

Pertemuan 3

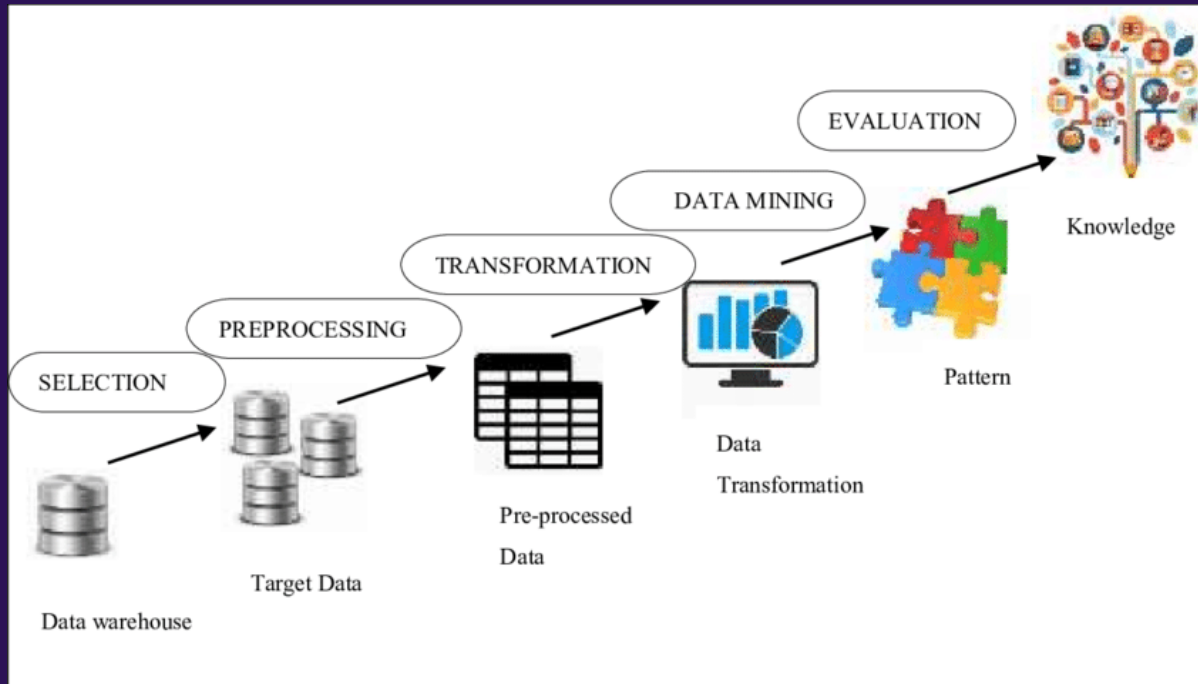
Knowledge Discovery in Database

Data Science

Pingky Dezar Zulkarnain, SE., M.Sc., Ph.D

Putri Cholidhazia, ST., Mkom

Proses Data Mining



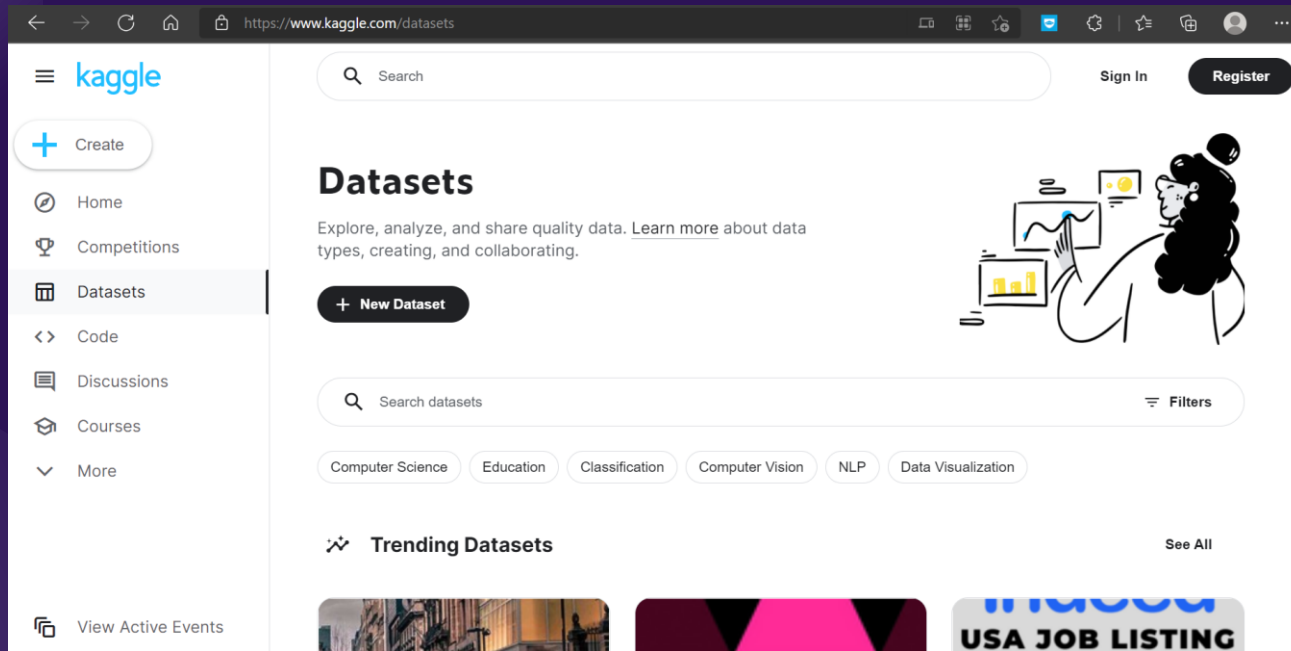
Data

"Data adalah fakta-fakta yang menggambarkan suatu kejadian yang sebenarnya pada waktu tertentu (Lubis, 2016)"

Informasi

Informasi adalah sekumpulan data atau fakta yang telah diproses dan diolah sedemikian rupa sehingga menghasilkan sesuatu yang bisa dipahami dan memberikan manfaat bagi penerimanya.

Akses Data



The screenshot shows the Kaggle Datasets homepage. The browser address bar displays <https://www.kaggle.com/datasets>. The page features a left sidebar with navigation links: Create, Home, Competitions, Datasets (selected), Code, Discussions, Courses, and More. The main content area is titled "Datasets" and includes a search bar, a "Sign In" button, and a "Register" button. Below the title, there is a description: "Explore, analyze, and share quality data. [Learn more](#) about data types, creating, and collaborating." A "+ New Dataset" button is also present. A search bar for datasets and a "Filters" button are located below. A row of category tags includes Computer Science, Education, Classification, Computer Vision, NLP, and Data Visualization. The "Trending Datasets" section is visible at the bottom, with a "See All" link. The first dataset thumbnail is titled "USA JOB LISTING".

