi - 1

Menghitung perubahan fungsi objektif

Fungsi Objektif (J) lama/sebelumnya = 22.401

Fungsi Objektif (J) baru/sekarang = $1 + 13.175 + 3.333 = 17.508$

Perubahan Fungsi Objektif = $|J \text{ baru} - J \text{ lama}|$
 $= |17.508 - 22.401| = 4.893$

Perubahan masih di atas ambang batas atau threshold (T); >0.1 , artinya pencarian centroid masih terus dilakukan.

Semester Genap - TA.2022/2023

32



Studi Kasus (lanjutan)

Iterasi - 1

Langkah 4: Alokasikan semua data ke centroid terdekat
(menggunakan jarak euclidean)

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	Terdekat	K Baru	K Lama
1	1	1	0.500	4.123	3.801	0.500	1	1
2	4	1	3.041	2.608	4.177	2.608	2	2
3	6	1	5.025	3.162	5.426	3.162	2	2
4	1	2	0.500	3.578	2.848	0.500	1	1
5	2	3	1.803	2.280	1.667	1.667	3	3
6	5	3	4.272	1	3.432	1	2	2
7	2	5	3.640	2.608	0.333	0.333	3	3
8	3	5	4.031	1.844	1.054	1.054	3	2
9	2	6	4.610	3.256	1.333	1.333	3	3
10	3	8	6.801	4.561	3.480	3.480	3	2

Masih terdapat data yang berpindah kelompok, artinya masih perlu dilakukan pencarian centroid baru (kembali ke langkah 3).

Semester Genap - TA.2022/2023

33



Studi Kasus (lanjutan)

Iterasi - 1

Langkah 4: Alokasikan semua data ke centroid terdekat

Data ke-i	Fitur x	Fitur y	Kelompok 1	Kelompok 2	Kelompok 3
1	1	1	*		
2	4	1		*	
3	6	1		*	
4	1	2	*		
5	2	3			*
6	5	3		*	
7	2	5			*
8	3	5			*
9	2	6			*
10	3	8			*

Masih terdapat data yang berpindah kelompok, artinya masih perlu dilakukan pencarian centroid baru (kembali ke langkah 3).

Semester Genap - TA.2022/2023

34



Studi Kasus (lanjutan)

Iterasi - 2

Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 1:

Data ke-i	Fitur x	Fitur y
1	1	1
4	1	2
M (Jumlah Data)		Jumlah y
2	2	3
Rata-rata		1.500

Centroid untuk kelompok 2:

Data ke-i	Fitur x	Fitur y
2	4	1
3	6	1
6	5	3
M (Jumlah Data)		Jumlah y
3	15	5
Rata-rata		1.667

Semester Genap - TA.2022/2023

35



Studi Kasus (lanjutan)

Iterasi - 2

Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 3:

Data ke-i	Fitur x	Fitur y
5	2	3
7	2	5
8	3	5
9	2	6
10	3	8
M (Jumlah Data)		Jumlah y
5	12	27
Rata-rata		5.400

Centroid ketiga untuk setiap kelompok (hasil iterasi - 2):

Kelompok	Fitur x	Fitur y
1	1	1.500
2	5	1.667
3	2.400	5.400

Semester Genap - TA.2022/2023

36

Studi Kasus
(lanjutan)

Iterasi - 2
Menghitung fungsi objektif baru

Jarak data ke centroid ketiga (menggunakan jarak *euclidean*):

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	K
1	1	1	0.500			1
2	4	1		1.202		2
3	6	1		1.202		2
4	1	2	0.500			1
5	2	3			2.433	3
6	5	3		1.333		2
7	2	5			0.566	3
8	3	5			0.721	3
9	2	6			0.721	3
10	3	8			2.668	3
			1.000	3.737	7.109	11.846
			Perubahan Fungsi Objektif			5.662

Semester Genap - TA.2022/2023 37

Studi Kasus
(lanjutan)

Iterasi - 2
Menghitung perubahan fungsi objektif

Fungsi Objektif (J) lama/sebelumnya = 17.508

Fungsi Objektif (J) baru/sekarang = $1 + 3.737 + 7.109 = 11.846$

Perubahan Fungsi Objektif = $|J \text{ baru} - J \text{ lama}|$
 $= |11.846 - 17.508| = 5.662$

Perubahan masih di atas ambang batas atau threshold (T); >0.1 , artinya pencarian centroid masih terus dilakukan.

Semester Genap - TA.2022/2023 38

Studi Kasus
(lanjutan)

Iterasi - 2
Langkah 4: Alokasikan semua data ke centroid terdekat
(menggunakan jarak *euclidean*)

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	Terdekat	K Baru	K Lama
1	1	1	0.500	4.055	4.617	0.500	1	1
2	4	1	3.041	1.202	4.682	1.202	2	2
3	6	1	5.025	1.202	5.685	1.202	2	2
4	1	2	0.500	4.014	3.677	0.500	1	1
5	2	3	1.803	3.283	2.433	1.803	1	3
6	5	3	4.272	1.333	3.538	1.333	2	2
7	2	5	3.640	4.484	0.566	0.566	3	3
8	3	5	4.031	3.887	0.721	0.721	3	3
9	2	6	4.610	5.270	0.721	0.721	3	3
10	3	8	6.801	6.641	2.668	2.668	3	3

Masih terdapat data yang berpindah kelompok, artinya masih perlu dilakukan pencarian centroid baru (kembali ke langkah 3).

Semester Genap - TA.2022/2023 39

Studi Kasus
(lanjutan)

Iterasi - 2
Langkah 4: Alokasikan semua data ke centroid terdekat

Data ke-i	Fitur x	Fitur y	Kelompok 1	Kelompok 2	Kelompok 3
1	1	1	*		
2	4	1		*	
3	6	1		*	
4	1	2	*		
5	2	3	*		
6	5	3		*	
7	2	5			*
8	3	5			*
9	2	6			*
10	3	8			*

Masih terdapat data yang berpindah kelompok, artinya masih perlu dilakukan pencarian centroid baru (kembali ke langkah 3).

Semester Genap - TA.2022/2023 40

Studi Kasus
(lanjutan)

Iterasi - 3
Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 1:

Data ke-i	Fitur x	Fitur y
1	1	1
4	1	2
5	2	3
M (Jumlah Data)	Jumlah x	Jumlah y
3	4	6
Rata-rata	1.333	2

Centroid untuk kelompok 2:

Data ke-i	Fitur x	Fitur y
2	4	1
3	6	1
6	5	3
M (Jumlah Data)	Jumlah x	Jumlah y
3	15	5
Rata-rata	5	1.667

Semester Genap - TA.2022/2023 41

Studi Kasus
(lanjutan)

Iterasi - 3
Langkah 3: Hitung pusat kelompok atau centroid (C) untuk setiap kelompok

Centroid untuk kelompok 3:

Data ke-i	Fitur x	Fitur y
7	2	5
8	3	5
9	2	6
10	3	8
M (Jumlah Data)	Jumlah x	Jumlah y
4	10	24
Rata-rata	2.500	6

Centroid keempat untuk setiap kelompok (hasil iterasi - 3):

Kelompok	Fitur x	Fitur y
1	1.333	2
2	5	1.667
3	2.500	6

Semester Genap - TA.2022/2023 42

Studi Kasus
(lanjutan)

Iterasi - 3
Menghitung fungsi objektif baru

Jarak data ke centroid ketiga (menggunakan jarak *euclidean*):

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	K
1	1	1	1.054			1
2	4	1		1.202		2
3	6	1		1.202		2
4	1	2	0.333			1
5	2	3	1.202			1
6	5	3		1.333		2
7	2	5			1.118	3
8	3	5			1.118	3
9	2	6			0.500	3
10	3	8			2.062	3
			2.589	3.737	4.798	11.124
			Perubahan Fungsi Objektif			0.732

Semester Genap - TA.2022/2023 43

Studi Kasus
(lanjutan)

Iterasi - 3
Menghitung perubahan fungsi objektif

Fungsi Objektif (J) lama/sebelumnya = 11.846

Fungsi Objektif (J) baru/sekarang = $2.589 + 3.737 + 4.798 = 11.124$

Perubahan Fungsi Objektif = $|J \text{ baru} - J \text{ lama}|$
 $= |11.124 - 11.846| = 0.732$

Perubahan masih di atas ambang batas atau threshold (T); >0.1 , artinya pencarian centroid masih terus dilakukan.

Semester Genap - TA.2022/2023 44



Studi Kasus (lanjutan)

Iterasi - 3

Langkah 4: Alokasikan semua data ke centroid terdekat
(menggunakan jarak euclidean)

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	Terdekat	K Baru	K Lama
1	1	1	1.054	4.055	5.220	1.054	1	1
2	4	1	2.848	1.202	5.220	1.202	2	2
3	6	1	4.773	1.202	6.103	1.202	2	2
4	1	2	0.333	4.014	4.272	0.333	1	1
5	2	3	1.202	3.283	3.041	1.202	1	1
6	5	3	3.801	1.333	3.905	1.333	2	2
7	2	5	3.073	4.484	1.118	1.118	3	3
8	3	5	3.432	3.887	1.118	1.118	3	3
9	2	6	4.055	5.270	0.500	0.500	3	3
10	3	8	6.227	6.641	2.062	2.062	3	3

Sudah tidak ada data yang berpindah kelompok, artinya pencarian centroid sudah bisa dihentikan.

Tetapi iterasi juga masih bisa dilakukan (kembali ke langkah 3), karena perubahan fungsi objektif masih di atas ambang batas atau threshold (T); $0.732 > 0.1$

Semester Genap - TA.2022/2023

45



Studi Kasus (lanjutan)

Iterasi - 4

Menghitung fungsi objektif baru

Data ke-i	Fitur x	Fitur y	K 1	K 2	K 3	K
1	1	1	1.054			1
2	4	1		1.202		2
3	6	1		1.202		2
4	1	2	0.333			1
5	2	3	1.202			1
6	5	3		1.333		2
7	2	5			1.118	3
8	3	5			1.118	3
9	2	6			0.500	3
10	3	8			2.062	3
			2.589	3.737	4.798	11.124
			Perubahan Fungsi Objektif			0.000

Semester Genap - TA.2022/2023

46



Studi Kasus (lanjutan)

Iterasi - 4

Menghitung perubahan fungsi objektif

Fungsi Objektif (J) lama/sebelumnya = 11.124

Fungsi Objektif (J) baru/sekarang = $2.589 + 3.737 + 4.798 = 11.124$

Perubahan Fungsi Objektif = $|J \text{ baru} - J \text{ lama}|$
 $= |11.846 - 11.124| = 0$

Perubahan sudah di bawah ambang batas atau threshold (T); $0 < 0.1$, artinya centroid sudah bisa digunakan.

Semester Genap - TA.2022/2023

47



Studi Kasus (lanjutan)

Nilai akhir centroid untuk setiap kelompok:

Kelompok	Fitur x	Fitur y
1	1.333	2
2	5	1.667
3	2.500	6

Hasil pengalokasian semua data latih ke centroid terdekat

Data ke-i	Fitur x	Fitur y	Kelompok 1	Kelompok 2	Kelompok 3
1	1	1	*		
2	4	1		*	
3	6	1		*	
4	1	2	*		
5	2	3	*		
6	5	3		*	
7	2	5			*
8	3	5			*
9	2	6			*
10	3	8			*

Semester Genap - TA.2022/2023

48



Referensi

- Ian H Witten, Eibe Frank, Mark A Hall. (2016). *Data Mining - Practical Machine Learning Tools and Techniques*, 4th Edition. Morgan-Kaufman.
- Hofmann, Markus., Klinkenberg, Ralf. (2016). *Data Mining Use Cases and Business Analytics Application*. CRC Press.
- Crisholm, Andrew. (2013). *Exploring Data With Rapidminer*. Packt Publishing.
- Mehmed Kantardzic. (2020). *Data Mining - Concepts, Models, Methods, and Algorithms*, 3rd edition. Wiley - IEEE Press.
- rapidminer.com, 2021., Rapidminer Manual and Documentation, <https://docs.rapidminer.com/latest/studio/operators/>.

Semester Genap - TA.2022/2023

49

hati merupakan sumber dari kebaikan atau keburukan seseorang, sebagaimana sabda Rasulullah Shallallahu 'alaihi wa sallam.

أَلَا وَإِنَّ فِي الْجَسَدِ مُضْغَةً، إِذَا صَلَحَتْ صَلَحَ الْجَسَدُ كُلُّهُ، وَإِذَا فَسَدَتْ فَسَدَ الْجَسَدُ كُلُّهُ، أَلَا وَهِيَ الْقَلْبُ .

... “Ketahuilah, sesungguhnya di dalam jasad ini ada segumpal daging, apabila ia (segumpal daging) tersebut baik, baiklah seluruh jasadnya, dan apabila ia (segumpal daging) tersebut rusak (buruk), maka rusaklah (buruklah) seluruh jasadnya. Ketahuilah, segumpal daging tersebut adalah hati”.

HR al Bukhari (1/28 no. 52), Muslim (3/1219 no. 1599), dan lain-lain, dari hadits an Nu'man bin Basyir

Terima Kasih



Data Mining



Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion



Session 12

PROGRAM STUDI TEKNIK INFORMATIKA

Sumber: <https://ft.uhamka.ac.id/prodi/informatika/>

- **VISI.**

- √ Menjadi program studi informatika yang menghasilkan karya unggul di bidang teknologi informasi melalui pembelajaran berlandaskan prophetic teaching untuk membentuk lulusan yang cerdas secara spiritual, intelektual, emosional, dan sosial.

- **MISI.**

1. Melaksanakan pendidikan di bidang informatika yang terintegrasi dengan nilai-nilai Al Islam Kemuhammadiyah.
2. Melaksanakan penelitian untuk menghasilkan karya teknologi informasi yang unggul untuk masyarakat berkembang
3. Melaksanakan pengabdian kepada masyarakat untuk kesejahteraan umat.
4. Menghasilkan mahasiswa yang berwawasan global dalam teknologi informasi.
5. Mendorong jiwa kewirausahaan mahasiswa untuk berkiprah di tengah masyarakat dalam meningkatkan kegiatan ekonomi.
6. Menerapkan standar mutu pendidikan informatika melalui layanan mutu yang berkesinambungan.
7. Membangun kerja sama terpadu dengan mitra yang berkontribusi untuk kemajuan pendidikan di bidang informatika.

- SKS : 3
 - Teori 3 SKS – (14 Kali Pertemuan)
- Keaktifan (bobot **10%**)
 - Terdiri dari: Aktivitas Perkuliahan (*Kehadiran, Etika, dan Sikap*).
- Project (bobot **25%**)
 - Terdiri dari: Tugas Besar / Project (*Penguasaan Keterampilan*) dengan cara observasi, mengimplementasikan dan menghasilkan Proyek Akhir, serta mempresentasikannya.
- Tes (UTS = bobot **25%** , UAS = bobot **40%**)
 - Terdiri dari: UTS dan UAS (*Penguasaan Pengetahuan*) dengan cara tes tertulis maupun tes praktikum.

Tujuan/Objectives (Session.12)

- Objectives (Tujuan Instruksional Umum).
 - RPS/SAP.
 - Klastering (part.2)
- Indikator Kompetensi.
 - Mahasiswa mampu mengetahui dan memahami tentang teknik klustering pada *data mining*.
 - Mahasiswa mampu mengetahui, memahami, serta dapat menjelaskan tentang metode dan teknik klustering serta algoritma K-means, studi kasus, serta studi kasus & latihan teknik klustering didalam *data mining*.

إِنَّ هَذَا الْقُرْآنَ يَهْدِي لِلَّتِي هِيَ أَقْوَمُ وَيُبَشِّرُ الْمُؤْمِنِينَ الَّذِينَ يَعْمَلُونَ الصَّالِحَاتِ
أَنَّ لَهُمْ أَجْرًا كَبِيرًا

- Qs. Al Israa, ayat 9
 - Sungguh, Al-Qur'an ini memberi petunjuk ke (jalan) yang paling lurus dan memberi kabar gembira kepada orang mukmin yang mengerjakan kebajikan, bahwa mereka akan mendapat pahala yang besar

Data Mining



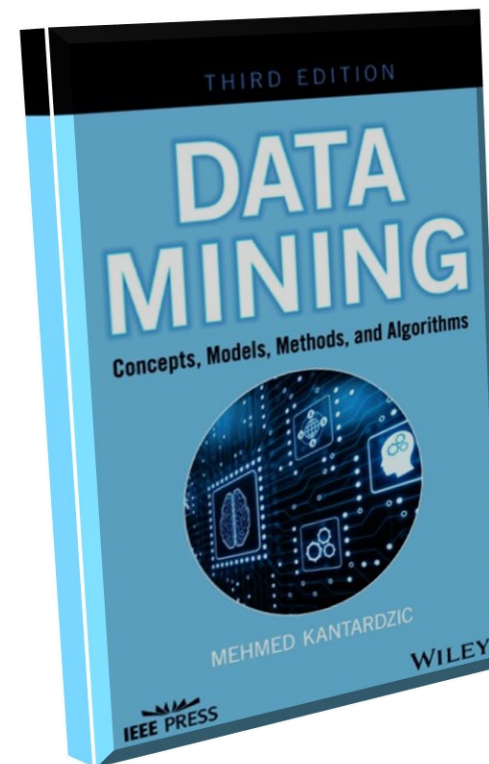
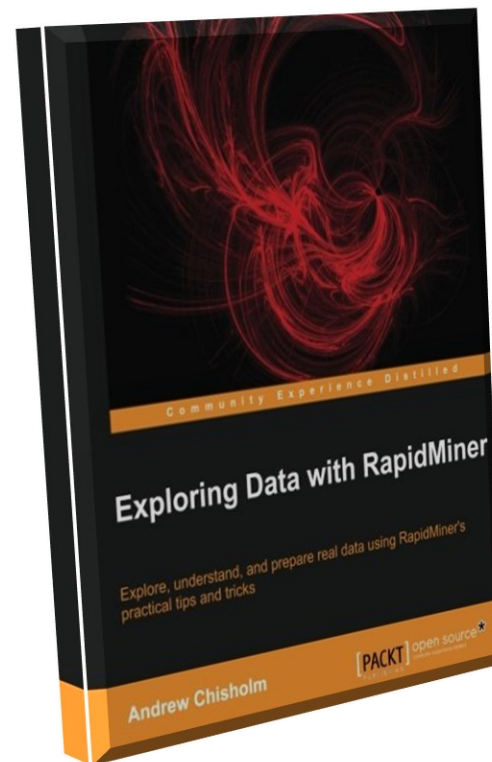
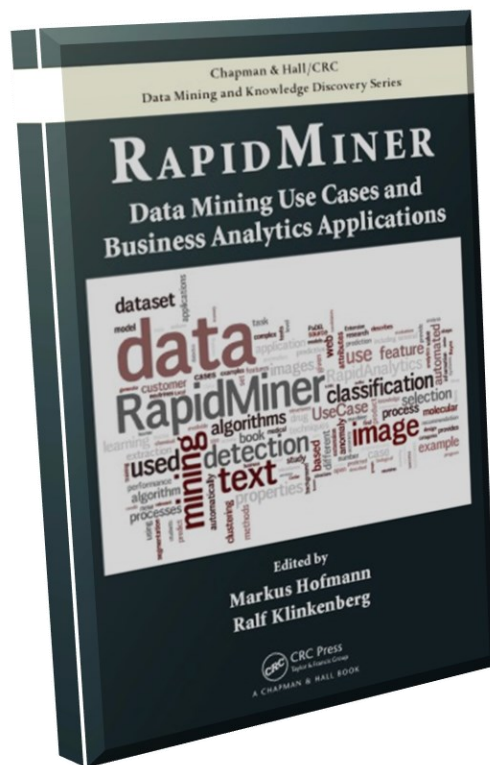
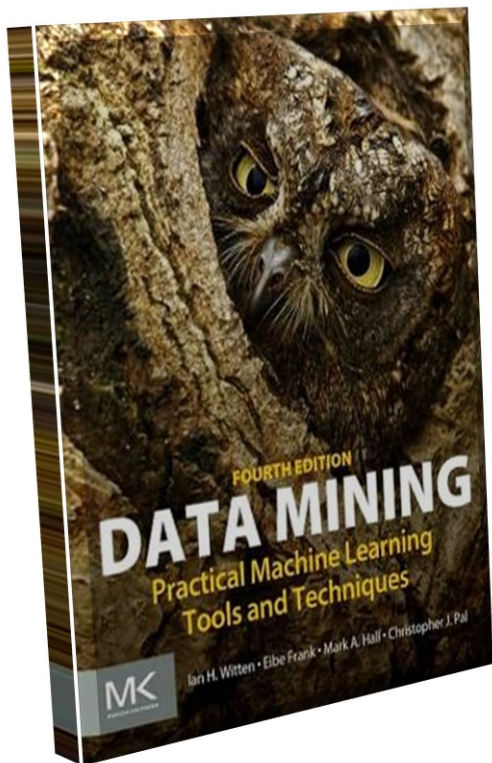
Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion

● ● ●
Session 12

Klustering (*part.2*)



Klustering (part.2)

Studi Kasus

- Studi kasus yang digunakan, adalah berdasarkan pada penelitian.
- Paper dapat di-*download* pada laman berikut.

