MATERI MODUL ONLINE DATA MINING
EKSPLORASI DATA DATA MINING
SESI ONLINE 2
Syefira Salsabila

Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogenous sources. Low-quality data will lead to low-quality mining results. *"How can the data be preprocessed in order to help improve the quality of the dataand, consequently, of the mining results? How can the data be preprocessed so as to improve the efficiency and ease of the mining process?"*

Kita mengenal istilah No quality data, no quality mining results! (Garbage in, garbage out). Keputusan yang baik harus berdasarkan data yang berkualitas pula (Quality decisions must be based on quality data). Data dalam dunia nyata "dirty" (tidak sempurna) selain itu data yang tidak komplit atau bisa juga berisi data yang hilang/kosong, kekurangan atribut yang sesuai, hanya berisi data aggregate. Data yang tidak berkualitas, akan menghasil kankualitas mining yang tidak baik pula.

Mengapa ada **data yang tidak lengkap**? Hal ini dapat terjadi karena saat dikumpulkan, nilai dari atribut tertentu tidak tersedia. Terjadi perbedaan pertimbangan sewaktu data dikumpulkan dengan sewaktu data dianalisa data problem yang disebabkan oleh manusia/hardware/software.mengapa ada **Noisy data**? Hal ini dapat terjadi karena faulty data collection instruments (kesalahan pada alat). Human atau komputer error pada saat entry data Terjadi error pada saat dikirim (errors in data transmission). Mengapa terjadi **Inconsistent data?,** hal ini dapat terjadi karena perbedaan sumber data (different data sources.

Eksplorasi data dilakukan sebagai langkah awal untuk memahami data dari sebelum dilakukan praproses. Tahapan ini bertujuan untuk menyeleksi teknik pemrosesan dan analisis data yang sesuai dengan dataset yang dimiliki. Dalam eksplorasi data, hal yang harus diperhatikan yaitu;

a. Tipe data
b. Kualitas data
c. Statistika ringkasan
d. Visualisasi

1. Tipe Data

Sebuah data set dapat dipandang sebagai sebuah koleksi dari objek-objek data. Nama lain dari sebuah objek data adalah *record,* titik, vector, pola, *event,* case, *sample,* observasi atau entitas. Objek digambarkan dengan sejumlah atribut yang menerangkan sifat atau karakteristik dari objek tersebut. Kumpulan dari objek data dan atributnya

Contoh: Informasi mahasiswa

Data set adalah sebuah file, dimana objek adalah *record-record* (baris) dalam file dan setiap *field* (kolom) berkaitan dengan sebuah atribut.

Tabel 1.1 *Data set* mahasiswa

Student ID	Year	Grade Point Average (GPA)	…..
….	….	….	….
1034262	Senior	3.24	….
1052663	Sophomore	3.51	….
1082246	Freshman	3.62	….
….	….	….	….

Skala pengukuran adalah aturan (fungsi) yang menghubungkan nilai numerik atau simbolik dengan sebuah atribut dari sebuah objek. Proses pengukuran adalah penggunaan skala pengukuran untuk mengubungkan sebuah nilai dengan sebuah atribut tertentu dari sebuah objek.

Contoh:

Menghitung banyaknya kursi dalam sebuah ruangan untuk melihat apakah terdapat cukup empat duduk untuk semua orang yang akan datang pada sebuah pertemuan. Dalam kasus tersebut, nilai fisik dari sebuah atribut dari sebuah objek dipetakan ke sebuah nilai numerik atau simbolik.

Atribut juga sering disebut variabel, *field*, fitur, atau dimensi. Atribut adalah sifat/properti/karakteristik objek yang nilainya bisa bermacam-macam dari satu objek dengan objek lainnya, dari satu waktu ke waktu yang lainnya. Atribut adalah sebuah sifat atau karakterik dari sebuah objek yang dapat bervariasi, baik dari satu objek ke objek lain atau dari satu waktu ke waktu yang lain. Sebuah atribut adalah sifat atau karakteristik dari sebuah objek.