Akses Data

The screenshot displays the official website of the Badan Pusat Statistik (BPS) of Indonesia. The browser address bar shows the URL <https://www.bps.go.id>. The website header includes the BPS logo and name in Indonesian, with language options for 'Indonesia' and 'English'. A search bar is present with a 'Cari' button. Below the header is a navigation menu with links: Beranda, Tentang Kami, Berita, Senarai Rencana Terbit, Publikasi, Berita Resmi Statistik, Layanan, and PPID. The main content area is divided into several sections. On the left, there are vertical filters for 'Sosial dan Kependudukan' (with sub-categories: Gender, Geografi, Iklim) and 'Ekonomi dan Perdagangan' (with sub-categories: Ekspor-Impor, Energi, Harga Eceran). Below these are filters for 'Pertanian dan Pertambangan'. The central part of the page features a 'Cari Data...' search bar and a 'Cari' button. Below this is a banner for 'TUJUAN PEMBANGUNAN BERKELANJUTAN | SUSTAINABLE DEVELOPMENT GOALS'. The 'BERITA RESMI STATISTIK' section lists recent news items with dates and download links for audio files. On the right side, there is a 'DATA SENSUS' section with a 'Website BPS Provinsi' dropdown menu. Below this are links for 'Informasi Seputar SP2020', 'Sensus Penduduk 2020', 'Survei C-VID-19 BPS', and 'Publikasi C-VID-19 BPS'. At the bottom right, there are links for 'DATA PERENCANAAN PEMBANGUNAN', 'Statistical Geographic Information System', and 'Satu Data'.