<https://jurnal.stmikroyal.ac.id/index.php/jurteksi/article/view/352>

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. V No. 2, Juni 2019, hlm. 97 – 104 ISSN 2550-0201 (Online)
DOI: <https://doi.org/10.33330/jurteksi.v5i2.352>
Available online at <http://jurnal.stmikroyal.ac.id/index.php/jurteksi>

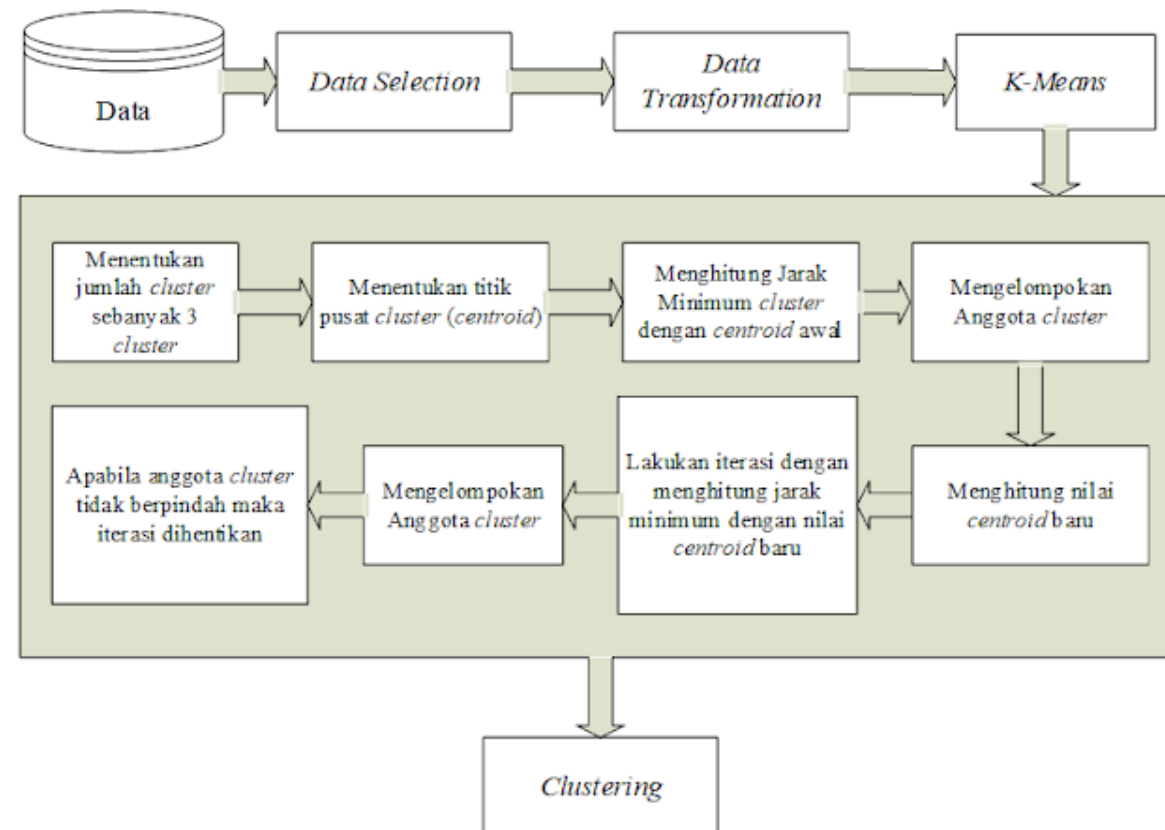
KLASTERISASI DATA JAMAAH UMROH PADA AULIYA TOUR & TRAVEL MENGUNAKAN METODE *K-MEANS CLUSTERING*

Muhammad Iqbal
Sistem Informasi, STMIK Royal Kisaran
email: muhammad.iqbal@gmail.com

Abstract: Religious tourism especially for Hajj and Umroh is increasingly in demand by the public. Auliya Tour & Travel, which is engaged in the travel agency, has diverse pilgrimage data so that the data collection is used to find new knowledge as a marketing strategy using Data Mining techniques. Data Mining is one of the KDD processes that has the function for grouping. *K-Means Clustering* is a data mining technique that aims to group data into a data subset. The grouping of data on Umroh pilgrims is conducted to find out the interest groups of pilgrims based on age. This study categorizes pilgrim data into three *clusters*, which are very popular, in high demand and less desirable. Attributes used in processing data include gender, age, and congregation package. Before the data calculation process is carried out, the data transformation process is carried out before data calculation process. Based on data calculations that have been done through the RapidMiner software, the members of the group were very interested, ranging in age from 56 to 83 years, interested groups ranging in age from 29 to 55 years and the group was less interested with an age range 2 to 22 years from 170 records.

Keywords: *Clustering*, Data Mining, K-Means, RapidMiner, Umroh.

- Sebelum mendapatkan hasil clustering, ada beberapa tahapan yang harus dipenuhi seperti pengumpulan data, seleksi data, transformasi data dan melakukan proses Data Mining.
- Tahapan-tahapan tersebut dapat dilihat pada gambar berikut:



- Dimulai dengan mengumpulkan data di mana data tersebut diperoleh, Misalnya, pada Layanan Jasa Tour & Travel.
- Kemudian data tersebut diseleksi berdasarkan keperluan.
- Data hasil seleksi dilakukan proses transformasi data agar dapat diolah pada metode K-Means.
- Data yang telah ditransformasi tersebut siap untuk diolah dengan metode K-Means yang dimulai dari menentukan jumlah cluster.
- Pada kasus ini menggunakan tiga cluster, yaitu:
 - Sangat Diminati,
 - Diminati, dan
 - Kurang Diminati.

- Kemudian menentukan nilai titik pusat cluster atau disebut dengan centroid yang diambil secara acak dari dataset.
- Setelah titik cluster ditentukan hitung jarak terdekat cluster lalu dikelompokkan berdasarkan jarak terdekat dengan cluster.
- Hitung kembali nilai centroid baru dengan cara menjumlahkan anggota kelompok dibagi total anggota.
- Lakukan iterasi untuk menghitung jarak minimum dengan nilai centroid baru dan mengelompokkan anggota cluster, apabila anggota cluster sudah tidak mengalami perubahan tempat maka iterasi dihentikan dan anggota kelompok telah ditemukan.
- Data yang dihasilkan dalam bentuk clustering atau kelompok.

- Pada tahap ini data yang dikumpulkan adalah data jamaah umroh pada tahun 2017 dari kantor Layanan Jasa Tour & Travel.
- Data yang diberikan masih dalam bentuk spreadsheet.
- Data yang telah dikumpulkan akan diseleksi terlebih dahulu guna mengetahui variabel apa saja yang akan mempengaruhi proses pengelompokan.
- Adapun hasil pengumpulan data pada penelitian ini lebih dari 200 record dan terdiri dari 11 kolom, yaitu:
 - Nama,
 - Jenis Kelamin,
 - Tanggal Lahir,
 - Tempat Lahir,
 - Nomor Pasport,
 - Tanggal Berlaku,
 - Tanggal Kadaluarsa,
 - Alamat,
 - Nama Mahram,
 - Manifest berdasarkan Paket

Penyeleksian Data

- *Data Selection* atau Penyeleksian data bertujuan untuk memilih data mana saja yang akan digunakan pada penelitian.
- Penyeleksian data atau pemilihan data yang akan dipakai pada penelitian ini terdiri dari empat kolom, yaitu:
 - Nama,
 - Jenis Kelamin,
 - Tanggal Lahir, dan
 - Paket. (yang dipilih oleh jamaah)
- Pada kasus kali ini hanya menampilkan data sampel sebanyak 26 *record*, di mana data tersebut belum dilakukan transformasi data dan sampel data tersebut dapat dilihat pada Tabel berikut:



Hasil Penyeleksian Data

NO.	NAMA	JENIS KELAMIN	TANGGAL LAHIR	PAKET
1	ABDUL HALIM MUHAMMAD NUH	M	10/12/1956	10 HARI
2	ARMIYANTI ANSARI ITAM	F	7/17/1979	10 HARI
3	ENNY HALAWATI	F	11/13/1974	10 HARI
4	ERFINA LEGIMIN YOSO	F	6/4/1974	10 HARI
5	HENNY SYARIF MUHAMMAD	F	5/24/1968	12 HARI
6	LOKOT UDIN PANJAITAN	M	1/5/1952	10 HARI
7	MAIMUNAH ABDUL MUNIR SAINAL	F	12/11/1959	13 HARI
8	MARTINI DARMAJI SENGUT	F	6/4/1958	10 HARI
9	MASNUN MUSA HASIBUAN	F	1/6/1967	10 HARI
10	MISRAN RASUL HASAN	M	2/25/1979	12 HARI
11	MUHAMMAD JUNAIDI SADINO	M	10/21/1960	10 HARI
12	NANING WINARSIH SUTARNO	F	11/9/1983	13 HARI
13	NEZA APRIDILLA	F	4/2/2006	12 HARI
14	PONI WIRYO SUMARTO	F	12/31/1953	13 HARI
15	RITAWATI NASIB GURUSINGA	F	7/30/1966	12 HARI
16	ROSDIANA MUCHTARAM NASUTION	F	8/15/1957	10 HARI
17	ROSLAINI ISMAIL FIRMAN	F	6/9/1953	10 HARI
18	SUHARTINI LASIMIN KROMO	F	3/20/1957	10 HARI
19	SUMARNI SUMARTO ABDULLAH	F	12/31/1959	12 HARI
20	SUNARMI OSERI ABDULLAH	F	4/3/1967	13 HARI
21	SUSILAWATI HUSIN TUMAN	F	7/1/1981	12 HARI
22	TAING MUHAMMAD SARI SITOMPUL	F	12/28/1968	10 HARI
23	TODUH KETEK ABDULLAH	F	12/19/1966	12 HARI
24	ZAHIR NAUFAL AR RAZAQ	M	3/17/2009	10 HARI
25	ZULJAWATI NGATIMAN MIKAR	F	1/21/1972	12 HARI
26	ZULISMAR SRI HARTATI	F	9/20/1980	10 HARI

Transformasi Data

- Pada tahap ini terdapat tiga parameter yang akan digunakan dalam pengelompokan data yaitu:
 - Jenis kelamin,
 - Usia, dan
 - Paket yang dipilih jamaah.
- Sebelum masuk pada tahapan Data Mining, data yang dikumpulkan mengalami proses transformasi.
- Proses transformasi dari data yang telah diseleksi pada kolom jenis kelamin diubah dari karakter menjadi angka dapat dilihat pada Tabel di bawah ini:

Proses Transformasi Data Kolom Jenis Kelamin

DATA REAL (JENIS KELAMIN)	TRANSFORMASI DATA (JENIS KELAMIN)	KETERANGAN
M	1	Laki-laki
F	0	Perempuan

- Sedangkan pada kolom **tanggal lahir** masih format ***date*** sehingga dilakukan **pengurangan dari tahun keberangkatan dengan tahun kelahiran untuk mendapatkan usia**.
- Setelah hasil usia diperoleh maka dilakukan proses transformasi data usia dengan ketentuan nilai berikut:

Ketentuan Nilai Untuk Transformasi Data Usia

KATEGORI UMUR	UMUR / USIA	NILAI TRANSFORMASI DATA
Balita	0 – 5 tahun	1
Kanak-kanak	5 – 11 tahun	2
Remaja Awal	12 – 16 tahun	3
Remaja Akhir	17 – 25 tahun	4
Dewasa Awal	26- 35 tahun	5
Dewasa Akhir	36- 45 tahun	6
Lansia Awal	46- 55 tahun	7
Lansia Akhir	56 – 65 tahun	8
Manula	65 – sampai atas	9



Proses Transformasi Data Kolom Usia

DATA REAL				TRANSFORMASI DATA (USIA)
TANGGAL LAHIR	TAHUN LAHIR	(TAHUN LAHIR) – (TAHUN KEBERANGKATAN)	USIA	
10/12/1956	1956	2017 -1956	61	8
7/17/1979	1979	2017 -1979	38	6
11/13/1974	1974	2017 -1974	43	6
6/4/1974	1974	2017 -1974	43	6
5/24/1968	1968	2017 -1968	49	7
1/5/1952	1952	2017 -1952	65	8
12/11/1959	1959	2017 -1959	58	8
6/4/1958	1958	2017 -1958	59	8
1/6/1967	1967	2017 -1967	50	7
2/25/1979	1979	2017 -1979	38	6
10/21/1960	1960	2017 -1960	57	8
11/9/1983	1983	2017 -1983	34	5
4/2/2006	2006	2017 -2006	11	2
12/31/1953	1953	2017 -1953	64	8
7/30/1966	1966	2017 -1966	51	7
8/15/1957	1957	2017 -1957	60	8
6/9/1953	1953	2017 -1953	64	8
3/20/1957	1957	2017 -1957	60	8
12/31/1959	1959	2017 -1959	58	8
4/3/1967	1967	2017 -1967	50	7
7/1/1981	1981	2017 -1981	36	6
12/28/1968	1968	2017 -1968	49	7
12/19/1966	1966	2017 -1966	51	7
3/17/2009	2009	2017 -2009	8	2
1/21/1972	1972	2017 -1972	45	6
9/20/1980	1980	2017 -1980	37	6

- Pada kolom **Paket**, proses transformasi dilakukan dengan mengubah nilai pada tiap jenis paket berdasarkan **besaran jumlah hari**, *semakin besar nilai hari pada paket semakin tinggi nilai transformasi data* seperti pada Tabel berikut ini:

Proses Transformasi Data Kolom Paket

DATA REAL (PAKET)	TRANSFORMASI DATA (PAKET)
10 HARI	1
12 HARI	2
13 HARI	3

- **Ketiga atribut** tersebut sudah melalui proses transformasi data dan **hasil transformasi tersebut** dikumpulkan menjadi ***dataset*** yang siap untuk diolah dengan **metode K-Means Clustering**. Dataset tersebut dapat dilihat pada tabel bawah ini:



Data Sampel Hasil Transformasi (Dataset yang siap diolah)

NO	NAMA	JENIS KELAMIN	USIA	PAKET
1	ABDUL HALIM MUHAMMAD NUH	1	8	1
2	ARMIYANTI ANSARI ITAM	0	6	1
3	ENNY HALAWATI	0	6	1
4	ERFINA LEGIMIN YOSO	0	6	1
5	HENNY SYARIF MUHAMMAD	0	7	2
6	LOKOT UDIN PANJAITAN	1	8	1
7	MAIMUNAH ABDUL MUNIR SAINAL	0	8	3
8	MARTINI DARMAJI SENGUT	0	8	1
9	MASNUN MUSA HASIBUAN	0	7	2
10	MISRAN RASUL HASAN	1	6	2
11	MUHAMMAD JUNAIDI SADINO	1	8	1
12	NANING WINARSIH SUTARNO	0	5	3
13	NEZA APRIDILLA	0	2	2
14	PONI WIRYO SUMARTO	0	8	3
15	RITAWATI NASIB GURUSINGA	0	7	2
16	ROSDIANA MUCHTARAM NASUTION	0	8	2
17	ROSLAINI ISMAIL FIRMAN	0	8	1
18	SUHARTINI LASIMIN KROMO	0	8	1
19	SUMARNI SUMARTO ABDULLAH	0	8	2
20	SUNARMI OSERI ABDULLAH	0	7	3
21	SUSILAWATI HUSIN TUMAN	0	6	2
22	TAING MUHAMMAD SARI SITOMPUL	0	7	1
23	TODUH KETEK ABDULLAH	0	7	2
24	ZAHIR NAUFAL AR RAZAQ	1	2	1
25	ZULIJAWATI NGATIMAN MIKAR	0	6	2
26	ZULISMAR SRI HARTATI	0	6	1

- Tabel di atas akan dilakukan proses **Data Mining** dengan membagi data menjadi tiga kelompok yaitu **Sangat Diminati**, **Diminati** dan **Kurang Diminati** dengan algoritma **K-Means Clustering**.
- **Data sampel** yang telah mengalami transformasi data pada Tabel sebelumnya, selanjutnya akan diolah dengan menggunakan metode **K-Means Clustering** agar mendapatkan hasil berupa pengelompokan data yang diharapkan sebelumnya.

