BRIN
BADAN RISET
DAN INOVASI NASIONAL

Modul
DATA MINING
Pelatihan Jabatan Fungsional
Analis Data Ilmiah

Fungsi Layanan Pengembangan Kompetensi Kedinasan
Direktorat Pengembangan Kompetensi
Deputi Sumber Daya Manusia
2022

BRIN
BADAN RISET
DAN INOVASI NASIONAL

MODUL
DATA MINING

PELATIHAN JABATAN FUNGSIONAL
ANALIS DATA ILMIAH

DIREKTORAT PENGEMBANGAN KOMPETENSI

BADAN RISET DAN INOVASI NASIONAL

TAHUN 2022

Penanggung Jawab:

1. Edy Giri Racman Putra, Ph.D.
2. Nining Setyowati Dwi Andayani, S.E., M.M.
3. Raden Arthur Ario Lelono, Ph.D.
4. Alpha Fadila Juliana Rahman, S.Pd., M.Pd.

Tim Penyusun Modul:

1. Hendro Subagyo, M.Eng.
2. Dr. Rifki Sadikin, M.Kom.
3. Slamet Riyanto, S.Kom., M.M.S.I.
4. Dr. Tr. Lindung Parningotan Manik
5. Dimas Sony Dewantara, S. Kom.
6. Ijang Pemana Sidik. M.Pd.

Diterbitkan oleh:

Direktorat Pengembangan Kompetensi - BRIN

Gedung B.J. Habiebie, Jalan M.H. Thamrin No. 8

Jakarta Pusat 10340

Diterbitkan pertama kali tahun 2022

KATA PENGANTAR

Pengembangan kompetensi Aparatur Sipil Negara (ASN) khususnya Pegawai Negeri Sipil (PNS) dalam mengembangkan karier jabatan fungsionalnya menjadi suatu tuntutan sehingga mampu menjalankan tugas dan fungsinya dengan sebaik – baiknya seingga mampu memberikan kontribusi yang nyata terhadap bangsa dan Negara Kesatuan Republik Indonesia (NKRI).

Badan Riset dan Inovasi Nasional (BRIN) sebagai Instansi pembina 11 (sebelas) jabatan fungsional yang meliputi peneliti, perekayasa, pengembangan teknologi nuklir, analis perkebunrayaan, analis pemanfaatan iptek, analis data ilmiah, kurator koleksi hayati, penata penerbitan ilmiah, teknisi perkebunrayaan, teknisi penelitian dan perekayasaan, dan pranata nuklir. BRIN melalui kedeputian Sumber Daya Manusia Ilmu Pengetauan dan Teknologi (SDMI) bertanggung jawab dalam penyelenggaraan pengembangan kompetensi 11 (sebelas) jabatan fungsional tersebut.

Kedeputian SDMI – BRIN melalui Direktorat Pengembangan Kompetensi sebagai penyelenggara pengembangan komptensi jabatan fungsional bertanggung jawab dalam menyiapkan kebutuhan tersebut baik berupa pengelolaan pembelajaran, fasilitator, modul, bahan ajar dan sebagainya, yang merujuk kepada regulasi peraturan BRIN nomor 28 tahun 2022 tentang pedoman pelatihan pembentukan jabatan fungsional peneliti, peraturan BRIN nomor 29 tahun 2022 tentang pedoman pelatihan jabatan fungsional kurator koleksi hayati, peraturan BRIN nomor 30 tahun 2022 tentang pedoman pelatihan jabatan fungsional analis pemanfaatan iptek, peraturan BRIN nomor 31 tahun 2022 tentang pedoman pelatihan jabatan fungsionalanalis data ilmiah, peraturan BRIN nomor 32 tahun 2022 tentang pedoman pelatihan jabatan fungsional analis perkebunrayaan, peraturan BRIN nomor 33 tahun 2022 tentang pedoman pelatihan jabatan fungsional teknisi perkebunrayaan, dan peraturan BRIN nomor 34 tahun 2022 tentang pedoman pelatihan jabatan fungsional penata penerbitan ilmiah.

