

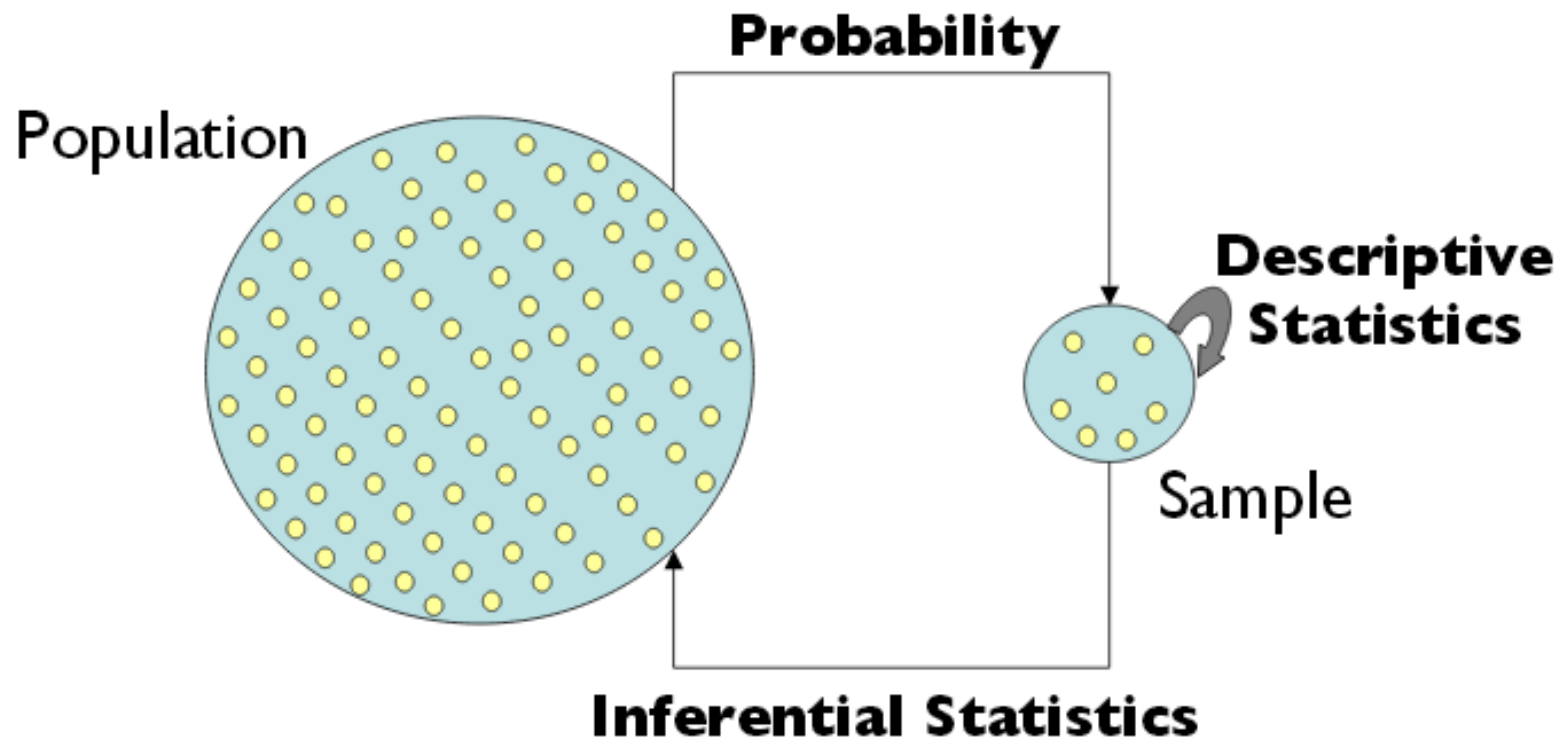
DATA MINING

Pertemuan 3: Exploratory Data Analysis

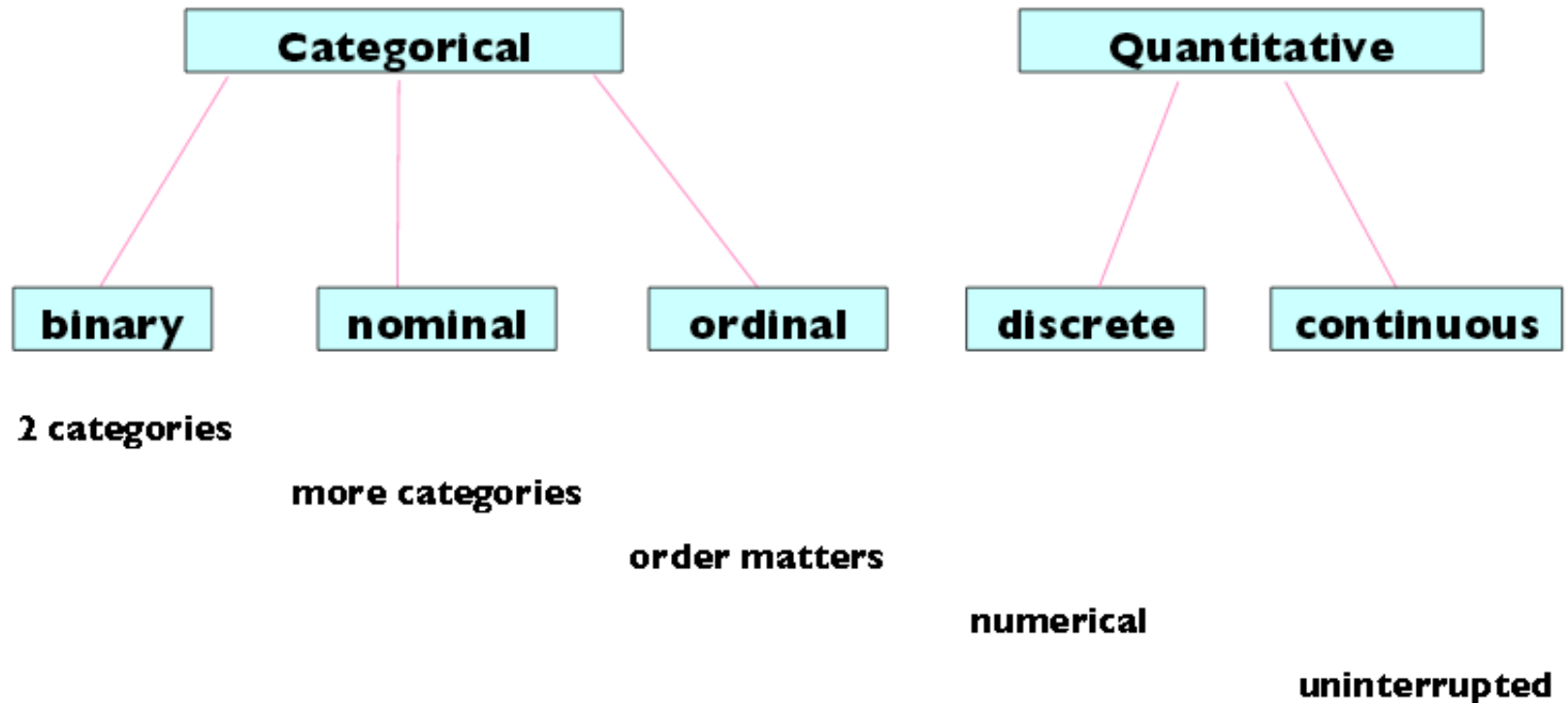
Deskripsi Data

- Kita perlu memahami data terlebih dahulu sebelum dapat membuat model prediksi data
 - Untuk melihat kesalahan pada data
 - Untuk melihat pola pada data
 - Untuk membangun hipotesis

Aturan Baku Statistik



Tipe Data



Dimensi Data Set

- **Univariate**

- Pengukuran didasarkan pada 1 variabel per subjek data

- **Bivariate**

- Pengukuran didasarkan pada 2 variabel per subjek data

- **Multivariate**

- Pengukuran didasarkan pada banyak variabel per subjek data

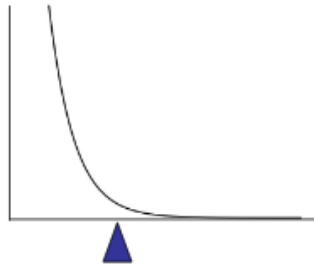
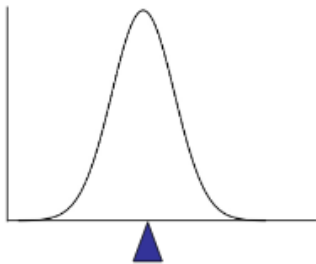
Bentuk Deskripsi Data

- Ringkasan (*Summary*)
 - Rata-rata (*mean*)
 - Rata-rata terbobot (*weighted mean*)
 - Nilai tengah (*median*)
 - *Quartile* dan *Percentile*
 - *Variance*
 - Standar deviasi
- Visualisasi
 - Univariate: Bar Plot, Histogram
 - Bivariate: Box Plot
 - Multivariate: Clustering, PCA

Rata-rata (Mean)