Contoh : warna mata dari sesorang, temperatur suhu

Atribut

Objek

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

a. Tipe dari atribut

Tabel 1.2 Tipe-tipe atribut yang berbeda

Tipe atribut		**Deskripsi**	**Contoh**
Kategori (Kualitatif)	Nominal	▪ Nilai dari atribut nominal adalah nama-nama yang berbeda, yaitu nilai nominal hanya menyediakan informasi yang cukup untuk membedakan satu objek dengan objek yang lain. (= dan ≠). ▪ Atribut yang nilainya berupa simbol-simbol atau nama-nama benda dan nilai nya tidak memiliki urutan yang memiliki arti. ▪ Nilai pada atribut tidak dapat melaukan operasi matematika.	Kode pos, ID Number karywan, warna mata, jenis kelamin , agama
	Ordinal	▪ Nilai dari atribut ordinal menyediakan informasi yang cukup mengurutkan objek (< , >) ▪ Atribut dengan nilai-nilai yang kemungkinan memiliki urutan yang mempunyai arti atau tingkatan (rangking), akan tetapi jarak antara nilai tidak diketahui.	Nomor antrian, *grade,* kecepatan prosesor {lambat, cepat, sangat cepat}
Numerik (Kuantitatif)	Interval	Dalam atribut interval perbedaan antar nilai merupakan sesuatu yang berarti, pada atribut interval terdapat unit pengukuran. Operator aritmatik yang dapat digunakan ialah sama dengan (=), tidak sama dengan (≠), lebih besar (>), lebih kecil (<), penambahan (+), pengurangan (-).	Tanggal pada kalender, temperature
	Ratio	Dalam atribut rasio, perbedaan rasio merupakan hal yang	Umur, berat badan, tinggi badan

Tipe atribut		Deskripsi	Contoh
		berarti. Operator atribut rasio, perbedaan rasio merupakan hal yang berarti. Operator numeric yang dapat digunakan ialah sama dengan (=), tidak sama dengan (≠), lebih besar (>) , lebih kecil (<), penambahan (+), pengurangan (-), perkalian (*), pembagian (/)	

2. Kualitas Data

Data Mining adalah proses menemukan pengetahuan dari sekian banyak data yang tersimpan dalam *database*; *data warehouse*; atau repositori lain.Usaha yang diperlukan pada pengolahan data yaitu *data understanding* sebesar 20%, 60% untuk *data preparation*, dan hanya 20% untuk *data mining* serta menganalisis *knowledge*. Ini menunjukkan bahwa *data preparation* membutuhkan usaha terbesar dalam pengolahan data. Data yang ada di dunia ini cenderung tidak lengkap dikarenakan adanya beberapa *missing value*. Hal ini bisa saja terjadi karena valuenya tidak relevan pada masalah, tidak terekam saat pengumpulan data, atau memang tidak dijawab oleh responden dikarenakan alasan privasi. Apabila ratarata *missing value* kurang dari 1 %, data yang missing ini tidak akan menimbulkan masalah untuk proses Knowledge Discovery in Database (KDD), 1-5% masih bias diolah, 5-15% dibutuhkan metode untuk menanganinya dan jika lebih dari 15% dapat menimbulkan interpretasi yang berbeda.

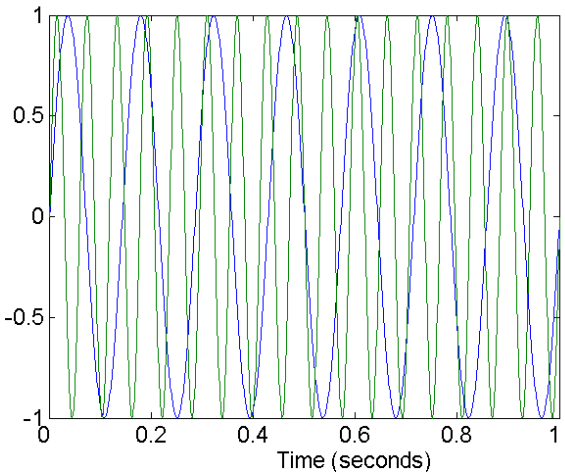
Penentuan dari Kualitas Data;

a. Apa jenis masalah dari kualitas data?
b. Bagaimana kita dapat mendeteksi masalah dalam data?
c. Apa yang dapat kita lakukan untuk menghadapi masalah tersebut?
d. Contoh dari masalah kualitas data :
    a) Noise dan outlier
    b) Missing Value
    c) Duplicate data

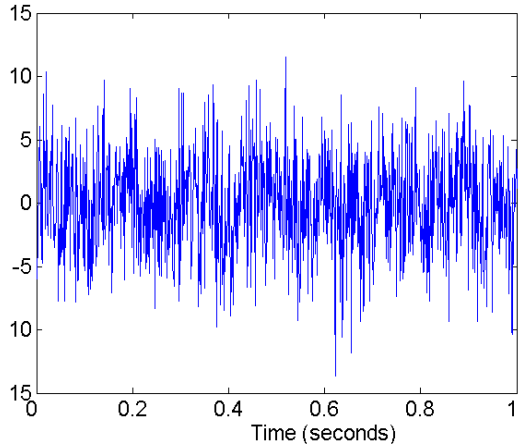
a) Noise

Noise mengarah kepada terjadinya modifikasi dari nilai yang sebenarnya. Merupakan komponen random dari suatu error pengukuran. Noise berkaitan dengan modifikasi dari nilai asli.