Kami megucapkan syukur ke hadirat Tuhan Yang Maha Esa, atas berkat rahmat-Nya, modul pelatihan jabatan fungsional analis data ilmiah/ penata

penerbitan ilmiahyang berjudul Data Mining dapat diselesaikan tepat waktu. Modul ini digunakan dalam pelatihan jabatan fungsional yang dibina oleh BRIN yang diselenggarakan oleh Kedeputian SDMI - BRIN melalui Direktorat Pengembangan Kompetensi. Kami berharap modul ini dapat memberikan manfaat dan kontribusi dalam meningkatkan dan mengembangkan kompetensi jabatan fungsional yang dibina BRIN.

Jakarta, Desember 2022

Plt. Deputi Sumber Daya Manusia
Ilmu Pengetauan dan Teknologi
Badan Riset dan Inovasi Nasional

(*Tanda tangan*)

Edy Giri Rachman Putra, Ph.D.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

PENDAHULUAN

A. Deskripsi Singkat

Mata pelatihan ini menjelaskan tentang konsep data mining, model data mining, pola data, data summarization, kategorisasi data, data rescalling, data scoring, missing value, data smoothing, dan transformasi data.

B. Alokasi Waktu

Pelatihan Jabatan Fungsional Analis Data Ilmiah dapat diselenggarakan dengan tiga metode. Setiap metode memiliki alokasi waktu yang berbeda. Berikut adalah alokasi waktu pembelajaran untuk mata pelatihan *data mining*.

Metode	*Syncronous*	*Asynchronous*	Total
Klasikal	12 JP	-	12 JP
Bauran	11 JP	3 JP	14 JP
Jarak Jauh	3 JP	12 JP	15 P

C. Tujuan Pembelajaran

1. **Hasil Belajar**

   Setelah mengikuti pembelajaran ini, peserta diharapkan mampu melakukan data mining sesuai dengan kaidah ilmiah yang benar.

2. **Indikator Hasil Belajar**

   Setelah selesai pembelajaran diharapkan peserta mampu:

   a) menjelaskan konsep penambangan data *(data mining)*;
   b) memahami metode-metode dalam *data mining:* klasifikasi, *clustering,* asosiasi;
   c) menganalisis pola data;
   d) merangkum data (data *summarization*);
   e) mengkategorikan data;

f) melakukan *data rescalling*;

g) melakukan *data scoring*;

h) mendeteksi dan mengatasi *missing value;*

i) mengoreksi *noise* yang ada dalam data;

j) melakukan transformasi data; dan

k) mengaplikasikan minimal 1 dari berbagai metode *data mining*.

D. Materi Pokok

Mata pelatihan ini terdiri dari sebelas Materi Pokok, yaitu:

a) Pengertian *data mining*;

b) Metode *data mining*;

c) Pola data;

d) *Data summarization*;

e) Kategorisasi data;

f) *Data rescalling*;

g) *Data scoring*;

h) *Missing value*;

i) *Data smoothing;*

j) Transformasi data;

k) Praktik data mining.

MATERI POKOK 1:

PENGERTIAN *DATA MINING*

Indikator Hasil Belajar:

Peserta mampu menjelaskan konsep penambangan data (data mining). Peserta dinyatakan lulus manakala mampu untuk :

1. menjelaskan pengertian *data mining,*
2. menjelaskan jenis data yang dapat ditambang,
3. menjelaskan pola data yang dapat ditambang, dan
4. menjelaskan tahapan *data mining.*