- Menghitung rata-rata \bar{x} dari sekumpulan observasi, menjumlahkan nilainya dan membagi dengan jumlah observasi

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$



Weighted Mean:

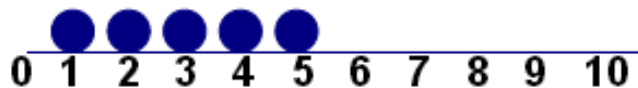
$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

Nilai Tengah (Median)

- Jika jumlah observasi ganjil, maka ambil 1 nilai tengah
 - Jika jumlah observasi genap, ambil 2 nilai tengah dan hitung rata-ratanya
 - Contoh:
 - Data umur: 17 19 21 22 23 23 23 38
- Median = $(22+23)/2 = 22.5$

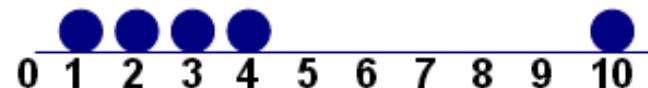
Mean vs Median

- Mean baik digunakan untuk distribusi yang simetris dan tanpa outlier
- Median baik digunakan untuk distribusi tidak seimbang atau data dengan outlier



Mean = 3

Median = 3



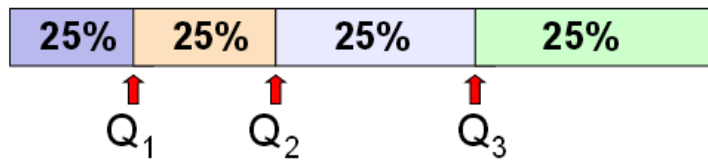
Mean = 4

Median = 3

Quartile dan Percentile

Quartile

- Pembagian seperempat-an dari data



- Q1: 25% data
- Q2: 50% data
- Q3: 75% data

Percentile

- Pembagian perseratus-an dari data
- Misal:
 - Q1 = persentil ke-25
 - Q2 = persentil ke-50
 - Q3 = persentil ke-75