Indonesia | English

Cari

Manual | Tautan | Peta Situs | S&K

Beranda | Tentang Kami | Berita | Senarai Rencana Terbit | Publikasi | Berita Resmi Statistik | Layanan | PPID

Beranda »

Sosial dan Kependudukan

Gender

Geografi

Iklim

Selengkapnya...

Ekonomi dan Perdagangan

Ekspor-Impor

Energi

Harga Eceran

Selengkapnya...

Pertanian dan Pertambangan

Cari Data..

Cari

TUJUAN PEMBANGUNAN BERKELANJUTAN | SUSTAINABLE DEVELOPMENT GOALS

INFORMASI TERBARU

Berita Resmi Statistik

15 Februari 2022 (Download Bahan Tayang)

- Ekspor Januari 2022 mencapai US\$19,16 miliar dan Impor Januari 2022 senilai US\$18,23 miliar
- Upah Nominal Harian Buruh Tani Nasional Januari 2022 Naik Sebesar 0,72 Persen

07 Februari 2022 (Download Bahan Tayang)

- Ekonomi Indonesia Triwulan IV 2021 Tumbuh 5,02 Persen (y-on-y)

02 Februari 2022 (Download Bahan Tayang)

DATA SENSUS

Website BPS Provinsi

Pilih Provinsi

Informasi Seputar SP2020

Sensus Penduduk 2020

Survei C-VID-19 BPS

Publikasi C-VID-19 BPS

DATA PERENCANAAN PEMBANGUNAN


Statistical Geographic Information System

Satu Data

Akses Data

← → ↻ 🏠 🔒 https://archive.ics.uci.edu/ml/datasets.php ☆ 📧 ⚙️ | ☆ 📁 👤 ⋮


UCI



Machine Learning Repository

Center for Machine Learning and Intelligent Systems





[About](#) [Citation Policy](#) [Donate a Data Set](#) [Contact](#)

☐ Repository ☐ Web 

[View ALL Data Sets](#)

Check out the [beta version](#) of the new UCI Machine Learning Repository we are currently testing! [Contact us](#) if you have any issues, questions, or concerns. [Click here to try out the new site.](#) ✕

Browse Through: **622 Data Sets** Table View [List View](#)

Default Task	Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Classification (466) Regression (151) Clustering (121) Other (56)	 Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Attribute Type Categorical (38) Numerical (422) Mixed (55)	 Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Data Type Multivariate (480) Univariate (30) Sequential (59) Time-Series (126) Text (69) Domain-Theory (23) Other (21)	 Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
	 Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998

Contoh Data

Data Tabular

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
5	6	5.4	3.9	1.7	0.4	Iris-setosa
6	7	4.6	3.4	1.4	0.3	Iris-setosa
7	8	5.0	3.4	1.5	0.2	Iris-setosa
8	9	4.4	2.9	1.4	0.2	Iris-setosa
9	10	4.9	3.1	1.5	0.1	Iris-setosa

Data Text



Clementine Ford 
@clementine_ford

Replying to @clementine_ford

Honestly, the corona virus isn't killing men fast enough.

4:33 PM · May 23, 2020 · [Twitter Web App](#)