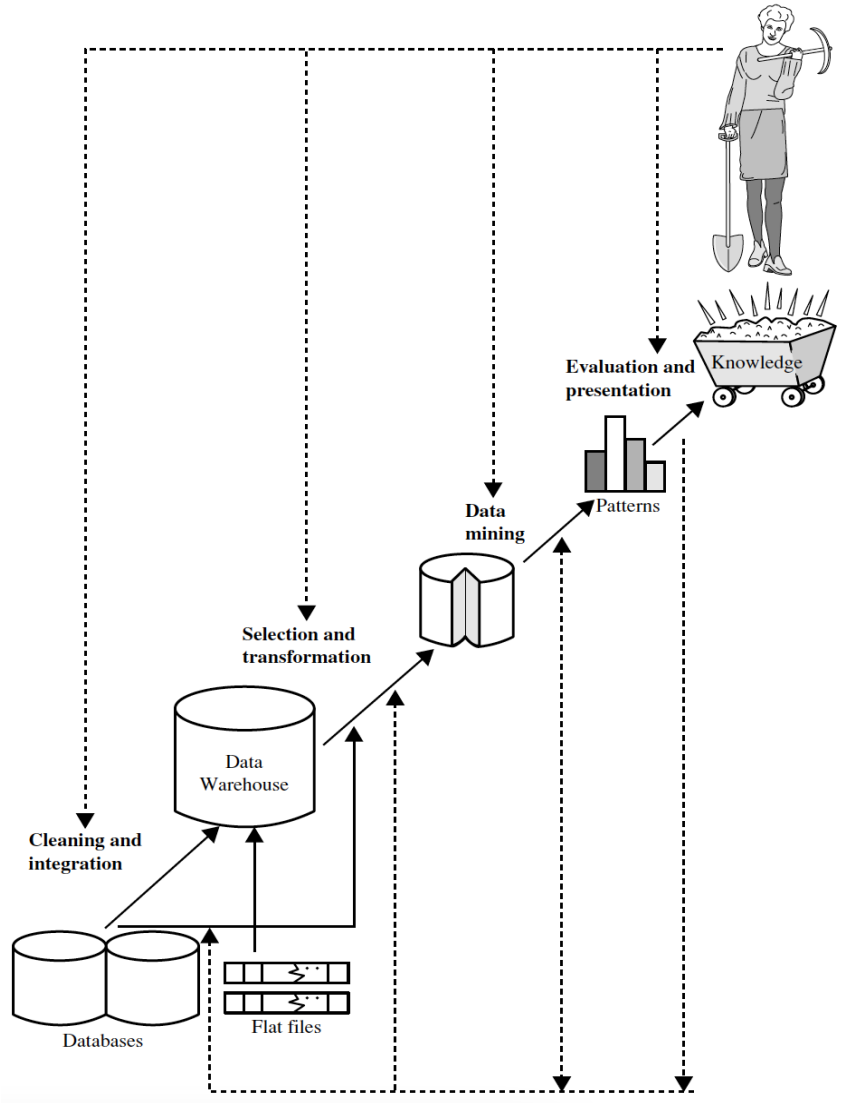
A. Konsep *Data Mining*

Penambangan data (*data mining*) dalam bahasa sederhana adalah menemukan pola yang berguna dalam data. *Data mining* juga disebut sebagai *machine learning* dan *predictive analytics*. Selain itu, banyak istilah lain yang memiliki arti serupa (sinonim) dengan *data mining*, misalnya: *knowledge mining from data*, *knowledge extraction*, *data/pattern analysis*, *data archaeology*, dan *data dredging*. Banyak orang menganggap *data mining* sinonim dengan *knowledge discovery from data* (KDD), namun sebagian lain melihat *data mining* hanya langkah penting dalam proses *knowledge discovery.*

Data mining diawali dengan data yang memuat larik sederhana dari beberapa observasi numerik hingga matriks kompleks dari jutaan observasi yang memiliki ribuan variabel. Pekerjaan dalam *data mining* menggunakan beberapa metode komputasi khusus untuk menemukan struktur bermakna dan berguna dalam sebuah data. Metode komputasi tersebut berasal dari bidang *statistics*, *machine learning*, dan *artificial intelligence*. Disiplin *data mining* berkaitan erat dengan bidang lain seperti: *database systems*, *data cleansing*, *visualization*, *exploratory data analysis*, dan *performance evaluation*.

Dalam *knowledge discovery* perlu urutan iteratif dari langkah-langkah yang

harus dilalui mencakup: *data cleansing*, *data integration*, *data selection*, *data transformation*, *data mining, pattern evaluation*, dan *knowledge presentation* yang ditunjukkan seperti pada Gambar 1. Oleh karena itu, modul ini mengadopsi pandangan luas tentang fungsi penambangan data yaitu: proses menemukan pola dan pengetahuan yang menarik dari sejumlah besar data.