Variance dan Standar Deviasi

Variance

- Rata-rata dari kuadrat simpangan nilai terhadap mean

$$\hat{\sigma}^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n - 1}$$

Standar Deviasi

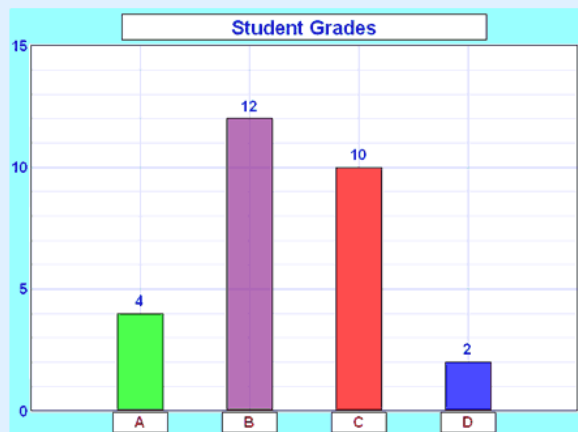
- Menstandarkan nilai dari variance
- Akar kuadrat dari Variance

$$\hat{\sigma} = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n - 1}}$$

Visualisasi Grafis dari Data: Bar Plot dan Histogram

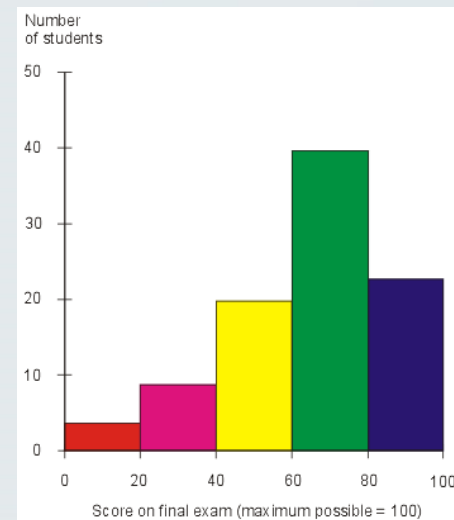
Bar Plot

- Untuk menunjukkan frekuensi/proporsi dari variabel bertipe kategori
- Menterjemahkan data frekuensi dari tabel ke dalam bentuk gambar



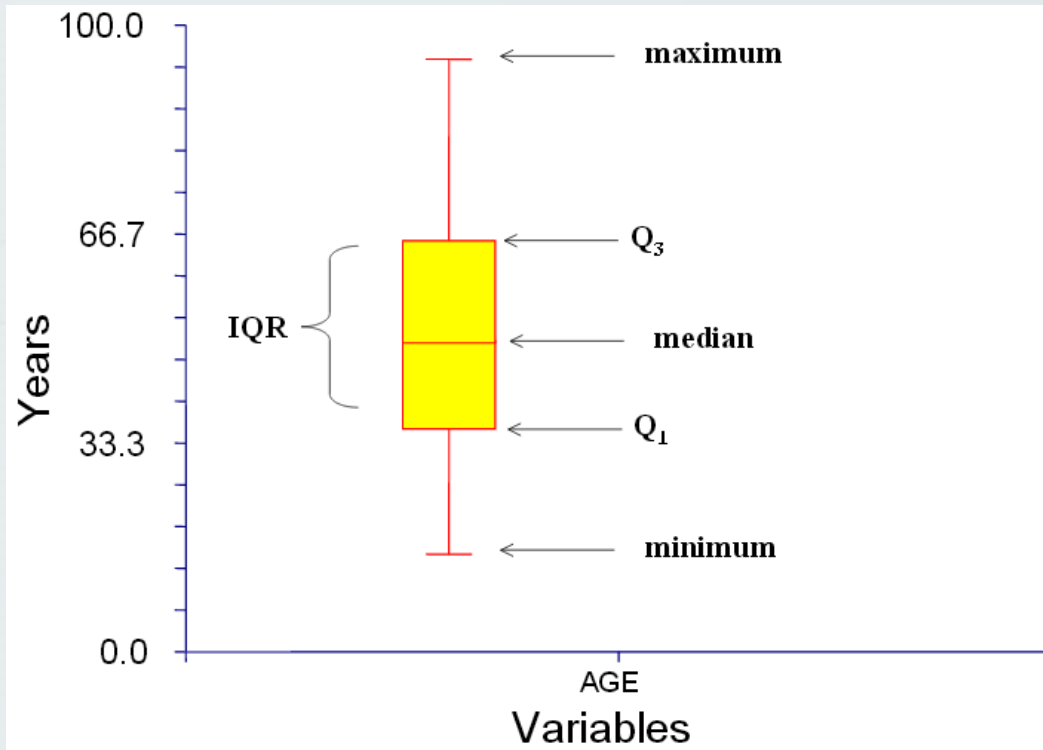
Histogram

- Untuk memvisualisasikan distribusi (bentuk, pusat, range, variasi) dari variabel bertipe kontinyu
- Ukuran bin sangat penting



Visualisasi Grafis dari Data: Box Plot

- Menggambarkan data numerik berdasarkan nilai quartile-nya



IQR: Range variasi yang mungkin

Pemilihan Atribut/Fitur Berbasis Statistik

- Fitur baik: fitur yang memiliki kemampuan tinggi dalam diskriminasi *class* (khususnya pada permasalahan *classification*)
- Pemilihan fitur baik dapat dilakukan dengan pengujian hipotesis:
 - H1: Nilai fitur berbeda secara signifikan
 - H0: Nilai fitur tidak berbeda secara signifikan