Contoh: Penyimpangan dari suara seseorang ketika berbicara dengan menggunakan jaringan sinyal telepon yang jelek. Distorsi atau penyimpangan dari suara orang saat berbicara di telepon yang jaringannya buruk.



Two Sine Waves

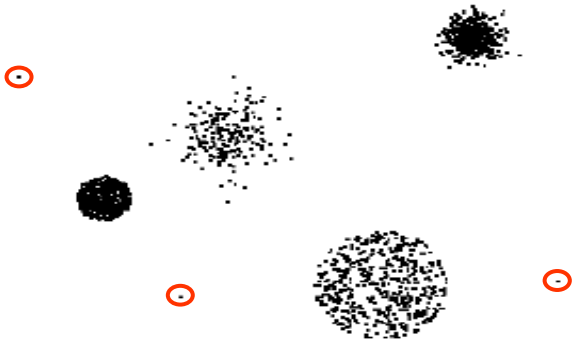


Two Sine Waves + Noise

b) ***Outlier***

Outlier adalah objek data dengan karakteristik berbeda dari karakterisktik sebagian besar objek pada set data.Merupakan objek data dengan sifat yang berbeda sekali dari kebanyakan objek data dalam data-set. Misalkan, terdapat data penelitian tentang tinggi anak siswa SMA yakni 160cm sampai 180cm. Tetapi dalam data tersebut terdapat anak yang mempunyai tinggi 40cm. Data anak dengan tinggi 140cm tersebut yang disebut data outlier, karena berbeda sangat jelas. Terdapat beberapa hal yang mempengaruhi munculnya data outlier antara lain:

1) Kesalahan dalam pemasukan data
2) Kesalahan dalam pengambilan *sample*
3) Memang ada data-data ekstrim yang tidak dapat dihindarkan keberadaannya.



c) Duplicate

Set data mungkin terdiri dari objek data yang ganda (duplikat), atau hampir selalu terjadi duplikasi antara satu dengan yang lainnya. Persoalan utama ketika menggabungkan data dari sumber-sumber yang bervariasi (heterogen). Contoh: orang yang sama dengan alamat email yang lebih dari satu. Pembersihan data (*data cleaning*) merupakan proses yang berkaitan dengan permasalahan data yang duplikat.

Duplikasi data dalam *data mining* dapat terjadi karena dua hal yaitu adanya *record* yang berulang dan adanya perbedaan identifikasi antara entitas yang sama dalam du-nia nyata. Adanya duplikasi data pada *dataset* dapat mempengaruhi kualitas perfor-ma *data mining*2. Duplikasi data pada *dataset* pasien dimungkinkan terjadi karena *input* data pasien dilakukan oleh orang atau waktu yang berbeda. Selain itu, duplikasi juga bisa terjadi karena adanya kesalahan penulisan pada saat proses input data se hingga terdapat beberapa data yang memiliki kemiripan atau bahkan sama dalam *dataset*.

Duplikasi terjadi karena ada perbedaan identifikasi antara entitas yang sama dalam dunia nyata misalnya duplikasi data pasien rumah sakit. Solusi dari permasalahan duplikasi adalah dengan melakukan deduplikasi. Deduplikasi dilakukan dengan mengeliminasi data yang memiliki kemiripan. Pendeteksian duplikasi data merupakan proses identifikasi *record* yang berbeda yang mengacu pada satu entitas atau objek yang memiliki kesamaan dalam dunia nyata.

Di dalam set data mungkin terdapat duplikasi objek data. Biasanya terjadi ketika terjadi penggabungan data dari sumber yang berbeda Contoh : Orang yang sama dengan banyak alamat email. Penghapusan Data:Proses yang dilakukan untuk menangani maslaah duplikasi data

d) Missing Value

Data pasien merupakan data yang sangat penting dalam dunia kesehatan. Data pasien yang disimpan secara terstruktur dapat memberikan informasi tentang riwayat penyakit pasien. Namun demikian, ada beberapa kendala yang dihadapi untuk memperoleh informasi tersebut terkait dengan *dataset* pasien yang cukup besar. Diantara permasalahan yang ditemui dalam *dataset* pasien tersebut adalah adanya duplikasi data dan adanya *missing value*.