Gambar 1 Data mining sebagai langkah dalam proses knowledge discovery
Sumber: Han, J. et al, 2011

Berdasarkan ilustrasi Gambar 1 dapat dijelaskan setiap tahap dalam proses *knowledge discovery* yaitu:

1. *data cleansing*: penghapusan terhadap data yang tidak lengkap, tidak konsisten, maupun *noise*
2. *data integration*: data yang berasal dari berbagai sumber dapat

dikombinasikan

3. *data selection*: pemilihan data relevan untuk analisis yang diperoleh dari database
4. *data transformation*: data diubah dan dikonsolidasikan ke dalam bentuk yang sesuai untuk penambangan dengan melakukan operasi ringkasan atau agregasi
5. *data mining*: proses penting di mana metode cerdas diterapkan untuk mengekstraksi pola data
6. *pattern evaluation*: tahap untuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan berdasarkan ukuran ketertarikan
7. *knowledge presentation*: tahap akhir, teknik representasi visualisasi dan pengetahuan digunakan untuk menyajikan pengetahuan yang ditambang kepada pengguna agar mudah dipahami.

B. Jenis Data Apa yang dapat Ditambang?

Bentuk paling umum dari data untuk penambangan data yang berasal dari database, data warehouse, dan data transaksi. Sehingga penambangan data dapat diterapkan pada semua jenis data selama data tersebut memuliki makna. Penambangan data juga dapat diterapkan ke dalam bentuk data lain, misalnya *data streams*, *sequence data*, *graph*, *spatial data*, *multimedia data* dan www.

1. Database

Database management system (DBMS) terdiri dari kumpulan data yang saling terkait (*database*) dan program perangkat lunak untuk mengelola dan mengakses data merupakan bagian dari sistem basis data. Program perangkat lunak berfungsi mendefinisikan struktur basis data serta penyimpanan data untuk menentukan dan mengelola akses data secara bersama, atau terdistribusi. Selain itu, program perangkat lunak juga berfungsi untuk memastikan konsistensi dan keamanan informasi yang disimpan meskipun tersendat atau upaya akses yang tidak sah.

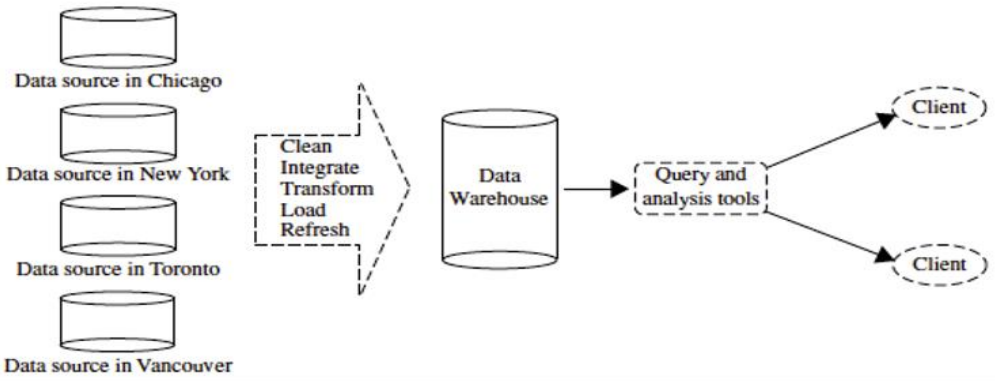
Basis data relasional adalah kumpulan tabel yang masing-masing diberi nama unik. Setiap tabel terdiri dari satu set atribut (*column* atau *field*) yang mampu menyimpan satu set tupel besar (*record* atau *row*).

Saat menambang basis data relasional, kita bisa melangkah lebih jauh dengan mencari tren atau pola data. Misalnya, sistem penambangan data dapat menganalisis data pelanggan untuk memprediksi risiko kredit pelanggan baru berdasarkan pendapatan, usia, dan informasi kredit sebelumnya. Sistem penambangan data juga dapat mendeteksi adanya penyimpangan, yaitu item dengan penjualan yang meningkat tajam dari yang diharapkan dibandingkan dengan tahun sebelumnya. Penyimpangan semacam itu kemudian dapat diselidiki lebih lanjut. Misalnya, penambangan data mungkin menemukan bahwa telah terjadi perubahan kemasan suatu barang atau penurunan harga yang signifikan.