1. Pengujian Hipotesis dengan T-Test

- T-Test melakukan pengujian fitur secara individu dan memeriksa ada atau tidaknya informasi diskriminasi data terhadap *class*
 - Jika tidak ada, fitur akan dibuang
- Tujuan T-Test: Menentukan mana di antara 2 hipotesis ini yang bernilai benar:
 - **H1**: Nilai rata-rata fitur dalam 2 class adalah berbeda
 - **H0**: Nilai rata-rata fitur dalam 2 class adalah sama
- Pengujian menggunakan nilai level signifikan α sesuai dengan kemungkinan kesalahan yang dilakukan dalam pengambilan keputusan
- Nilai α yang umum digunakan adalah $\alpha=0,05$ atau $\alpha=0,001$

Contoh perhitungan

- Hasil pengukuran yang didapat oleh sebuah fitur dari 2 *class* sbb:

Kelas	Data ke-									
	1	2	3	4	5	6	7	8	9	10
C1	3.5	3.7	3.9	4.1	3.4	3.5	4.1	3.8	3.6	3.7
C2	3.2	3.6	3.1	3.4	3.0	3.4	2.8	3.1	3.3	3.6

- Apakah fitur tersebut dapat digunakan untuk mendiskriminasi 2 class tersebut?
 - Digunakan $\alpha=0,05$ dan $\alpha=0,001$

Perhitungan t-test dengan Excel

- Aktifkan add-in Analysis ToolPak

t-Test: Two-Sample Assuming Unequal Variances		
	Class C1	Class C2
Mean	3,73	3,25
Variance	0,060111	0,067222
Observations	10	10
Hypothesized Mean Difference	0	
df	18	
t Stat	4,253733	
P(T<=t) one-tail	0,000239	
t Critical one-tail	1,734064	
P(T<=t) two-tail	0,000478	
t Critical two-tail	2,100922	

- Menggunakan two-tail
- Jika nilai $p < \alpha$, tolak H_0
- $\alpha=0,05$: $p < \alpha$, maka tolak H_0
- $\alpha=0,001$: $p < \alpha$, maka tolak H_0

Kesimpulan: H_1 diterima, artinya nilai rata-rata dari kedua kelas berbeda secara signifikan, dan fitur tersebut dapat dipilih sebagai diskriminan class

Jarak Antar-objek Data

- Jarak antara dua objek data dihitung menggunakan tingkat kemiripan (*similarity*) atau ketidakmiripan (*dissimilarity*)
- Contoh: Data dengan n objek dan p atribut di mana x_{ij} adalah nilai untuk objek i pada atribut ke- j
- **$d(i,j)$** menyatakan *dissimilarity* antara objek i dan objek j dan merupakan bilangan non-negative:
 - Nilai mendekati 0: tingkat kemiripan tinggi
 - Nilai mendekati 1: tingkat kemiripan rendah

$$\bullet \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

1. Jarak untuk Atribut Numerik

- Euclidean Distance (L_2 - 2-norm)

- $d(i, j) =$

$$\sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

- Manhattan Distance (L_1 - 1-norm)

- $d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$

2. Jarak untuk Atribut Nominal

- *Dissimilarity* antara objek i dan j dihitung berdasarkan rasio ketidaksamaan status:

- Misal jumlah status dalam suatu atribut nominal disimbolkan dengan M (Contoh: atribut Usia memiliki 3 status: muda, paruh baya, tua, maka $M=3$)
- $d(i, j) = \frac{p-m}{p}$, di mana $m \in M$ dan m menyatakan jumlah status yang bernilai sama dan p adalah jumlah atribut pada objek

Nama	Pekerjaan	Lokasi Rumah
Andi	Analisis	A
Budi	Dokter	A
Citra	Guru	B
Dedi	Analisis	A
Evan	Dokter	C

Matriks dissimilarity-nya:

$$\begin{bmatrix} 0 & & & & \\ 0.5 & 0 & & & \\ 1 & 1 & 0 & & \\ 0 & 1 & 1 & 0 & \\ 1 & 0.5 & 1 & 1 & 0 \end{bmatrix}$$

Dissimilarity antara objek 1 (Andi) dan objek 2 (Budi):

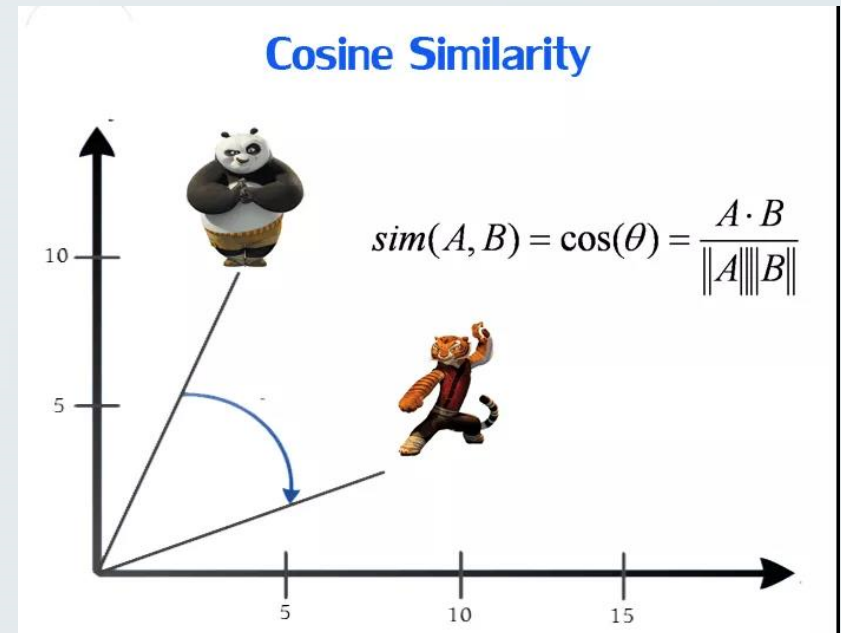
$$d(1,2) = d(2,1) = \frac{2-1}{2} = 0.5$$