Data Image



Data Suara dan Video

Bentuk Data

Structured data

Databases

Semi-structured data

XML / JSON data

Email

Web pages

Unstructured data

Audio

Video

Image data

Natural language

Documents

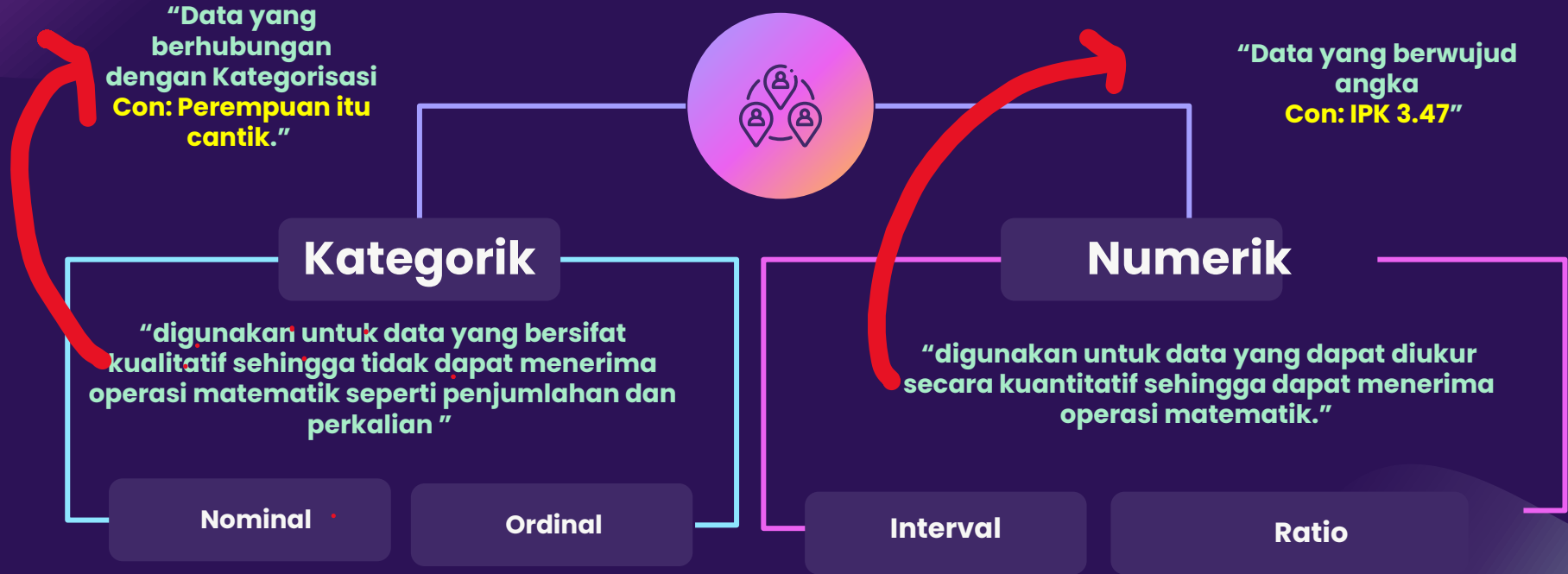
Contoh Struktur Data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
  </Student>
  ....
</University>
```

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

PENGELOMPOKAN DATA



Data Kategorik

Nominal

“Data yang berbentuk kelompok / kategori, dimana masing-masing kategori memiliki derajat yang sama ”

Contoh: Suku Jawa, Minang, Madura, Sunda

Ordinal

“Data yang berbentuk kelompok / kategori, dimana masing-masing kategori memiliki derajat yang berbeda ”

Contoh: hasil survey: Setuju, netral, tidak setuju

Data Numerik

Interval

“Data yang diukur dengan jarak diantara dua titik pada skala yang sudah diketahui”

Contoh: Suhu udara dalam Celsius antara 0 derajat hingga 100 derajat

Ratio

“Data yang diukur suatu proporsi dan mempunyai jarak yang sama”

Contoh: Berat bayi dimana bayi A beratnya adalah 3, B adalah 2, dan C adalah 1, jika dilihat menggunakan skala rasio berat badan bayi A tiga kali lipat dari berat badan bayi C, berat badan bayi B dua kalilipat dari C

JENIS DATA

	Data Kategorik		Data Numerik	
	Nominal	Ordinal	Interval	Ratio
Perbedaan	✓	✓	✓	✓
Urutan		✓	✓	✓
Jarak yang sama			✓	✓
Nol Mutlak				✓

Contoh

Contoh Data	Jenis Data
Tingkatan Pendidikan: TK, SD, SMP, SMA, Diploma & Sarjana	Data Ordinal
Jumah produksi jagung	Data Ratio
Pekerjaan : Guru, Programmer, Petani	Data Nominal
Rata – rata tinggi badan	Data Interval

Pre Processing

Data Cleaning

Data dapat memiliki banyak bagian yang tidak relevan dan hilang. Untuk menangani data noise.

Delete Data Duplikat

Memperbaiki kesalahan pada data.

Memeriksa data yang
inkonsisten

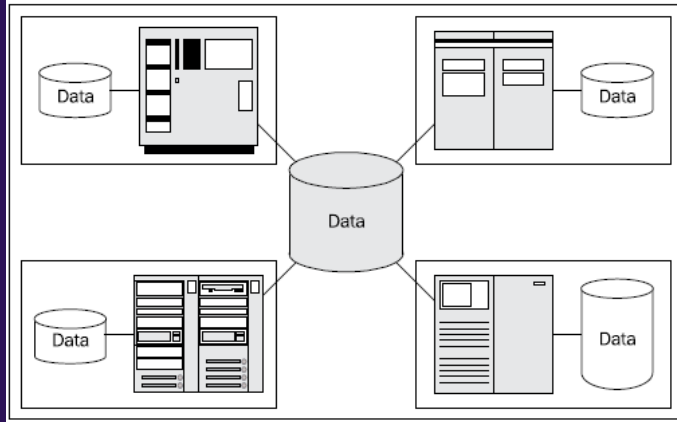
Delete / Isi Data Hillang (Null)

Delete Data Outlier

Pre Processing

Data Integrasi

Mengkombinasikan data dari banyak sumber kedalam suatu simpanan terpadu



Data Transformasi

Data Transformasi