2. *Data Warehouses*

Data warehouse (gudang data) adalah gudang informasi yang dikumpulkan dari berbagai sumber, disimpan di bawah skema terpadu, dan biasanya berada di satu situs. Gudang data dibangun melalui proses pembersihan data, integrasi data, transformasi data, pemuatan data, dan penyegaran data berkala.

Gudang data biasanya dimodelkan dengan struktur data multidimensi, yang disebut kubus data, di mana setiap dimensi sesuai dengan atribut atau sekumpulan atribut dalam skema, dan setiap sel menyimpan nilai dari beberapa ukuran agregat seperti jumlah atau jumlah penjualan. Kubus data memberikan tampilan data multidimensi dan memungkinkan prakomputasi dan akses cepat dari data yang diringkas.



Gambar 2 Contoh kerangka kerja data warehouses "Electronics"
Sumber: Han, J. et al, 2011

Meskipun gudang data membantu mendukung analisis data, namun untuk analisis mendalam membutuhkan alat tambahan untuk penambangan data. Penambangan data multidimensi (juga disebut penambangan data multidimensi eksplorasi) melakukan penambangan data dalam ruang multidimensi melalui *Online Analytical Processing* (OLAP). Artinya, ini memungkinkan eksplorasi berbagai kombinasi dimensi pada berbagai tingkat perincian dalam penambangan data, dan dengan demikian memiliki potensi lebih besar untuk menemukan pola menarik yang merepresentasikan pengetahuan.

3. Transactional Data

Secara umum, setiap *record* dalam basis data transaksional menangkap sebuah transaksi, misalnya pemesanan pesawat, pembelian produk, atau mengubah password. Setiap transaksi mencakup nomor identitas transaksi yang unik dan daftar transaksi (misalnya pembayaran). Kegiatan transaksi dapat disimpan dalam sebuah tabel dengan satu cantuman (*record*) per transaksi.

Sebagai seorang analis harus memiliki kemampuan untuk membaca data, hal ini karena tabel yang satu berkaitan dengan tabel yang lain. Biasanya, identitas transaksi hanya mengandung kode barang atau kode transaksi dalam bentuk angka unik. Sehingga bisa jadi sebuah transaksi memuat beberapa daftar kode transaksi. Fragmentasi basis data transaksional ditunjukkan dalam Tabel 1.1.

Tabel 1 Fragmentasi basis data transaksional

trains_id	*list_of_items_id*
T1000	I4, I6, I7, I70
T2000	I3, I8, I20
T3000	I2, I9, I12
…	…

Mungkin kita akan bertanya, “produk mana yang terjual berbarengan?” Ini bagian dari *market basket data analysis* yang memungkinkan untuk mengelompokkan produk bersama-sama sebagai strategi untuk meningkatkan penjualan. Misalnya, printer biasanya dibeli bersama dengan computer,

4. **Jenis Data Lainnya**

Selain data berasal dari basis data relasional, gudang data, dan data transaksi, ada beberapa jenis data lain yang memiliki bentuk dan struktur serbaguna dan makna semantik. Jenis data tersebut diantaranya: *sequence data* (catatan sejarah, data urutan biologis, dan deret waktu), *data stream* (cctv dan data sensor yang terus menerus ditransmisikan), *spatial data* (peta), *engineering design data* (rancang bangun, komponen sistem), *media data* (termasuk teks, video, image, dan audio), *graph and networked data* (jaringan informasi media sosial), dan Web (gudang informasi yang sangat besar dan tersebar luas di Internet).

Berbagai jenis pengetahuan dapat ditambang dari jenis data tersebut. Kita dapat menambang *data stream* jaringan komputer untuk mendeteksi intrusi berdasarkan anomali aliran pesan, yang dapat ditemukan dengan pengelompokan, konstruksi dinamis model aliran atau dengan membandingkan arus yang sering pola dengan mereka pada waktu sebelumnya. *Dengan spatial data*, kita dapat mencari pola yang menggambarkan perubahan tingkat kemiskinan metropolitan berdasarkan jarak kota dari jalan raya utama.