Pemberian bobot untuk tiap atribut juga dapat dilakukan untuk membedakan derajat tiap atribut

$d(i, j) = d(j, i) = \frac{p - \sum_{a=1}^p w_a b_a}{p}$, dimana w_a adalah bobot atribut ke- a dan b_a adalah status kesamaan atribut ke- a

3. Jarak Vektor/Jarak Dokumen/Cosine Similarity

- Sebuah dokumen dapat dianggap sebagai data vector dengan ratusan bahkan ribuan atribut, di mana atributnya berupa istilah kata (*term*) yang nilainya berupa frekuensi kemunculan istilah tersebut dalam dokumen
- Vektor dokumen juga sering disebut vector frekuensi-istilah (*term frequency vector*)
- Jarak Vektor juga dapat diterapkan pada image dan video



Sumber: <http://i0.wp.com/technipink.com/wp-content/uploads/2017/07/cosine.png>

Cosine Similarity (lanjutan)

$$\text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

- Contoh data 5 vector dokumen dengan 10 kata beserta frekuensinya

Dok	Agama	Aksi	Bela	Calon	Gubernur	Islam	Monas	Pemilihan	Penista	Presiden
D1	3	4	2	0	0	1	1	0	0	0
D2	1	5	2	0	0	4	3	0	0	0
D3	0	3	2	2	2	2	0	0	0	0
D4	0	0	0	4	0	0	0	3	0	2
D5	0	0	0	4	0	0	0	5	0	6

- Vektor dokumen merupakan matriks jarang (*sparse matrix*), artinya banyak elemen yang memiliki nilai 0
 - Kesamaan dokumen D1 dengan D2:
 - $D1 \cdot D2 = 3 * 1 + 4 * 5 + 2 * 2 + 0 * 0 + 0 * 0 + 1 * 4 + 1 * 3 + 0 * 0 + 0 * 0 + 0 * 0 = 34$
 - $\|D1\| = \sqrt{3^2 + 4^2 + 2^2 + 0^2 + 0^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2} = 5.57$
 - $\|D2\| = \sqrt{1^2 + 5^2 + 2^2 + 0^2 + 0^2 + 4^2 + 3^2 + 0^2 + 0^2 + 0^2} = 7.42$
 - $\text{sim}(D1, D2) = \frac{D1 \cdot D2}{\|D1\| \|D2\|} = \frac{34}{5.57 * 7.42} = 0.82$
- ∴ D1 dan D2 memiliki kemiripan yang tinggi**

SUMBER DATA SET

UCI Dataset

- UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets.html>) adalah kumpulan database, domain teori, dan data generator yang digunakan oleh komunitas pembelajaran mesin (*machine learning*) untuk analisis empirik dari algoritma pembelajaran mesin
- Arsipnya dikumpulkan sejak tahun 1987 oleh David Aha dan rekan kuliahnya di UC Irvine dan menjadi sumber utama dari dataset pembelajaran mesin dengan jumlah rujukan dokumen lebih dari 1000 kali, menjadikannya sebagai salah satu dari top 100 dokumen yang paling banyak dirujuk di bidang *computer science*

UCI Dataset




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






☒ Repository
 ☐ Web
 

[View ALL Data Sets](#)

Machine Learning Repository
 Center for Machine Learning and Intelligent Systems

Browse Through: 418 Data Sets

Table View [List View](#)

Default Task	Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Classification (309) Regression (79) Clustering (69) Other (54)	 Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Attribute Type Categorical (37) Numerical (266) Mixed (55)	 Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Data Type Multivariate (319) Univariate (18) Sequential (42) Time-Series (78) Text (43) Domain-Theory (22) Other (21)	 Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Area Life Sciences (97) Physical Sciences (47)	 Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
	 Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998

Contoh: Iris Data Set

Iris Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: Famous database; from Fisher, 1936



Data Set Characteristics:	Multivariate	Number of Instances:	150	Area:	Life
Attribute Characteristics:	Real	Number of Attributes:	4	Date Donated	1988-07-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	1217125

Source:

Creator:


R.A. Fisher

Donor:


Michael Marshall (MARSHALL%PLU '@' io.arc.nasa.gov)

Iris Data Set

- Terdiri dari 2 file: iris.data dan iris.names



```
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
```