Algoritma yang terdapat pada K-Means adalah sebagai berikut:

1. Menentukan Jumlah Cluster

- Jumlah cluster atau kelompok pada pengolahan data jamaah umroh diberikan sebanyak **tiga cluster** yaitu Sangat Diminati, Diminati dan Kurang Diminati.

2. Menentukan Centroid

- Penentuan pusat awal cluster (centroid) ditentukan secara **random** atau acak yang diambil dari data yang ada.
 - Nilai centroid pertama diambil dari Tabel Hasil Transformasi tepatnya baris ke-13,
 - Nilai centroid kedua baris ke-16 dan
 - Nilai centroid terakhir baris ke-25.

Centroid Awal

NO.	NAMA	JENIS KELAMIN	USIA	PAKET
13	NEZA APRIDILLA	0	2	2
16	ROSDIANA MUCHTARAM NASUTION	0	8	2
25	ZULJAWATI NGATIMAN MIKAR	0	6	2

3. Menghitung Jarak Minimum dari Dataset dengan Centroid

- Formula yang dipakai untuk menghitung jarak antara titik centroid dengan titik tiap objek yaitu menggunakan ***Euclidean Distance***.
- Berikut ini adalah proses perhitungan jarak minimum dataset dengan titik cluster dimulai dari:

- Data sampel baris ke-1 atas nama **Abdul Halim Muhammad Nuh** dihitung dengan titik centroid ke-1 sebagai DC1, centroid ke-2 sebagai DC2 dan centroid ke-3 sebagai DC3, di mana:

$$DC1 = \sqrt{(1 - 0)^2 + (8 - 2)^2 + (1 - 2)^2} = 6.164414003$$

$$DC2 = \sqrt{(1 - 0)^2 + (8 - 8)^2 + (1 - 2)^2} = 1.414213562$$

$$DC3 = \sqrt{(1 - 0)^2 + (8 - 6)^2 + (1 - 2)^2} = 2.449489743$$

- Jadi, nilai terkecil baris ke-1 terdapat pada DC2 maka data jamaah baris ke-1 ditempatkan pada kelompok C2.

- Data sampel baris ke-2 atas nama Armiyanti Ansari Itam dihitung dengan titik centroid ke-1 sebagai DC1, centroid ke-2 sebagai DC2 dan centroid ke-3 sebagai DC3, di mana:

$$DC1 = \sqrt{(0 - 0)^2 + (6 - 2)^2 + (1 - 2)^2} = 4.123105626$$

$$DC2 = \sqrt{(0 - 0)^2 + (6 - 8)^2 + (1 - 2)^2} = 2.236067977$$

$$DC3 = \sqrt{(0 - 0)^2 + (6 - 6)^2 + (1 - 2)^2} = 1$$

- Jadi, nilai terkecil pada baris ke-2 terdapat pada DC3 maka data jamaah baris ke-2 ditempatkan pada kelompok C3.

- Data sampel baris ke-3 atas nama Enny Halawati dihitung dengan titik centroid ke-1 sebagai DC1, centroid ke-2 sebagai DC2 dan centroid ke-3 sebagai DC3, di mana:

$$DC1 = \sqrt{(0 - 0)^2 + (6 - 2)^2 + (1 - 2)^2} = 4.123105626$$

$$DC2 = \sqrt{(0 - 0)^2 + (6 - 8)^2 + (1 - 2)^2} = 2.236067977$$

$$DC3 = \sqrt{(0 - 0)^2 + (6 - 6)^2 + (1 - 2)^2} = 1$$

- Jadi, nilai terkecil pada baris ke-3 terdapat pada DC3 maka data jamaah baris ke-3 ditempatkan pada kelompok C3

- Perhitungan dilanjutkan sampai baris data ke-26 dan mendapatkan hasil penempatan anggota cluster berdasarkan kelompok.
- Hasil pengelompokan disajikan ke dalam bentuk tabel yang dapat dilihat pada Tabel berikut ini:
- Di mana:
 - DC1 = Data hasil perhitungan jarak minimum dari dataset ke Centroid 1
 - DC2 = Data hasil perhitungan jarak minimum dari dataset ke Centroid 2
 - DC3 = Data hasil perhitungan jarak minimum dari dataset ke Centroid 3



Hasil Perhitungan Jarak dan Pengelompokan Pada Iterasi Ke-1

NO	NAMA	DC1	DC2	DC3	C1	C2	C3
1	ABDUL HALIM MUHAMMAD NUH	6.164414003	1.414213562	2.449489743		1	
2	ARMİYANTI ANSARI ITAM	4.123105626	2.236067977	1			1
3	ENNY HALAWATI	4.123105626	2.236067977	1			1
4	ERFINA LEGIMIN YOSO	4.123105626	2.236067977	1			1
5	HENNY SYARIF MUHAMMAD	5	1	1		1	
6	LOKOT UDIN PANJAITAN	6.164414003	1.414213562	2.449489743		1	
7	MAIMUNAH ABDUL MUNIR SAINAL	6.08276253	1	2.236067977		1	
8	MARTINI DARMAJI SENGUT	6.08276253	1	2.236067977		1	
9	MASNUN MUSA HASIBUAN	5	1	1		1	
10	MISRAN RASUL HASAN	4.123105626	2.236067977	1			1
11	MUHAMMAD JUNAIDI SADINO	6.164414003	1.414213562	2.449489743		1	
12	NANING WINARSIH SUTARNO	3.16227766	3.16227766	1.414213562			1
13	NEZA APRIDILLA	0	6	4	1		
14	PONI WIRYO SUMARTO	6.08276253	1	2.236067977		1	
15	RITAWATI NASIB GURUSINGA	5	1	1		1	
16	ROSDIANA MUCHTARAM NASUTION	6	0	2		1	
17	ROSLAINI ISMAIL FIRMAN	6.08276253	1	2.236067977		1	
18	SUHARTINI LASIMIN KROMO	6.08276253	1	2.236067977		1	
19	SUMARNI SUMARTO ABDULLAH	6	0	2		1	
20	SUNARMİ OSERİ ABDULLAH	5.099019514	1.414213562	1.414213562		1	
21	SUSILAWATI HUSIN TUMAN	4	2	0			1
22	TAING MUHAMMAD SARI SITOMPUL	5.099019514	1.414213562	1.414213562		1	
23	TODUH KETEK ABDULLAH	5	1	1		1	
24	ZAHİR NAUFAL AR RAZAQ	1.414213562	6.164414003	4.242640687	1		
25	ZULJAWATI NGATIMAN MIKAR	4	2	0			1
26	ZULISMAR SRI HARTATI	4.123105626	2.236067977	1			1

4. Mengelompokkan objek untuk menentukan anggota cluster

- dengan menghitung jarak minimum objek ke pusat cluster.
- Perhitungan dilakukan sampai data ke-26 terhadap pusat cluster dengan memberikan kode “1” jika hasil cluster mendekati nol.

- Setelah dilakukan perhitungan jarak dan pengelompokan data pada Tabel (Hasil Perhitungan Jarak dan Pengelompokan Pada Iterasi Ke-1) maka didapatkan anggota cluster baru di mana,
- C1 memiliki 2 anggota, yaitu:

Anggota Cluster 1 Pada Iterasi Ke-1

No.	NAMA	JENIS KELAMIN	USIA	PAKET
1	NEZA APRIDILLA	0	2	2
2	ZAHIR NAUFAL AR RAZAQ	1	2	1

- C2 memiliki 16 anggota, yaitu:

Anggota Cluster 2 Pada Iterasi Ke-1

No.	NAMA	JENIS KELAMIN	USIA	PAKET
1	ABDUL HALIM MUHAMMAD NUH	1	8	1
2	HENNY SYARIF MUHAMMAD	0	7	2
3	LOKOT UDIN PANJAITAN	1	8	1
4	MAIMUNAH ABDUL MUNIR SAINAL	0	8	3
5	MARTINI DARMAJI SENGUT	0	8	1
6	MASNUN MUSA HASIBUAN	0	7	2
7	MUHAMMAD JUNAIDI SADINO	1	8	1
8	PONI WIRYO SUMARTO	0	8	3
9	RITAWATI NASIB GURUSINGA	0	7	2
10	ROSDIANA MUCHTARAM NASUTION	0	8	2
11	ROSLAINI ISMAIL FIRMAN	0	8	1
12	SUHARTINI LASIMIN KROMO	0	8	1
13	SUMARNI SUMARTO ABDULLAH	0	8	2
14	SUNARMi OSERI ABDULLAH	0	7	3
15	TAING MUHAMMAD SARI SITOMPUL	0	7	1
16	TODUH KETEK ABDULLAH	0	7	2

- C3 memiliki 8 anggota, yaitu:

Anggota Cluster 3 Pada Iterasi Ke-1

No.	NAMA	JENIS KELAMIN	USIA	PAKET
1	ARMIYANTI ANSARI ITAM	0	6	1
2	ENNY HALAWATI	0	6	1
3	ERFINA LEGIMIN YOSO	0	6	1
4	MISRAN RASUL HASAN	1	6	2
5	NANING WINARSIH SUTARNO	0	5	3
6	SUSILAWATI HUSIN TUMAN	0	6	2
7	ZULIJAWATI NGATIMAN MIKAR	0	6	2
8	ZULISMAR SRI HARTATI	0	6	1

5. Lakukan iterasi untuk melihat apakah jumlah anggota cluster berpindah tempat atau tetap.

- Jika diiterasi kedua anggota cluster tidak berpindah tempat maka iterasi dihentikan dan anggota tiap cluster telah ditetapkan.
- Iterasi kedua dilakukan dengan cara menentukan nilai centroid baru dari hasil rata-rata anggota cluster yang berada pada anggota cluster yang sama.

Studi Kasus (lanjutan)

$$C1 = \frac{(1+0)}{2} = 0.5$$

$$\frac{(2+2)}{2} = 2$$

$$\frac{(1+2)}{2} = 1.5$$

$$C1 = 0.5, 2, 1.5$$

$$\mathbf{C2} = \frac{(1+0+1+0+0+0+1+0+0+0+0+0+0+0+0+0)}{16} = 0.1875$$

$$= \frac{(8+7+8+8+8+7+8+8+7+8+8+8+8+7+7+7)}{16} = 7.625$$

$$= \frac{(1+2+1+3+1+2+1+3+2+2+1+1+2+3+1+2)}{16} = 1.75$$

$$\mathbf{C2} = \mathbf{0.1875, 7.625, 1.75}$$

$$C3 = \frac{(0+0+0+1+0+0+0+0)}{8} = 0.125$$

$$\frac{(6+6+6+6+5+6+6+6)}{8} = 5.875$$

$$\frac{(1+1+1+2+3+2+2+1)}{8} = 1.625$$

$$C3 = 0.125, 5.875, 1.625$$

- Dari hasil rata-rata di atas didapatkan nilai pusat cluster baru yang ditampilkan dalam bentuk Tabel berikut:

Nilai *Centroid* Baru Untuk Iterasi Ke-2

CLUSTER	NILAI CENTROID BARU		
C1	0.5	2	1.5
C2	0.1875	7.625	1.75
C3	0.125	5.875	1.625

- Ian H Witten, Eibe Frank, Mark A Hall. (2016). *Data Mining - Practical Machine Learning Tools and Techniques*, 4th Edition. Morgan-Kaufman.
- Hofmann, Markus., Klinkenberg, Ralf. (2016). *Data Mining Use Cases and Business Analytics Application*. CRC Press.
- Crisholm, Andrew. (2013). *Exploring Data With Rapidminer*. Packt Publishing.
- Mehmed Kantardzic. (2020). *Data Mining - Concepts, Models, Methods, and Algorithms*, 3rd edition. Wiley - IEEE Press.
- rapidminer.com, 2021,. Rapidminer Manual and Documentation, <https://docs.rapidminer.com/latest/studio/operators/>.



TERBATA-BATA? **DAPAT 2 PAHALA** **LANCAR?** **BERSAMA MALAIKAT**

”Orang yang pandai dalam membaca Al Qur’an, maka ia beserta para Malaikat yang mulia. Orang yang terbata-bata dan merasa kesulitan, maka ia akan memperoleh dua pahala.” (HR. Ibnu Majah: 3769)

Terima Kasih



Data Mining



Uhamka

UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion



Session 13

PROGRAM STUDI TEKNIK INFORMATIKA

Sumber: <https://ft.uhamka.ac.id/prodi/informatika/>

- **VISI.**

- √ Menjadi program studi informatika yang menghasilkan karya unggul di bidang teknologi informasi melalui pembelajaran berlandaskan prophetic teaching untuk membentuk lulusan yang cerdas secara spiritual, intelektual, emosional, dan sosial.

- **MISI.**

1. Melaksanakan pendidikan di bidang informatika yang terintegrasi dengan nilai-nilai Al Islam Kemuhammadiyah.
2. Melaksanakan penelitian untuk menghasilkan karya teknologi informasi yang unggul untuk masyarakat berkembang
3. Melaksanakan pengabdian kepada masyarakat untuk kesejahteraan umat.
4. Menghasilkan mahasiswa yang berwawasan global dalam teknologi informasi.
5. Mendorong jiwa kewirausahaan mahasiswa untuk berkiprah di tengah masyarakat dalam meningkatkan kegiatan ekonomi.
6. Menerapkan standar mutu pendidikan informatika melalui layanan mutu yang berkesinambungan.
7. Membangun kerja sama terpadu dengan mitra yang berkontribusi untuk kemajuan pendidikan di bidang informatika.

- SKS : 3
 - Teori 3 SKS – (14 Kali Pertemuan)
- Keaktifan (bobot **10%**)
 - Terdiri dari: Aktivitas Perkuliahan (*Kehadiran, Etika, dan Sikap*).
- Project (bobot **25%**)
 - Terdiri dari: Tugas Besar / Project (*Penguasaan Keterampilan*) dengan cara observasi, mengimplementasikan dan menghasilkan Proyek Akhir, serta mempresentasikannya.
- Tes (UTS = bobot **25%** , UAS = bobot **40%**)
 - Terdiri dari: UTS dan UAS (*Penguasaan Pengetahuan*) dengan cara tes tertulis maupun tes praktikum.

Tujuan/Objectives (Session.13)

- Objectives (Tujuan Instruksional Umum).
 - RPS/SAP.
 - Data Mining & BI
- Indikator Kompetensi.
 - Mahasiswa mampu mengetahui dan memahami tentang keterhubungan *data mining* dan *business intelligence* pada *data mining*.
 - Mahasiswa mampu mengetahui, memahami, serta dapat menjelaskan tentang penelitian yang berhubungan dengan topik *business intelligence* dan *data mining* didalam *data mining*.

وَنُزِّلُ مِنَ الْقُرْآنِ مَا هُوَ شِفَاءٌ وَرَحْمَةٌ لِّلْمُؤْمِنِينَ
وَلَا يَزِيدُ الظَّالِمِينَ إِلَّا خَسَارًا

- Qs. Al Israa, ayat 82
 - Dan Kami turunkan dari Al-Qur'an (sesuatu) yang menjadi penawar dan rahmat bagi orang yang beriman, sedangkan bagi orang yang zalim (Al-Qur'an itu) hanya akan menambah kerugian.



Data Mining



Uhamka

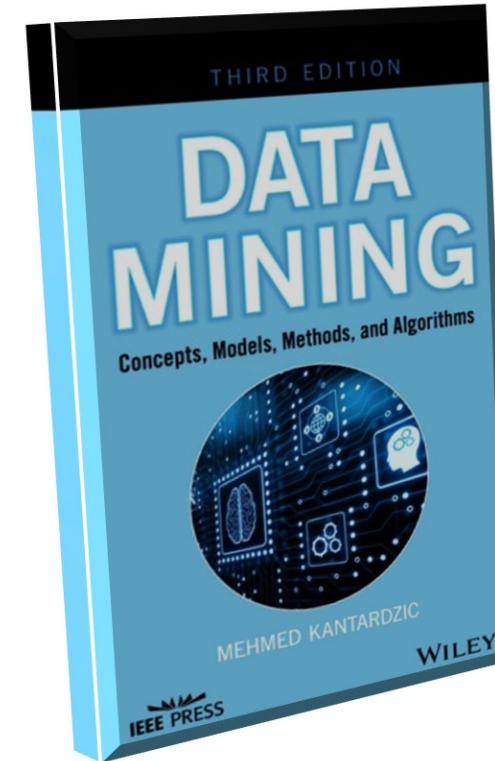
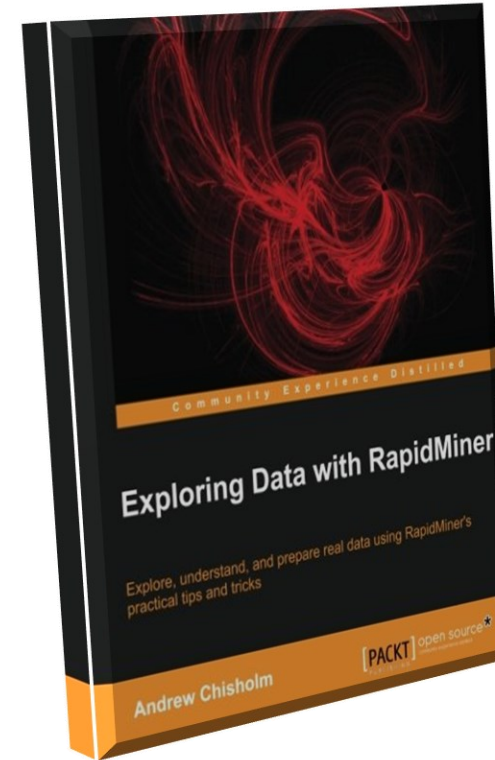
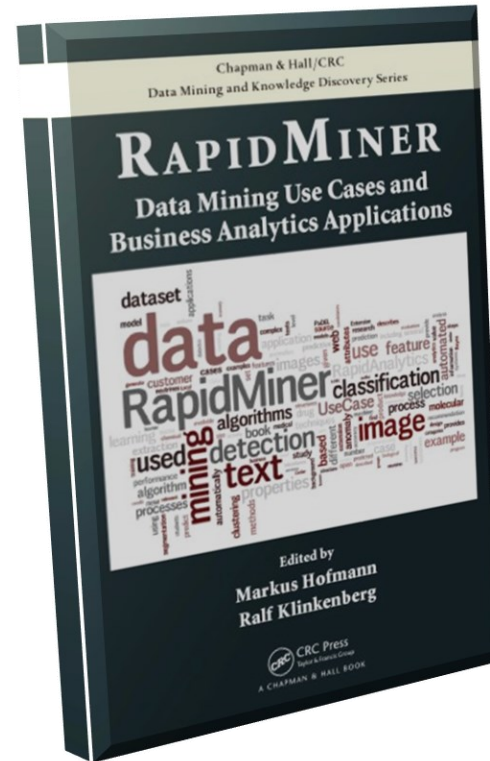
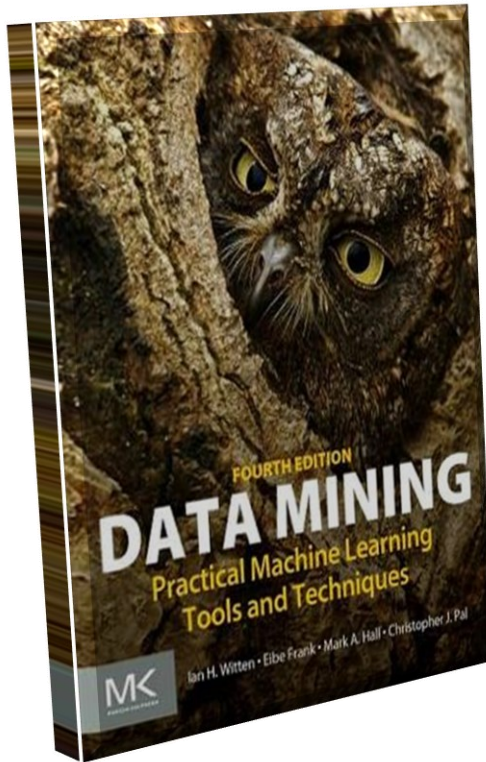
UNIVERSITAS MUHAMMADIYAH PROF. DR. HAMKA

integrity, trust, compassion



Session 13

Data Mining & Business Intelligence



Data Mining & Business Intelligence

Paper Review

- Studi kasus yang digunakan, adalah berdasarkan pada penelitian.
- *Paper* dapat di-*download* pada laman berikut.