Missing value terjadi jika ada nilai dari suatu atribut yang tidak ditemukan. Atribut yang mengandung *missing value* diganti dengan nilai rata-rata seluruh data dalam setiap atribut. *Missing value* dalam *dataset* pasien berasal dari data-data yang atributnya tidak memiliki nilai. Informasi ini tidak diperoleh dimungkinkan karena adanya data pasien yang tidak lengkap seperti jenis kelamin, nama belakang pasien, tanggal lahir pasien, dan sebagainya. Keberadaan *missing value* ini juga dapat menyebabkan duplikasi data karena ada lebih dari satu data dengan nama yang sama dan memiliki kelengkapan data yang berbeda.

Missing value dapat terjadi karena nilainya tidak relevan untuk kasus tertentu, tidak bisa dicatat pada saat data dikumpulkan, atau disebabkan adanya privasi.Untuk mengatasi *missing value*, dapat dilakukan beberapa hal seperti melakukan pengurangan objek data, memperkirakan nilai *missing values*, tidak melibatkan *missing values* dalam analisis data, dan mencari nilai rata-rata pada atribut yang memiliki *missing value*. Selain itu untuk mengatasi *missing value* terdapat beberapa cara, diantaranya adalah dengan menghapus data yang mengandung *missing value*, namun cara ini lebih cocok untuk data yang memiliki *missing value* sedikit. Cara lain dengan mengganti nilainya dengan nilai konstan seperti 999 atau ”unknown” atau dengan nilai rata-rata dari atribut. Namun ini dirasakan tidak efektif dan bias menyebabkan keakuratan data berkurang. Oleh karena itu dibutuhkan suatu metode yang mampu melakukan imputasi terhadap *missing value* dengan nilai yang masuk akal.

Cara menangani missing value dengan **e**liminasi objek data tersebut, estimasi nilai dari missing value,abaikan missing value tersebut selama proses analisis. Misalkan objek tersebut akan digunakan pada proses clustering. Jarak kedekatan yang diperlukan dalam proses clustering dapat dihitung dengan menggunakan atribut lain yang tidak hilang. Asalannya terjadinya *missing value* adalah:

1) Informasi tidak diperoleh (misal, orang-orang menolak untuk memberikan data umur dan berat bedan)
2) 2) Atribut yang mungkin tidak bisa diterapkan ke semua kasus (misal, pendapatan tahunan tidak bisa diterapkan pada seseorang yang pengangguran)

3. **Visualisasi**

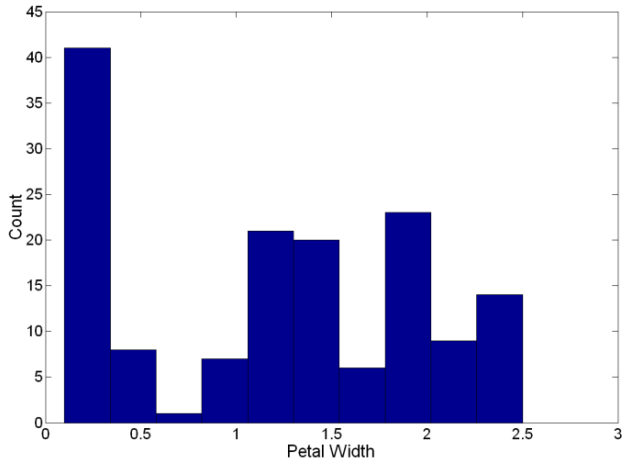
Visualisasi adalah konversi dari data menjadi sebuah format visual atau tabular sehingga karakteristik data dan hubungan antar data atau atribut dapat dianalisis. Visualisasi dari data adalah salah satu teknik yang tepat untuk ekplorasi data. Dapat mendeteksi pola umum dan trend data.Dapat mendeteksi outlier dan pola yang tidak biasa. Informasi berupa data numerik memang diperlukan, tetapi terkadang sulit memahami atau menggambarkannya. Sehingga, seringkali visualisasi data juga dilakukan untuk lebih mudah memahami data.

Baik **histogram** dan **stem-and-leaf plots** berguna untuk memberikan gambaran ukuran tendensi sentral dan kesimetrisan data pengamatan. Penyajian grafis lainnya yang bisa merangkum informasi lebih detail mengenai distribusi nilai-nilai data pengamatan adalah **Box and Whisker Plots** atau lebih sering disebut dengan **BoxPlot** atau **Box-Plot** (kotak-plot) saja. Seperti namanya, *Box and Whisker*, bentuknya terdiri dari **Box** (kotak) dan **whisker**.