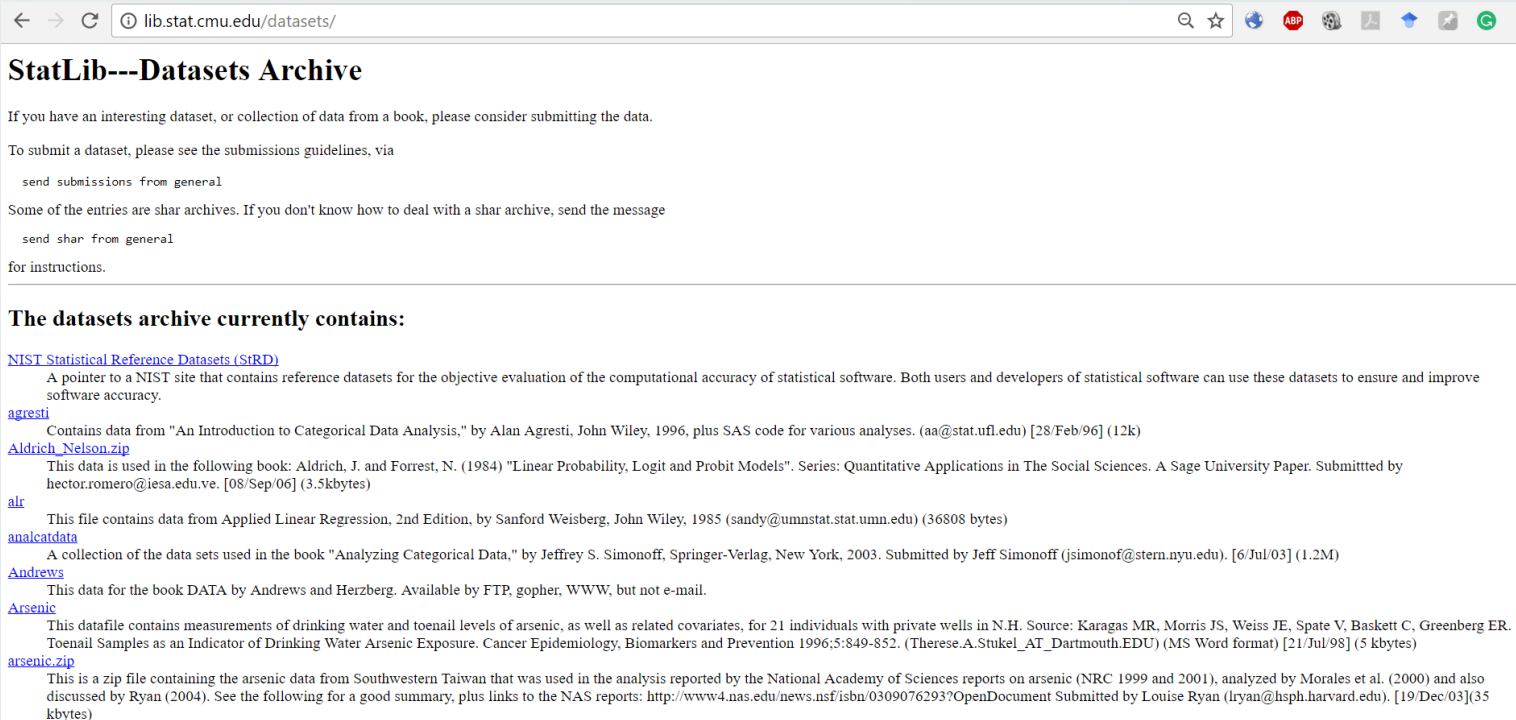


Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
 - Iris Setosa
 - Iris Versicolour
 - Iris Virginica

Data Set Carnegie Mellon University

- <http://lib.stat.cmu.edu/datasets/>
- [!] Last modified tahun 2005



The screenshot shows a web browser window with the address bar displaying lib.stat.cmu.edu/datasets/. The page title is "StatLib---Datasets Archive". The main content area contains the following text:

If you have an interesting dataset, or collection of data from a book, please consider submitting the data.

To submit a dataset, please see the submissions guidelines, via

send submissions from general

Some of the entries are shar archives. If you don't know how to deal with a shar archive, send the message

send shar from general

for instructions.

The datasets archive currently contains:

[NIST Statistical Reference Datasets \(StRD\)](#)
A pointer to a NIST site that contains reference datasets for the objective evaluation of the computational accuracy of statistical software. Both users and developers of statistical software can use these datasets to ensure and improve software accuracy.

[agresti](#)
Contains data from "An Introduction to Categorical Data Analysis," by Alan Agresti, John Wiley, 1996, plus SAS code for various analyses. (aa@stat.ufl.edu) [28/Feb/96] (12k)

[Aldrich_Nelson.zip](#)
This data is used in the following book: Aldrich, J. and Forrest, N. (1984) "Linear Probability, Logit and Probit Models". Series: Quantitative Applications in The Social Sciences. A Sage University Paper. Submitted by hector.romero@iesia.edu.ve. [08/Sep/06] (3.5kbytes)

[air](#)
This file contains data from Applied Linear Regression, 2nd Edition, by Sanford Weisberg, John Wiley, 1985 (sandy@umnstat.stat.umn.edu) (36808 bytes)

[analcata](#)
A collection of the data sets used in the book "Analyzing Categorical Data," by Jeffrey S. Simonoff, Springer-Verlag, New York, 2003. Submitted by Jeff Simonoff (jsimonof@stern.nyu.edu). [6/Jul/03] (1.2M)

[Andrews](#)
This data for the book DATA by Andrews and Herzberg. Available by FTP, gopher, WWW, but not e-mail.

[Arsenic](#)
This datafile contains measurements of drinking water and toenail levels of arsenic, as well as related covariates, for 21 individuals with private wells in N.H. Source: Karagas MR, Morris JS, Weiss JE, Spate V, Baskett C, Greenberg ER. Toenail Samples as an Indicator of Drinking Water Arsenic Exposure. Cancer Epidemiology, Biomarkers and Prevention 1996;5:849-852. (Therese.A.Stukel_AT_Dartmouth.EDU) (MS Word format) [21/Jul/98] (5 kbytes)

[arsenic.zip](#)
This is a zip file containing the arsenic data from Southwestern Taiwan that was used in the analysis reported by the National Academy of Sciences reports on arsenic (NRC 1999 and 2001), analyzed by Morales et al. (2000) and also discussed by Ryan (2004). See the following for a good summary, plus links to the NAS reports: <http://www4.nas.edu/news.nsf/isbn/0309076293?OpenDocument> Submitted by Louise Ryan (lryan@hsph.harvard.edu). [19/Dec/03](35 kbytes)

Data Set Indonesia

- www.data.go.id
 - Bidang Pangan, Energi, Infrastruktur, Maritim, Kesehatan, Pendidikan, Ekonomi, Industri, Pariwisata, Reformasi Birokrasi



Data Set Indonesia

Kedaulatan Pangan

Mencakup pertanian dan sektor pangan secara umum [read more](#)

Pengikut
0
Kumpulan data
75

Organisasi

Pemerintah Provinsi... (28)
Badan Pusat Statistik (26)
Kementerian Pertanian (10)
Pemerintah Kabupate... (7)
Kementerian Kelaut... (2)
Lainnya (1)
Badan Nasional Pena... (1)

☐ Kumpulan data ☐ Activity Stream ☐ Tentang

Search datasets...