Langkah ini dilakukan untuk mengubah data dalam bentuk yang sesuai



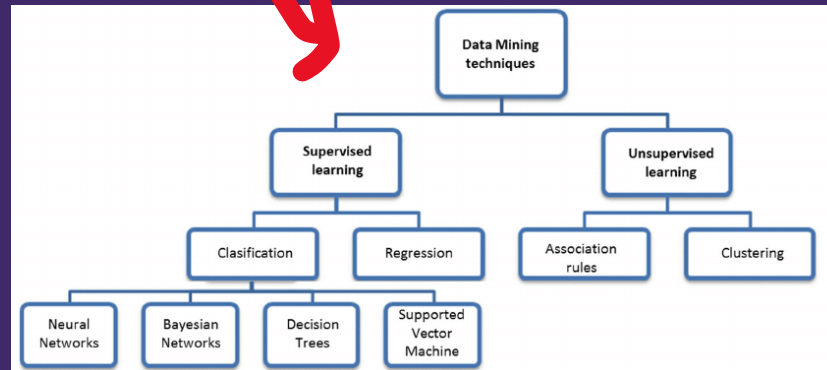
Normalisasi

Merubah skala data

Data Mining

Data Mining

Proses Data mining yaitu proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu.



Evaluation

Evaluation

- ❑ Penerjemahan pola-pola yang dihasilkan dari data mining.
- ❑ Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan.
- ❑ Tahap ini merupakan bagian dari proses KDD yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

Contoh Penerapan KDD

Data

Loan ID	Nama	Age	JK	Gaji	Pendidikan	Status	Jumlah Pinjaman	Riwayat Credit	Status Pinjaman
15678	Bunga	45	P	5 – 10 Jt	SMA	Sudah Menikah	50.000.000	Tidak	Ya
19878	Irfan	25	L	10 – 15 Jt	S1	Lajang	60.000.000	Ya	Tidak
17687	Fajar	30	L	5 – 10 Jt	S2	Lajang	45.000.000	Null	Tidak
29389	Winda	27	P	5 – 10 Jt	S1	Menikah	70.000.000	ya	Ya
12345	Bayu	19	L	Null	TK	Lajang	35.000.000	tidak	Ya
29389	Winda	27	P	10 – 15 Jt	S1	Menikah	70.000.000	ya	Ya

Contoh Penerapan KDD

Pre-processing

Loan ID	Nama	Age	JK	Gaji	Pendidikan	Status	Jumlah Pinjaman	Riwayat Credit	Status Pinjaman
15678	Bunga	45	P	5 – 10 Jt	SMA	Sudah Menikah	50.000.000	Tidak	Ya
19878	Irfan	25	L	10 – 15 Jt	S1	Lajang	60.000.000	Ya	Tidak
17687	Fajar	30	L	5 – 10 Jt	S2	Lajang	45.000.000	Null	Tidak
29389	Winda	27	P	5 – 10 Jt	S1	Menikah	70.000.000	ya	Ya
12345	Bayu	19	L	Null	TK	Lajang	Null	tidak	Ya
29389	Winda	27	P	10 – 15 Jt	S1	Menikah	70.000.000	ya	Ya

Data Cleaning

1. Hapus Atribut yang tidak digunakan

Age	JK	Gaji	Pendidikan	Status	Jumlah Pinjaman	Riwayat Credit	Status Pinjaman
45	P	5 – 10 Jt	SMA	Sudah Menikah	50.000.000	Tidak	Ya
25	L	10 – 15 Jt	S1	Lajang	60.000.000	Ya	Tidak
30	L	5 – 10 Jt	S2	Lajang	45.000.000	Null	Tidak
27	P	5 – 10 Jt	S1	Menikah	70.000.000	ya	Ya
19	L	Null	TK	Lajang	Null	tidak	Ya
27	P	10 – 15 Jt	S1	Menikah	70.000.000	ya	Ya

Data Cleaning

1. Update data tidak konsisten

Age	JK	Gaji	Pendidikan	Status	Jumlah Pinjaman	Riwayat Credit	Status Pinjaman
45	P	5 – 10 Jt	SMA	Menikah	50.000.000	Tidak	Ya
25	L	10 – 15 Jt	S1	Lajang	60.000.000	Ya	Tidak
30	L	5 – 10 Jt	S2	Lajang	45.000.000	Null	Tidak