Dengan menambang data teks, seperti literatur tentang *data mining* dari sepuluh tahun terakhir, kita dapat mengidentifikasi evolusi topik hangat di lapangan. Dengan menambang komentar atau ulasan pengguna tentang produk (biasanya dikirimkan sebagai pesan teks singkat atau rating), kita dapat menilai sentimen pelanggan dan memahami seberapa baik suatu produk dianut oleh pasar

C. Jenis Pola Apa yang Dapat Ditambang

Ada beberapa fungsi penambangan data yang terbagi dalam karakterisasi dan diskriminasi (penambangan pola, asosiasi, korelasi, klasifikasi, regresi, analisis pengelompokan, dan analisis outlier). Secara umum, tugas-tugas tersebut dapat diklasifikasikan menjadi dua kategori, yaitu: deskriptif dan prediktif. Tugas deskriptif mencirikan properti data dalam kumpulan data target, sedangkan tugas prediktif melakukan induksi pada data saat ini untuk membuat prediksi.

1. Deskripsi Kelas/Konsep: Karakterisasi dan Diskriminasi

Entri data dapat direlasikan dengan kelas atau konsep, misalnya sebuah toko elektronik memiliki kelas barang yang dijual mencakup komputer, printer, serta keyboard sedangkan konsep pelanggan mencakup Pembelanja besar dan Pembelanja anggaran. Hal ini dapat bermanfaat untuk menggambarkan kelas dan konsep secara individu dalam istilah yang diringkas, singkat dan tepat. Deskripsi kelas atau konsep seperti itu disebut deskripsi kelas/konsep.

Deskripsi ini dapat diturunkan dengan menggunakan tiga karakteristik yaitu: (1) karakterisasi data yaitu meringkas data kelas yang diamati (sering disebut kelas target), atau (2) diskriminasi data yaitu membandingkan kelas target dengan satu atau sekumpulan kelas komparatif (sering disebut kelas kontras), atau (3) karakterisasi dan diskriminasi data.

Karakterisasi data adalah ringkasan dari karakteristik umum atau fitur dari kelas data target. Data yang sesuai dengan kelas ditentukan oleh pengguna biasanya dikumpulkan melalui kueri. Misalnya, produk perangkat keras mengalami peningkatan penjualan 20%, sehingga data terkait dengan produk tersebut dapat ditampilkan melalui eksekusi *query* SQL pada basis data penjualan. Keluaran karakterisasi data dapat disajikan dalam berbagai bentuk. Contohnya termasuk diagram lingkaran, diagram batang, kurva, kubus data multidimensi, dan tabel multidimensi, termasuk tab silang.

Diskriminasi data adalah perbandingan fitur umum dari objek data kelas target terhadap fitur umum objek dari satu atau beberapa kelas yang kontras. Kelas target dan kontras dapat ditentukan oleh pengguna, dan

objek data yang sesuai dapat diambil melalui kueri basis data. Misalnya, pengguna mungkin ingin membandingkan fitur umum produk perangkat keras dengan penjualan yang meningkat sebesar 20% tahun lalu dengan penjualan yang menurun setidaknya 10% selama periode yang sama. Metode yang digunakan untuk diskriminasi data mirip dengan yang digunakan untuk karakterisasi data.

Bentuk penyajian keluaran mirip dengan deskripsi karakteristik, meskipun deskripsi diskriminasi harus menyertakan ukuran komparatif yang membantu membedakan antara kelas sasaran dan kelas yang kontras. Deskripsi diskriminasi yang dinyatakan dalam bentuk aturan disebut sebagai aturan diskriminan.