<https://journal.uhamka.ac.id/index.php/teknoka/article/view/3943>



Vol. 4, 2019
ISSN No. 2502-8782

Implementasi Sistem *Business Intelligence* Untuk Data Penelitian di Perguruan Tinggi

Firman Noor Hasan
Teknik Informatika Universitas Muhammadiyah Prof. Dr. Hamka
Jl. Tanah Merdeka No.6, Pasar Rebo, Jakarta Timur Telp: 021-8778.2739 , Mobile: 0813.1140.6664
Website: www.ft.uhamka.ac.id , E-mail: firmam.noorhasan@uhamka.ac.id

Abstrak - Data didalam sebuah perguruan tinggi merupakan aset, aset yang senantiasa berkembang dan membutuhkan pengelolaan secara khusus, baik dari sisi pemanfaatannya maupun dari sisi penyimpanannya. Pembangunan business intelligence merupakan salah satu cara untuk dapat mengekstrak informasi penting di sebuah perguruan tinggi. Kondisi sistem yang berjalan saat ini di Perguruan Tinggi XYZ yaitu laporan tentang data-data penelitian belum mempunyai database dan hanya dihasilkan langsung dengan file yang berbentuk excel, dan setiap program studi mempunyai format yang berbeda-beda. Oleh karena itu dibutuhkan sistem business intelligence yang perancangannya meliputi perancangan data warehouse, perancangan olap analysis, perancangan report, dan perancangan dashboard. Penelitian ini menghasilkan sebuah sistem business intelligence di Perguruan Tinggi XYZ yang sangat membantu didalam mengumpulkan data-data penelitian sehingga dapat dianalisis. Report yang dirancang sangat membantu didalam membuat laporan penelitian dan dapat disesuaikan dengan kebutuhan, sehingga masalah yang sering terjadi dalam hal kebergantungan dalam perolehan laporan diharapkan tidak terjadi kembali. Dashboard penelitian yang dirancang membantu pimpinan didalam menganalisis data untuk mempelajari tren penelitian yang dilakukan di Perguruan Tinggi XYZ, serta dapat dijadikan untuk mendukung pengambilan keputusan dan nantinya dapat juga sebagai pengukur kinerja dosen.

Kata kunci: business intelligence, data warehouse, olap analysis, dashboard.



Implementasi Sistem *Business Intelligence* Untuk Data Penelitian di Perguruan Tinggi

Firman Noor Hasan¹

¹Teknik Informatika Universitas Muhammadiyah Prof. Dr. Hamka



Abstrak

Data didalam sebuah perguruan tinggi merupakan aset, aset yang senantiasa berkembang dan membutuhkan pengelolaan secara khusus, baik dari sisi pemanfaatannya maupun dari sisi penyimpanannya. Pembangunan *business intelligence* merupakan salah satu cara untuk dapat mengekstrak informasi penting di sebuah perguruan tinggi.

Kondisi sistem yang berjalan saat ini di Perguruan Tinggi XYZ yaitu laporan data-data penelitian belum mempunyai database dan hanya dihasilkan langsung dengan file yang berbentuk excel, dan setiap program studi mempunyai format laporan yang berbeda-beda. Karena itu dibutuhkan sebuah sistem *business intelligence* diharapkan dapat mengatasi permasalahan kebutuhan laporan yang sering terjadi.

Keyword: *business intelligence*, *data warehouse*, *dashboard*, *on-line analytical processing*,

Pendahuluan

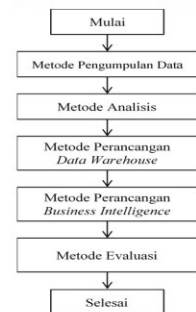
Data merupakan sebuah aset yang sangat berharga didalam sebuah perguruan tinggi, yang membutuhkan pengelolaan secara khusus, baik dari sisi pemanfaatan maupun dari sisi penyimpanannya. Pembangunan *business intelligence* merupakan salah satu cara untuk dapat mengekstrak informasi penting di sebuah perguruan tinggi.

Penelitian ini bertujuan untuk melakukan implementasi sistem *business intelligence* untuk data-data penelitian yang ada di sebuah Perguruan Tinggi swasta di Jakarta. Perancangan sistem *business intelligence* ini mengacu pada standar yang dijelaskan oleh Vercellis, sedangkan perancangan *data warehouse* dilakukan dengan menerapkan *nine-step methodology* oleh Ralph Kimball. Perangkat lunak yang digunakan untuk membangun sistem *business intelligence* didalam penelitian ini yaitu dengan memanfaatkan perangkat lunak berbasis *open source* (kode terbuka), yaitu *Pentaho Data Integration* (PDI) yang digunakan untuk perancangan *data warehouse* data-data penelitian, *Mondrian OLAP* (*Online Analytical Processing*) digunakan untuk perancangan OLAP, *Pentaho Report Designer* (PRD) digunakan untuk perancangan laporan data-data penelitian, *Pentaho Community Dashboard Editor* (Pentaho CDE) merupakan perangkat lunak berbasis *open source* yang digunakan merancang *dashboard* untuk menampilkan data-data penelitian.

Kontak Penulis

Firman Noor Hasan
Universitas Muhammadiyah Prof. Dr. Hamka
Email: firman.noorhasan@uhamka.ac.id
Website: www.uhamka.ac.id
Phone: 0813.1140.6664

Metodologi Penelitian



Gambar 1. Diagram Alir

Melalui pengembangan dan implementasi dari sistem *business intelligence* merupakan salah satu solusi yang dinilai cukup potensial untuk dapat memecahkan masalah tersebut diatas yang akan menjadi salah satu tools yang bisa dipakai untuk melakukan analisis dan monitoring terhadap data-data yang di perlukan untuk pencapaian target nilai akreditasi.

Gambar.1 menunjukkan diagram alir dari penelitian yang dilakukan. Sedangkan gambar.2 menunjukkan skema bintang dari *datawarehouse* yang dibangun.

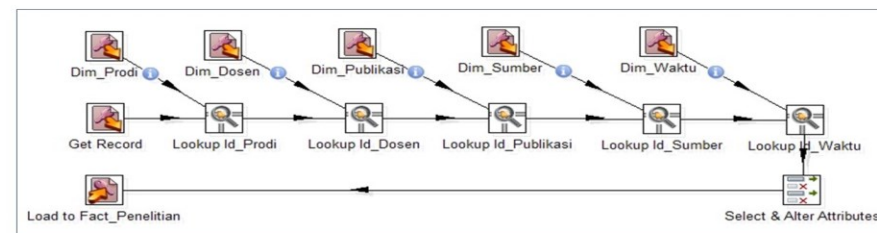
Terdapat beberapa metode yang dilakukan didalam penelitian ini, diantaranya adalah:

- Metode Pengumpulan Data
- Metode Analisis
- Metode Perancangan *Data Warehouse*
- Metode Perancangan *Business Intelligence*
- Metode Evaluasi

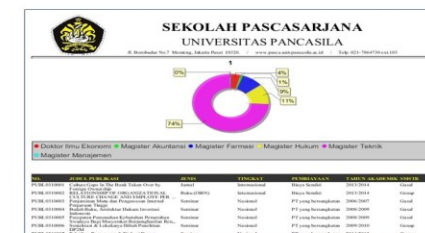
Hasil dan Pembahasan

Data-data penelitian untuk perancangan *data warehouse* berasal dari data yang dikumpulkan melalui data borang standar 7 yang dibuat dengan Ms.Excell & Ms.Access. Setelah itu dibuat *database* operasional yang dimasukkan oleh sebuah *job* yang berisi kumpulan proses ETL (*Extract, Transform, Load*) dari masing-masing tabel dimensi dan tabel fakta, seperti yang ditampilkan oleh gambar.3. Proses perancangan *OLAP Analysis* nantinya digunakan untuk membuat laporan di *OLAP* (*On-Line Analytical Processing*) dengan menghubungkan antara tabel fakta dan dimensi, seperti ditunjukkan pada gambar.4. *Cube PenelitanModels* dibuat dengan tujuan untuk memodelkan data, sehingga data penelitian dapat dianalisis dari berbagai dimensi. Perancangan *dashboard SPs* menggunakan *CDE* (*Community Dashboard Editor*), seperti ditunjukkan pada gambar.5. Untuk membuat *dashboard* memerlukan penggabungan beberapa tabel untuk membentuk sebuah relasi, yaitu dari tabel-tabel dimensi dan tabel fakta. Sistem *business intelligence* untuk data penelitian yang telah dibangun ini dapat memberikan rekomendasi kepada pimpinan Perguruan Tinggi XYZ sebagai berikut:

- Perguruan Tinggi XYZ harus mendorong program studi lain agar dapat meningkatkan jumlah penelitiannya.
- Perguruan Tinggi XYZ harus menambah anggaran untuk membiayai penelitian-penelitian, sehingga dapat meningkatkan kualitas dan kuantitas penelitian yang ada.
- Perguruan Tinggi XYZ harus mendorong kepada dosen agar meningkatkan publikasi pada level tingkat internasional.
- Perguruan Tinggi XYZ harus memotivasi kepada dosen-dosen yang masih mempunyai jenjang jabatan akademik sebagai Asisten Ahli dan Lektor agar terus meningkatkan kepengkatannya salah satunya yaitu melalui penelitian yang luaran nya menghasilkan publikasi.

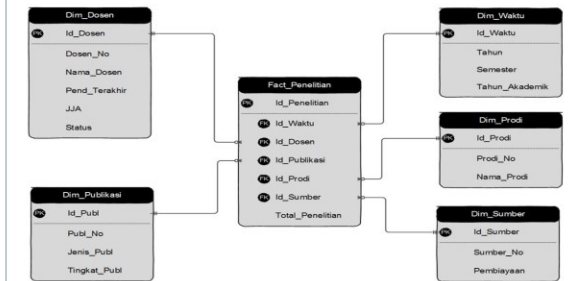


Gambar 4. Proses ETL Fact_Penelitian



Gambar 4. Tampilan Report Data Penelitian

Skema Bintang Data Warehouse

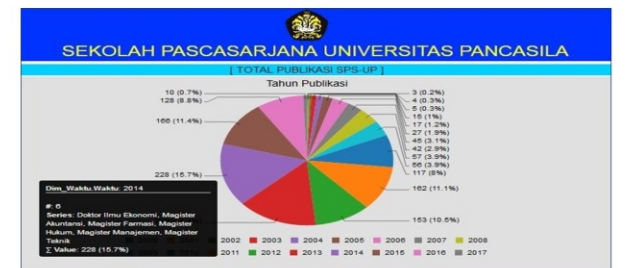


Gambar 2. Star Schema Data Warehouse Penelitian

Simpulan

Penelitian ini menghasilkan sebuah sistem *business intelligence* di Perguruan Tinggi XYZ yang sangat membantu didalam mengumpulkan data-data penelitian sehingga dapat dianalisis. Selain itu *Report* yang dirancang sangat membantu didalam membuat laporan penelitian dan dapat disesuaikan dengan kebutuhan, sehingga masalah yang sering terjadi dalam hal ketergantungan dalam perolehan laporan diharapkan tidak terjadi kembali.

Dashboard penelitian yang dirancang membantu pimpinan didalam menganalisis data untuk mempelajari tren penelitian yang dilakukan di Perguruan Tinggi XYZ, serta dapat dijadikan untuk mendukung pengambilan keputusan dan nantinya dapat juga untuk dijadikan sebagai pengukur kinerja dosen.



Gambar 5. Dashboard Data Penelitian

Kepustakaan

1. B. M. Drake and A. Walz, "Evolving Business Intelligence and Data Analytics in Higher Education," *New Dir. Institutional Res.*, vol. 2018, no. 178, pp. 39-52, 2018.
2. Henderi and E. Winarko, "Rancangan Framework Business Intelligence Pada Perguruan Tinggi," *Semin. Nas. Mat. dan Apl.*, 2013, no. June 2013, pp. 406-411, 2013.
3. H. P. Poerbo, "Purwarupa Data Warehouse Pada Sistem Informasi Manajemen Perguruan Tinggi Studi Kasus: STIKOM Surabaya," pp. 1-13.
4. B. A. N. P. T. BAN PT, "Buku II: Standar dan Prosedur," 2010.
5. Henderi, I. Handayani, and M. A. Dewi, "Business Intelligence Development Model Using Star Schema Methodology," vol. 5, no. 3, pp. 233-250, 2012.
6. Salman, "Rancangan Sistem Business Intelligence Sebagai Pendukung Pengambilan Keputusan Di Anukom Mataram," *CommIT (Communication Inf. Technol. J.)*, no. 2017, 2017.
7. Z. Arifin and A. Sugiharto, "Rancang Bangun Sistem Business Intelligence Universitas Sebagai Pendukung Pengambilan Keputusan Akademik," *J. Sist. Inf. Bisnis*, vol. 3, no. 1, pp. 30-40, 2013.
8. C. Vercellis, *Business Intelligence: Data Mining and Optimization for Decision Making*, 1st ed. Italy: Wiley Publishing, Inc., 2009.
9. R. Kimball and M. Ross, *The Data Warehouse Lifecycle Toolkit, Relentlessly Practical Tools for Data Warehousing and Business Intelligence*. Wiley Publishing, Inc., 2010.
10. C. D. Aquila, F. Di Tria, E. Lefors, and F. Tangorra, "Business Intelligence Applications for University Decision Makers," *WSEAS Trans. Comput.*, vol. 7, no. 7, pp. 1010-1019, 2017.
11. W. H. Inmon, *Building The Data Warehouse*, Fourth Ed. Indiana: Wiley Publishing, Inc., 2005.
12. S. Williams and N. Williams, *The Profit Impact of Business Intelligence*. San Fransisco: Morgan Kaufmann Publishers, 2007.

Acknowledgements

Paper ini dibuat dalam rangka Seminar Nasional Teknologi, Kualitas, dan Aplikasi 2019 (TEKNOKA 2019) Sabtu, 30 November 2019 Jakarta, Indonesia

Abstrak

Data didalam sebuah perguruan tinggi merupakan aset, aset yang senantiasa berkembang dan membutuhkan pengelolaan secara khusus, baik dari sisi pemanfaatannya maupun dari sisi penyimpanannya. Pembangunan business intelligence merupakan salah satu cara untuk dapat mengekstrak informasi penting di sebuah perguruan tinggi.

Kondisi sistem yang berjalan saat ini di Perguruan Tinggi XYZ yaitu laporan tentang data-data penelitian belum mempunyai database dan hanya dihasilkan langsung dengan file yang berbentuk excel, dan setiap program studi mempunyai format yang berbeda-beda. Oleh karena itu dibutuhkan sistem business intelligence yang perancangannya meliputi perancangan data warehouse, perancangan olap analysis, perancangan report, dan perancangan dashboard.

Penelitian ini menghasilkan sebuah sistem business intelligence di Perguruan Tinggi XYZ yang sangat membantu didalam mengumpulkan data-data penelitian sehingga dapat dianalisis. Report yang dirancang sangat membantu didalam membuat laporan penelitian dan dapat disesuaikan dengan kebutuhan, sehingga masalah yang sering terjadi dalam hal kebergantungan dalam perolehan laporan diharapkan tidak terjadi kembali. Dashboard penelitian yang dirancang membantu pimpinan didalam menganalisis data untuk mempelajari tren penelitian yang dilakukan di Perguruan Tinggi XYZ, serta dapat dijadikan untuk mendukung pengambilan keputusan dan nantinya dapat juga sebagai pengukur kinerja dosen.

Kata kunci: *business intelligence, data warehouse, olap analysis, dashboard.*

Pendahuluan

- Rumusan Masalah
 - Untuk mencapai nilai akreditasi dari BAN-PT, maka diperlukan data dan informasi yang dapat digunakan dalam mendukung pengambilan keputusan.
 - Yaitu adanya sistem yang dapat menyediakan informasi dalam bentuk dashboard, sehingga memudahkan pimpinan didalam menganalisis dan membuat keputusan yang strategis.
- Tujuan Penelitian
 - Menghasilkan rancangan sistem business intelligence yang dapat digunakan dalam pengambilan keputusan didalam borang standar 7, yaitu Penelitian di Perguruan Tinggi XYZ.