27	P	5 – 10 Jt	S1	Menikah	70.000.000	ya	Ya
19	L	Null	TK	Lajang	Null	tidak	Ya
27	P	10 – 15 Jt	S1	Menikah	70.000.000	ya	Ya

Data Cleaning

1. Update data tidak konsisten

Age	JK	Gaji	Pendidikan	Status	Jumlah Pinjaman	Riwayat Credit	Status Pinjaman
45	P	5 – 10 Jt	SMA	Menikah	50.000.000	Tidak	Ya
25	L	10 – 15 Jt	S1	Lajang	60.000.000	Ya	Tidak
30	L	5 – 10 Jt	S2	Lajang	45.000.000	Null	Tidak
27	P	5 – 10 Jt	S1	Menikah	70.000.000	ya	Ya
19	L	Null	TK	Lajang	Null	tidak	Ya
27	P	10 – 15 Jt	S1	Menikah	70.000.000	ya	Ya

Data Cleaning

1. delete data Outlier / Agregasi Data

Age	JK	Gaji	Pendidikan	Status	Jumlah Pinjaman	Riwayat Credit	Status Pinjaman
45	P	5 – 10 Jt	SMA	Menikah	50.000.000	Tidak	Ya
25	L	10 – 15 Jt	S1	Lajang	60.000.000	Ya	Tidak
30	L	5 – 10 Jt	S2	Lajang	45.000.000	Null	Tidak
27	P	5 – 10 Jt	S1	Menikah	70.000.000	ya	Ya
19	L	Null	→ S1 modus	Lajang	Null	tidak	Ya
27	P	10 – 15 Jt	S1	Menikah	70.000.000	ya	Ya

Data Cleaning

1. delete Data Duplikat

Age	JK	Gaji	Pendidikan	Status	Jumlah Pinjaman	Riwayat Credit	Status Pinjaman
45	P	5 – 10 Jt	SMA	Menikah	50.000.000	Tidak	Ya
25	L	10 – 15 Jt	S1	Lajang	60.000.000	Ya	Tidak
30	L	5 – 10 Jt	S2	Lajang	45.000.000	Null	Tidak
27	P	5 – 10 Jt	S1	Menikah	70.000.000	ya	Ya
19	L	Null	S1	Lajang	Null	tidak	Ya

Data Cleaning

1. Delete data null / Update dengan Agregasi Data

Age	JK	Gaji	Pendidikan	Status	Jumlah Pinjaman	Riwayat Credit	Status Pinjaman
45	P	5 – 10 Jt	SMA	Menikah	50.000.000	Tidak	Ya
25	L	10 – 15 Jt	S1	Lajang	60.000.000	Ya	Tidak
30	L	5 – 10 Jt	S2	Lajang	45.000.000	⇒ Null modus	Tidak
27	P	5 – 10 Jt	S1	Menikah	70.000.000	ya	Ya
19	L	⇒ Null modus	S1	Lajang	⇒ Null mean	tidak	Ya

Data Cleaning

Gaji :

- > Level 1 = 5 – 10 Jt
- > Level 2 = 10 – 15 Jt

Jumlah Pinjaman:

$50.000.000 + 60.000.000 + 45.000.000 + 70.000.000 = 56.250.000$

Riwayat Kredit:

- > Ya muncul 2 kali
- > Tidak muncul 2 kali

Kesimpulan bisa input ya / tidak

Data Cleaning

Hasil

Age	JK	Gaji	Pendidikan	Status	Jumlah Pinjaman	Riwayat Credit	Status Pinjaman
45	P	Lv1	SMA	Menikah	50.000.000	Tidak	Ya
25	L	Lv2	S1	Lajang	60.000.000	Ya	Tidak
30	L	Lv1	S2	Lajang	45.000.000	Ya	Tidak