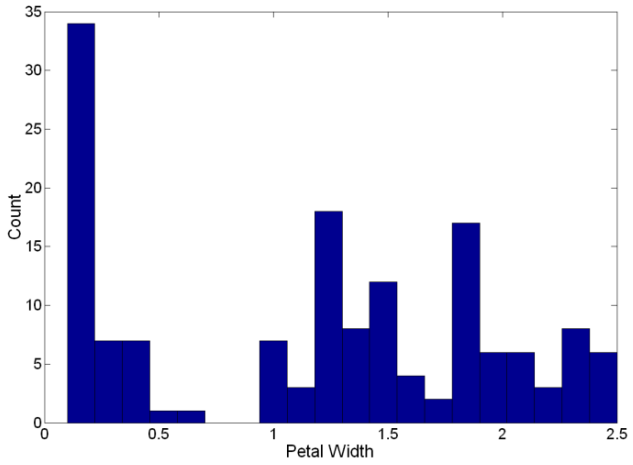
a. Histogram

Histogram adalah grafik yang digunakan untuk menyajikan data kontinu. Grafik ini merupakan areal diagram sehingga kalau interval kelas tidak sama, dilakukan pendataan dengan membandingkan nilai interval kelas dengan frekuensi kelas. Contoh: histogram distribusi volume ekspirasi paru dari 57 orang mahasiswa

Biasanya menunjukkan distribusi nilai dari sebuah single variable. Membagi nilai menjadi beberapa bagian. Tinggi dari setiap bar menunjukkan jumlah dari objek. Gambar dibawah ini menunjukkan histogram dengan 10 bin untuk lebar petal. Bentuk dari histogram dapat tergantung pada banyaknya bin. Histogram untuk data yang sama, tetapi dengan 20 bin



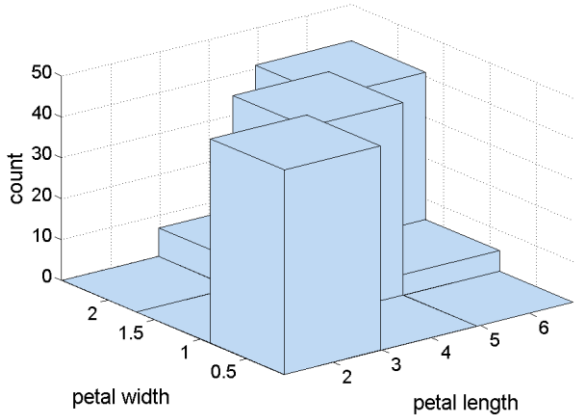
Gambar Histogram untuk lebar petal dengan 10 bin



Gambar Histogram untuk lebar petal dengan 20 bin

Terdapat variasi dari plot histogram. Histogram (frekuensi) relative menggantikan count dengan frekuensi relative. Dalam histogram ini skala dalam sumbu y berubah, dan bentuk dari histogram tidak berubah. Variasi lainnya, khususnya untuk data kategori yang tidak terutut, adalah histogram Pareto. Histogram ini sama dnegan histogram biasa, hanya saja dalam histogram Paretokategori diurut oleh count sedemikian sehingga count menurun dari kiri ke kanan.

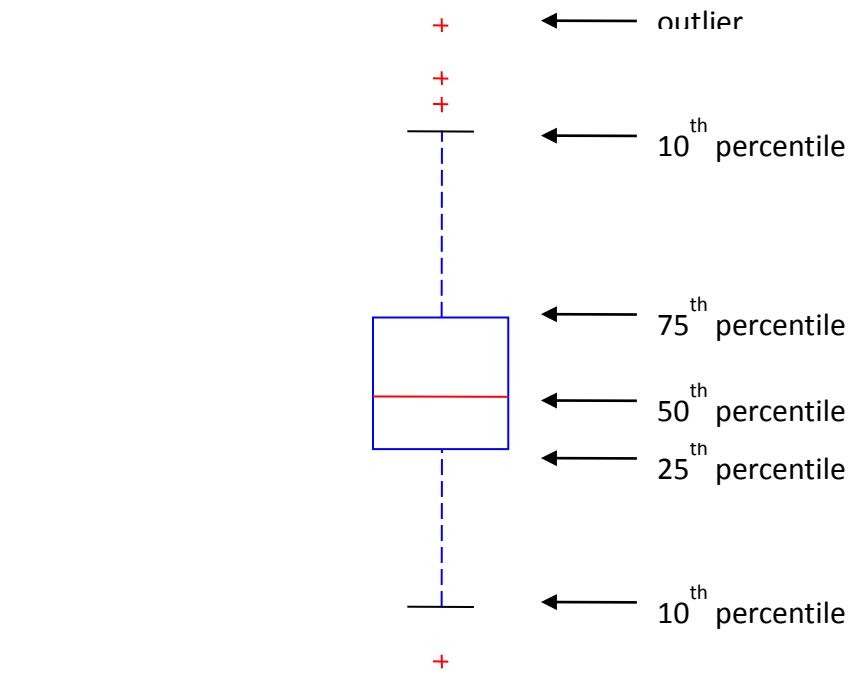
Two-Dimensional Histograms
Menunjukkan distribusi gabungan dari dua atribut. Example: petal width and petal length