Manfaat Penelitian

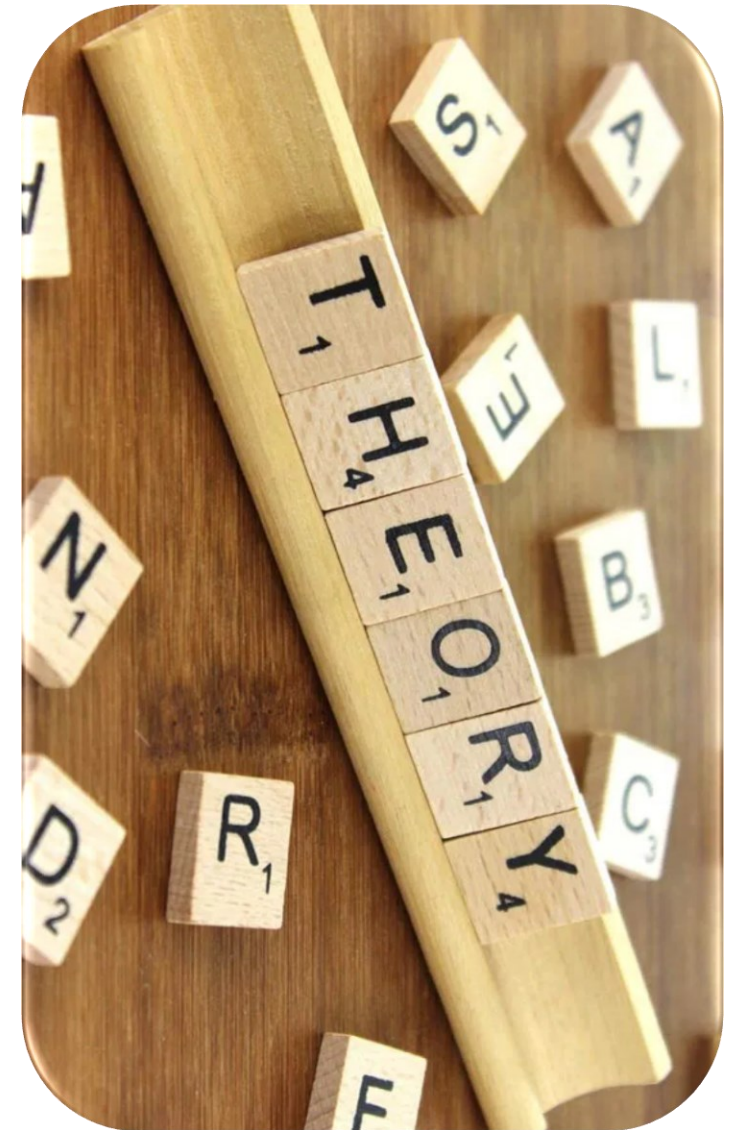
- Sistem business intelligence yang dibangun memiliki dashboard dan laporan analisis yang dapat digunakan untuk mendukung didalam pengambilan keputusan yang menyangkut borang standar 7 di SPs – UP.
- Sistem business intelligence yang dibangun dapat menjadi tools analisis data sehingga menunjang kualitas pendidikan di SPs – UP.
- Membantu tim borang tentang data penelitian dosen, sehingga memudahkan pekerjaan dan menghemat waktu.
- Bagi peneliti, dapat digunakan sebagai bahan referensi tentang business intelligence.



Landasan Teori

- Data Warehouse (DW).
 - Structure, ETL process, Star Schema.
 - DW development with nine steps methodology (Kimball & Ross, 2010).
- Business Intelligence (BI).
 - Benefit, Characteristize, Architecture.
 - BI development phase (Vercellis, 2009).
- Decission Support Systems (DSS).
- On-Line Analytical Processing (OLAP).
- Badan Akreditasi Nasional Perguruan Tinggi (BAN – PT).
 - Borang Standar 7 (Penelitian, Pelayanan Pengabdian kepada masyarakat, dan kerjasama).

Paper Review (lanjutan)



Software Pendukung

Paper Review (lanjutan)

- XAMPP (*versi 3.2.2*)
 - Apache Web Server.
 - MySQL.
- Pentaho (*versi ce-5.0.1-stable*)
 - Business Intelligence Server.
 - Data Integration (PDI) / Kettle.
 - Schema Workbench
 - Data Analysis / Mondrian OLAP Engine.
 - Report Designer.
 - Community Dashboard Framework.
- Ms.Office (*versi 2010*)
 - Ms.Word, Ms.Excell, Ms.Access.





User Console

User Name:

Password:

Login

Login as an Evaluator ▾

Administrator

User Name: Admin

Password: password

GO

Business User

User Name: Suzy

Password: password

GO

© 2005-2017 Pentaho Corporation. All rights reserved.

Metode Pengumpulan Data

Paper Review (lanjutan)

- Observasi
 - Terhadap bisnis proses yang ada dan sedang berjalan.
- Wawancara
 - Terhadap user agar memahami kebutuhan informasi yang diperlukan.
- Studi Literatur
 - Buku, Jurnal, Artikel ilmiah, dan penelitian terkait yang telah dilakukan sebelumnya.
- Dokumen Data Borang
 - Dokumen data borang akreditasi Standar 7.



Metode Analisis

Paper Review (lanjutan)

“Business intelligence opportunity analysis, diperlukan untuk mengidentifikasi dan memberi prioritas kebutuhan business information mana yang akan membawa keuntungan maksimal bagi perusahaan” (*Williams & Williams, 2007, p.26*)

Business intelligence opportunity analysis, terdiri dari :

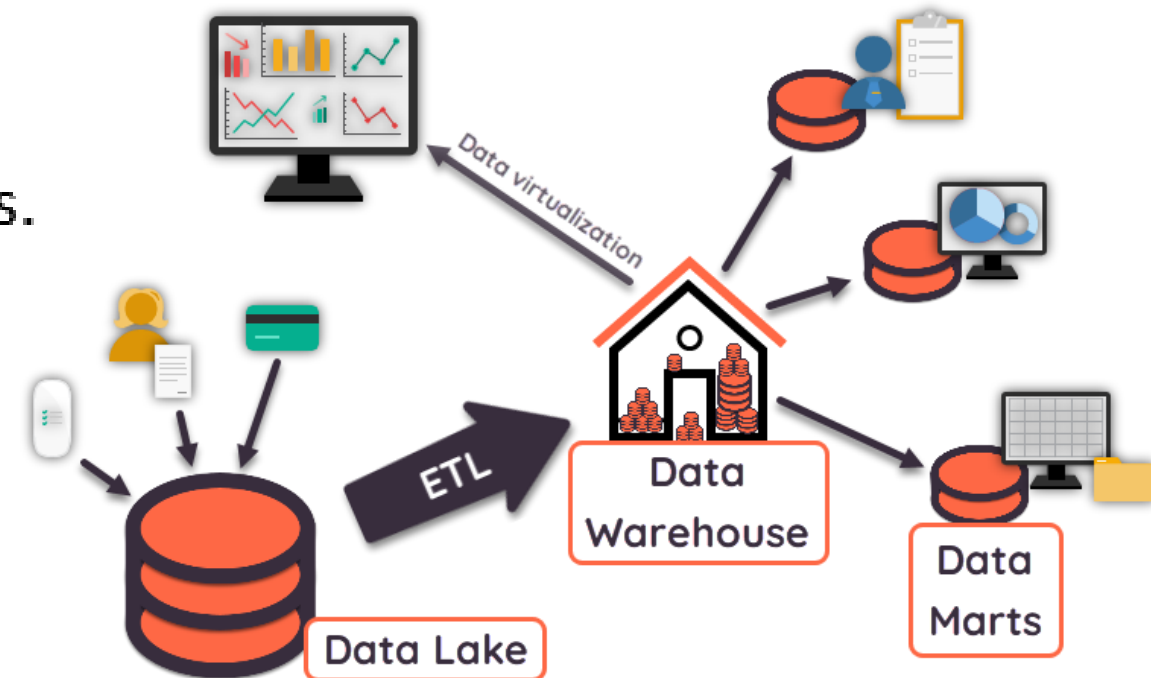
- Business Driver.
- Business Strategy, Goal, and Objectives.
- Value Disciplines.
- Core Business Process.
- Business Value.



Metode Perancangan DW

- Nine step methodology for Data Warehouse (DW)
(Kimball & Ross, 2010, p.210).

1. Choosing the process.
2. Choosing the grain.
3. Identifying and conforming the dimensions.
4. Choosing the facts.
5. Storing pre-calculation in the fact table.
6. Rounding out the dimensions table.
7. Choosing the duration of the database.
8. Tracking slowly changing dimensions.
9. Deciding the physical design.



Metode Perancangan BI

Paper Review
(lanjutan)

“Untuk mengembangkan sebuah Business Intelligence (BI) didalam sebuah perusahaan, terdapat 4 (empat) fase atau tahapan-tahapan utama yang harus dilakukan.”

(Business Intelligence: Data Mining and Optimization for Decission Making, Carlo Vercellis, 2009, p.14).

Empat (4) fase utama tersebut yaitu:

1. Analysis.
2. Design.
3. Planning.
4. Implementation and Control.

Paper Review

(lanjutan)

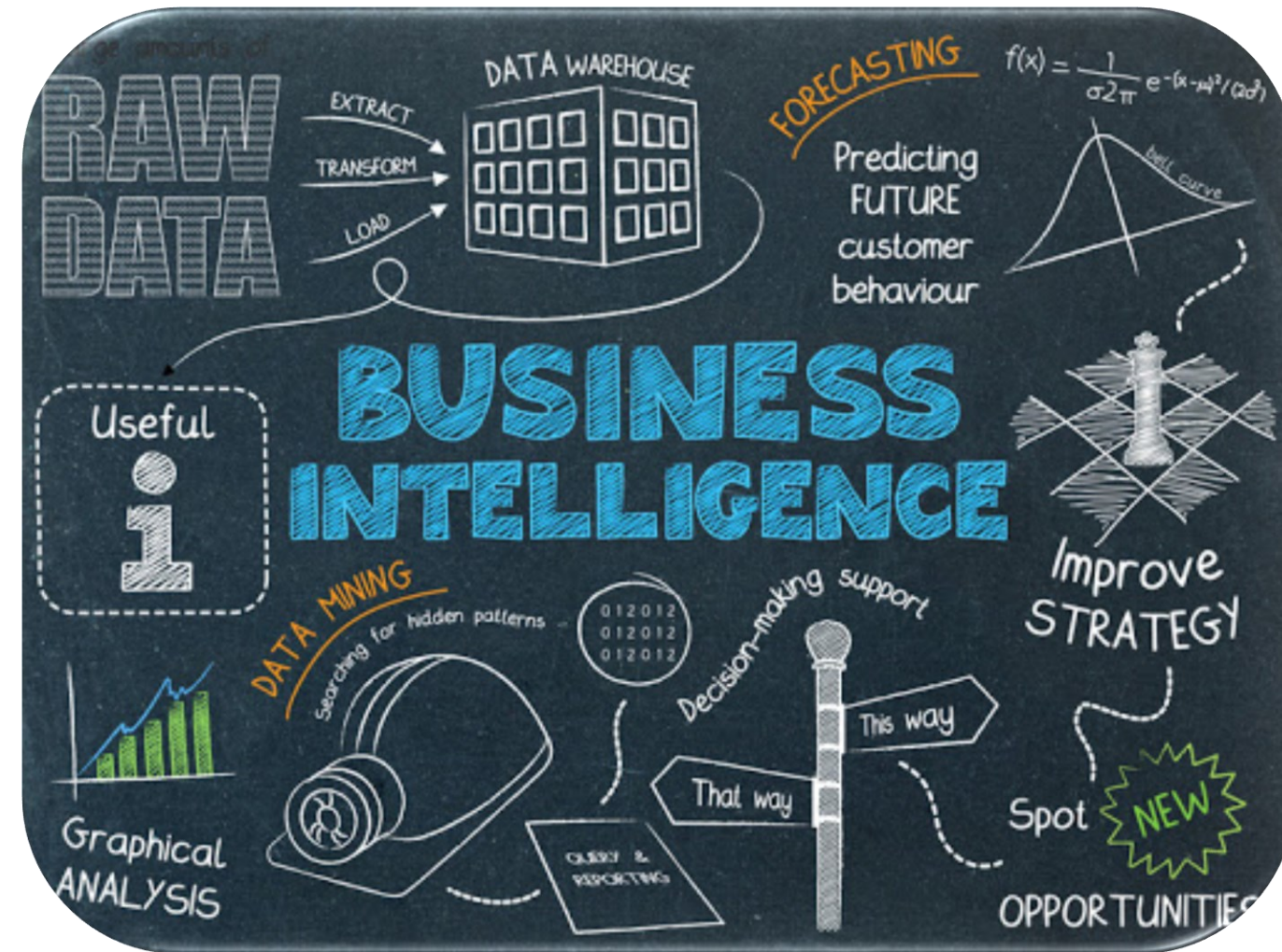
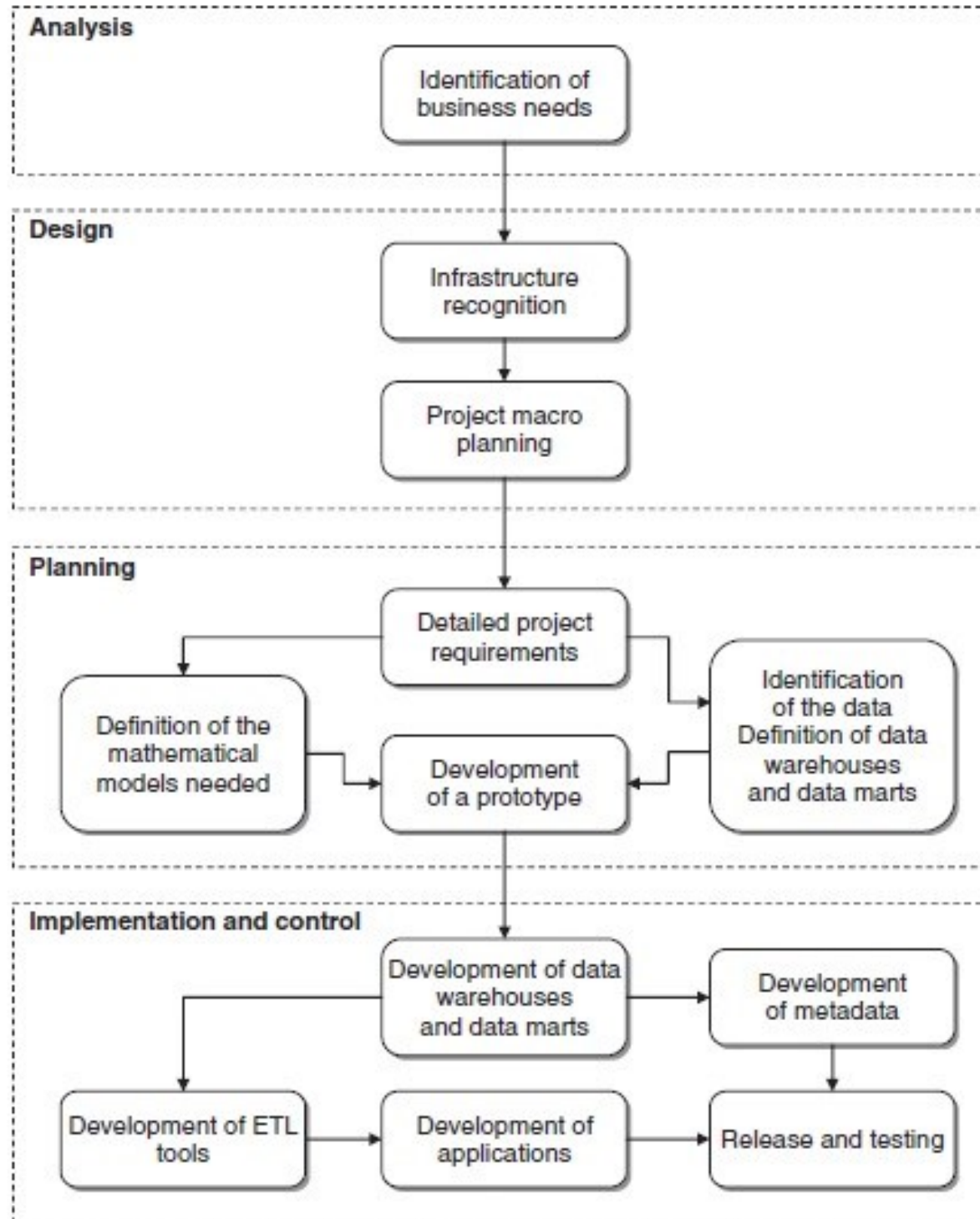


Figure 1.6 Phases in the development of a business intelligence system

1. Fase Analysis

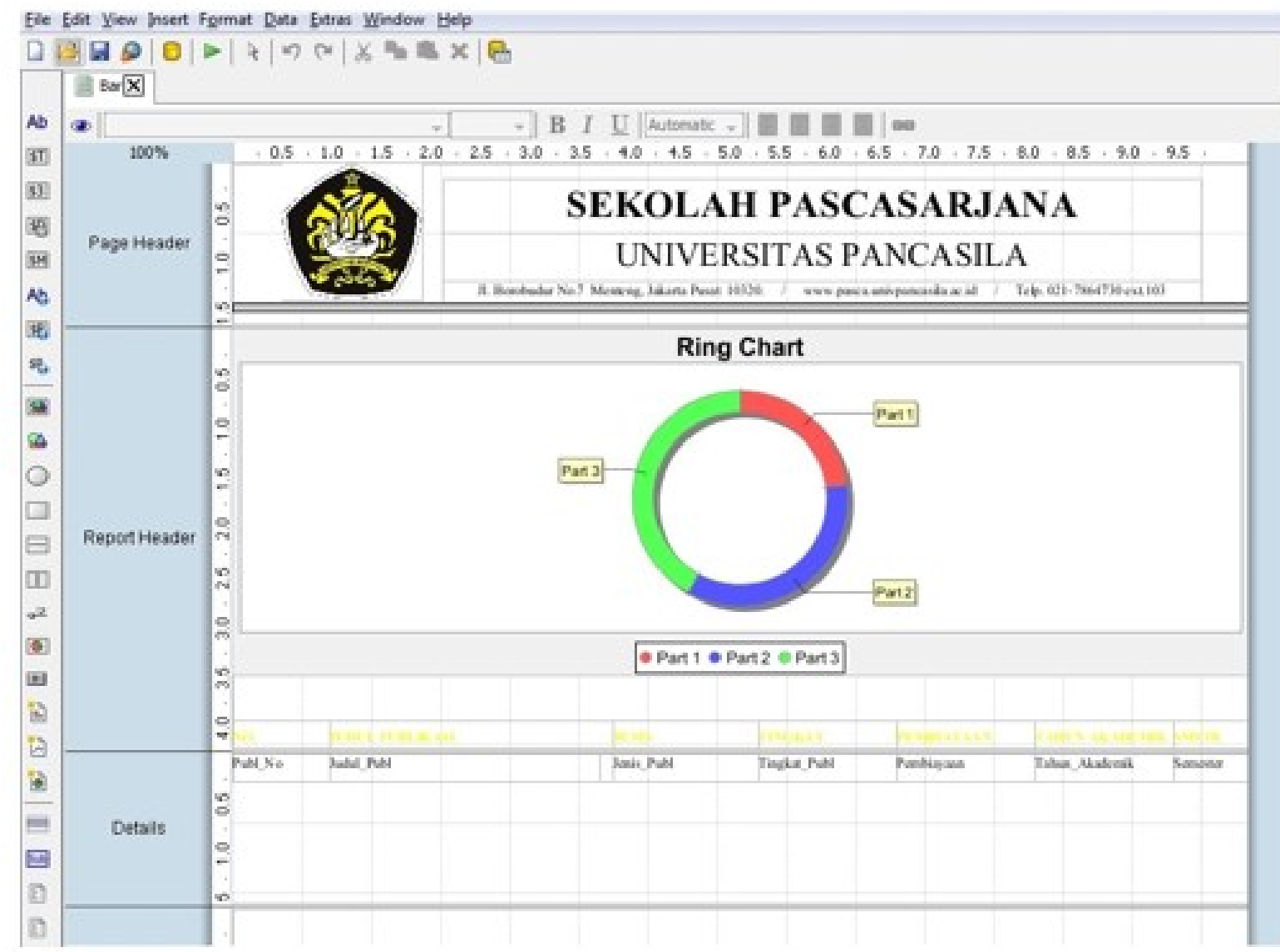
- Identification of business needs
 - Tujuan umum.
 - Prioritas dari project.
 - Manfaat dari project.