2. Menambang Pola, Asosiasi, dan Korelasi

Frequent pattern adalah pola yang sering muncul dalam data. Ada banyak jenis *frequent pattern*, diantaranya *frequent itemset*, *frequent subsequences* (juga dikenal sebagai *sequential pattern*), dan *frequent substructures*. Kumpulan item yang sering biasanya mengacu pada kumpulan item yang sering muncul bersama dalam kumpulan data transaksi. Misalnya, indomie dan saos, yang sering dibeli bersama dalam toko makanan oleh banyak pelanggan. Urutan yang sering terjadi, seperti pelanggan cenderung membeli laptop terlebih dulu baru diikuti printer kemudian kartu memori. Pola ini disebut *sequent pattern*.

Sebuah substruktur dapat mengacu pada bentuk struktural yang berbeda (misalnya, grafik, pohon, atau kisi) yang dapat digabungkan dengan kumpulan item atau urutan. Jika substruktur sering terjadi disebut *structured pattern* (*frequent*). Menambang pola sering mengarah pada penemuan asosiasi dan korelasi yang menarik di dalam data.

Analysis Asosiasi. Seandainya sebagai seorang manajer pemasaran pada "Electronics", Anda ingin mengetahui produk mana yang sering dibeli bersama-sama (transaksi yang sama). Contoh aturan seperti berikut, yang ditambang dari basis data "Electronics".

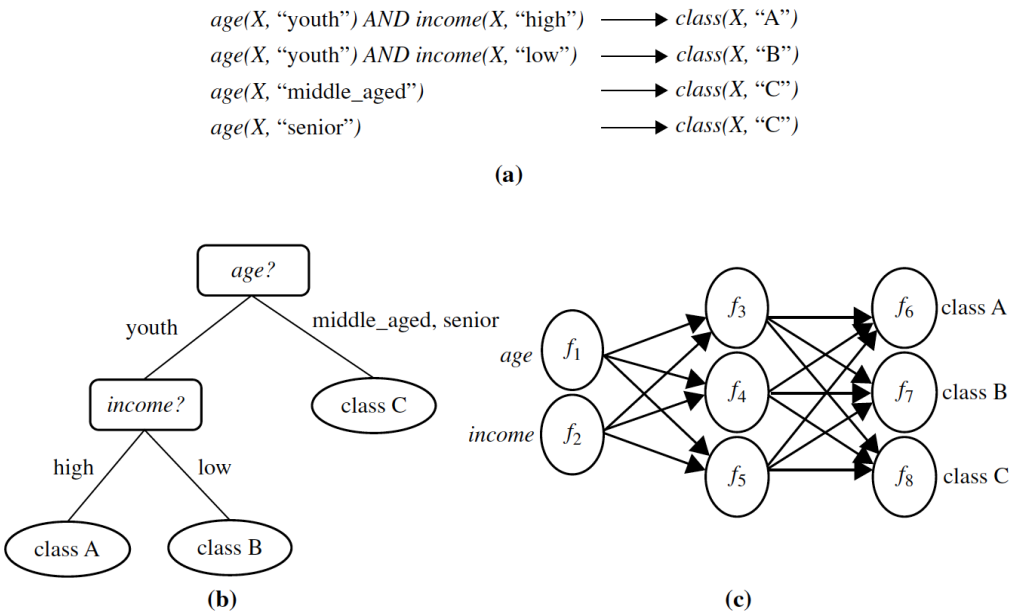
buys(X, notebook)=>buys(X, printer) [support => 3%, confidence => 60%],

di mana X adalah variabel yang mewakili pelanggan. Keyakinan, atau kepastian, sebesar 60% berarti jika pelanggan membeli *notebook*, ada kemungkinan 60% akan membeli *printer* juga. Dukungan 3% berarti bahwa 3% dari semua transaksi yang dianalisis menunjukkan bahwa *notebook* dan *printer* dibeli bersama. Aturan asosiasi ini melibatkan satu atribut atau predikat (yaitu, membeli) yang berulang. Aturan asosiasi yang berisi predikat tunggal disebut sebagai aturan asosiasi dimensi tunggal. Menjatuhkan notasi predikat, aturan dapat ditulis hanya sebagai "printer notebook [3%, 60%]."