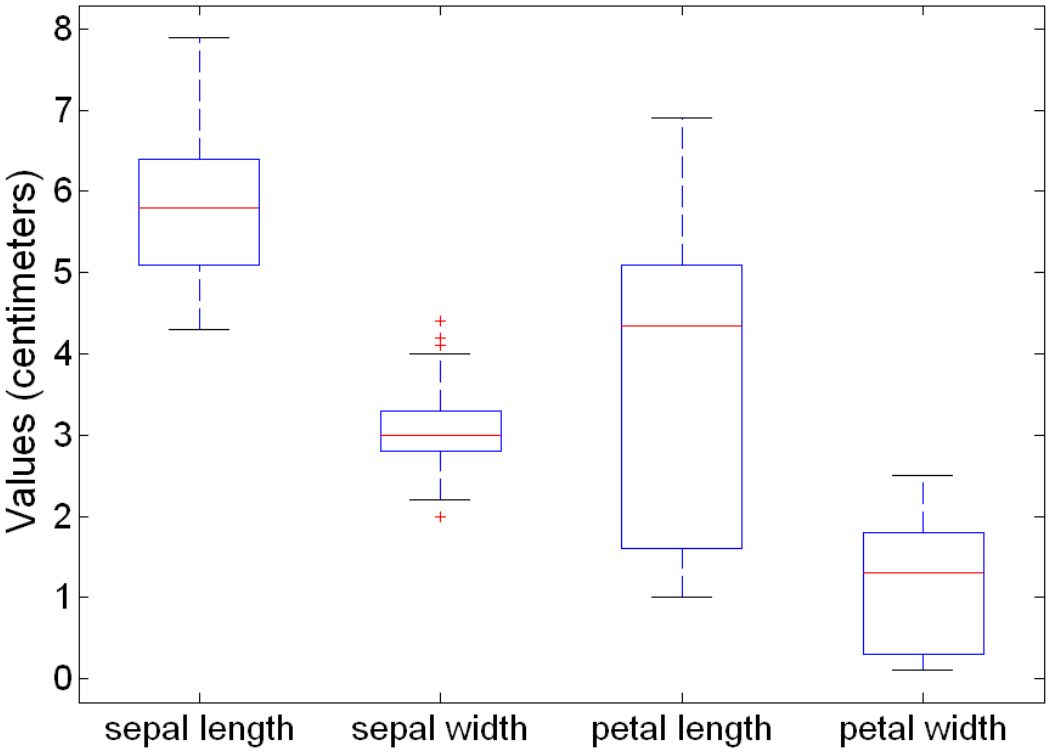
b. **Box Plots**

Box-Plot merupakan ringkasan distribusi sampel yang disajikan secara grafis yang bisa menggambarkan bentuk distribusi data (*skewness*), ukuran tendensi sentral dan ukuran penyebaran (keragaman) data pengamatan.

Metode lain untuk menunjukkan distribusi nilai dari sebuah atribut numeric adalah box plot. Gambar dibawah ini menunjukkan sebuah box plot berlabel untuk panjang sepal. Ujung paling bawah dan paling atas berturut-turut menunjukkan persentil ke 25 dan ke 75, sedangkan garis di dalam kotak menunjukkan persentil ke 50. Garis bawah dan atas dari ekor menunjukkan persentil ke 10 dan ke 90. Outlier ditunjukkan dengan tanda “+”