2. Fase Design

a. Infrastructure recognition.

Tabel 3.16 – Tabel *Server Specification* di SPs

Model Name	Server
Processor	Intel Core i3-4160 CPU @ 3.60GHz
RAM	10 Gigabyte
Harddisk	1 Terrabyte
Operating System	Windows 7
DBMS	MySQL
ETL	Pentaho Data Integration (PDI)
Analysis	Pentaho Data Analysis (Mondrian OLAP Engine)
Report	Pentaho Report Designer
Dashboard	Pentaho Community Dashboard Framework (CDF)

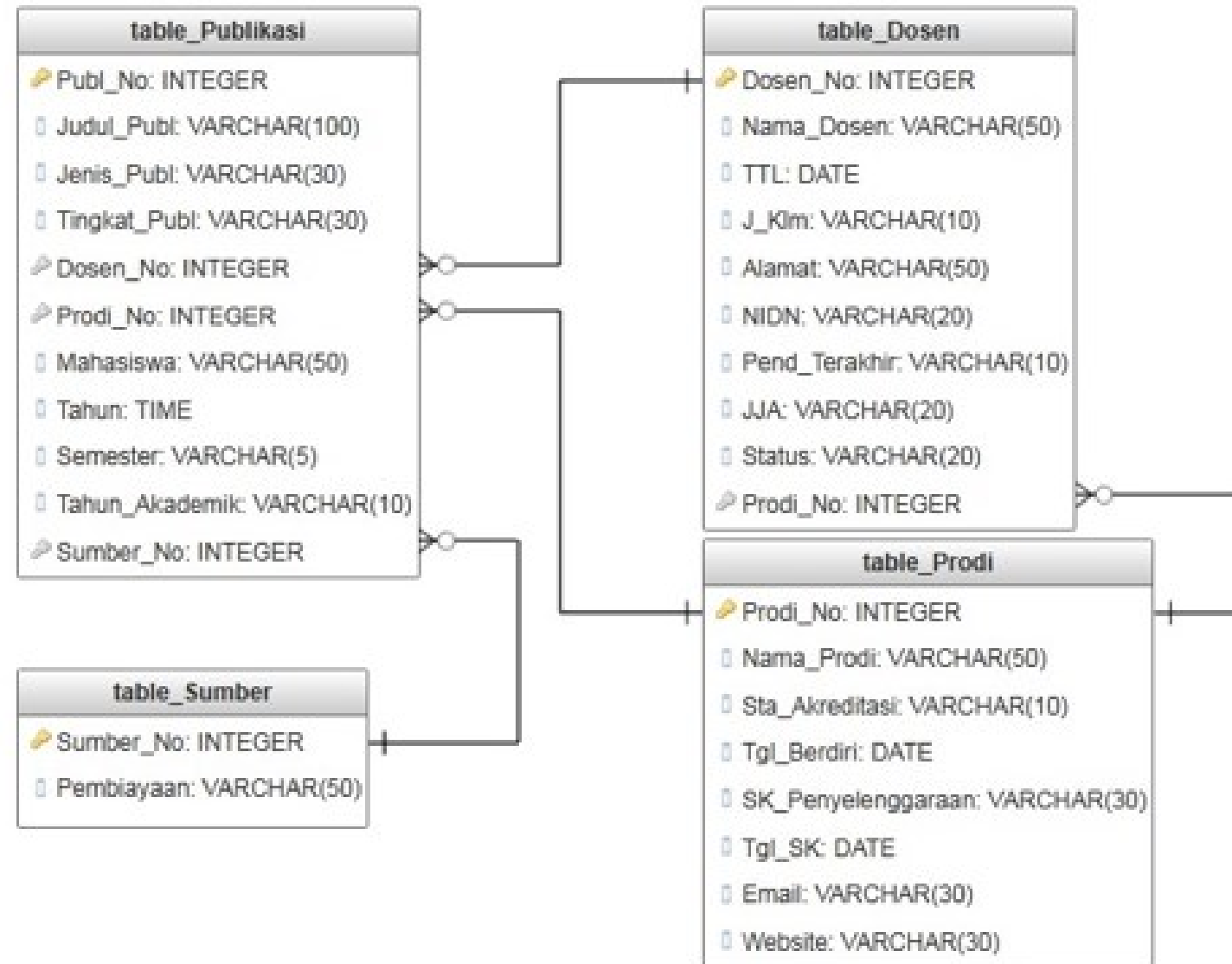


Gambar 4.106 – Desain *Report*.
(Sumber: SPs-UP)

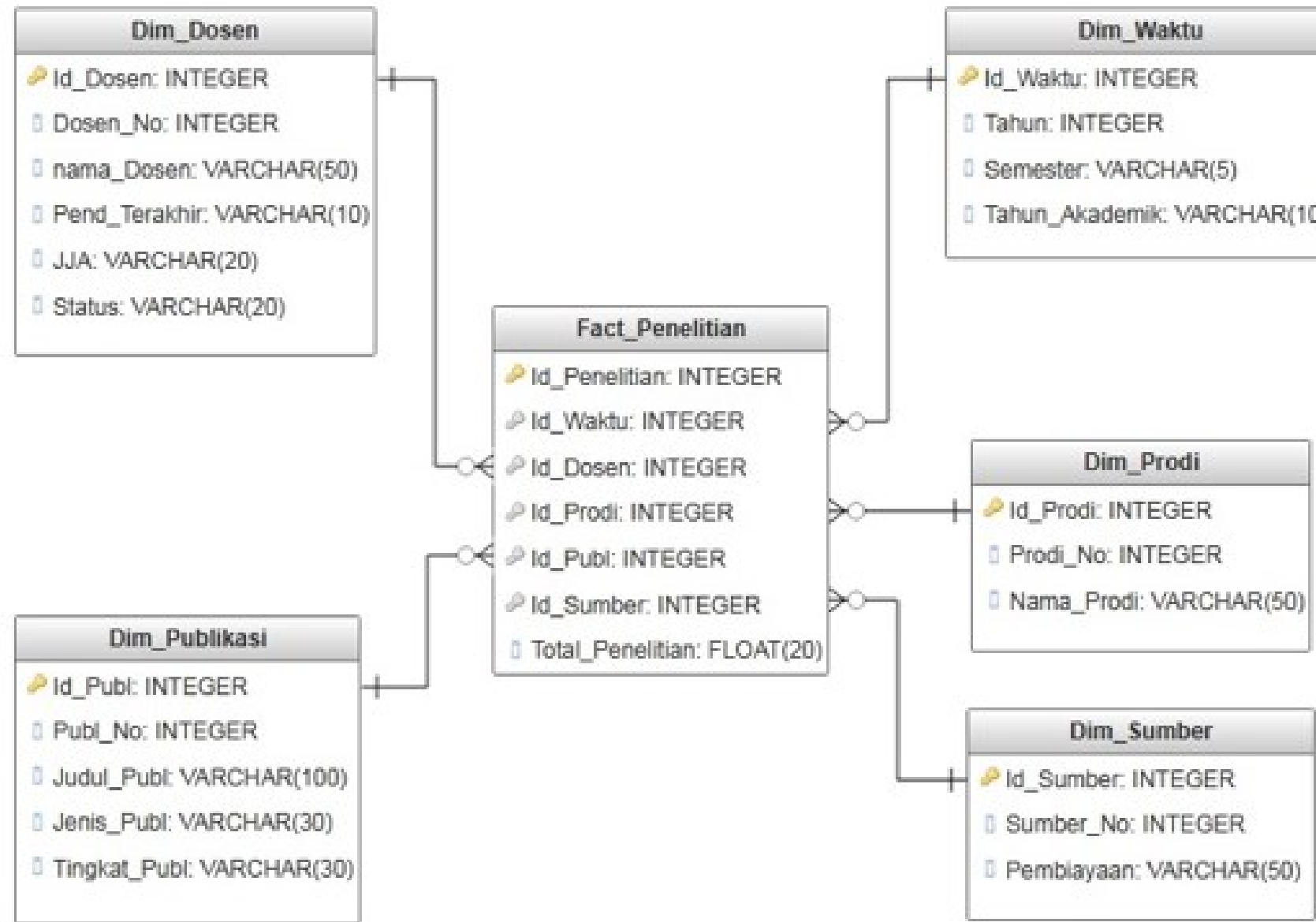
3. Fase Planning

- a. Detailed project requirements.
- b. Definition of the mathematical models needed.
- c. Identification of the data definition of data warehouses and data marts.
- d. Development of a prototype.

ERD db_Penelitian

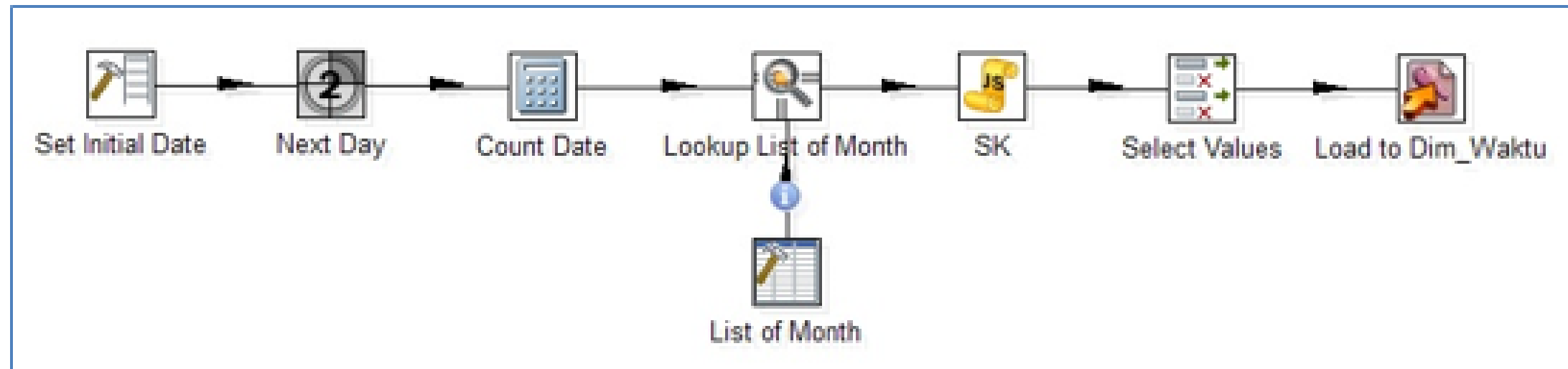


Model Star Schema DWH Penelitian

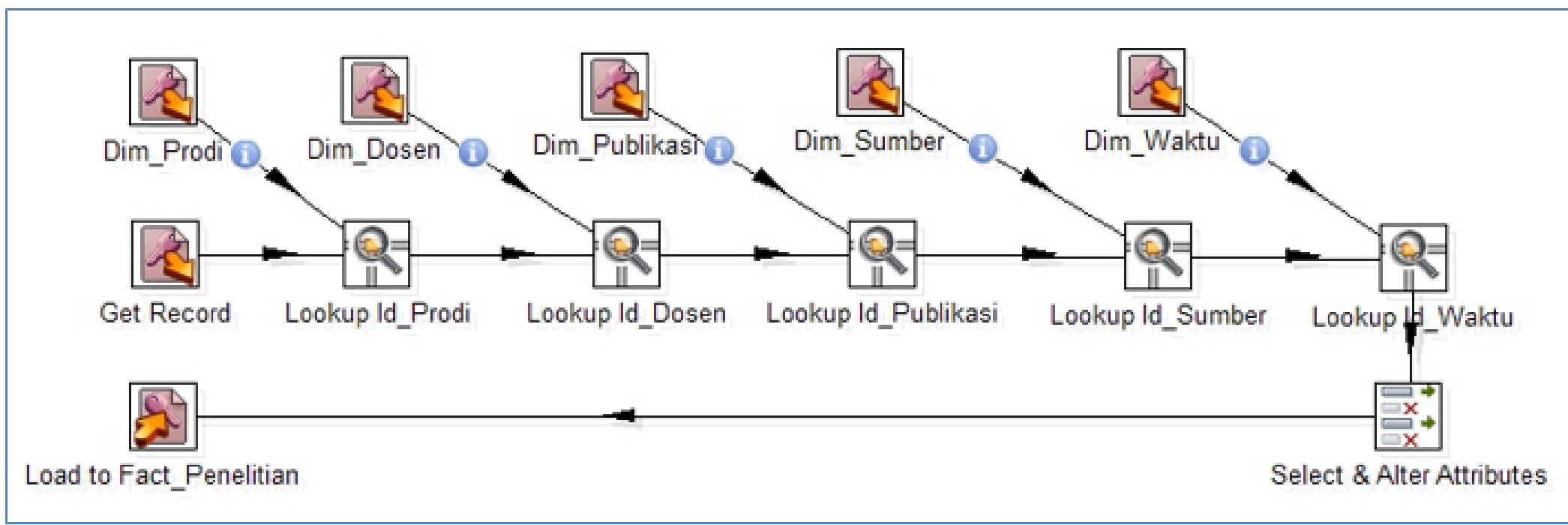


4. Fase Implementation & Control

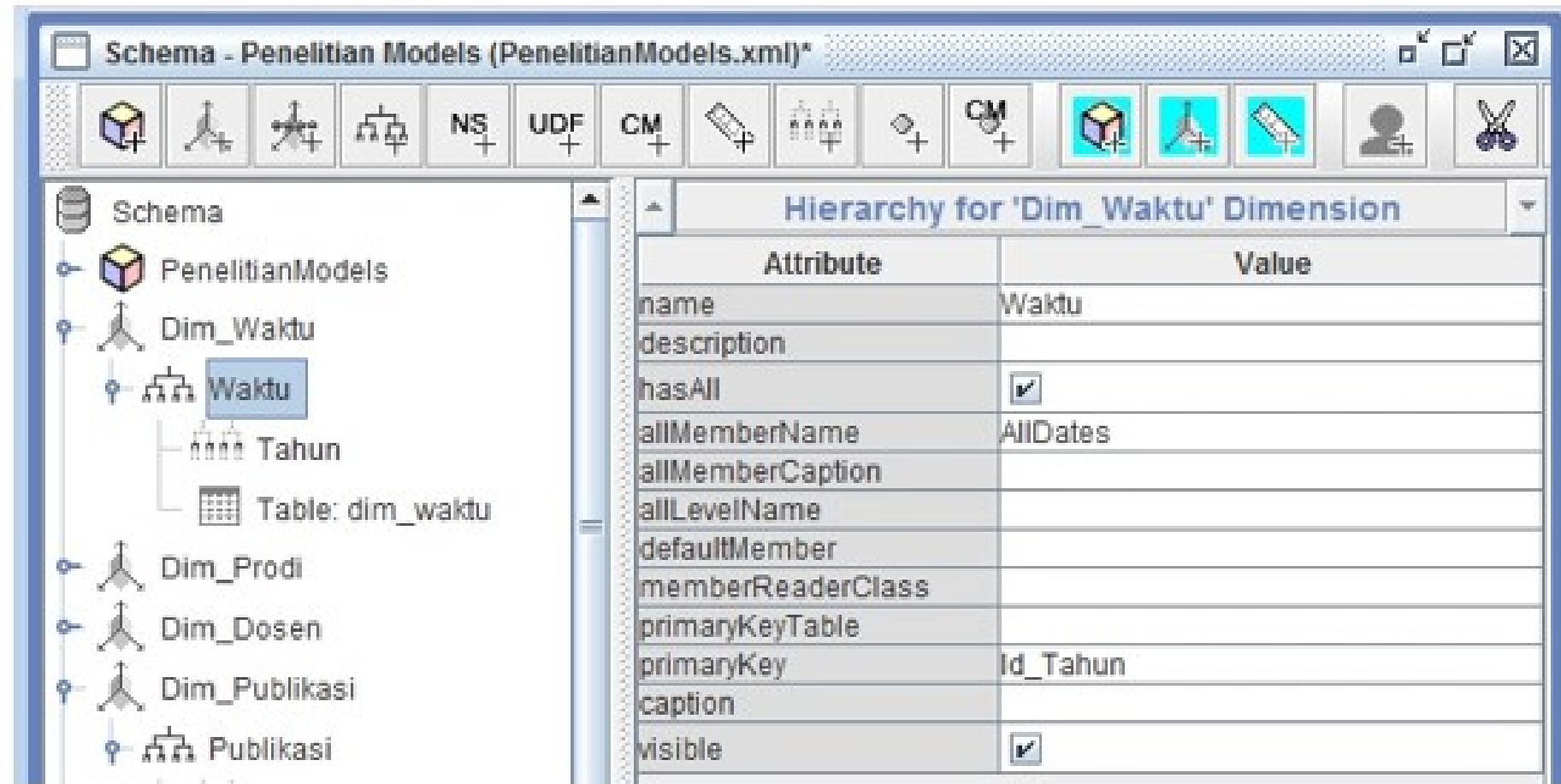
- a. Development of data warehouses and data marts.
- b. Development of metadata.
- c. Development of ETL tools.
- d. Development of applications.
- e. Release and Testing.