75 dataset found

Order by:

Data Bencana Kekeringan

Dataset ini berisi mengenai data mengenai Variabel pada Datas...

[CSV](#) [xlsx](#)

Jumlah Rumah Tangga Usa

Dataset ini berisi informasi men... lahan yang dikelola berdasarkan

[CSV](#)

Jumlah Usaha Tani berdas

Dataset ini berisi informasi men...

Kesehatan

Mencakup sektor kesehatan dan pengelolaan kesehatan di Indonesia [read more](#)

Pengikut
1
Kumpulan data
120

Organisasi

Kementerian Kesehatan (59)
Pemerintah Kabupate... (24)
Pemerintah Provinsi... (15)
Badan Pusat Statistik (8)
Kementerian Perenca... (6)
Tim Nasional Percep... (5)

☐ Kumpulan data ☐ Activity Stream ☐ Tentang

Search datasets...



120 dataset found

Order by:

Tempat Pembuangan Akhir Tinja Rumah Tangga Miskin

Dataset ini berisi data mengenai jumlah rumah tangga dengan status kesejahteraan 30% terendah berdasarkan jenis tempat pembuangan akhir tinja per Provinsi Penjelasan...

[CSV](#)

Status Kepemilikan Fasilitas Tempat Buang Air Besar

Dataset ini berisi data mengenai jumlah rumah tangga dengan status kesejahteraan 30% terendah berdasarkan status kepemilikan jamban per Provinsi Penjelasan mengenai...

[CSV](#)

Sumber Air Minum Rumah Tangga Miskin

Dataset ini berisi data mengenai jumlah rumah tangga dengan status kesejahteraan 30% terendah berdasarkan sumber air minum per provinsi Penjelasan mengenai Variabel pada...

Format Data Set

- Ekstensi .arff
 - Attribute-Relation File Format
 - format data default untuk Weka
 - Menggunakan keterangan @relation dan @attribute
- Ekstensi .csv
 - Comma-Separated Values
 - baris pertama adalah judul atribut/kolom
- Ekstensi .data
 - mirip dengan CSV
 - Tidak memiliki baris judul atribut/kolom
 - Keterangan judul atribut/kolom ada di file dengan ekstensi .names (.data dan .names merupakan satu paket data set)

Pengenalan Tools untuk Data Mining: MATLAB

- MATrix LABoratory (MATLAB) adalah sebuah bahasa pemrograman untuk komputasi numerik milik MathWorks™
- MATLAB memiliki kemampuan untuk melakukan manipulasi data matriks, plot fungsi dan data, implementasi algoritma, pembuatan *user interface*, dan menjembatani dengan bahasa pemrograman lain seperti C, C++, C#, Java, Fortran, dan Python



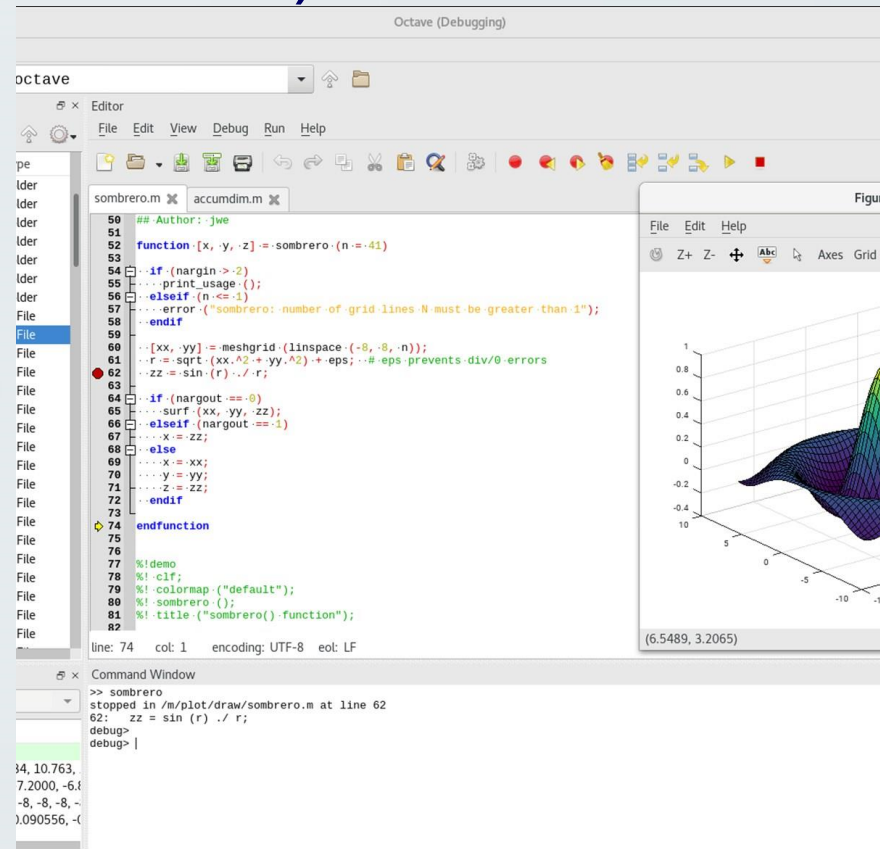
MATLAB untuk Data Mining

- MATLAB memiliki kemampuan untuk analisis data yang merupakan bagian dari proses data mining, seperti:
 - Preprocessing
 - Scaling, averaging, interpolating, decimating, clipping, thresholding, extracting section of data, smoothing, filtering
 - Penerapan operasi numerik dan matematika
 - Correlation, basic statistics, curve fitting, Fourier analysis and filtering, matrix analysis, differential equation solvers