Gambar Deskripsi box plot untuk panjang sepal



Gambar Box plot untuk atribut-atribut dari set Iris

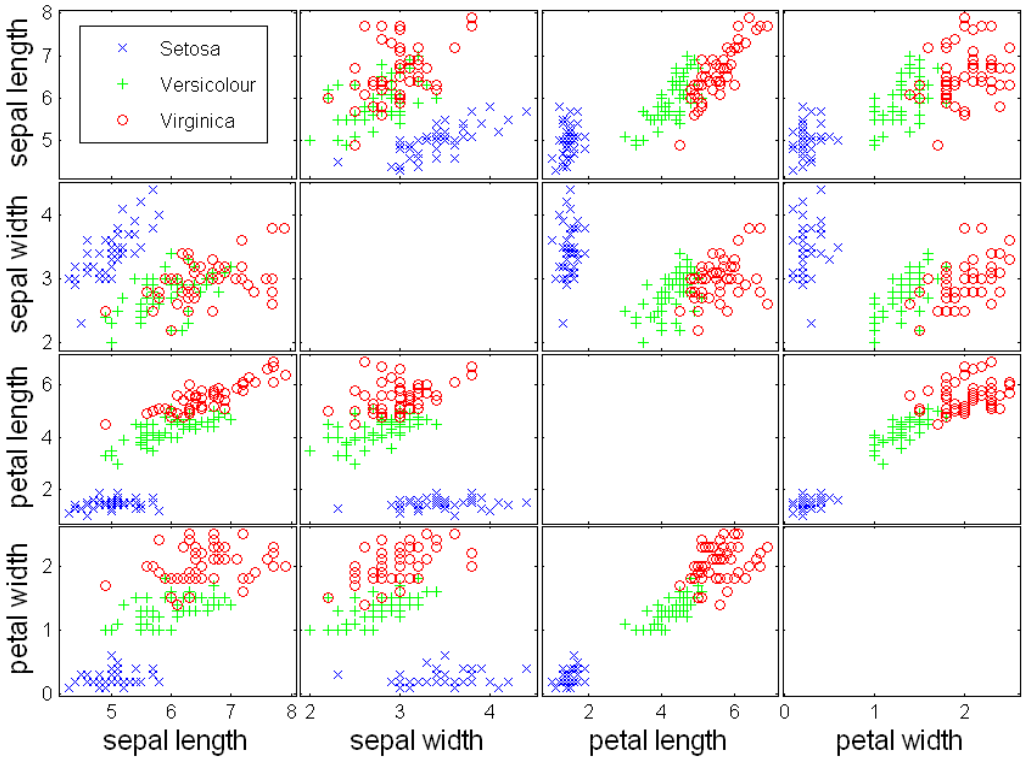
c. Scatter plots

Nilai atribut menjelaskan posisi. Scatter plot berguna untuk mendapatkan ringkasan data hubungan antara beberapa pasangan atribut. Terdapat tiga jenis analisa yang dapat dilakukan dengan menggunakan *scatter plot*:

a) Scatter plot dapat menunjukkan hubungan (korelasi) antara dua variabel/atribut dan juga dapat digunakan untuk mendeteksi hubungan non linier antar dua variabel/atribut.
b) Ketika label dari kelas tersedia *scatter plot* dapat digunakan untuk menyelidiki derajat kedua atribut dalam memisahkan kelas
c) Menganalisa pencilan/*outlier.*

Grafik scatter plot digunakan untuk menampilkan hubungan antara dua variabel. grafik scatter plot menempatkan satu variabel pada sumbu vertikal dan variabel yang berbeda pada sumbu horizontal. setiap potongan data kemudian diplot sebagai titik diskrit pada grafik. dalam grafik scatter plot, kedua sumbu X dan Y menampilkan nilai - sebuah grafik XY tidak memiliki sumbu kategori. Sumbu X mewakili nilai abstrak yang tidak bergantung pada variabel lain, yang disebut sebagai variabel independen. nilai Y ditempatkan pada sumbu vertikal dan mewakili variabel dependen.

Dalam scatter plot, setiap objek data diplot sebagai titik dalam bidang dengan menggunakan nilai-nilai dari dua atribut sebagai koordinat x dan y. diasumsikan bahwa atribut adalah bernilai interger atau real. Gambar dibawah ini menunjukkan scatter plot untuk setiap pasang atribut dari data set Iris. Spesies yang berbeda dari Iris ditunjukkan dengan tanda yang berbeda. Penyusunan scatter plot dari pasangan atribut dalam format tabu;ar ini, yang dikenal sebagai scatter plot matrix, memberikan cara yang terorganisasi untuk mengevaluasi sejumlah scatter plot secara simultan.



d. Statistika Ringkas

Nilai Tengah : dari sekumpulan data (distribusi), ada beberapa harga/nilai yang dapat kita anggap sebagai wakil dari kelompok data tersebut. Nilai-nilai yang biasa digunakan untuk mewakili data tersebut adalah mean, median, dan modus. nilai-nilai tersebut disebut sebagai nilai tengah.

a) Mean

Rata-rata hitung atau mean adalah nilai yang baik mewakili suatu data. Nilai ini sangat sering dipakai dan bahkan yang paling banyak dikenal dalam menyimpulkan sekelompok data.

Misalnya kalau kita mempunyai n pengamat yang terdiri dari x1, x2, x3.....xn, maka nilai rata-rata adalah:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

Sifat dari mean:

1) merupakan wakil dari keselurahan nilai;

2) mean sangat dipengaruhi nilai ekstrem baik ekstrem kecil maupun ekstrem besar;

3) nilai mean berasal dari semua nilai pengamatan

b) Median

Adalah nilai tengah dari nilai-nilai pengamatan setelah disusun secara teratur menurut besarnya data. Nilai ini dipengaruhi oleh letak data dalam urutannya, sehingga nilai ini sering disebut dengan “rata-rata posisi”. Karena nilai median berada di tengah-tengah dari suatu gugus data (yang disusun berurutan), maka akan terdapat 50% dari jumlah data yang letaknya di bawah median, dan 50% dari jumlah yang lain ada di atas median. Untuk data yang tidak dikelompokkan, median adalah nilai yang terletak pada posisi:
(N + 1) / 2; dimana N menunjukkan jumlah observasi keseluruhan.