27	P	Lv1	S1	Menikah	70.000.000	ya	Ya
19	L	Lv1	S1	Lajang	56.250.000	tidak	Ya

Data Transformasi

Age	JK	Gaji	Pendidikan	Status	Jumlah Pinjaman	Riwayat Credit	Status Pinjaman
45	P	Lv1	SMA	Menikah	50.000.000	Tidak	Ya
25	L	Lv2	S1	Lajang	60.000.000	Ya	Tidak
30	L	Lv1	S2	Lajang	45.000.000	Ya	Tidak
27	P	Lv1	S1	Menikah	70.000.000	ya	Ya
19	L	Lv1	S1	Lajang	56.250.000	tidak	Ya

JK:

1 = L
2 = P

Gaji :

1 = Lv1
2 = Lv2

Pendidikan:

1 = SMA
2 = S1
3 = S2

Riwayat Credit:

1 = Tidak
2 = ya

Status Pinjaman:

1 = Tidak
2 = ya

Data Transformasi

Age	JK	Gaji	Pendidikan	Status	Jumlah Pinjaman	Riwayat Credit	Status Pinjaman
45	2	1	1	1	50.000.000	1	2
25	1	2	1	2	60.000.000	2	1
30	1	1	2	2	45.000.000	2	1
27	2	1	1	1	70.000.000	2	2
19	1	1	1	2	56.250.000	1	2

Data Mining

Age	JK	Gaji	Pendidikan	Status	Jumlah Pinjaman	Riwayat Credit	Status Pinjaman
45	2	1	1	1	50.000.000	1	2
25	1	2	1	2	60.000.000	2	1
30	1	1	2	2	45.000.000	2	1
27	2	1	1	1	70.000.000	2	2
19	1	1	1	2	56.250.000	1	2

Latihan Data Histori Pasien Penyakit Darah Tinggi

No	Nama	Atribut								Kelas Data
		Jenis Kelamin	Umur	Jumlah Saudara	Tensi	Berat Badan	Golongan darah	Konsumsi Obat	Pendidikan	Kemungkinan Darah Tinggi
1.	Agusta Yusalendra	Laki-Laki	28	2	Rendah	55	O	Ya	S1	Tidak
2.	Hidayah Ramadhan	Laki-Laki	30	4	Tinggi	48	A	Ya	S1	Ya
3.	Via Vidya	Perempuan	19	1	Rendah	45	A	Tidak	S1	Tidak
4.	Setiani Dian	Perempuan	29	4	Sedang	50	O	Ya	S1	Ya
5.	Ikhsanul Wardana	Laki-Laki	20	3		53	O		S2	Tidak
6.	Nikita Putri	Perempuan	32	3	Tinggi	52	A	Tidak	S2	Ya
7.	Ivo Dwi Ananda	Perempuan	21	6	Rendah	55	AB	Tidak	S1	Tidak
8.	Jeri Habibi Pradana	Laki-Laki	17	5	Rendah	45	O	Ya	S1	Ya
9.	Jozu Dermawan	Laki-Laki	32	2	Tinggi	60	AB	Ya	S1	Ya
10.	Kurnia Dewi	Perempuan	17	3	Sedang	50	O		S2	Tidak
11.	Lutfi Aristo Ahmad		18	6		43	B	Tidak	SMA	Tidak
12.	M. Ikhsan Maulana	Laki-Laki	20	2	Tinggi		AB	Tidak	S1	Tidak
13.	M. Alif	Laki-Laki		4	Rendah	48	A	Ya	S2	Ya
14.	M. Hamdan Habib	Laki-Laki	34	4	Tinggi	45	A	Ya	S1	Ya
15.	M. Ikhsan Maulana	Laki-Laki	20	2	Tinggi		AB	Tidak	S1	Tidak
16.	Dina Pritami	Perempuan	21	4	Tinggi	46	O	Ya	SMA	Ya

Thanks!

Sampai Ketemu dipertemuan Selanjutnya