MATLAB untuk Data Mining

- Memiliki berbagai toolbox untuk advanced analysis:
 - Curve Fitting, Filter design, statistics, Communications, Optimization, Wavelets, Spline, Image processing, Symbolic math, control system design, partial differential equations, neural networks, signal processing, fuzzy logic

- OCTAVE (Open Source)



Pengenalan Tools untuk Data Mining: WEKA



- Waikato Environment for Knowledge Analysis (WEKA) adalah software paket *machine learning* yang ditulis dengan bahasa Java, dibuat oleh University of Waikato, New Zealand
- WEKA adalah software gratis di bawah lisensi GNU General Public License

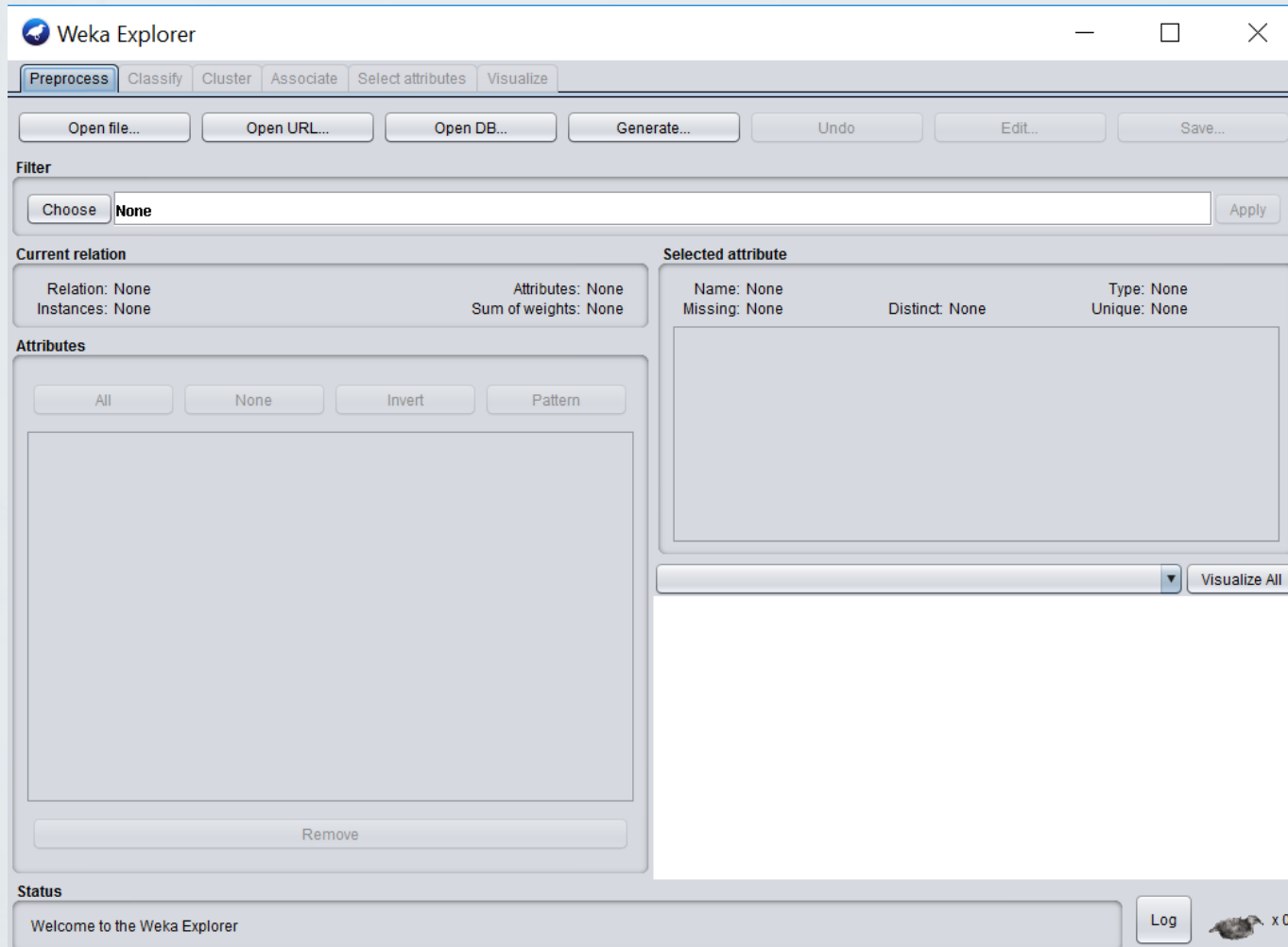
WEKA untuk Data Mining

- WEKA mendukung sejumlah tugas standar dalam data mining seperti:
 - Data preprocessing, clustering, classification, regression, visualization, feature selection
- WEKA menyediakan akses ke database SQL menggunakan Java Database Connectivity (JDBC)
 - Tidak mendukung untuk multi-relational data mining, tetapi ada software converter terpisah untuk mengubah *linked database tables* menjadi 1 table sehingga dapat diproses oleh WEKA

WEKA untuk Data Mining

- WEKA dapat digunakan dengan berbagai cara:
 - Dengan *user interface* bawaan, yakni Explorer
 - Dengan *command line*
 - Dengan Experimenter, yakni membandingkan kinerja beberapa algoritma dalam menjalankan sejumlah dataset
 - Dengan menggunakan library `weka.jar` dalam program Java kita

GUI WEKA



Next Week

- Classification