Gambar 4.39 – Proses ETL *Dim_Waktu*.
(Sumber: SPs-UP)



Gambar 4.56 – Proses ETL *Fact_Penelitian*.
(Sumber: SPs-UP)



Gambar 4.69 – Attribute Hierarchy Dim_Waktu.
(Sumber: SPs-UP)

Paper Review

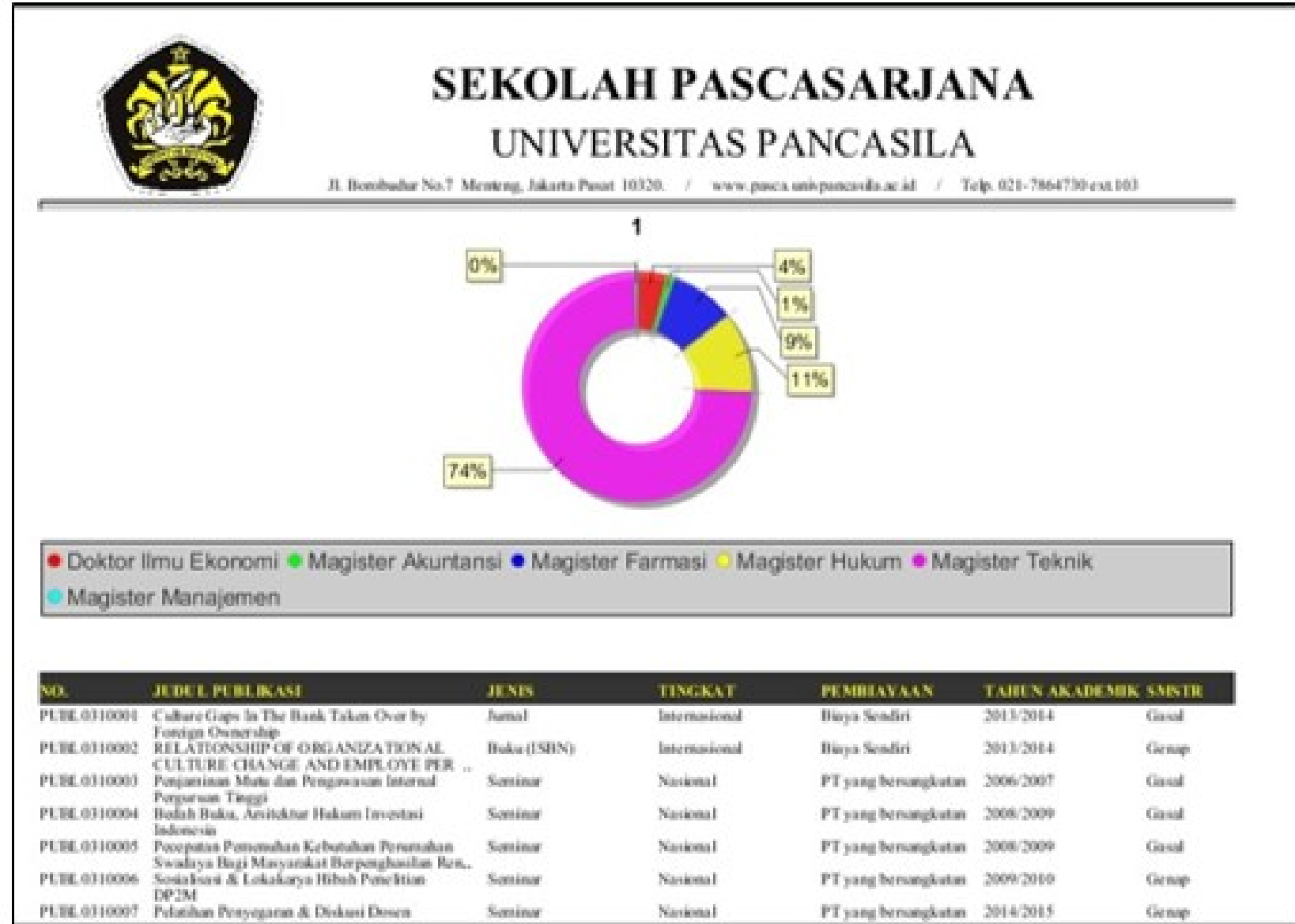
(lanjutan)

Cube PenelitianModels.xjpivot

		Measures
Waktu	Prodi	● Total Penelitian
2012	+ all_Prodi	153
2013	- all_Prodi	219
	Doktor Ilmu Ekonomi	55
	Magister Akuntansi	1
	Magister Farmasi	9
	Magister Hukum	43
	Magister Kenotariatan	
	Magister Manajemen	
	Magister Teknik	111
2014	+ all_Prodi	228
2015	+ all_Prodi	166
2016	+ all_Prodi	128

Gambar 4.100 – OLAP Analysis Dimensi Waktu dan Prodi.
(Sumber: SPs-UP)

Report



Metode Evaluasi

- Yaitu mengevaluasi sistem BI yang telah dibuat.
- Tujuannya :
 - Mengevaluasi apakah tujuan dari penelitian tercapai.
 - Untuk membandingkan sistem yang sedang berjalan (*existing*) terhadap sistem baru (BI) yang telah dibuat.
- Dilakukan dengan cara interview terhadap pengguna.
- Alat pengukuran yang digunakan adalah kuesioner.
- Berdasarkan kuesioner yang diujikan kepada 10 responden, menunjukkan bahwa sistem BI yang dibuat “MEMUASKAN”.

Kuesioner Penelitian.

No.	Pertanyaan	Pilihan Jawaban				
		SM	M	R	KM	TM
Functionality						
1	Bagaimana tanggapan <i>user</i> terhadap penerapan sistem <i>business intelligence</i> mengenai borang akreditasi standar 7 yaitu tentang penelitian ?					
2	Bagaimana tanggapan <i>user</i> terhadap penerapan sistem <i>business intelligence</i> yang menyajikan informasi dalam bentuk dashboard ?					
3	Bagaimana tanggapan <i>user</i> terhadap penerapan sistem <i>business intelligence</i> dalam menyajikan informasi yang secara baik, tepat dan tepat ?					
Useability						
4	Apakah sistem <i>business intelligence</i> yang dibuat dapat mengintegrasikan seluruh data penelitian yang ada pada program studi ke dalam <i>data warehouse</i> terkait dengan borang standar 7 ?					
5	Apakah hasil OLAP dapat dilakukan Roll-Up, Dril-down dan Pivoting, sehingga dapat membantu dalam menganalisis untuk mendukung pengambilan keputusan ?					
6	Bagaimana tanggapan <i>user</i> terhadap penerapan sistem <i>business intelligence</i> untuk mendukung dalam pengambilan keputusan ?					
7	Apakah sistem <i>business intelligence</i> dapat menyelesaikan permasalahan penelitian yang ada di program studi ?					
8	Apakah manfaat dari sistem <i>business intelligence</i> sudah sesuai dengan kebutuhan program studi untuk borang akreditasi standar 7 khususnya tentang data penelitian ?					

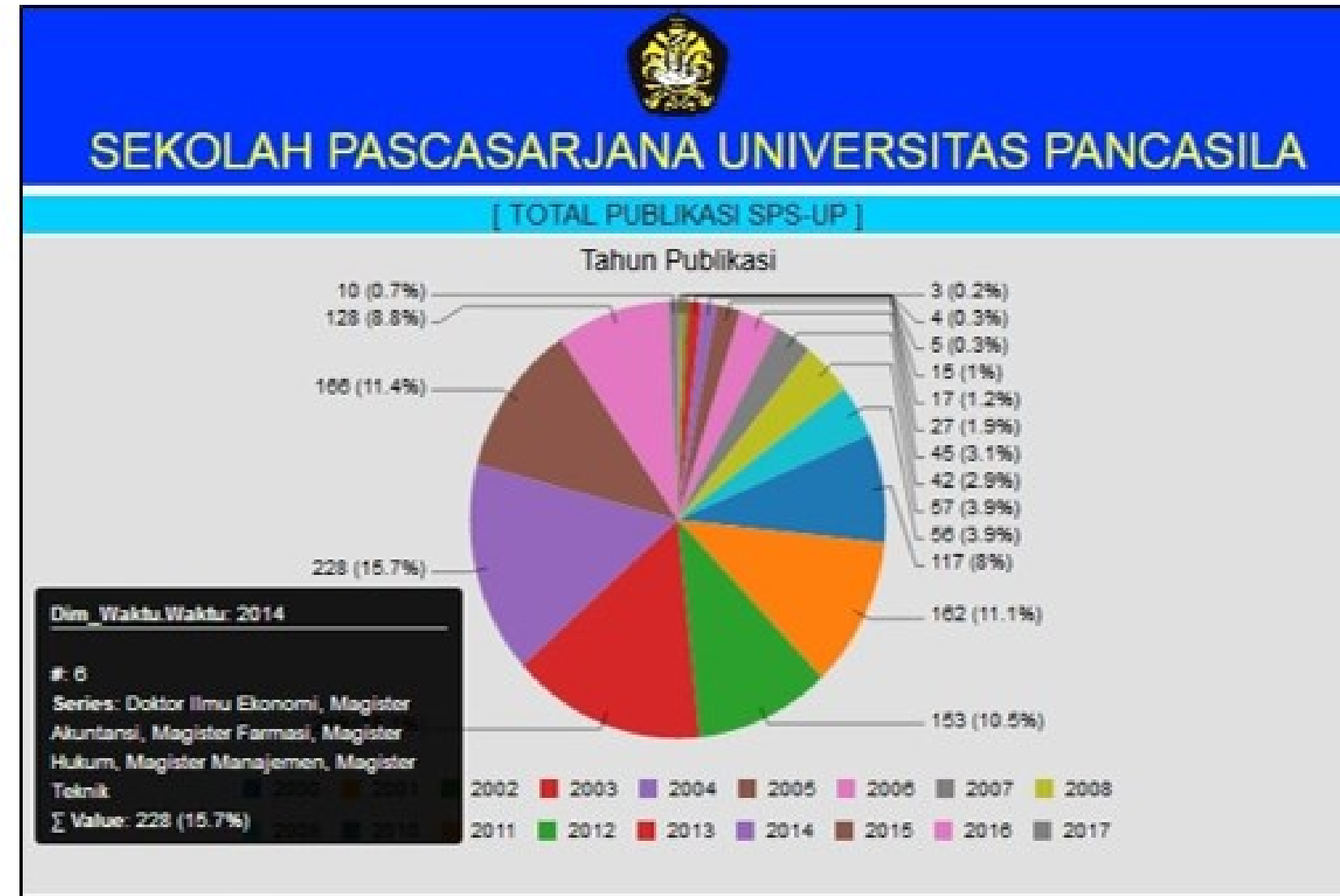
Paper Review

(lanjutan)

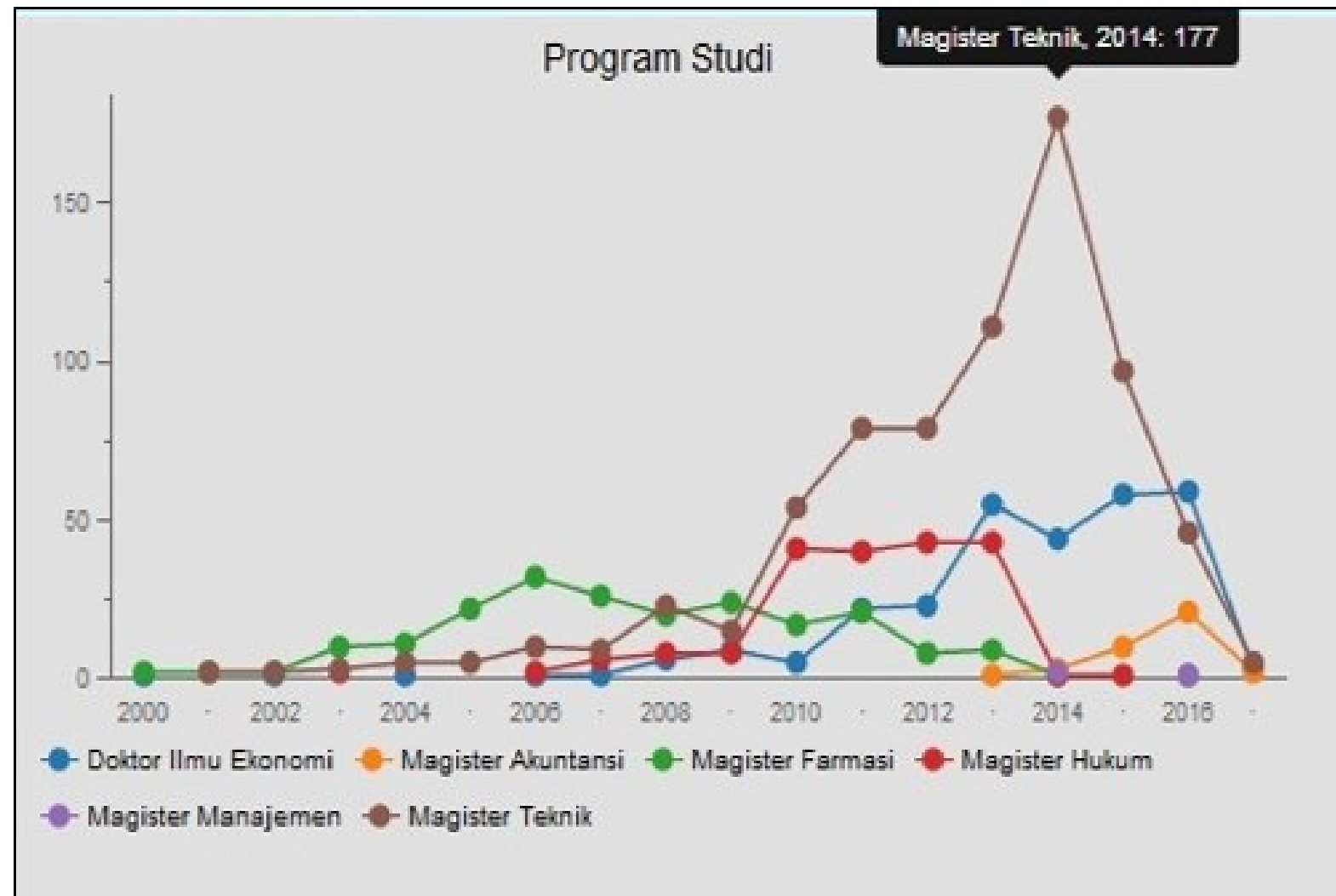
Pilihan jawaban yang tersedia adalah:

- SM** = apabila Anda merasa **Sangat Memuaskan**
- M** = apabila Anda merasa **Memuaskan**
- R** = apabila anda merasa **Ragu-ragu**
- KM** = apabila Anda merasa **Kurang Memuaskan**
- TM** = apabila Anda merasa **Tidak Memuaskan**