Median adalah nilai yang terletak pada observasi yang di tengah, kalau data tersebut telah disusun (array). Nilai median disebut juga nilai letak. Nilai media adalah nilai pada posisi tersebut.
contoh: kalau berat badan lima orang dwasa di atas disusun menurut besar kecilnya nilai, maka diapatkan susunan seperti berikut 48, 52, 56, 62, 67 kg.

Nilai observasi ketiga adalah 56, maka dikatakan median adalah 56 kg. Kalau datanya genap, posisi media terletak antara dua nilai, maka nilai median adalah rata-rata dari kedua nilai tersebut.
Contoh : pengamatan di atas tidak lima orang, tetapi enam orang, 48, 52, 56, 62, 67, 70 kg.

Posisi median adalah pengamatan ke 3.4. maka nilai median adalah jumlah pengamatan ketida dan keempat dibagi dua.

Sifat dari median: Nilai Median tidak terpengaruh oleh data ekstrim.

c) Modus (Mode)

Adalah nilai yang mempunyai frekuensi terbanyak dalam kumpulan data. Ukuran ini biasanya digunakan untuk mengetahui tingkat seringnya terjadi suatu peristiwa. (ukuran ini (sebenarnya) cocok digunakan untuk data berskala nominal.
Pada data yang tidak dikelompokkan, modus diperoleh dengan menghitung frekuensi dari masing-masing niai pengamatan, dan kemudian dicari nilai pengamatan yang mempunyai frekuensi observasi paling banyak (nilai data yang paling sering muncul).

d) Range

Adalah ukuran variasi yang dihitung dari selisih antara nilai yang terbesar dengan nilai terkecil. Range sangat mudah dihitung tetapi memang sangat jarang digunakan sebagai ukuran penyimpangan. Biasanya range digunakan

dalam pengendalian mutu, atau dalam melihat fluktuasi harga, dan ramalan cuaca. Range = Xh – Xl,
dimana :
Xh = data tertinggi
Xl = data terendah
Misalnya dari contoh gugus data di depan, kita ketahui bahwa data tertinggi adalah 53 dan data terkecil adalah 15, berarti range dari gugus data kita adalah:
Range = 53 – 15
= 38

e) Variasi dan Standar Deviasi
Standar deviasi ini merupakan ukuran variasi yang paling banyak digunakan, karena nilainya paling memenuhi kriteria statistika.Standar deviasi adalah akar kuadrat dari variasi. Variasi dicari dengan menghitung selisih dari setiap elemen data dengan rata-rata. Variasi dibedakan antara Variasi populasi (σ^2) dengan variasi sampel (S^2), demikian juga kita mengenal standar deviasi populasi (σ) dan standar deviasi sampel (S).

Rumus Variasi untuk sampel dan populasi adalah sebagai berikut:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^{\,2} - \mu^2$$

Sedangkan standar deviasi populasi dan sampel adalah:

$$\sigma = \sqrt{\sigma^2} \qquad\qquad s = \sqrt{s^2}$$

DAFTAR PUSTAKA

Acuna E. Dan Rodriguez C., *The Treatment of Missing values and its Effect in The Classifier Accuracy*. In D. Banks, L. Springer-Verlag Berlin- Heidelberg, 639-648. 2004.

Agrawal dan R. Srikant, ”*Privacy Preserving Data Mining*”, ACM SIGMOD, 2000.

Jermyn, P., Dixon, M., & Read, B. J. (1999). Preparing Clean Views of Data For Data Mining. *ERCIM Work on Database Res*.

Kolcz, A., Chowdury, A., & Alspector, J. (2003). Data Duplication : An Imbalance Problem? *Workshop on Learning from Imbalanced Datasets II, ICML*.

Cios K.J dan Kurgan L., *Trends in Data Mining and Knowledge*

Discovery. In N.R. Pal, L.C. Jain, dan Teoderesku N., editor, *Knowledge*

Discovery in Advanced Infromation System. Springer, 2002.

Han J & Kamber M. 2006. *Data mining – Concept and Techniques.*Morgan-Kauffman, San Diego

Tan P., Michael S., & Vipin K. 2006. *Introduction to Data mining*. Pearson Education, Inc.