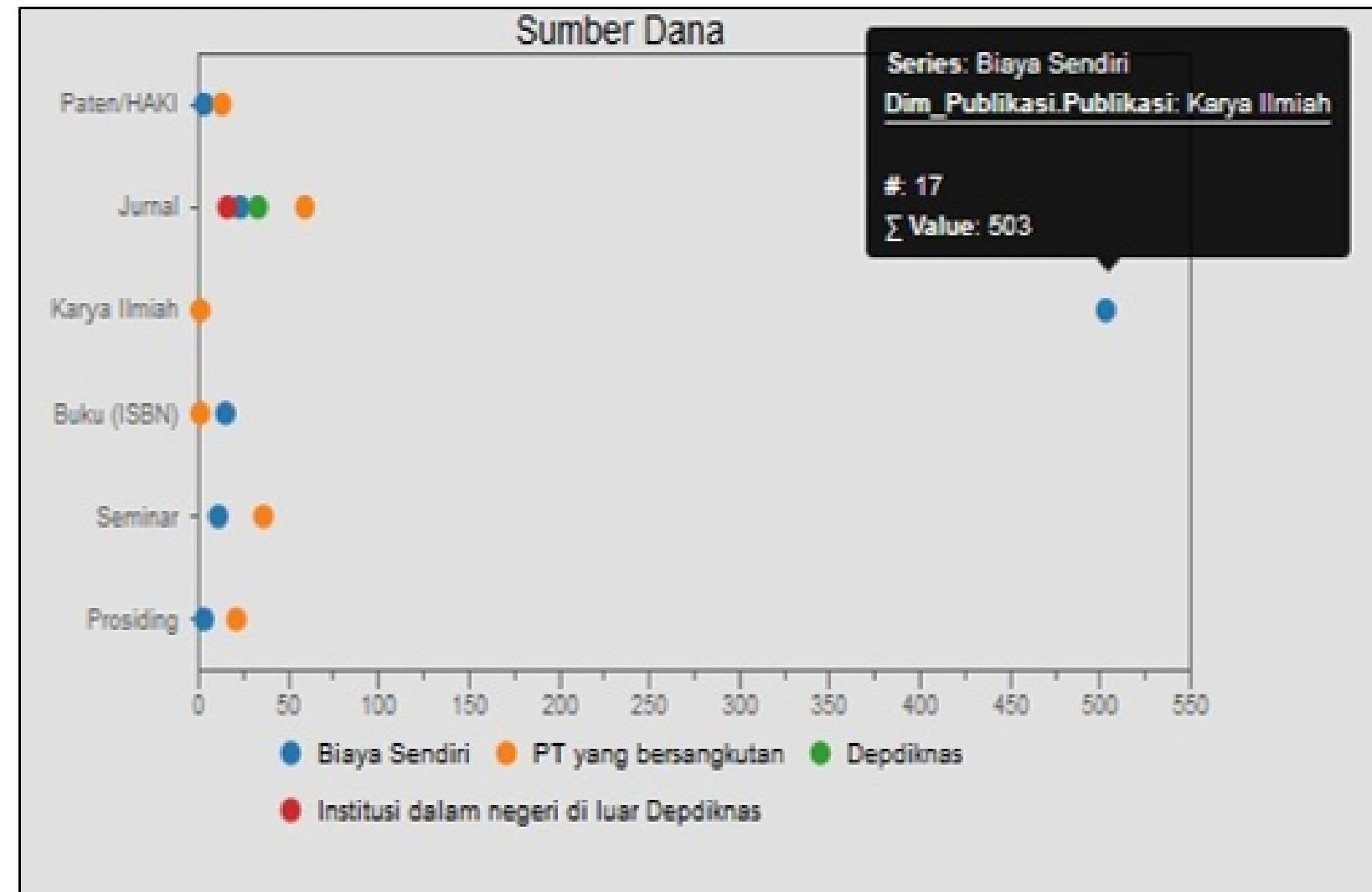
Dashboard Total Publikasi SPS-UP



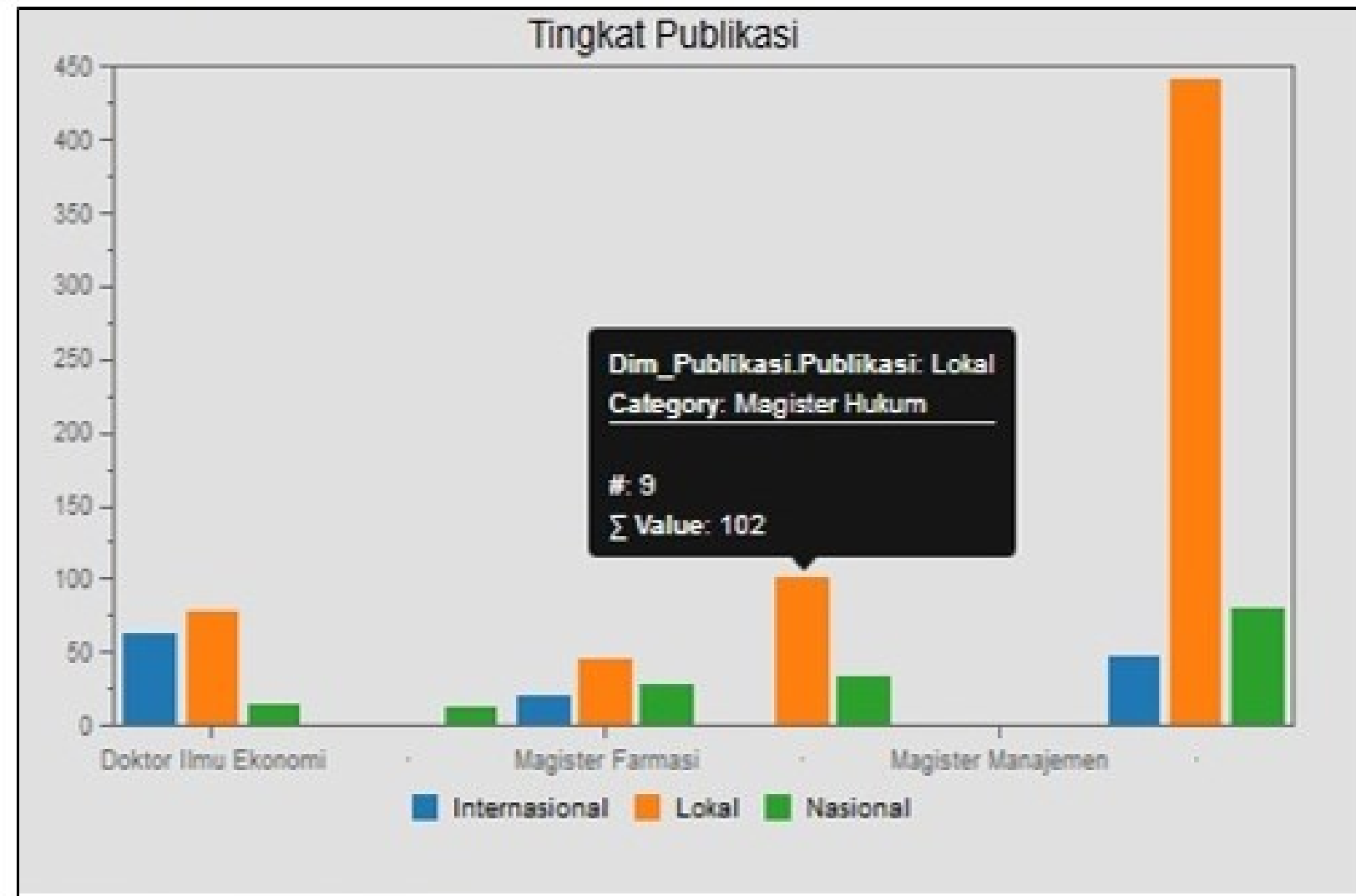
Dashboard Publikasi Berdasarkan Program Studi



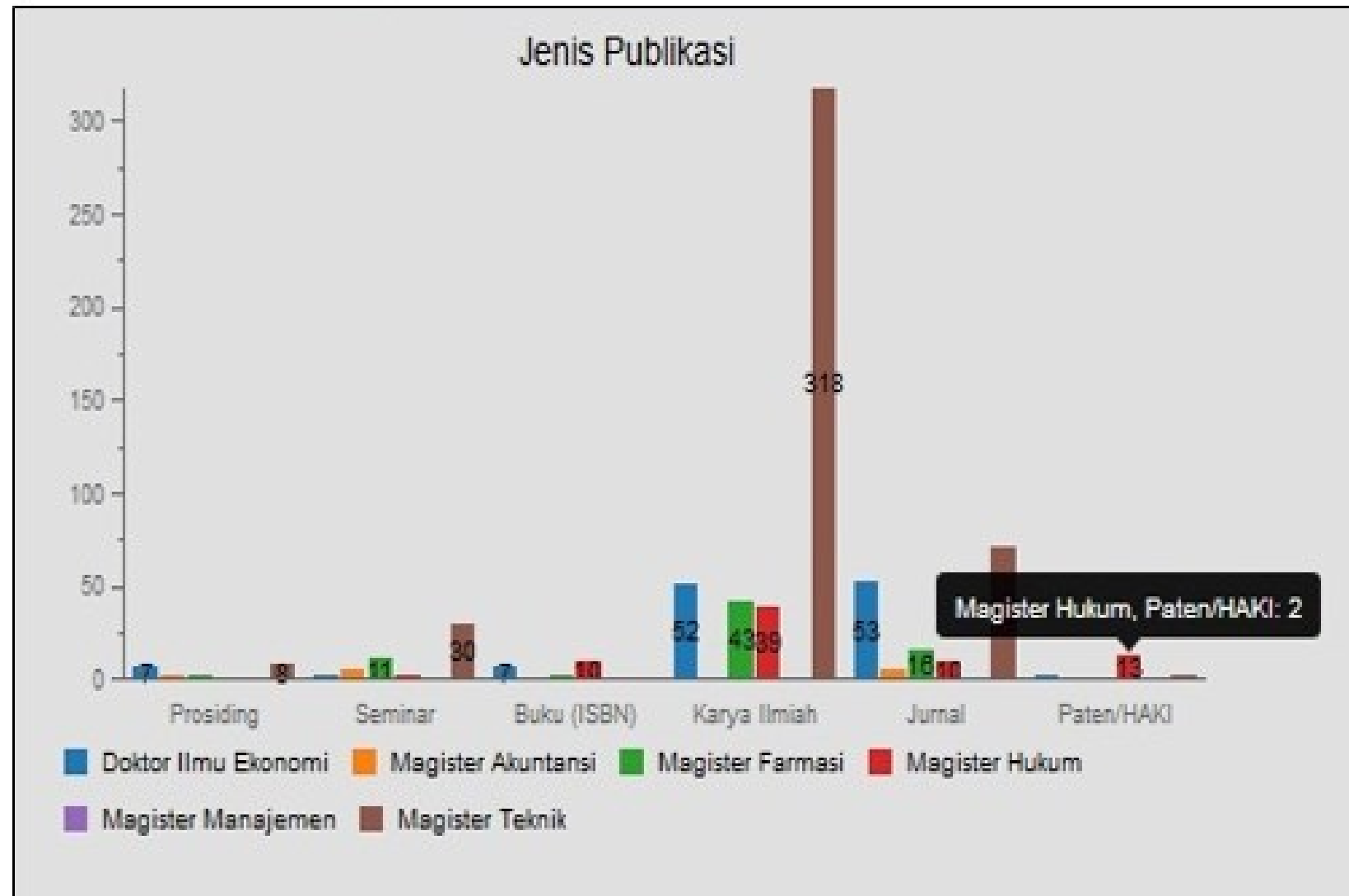
Dashboard Publikasi Berdasarkan Sumber Dana



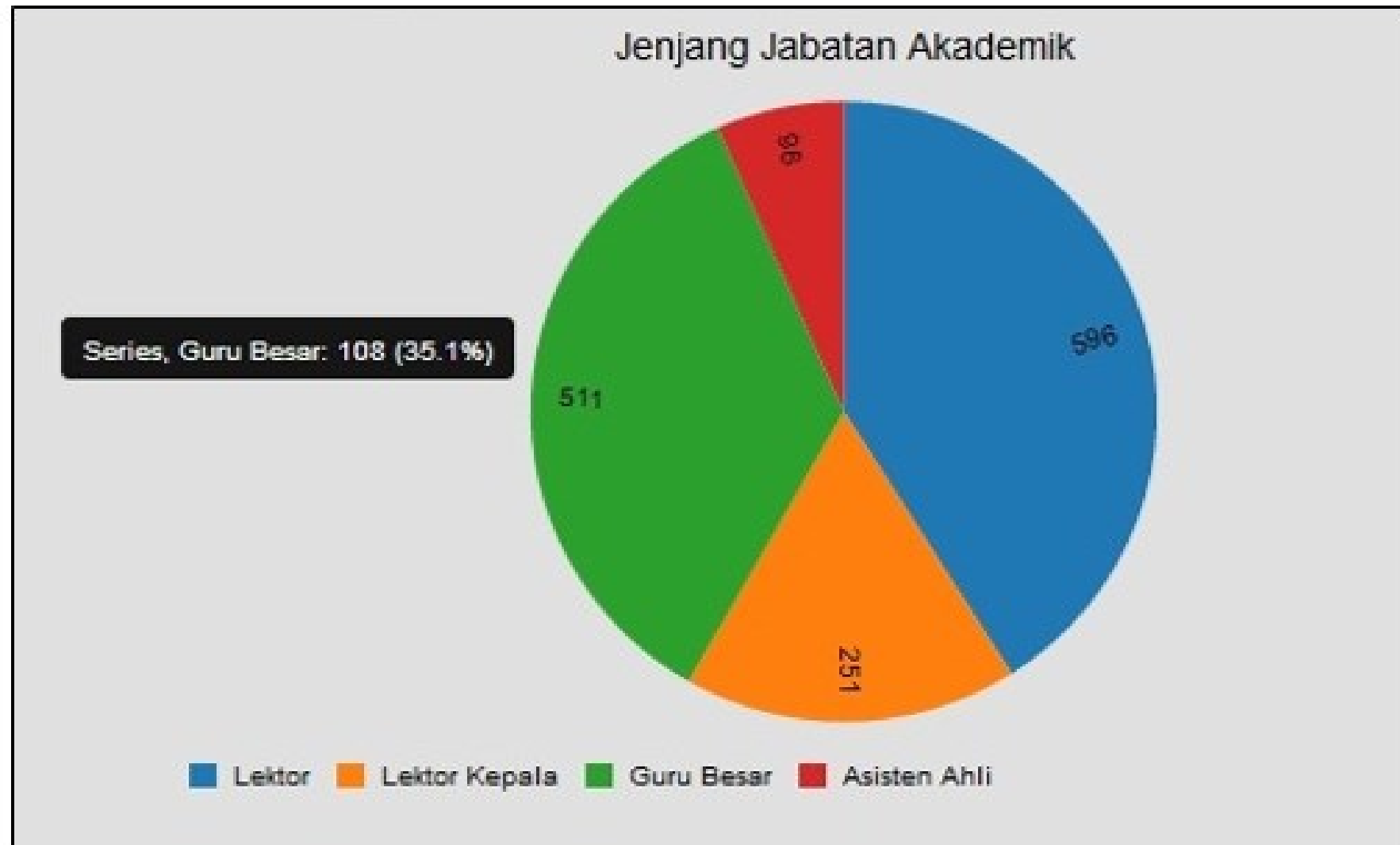
Dashboard Berdasarkan Tingkat Publikasi



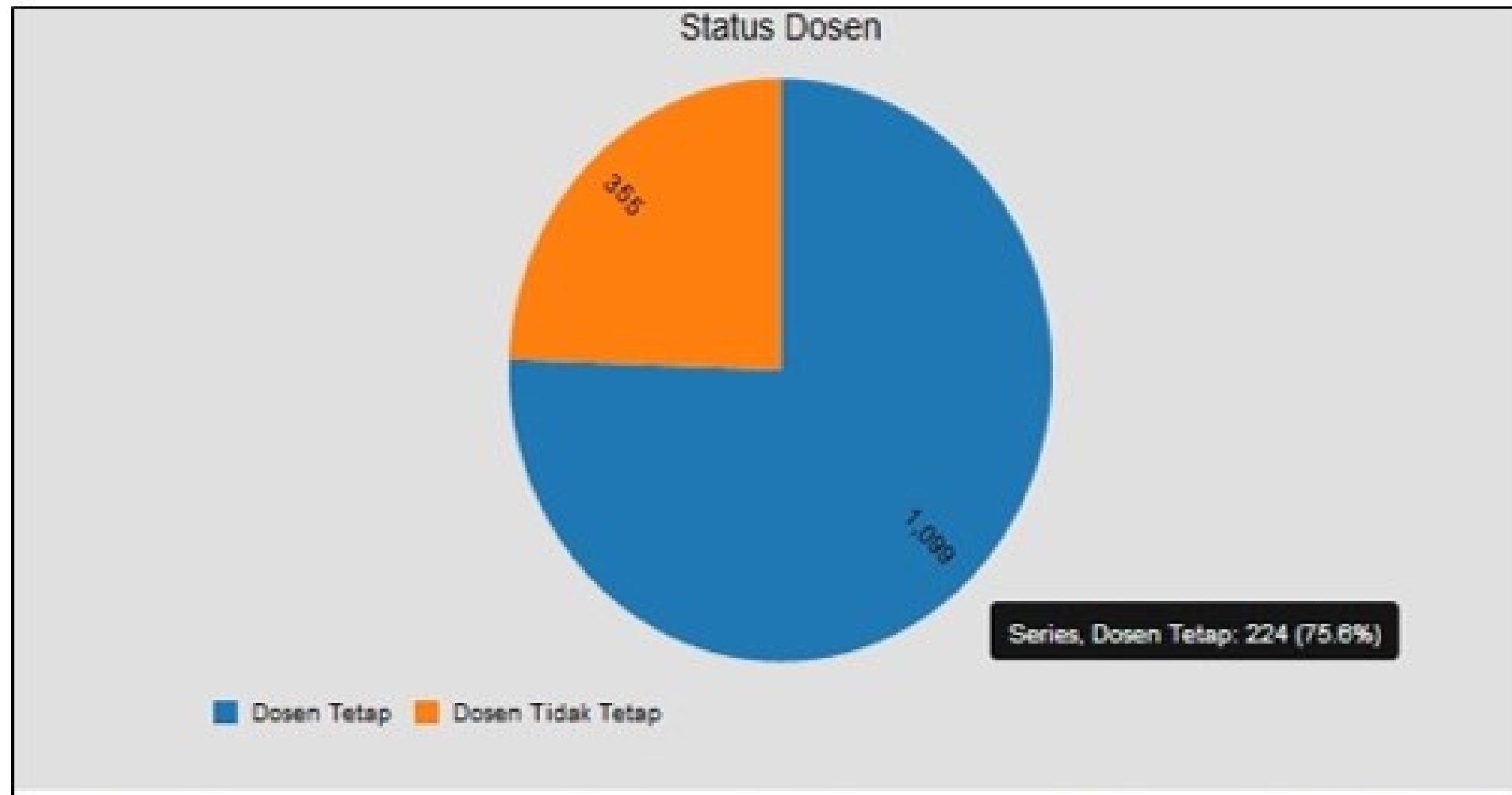
Dashboard Berdasarkan Jenis Publikasi



Dashboard Publikasi Berdasarkan JJA (Jenjang Jabatan Akademik)



Dashboard Publikasi Berdasarkan Status Dosen



Simpulan

Paper Review (lanjutan)

- ✓ Dengan adanya *data warehouse* penelitian yang dirancang dengan menggunakan *PDI (Pentaho Data Integration)*, sangat membantu didalam mengumpulkan data-data penelitian dari yang semula data-data penelitian tersebut belum mempunyai *database* dan hanya sebagai data *OLTP (Online Transactional Processing)* sehingga sekarang dapat dijadikan sebagai data *OLAP (Online Analysis Processing)* dan dapat dianalisis dengan menggunakan *Mondrian OLAP*.
- ✓ Dengan adanya *reporting* penelitian yang dirancang dengan menggunakan *PRD (Pentaho Report Designer)*, sangat membantu didalam membuat laporan penelitian dan dapat disesuaikan dengan kebutuhan, sehingga masalah yang sering terjadi dalam hal kebergantungan dalam perolehan laporan diharapkan tidak terjadi kembali.
- ✓ Dengan adanya *dashboard* penelitian yang dirancang dengan menggunakan *Pentaho CDE (Community Dashboard Editor)*, sangat membantu pimpinan didalam menganalisis data untuk mendukung pengambilan keputusan, dan juga dapat dijadikan untuk mempelajari tren penelitian yang ada pada SPs-UP.

Saran

- Sistem *business intelligence* yang dibuat tidak hanya untuk borang standar 7 saja yang saat ini dibuat hanya yang terkait dengan data-data penelitian, melainkan juga dapat dibuat berdasarkan data lainnya seperti akademik, lulusan, sarana & prasarana, mahasiswa, dan lain-lain.
- Pada pengembangan selanjutnya dapat dibuat *data mining* yang terintegrasi ke dalam sistem *business intelligence* yang dapat berguna untuk memprediksi serta mengetahui pola karakteristik dari penelitian dan juga publikasi dari dosen tersebut.

- Ian H Witten, Eibe Frank, Mark A Hall. (2016). *Data Mining - Practical Machine Learning Tools and Techniques*, 4th Edition. Morgan-Kaufman.
- Hofmann, Markus., Klinkenberg, Ralf. (2016). *Data Mining Use Cases and Business Analytics Application*. CRC Press.
- Crisholm, Andrew. (2013). *Exploring Data With Rapidminer*. Packt Publishing.
- Mehmed Kantardzic. (2020). *Data Mining - Concepts, Models, Methods, and Algorithms*, 3rd edition. Wiley - IEEE Press.
- rapidminer.com, 2021,. Rapidminer Manual and Documentation, <https://docs.rapidminer.com/latest/studio/operators/>.

The Power Of Tilawah **MENGHILANGKAN BAU MULUT**

Terima Kasih



عَنْ أَبِي هُرَيْرَةَ رَضِيَ اللَّهُ عَنْهُ قَالَ رَسُولُ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ: ...
وَالَّذِي نَفْسُ مُحَمَّدٍ بِيَدِهِ، لَخُلُوفٌ فِيمَ الصَّائِمِ أَطْيَبُ عِنْدَ اللَّهِ مِنْ رِيحِ الْمِسْكِ
(حديث صحيح أخرجه البخاري و مسلم)

Dari Abu Hurairah RA, Rasulullah SAW bersabda: "...Dan demi jiwa Muhammad pada kekuasaan-Nya, sungguh bau mulut orang yang berpuasa lebih wangi di sisi Allah daripada aroma minyak kasturi." (Hadis Shahih diriwayatkan oleh Bukhari dan Muslim)