3. Klasifikasi dan Regresi untuk Analisis Prediktif

Klasifikasi adalah proses menemukan model (atau fungsi) yang menggambarkan dan membedakan kelas data atau konsep. Model diturunkan berdasarkan analisis sekumpulan data pelatihan (yaitu, objek data yang label kelasnya diketahui). Model ini digunakan untuk memprediksi label kelas objek yang label kelasnya tidak diketahui.

“Bagaimana model turunan disajikan?” Model turunan dapat direpresentasikan dalam berbagai bentuk, seperti aturan klasifikasi (yaitu, aturan IF-THEN), pohon keputusan, rumus matematika, atau jaringan saraf. Pohon keputusan adalah struktur pohon seperti diagram alir, di mana setiap node menunjukkan uji pada nilai atribut, setiap cabang mewakili hasil uji, dan daun pohon mewakili kelas atau distribusi kelas. Pohon keputusan dapat dengan mudah diubah menjadi aturan klasifikasi. Jaringan saraf, bila digunakan untuk klasifikasi, biasanya merupakan kumpulan unit pemrosesan seperti neuron dengan koneksi berbobot antar unit. Ada banyak metode lain untuk membangun model klasifikasi, seperti klasifikasi *Naive Bayesian classification*, *Support Vector Machines*, dan *k-nearest-neighbor classification*.



Gambar 3 Model klasifikasi dapat ditunjukkan dalam berbagai bentuk (a) aturan IF-THEN, (b) pohon keputusan, atau (c) jaringan syaraf.
Sumber: Han, J. et al, 2011

Sedangkan klasifikasi memprediksi label kategori (diskrit, tidak berurutan), model regresi fungsi bernilai kontinu. Artinya, regresi digunakan untuk memprediksi nilai data numerik yang hilang atau tidak tersedia daripada label kelas (diskrit). Istilah prediksi mengacu pada prediksi numerik dan prediksi label kelas. Analisis regresi adalah metodologi statistik yang paling sering digunakan untuk prediksi numerik, meskipun ada juga metode lain. Regresi juga mencakup identifikasi tren distribusi berdasarkan data yang tersedia. Klasifikasi dan regresi mungkin perlu didahului oleh analisis relevansi, yang berupaya mengidentifikasi atribut secara signifikan relevan dengan proses klasifikasi dan regresi. Atribut tersebut akan dipilih untuk proses klasifikasi dan regresi. Atribut lain yang tidak relevan, kemudian dapat dikecualikan dari pertimbangan.

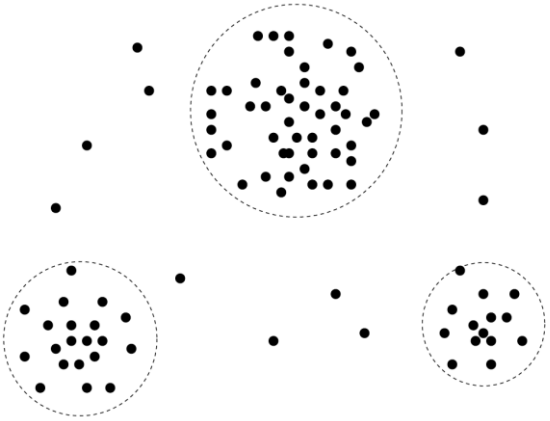
Contoh klasifikasi dan regresi, misalkan sebagai manajer penjualan "Electronics", Anda ingin mengklasifikasikan barang-barang di toko berdasarkan tiga jenis tanggapan terhadap kampanye penjualan: baik, ringan, dan tidak ada tanggapan. Anda ingin memperoleh model untuk masing-masing dari ketiga kelas ini berdasarkan fitur deskriptif item, seperti harga, merek, warna, tempat dibuat, tipe, dan kategori. Klasifikasi yang

dihasilkan harus secara maksimal membedakan setiap kelas dari yang lain. Misalkan klasifikasi yang dihasilkan dinyatakan sebagai pohon keputusan. Pohon keputusan dapat mengidentifikasi harga sebagai faktor tunggal yang paling membedakan ketiga kelas tersebut. Pohon tersebut dapat mengungkapkan bahwa, selain harga, ciri-ciri lain yang membantu untuk lebih jauh membedakan objek dari setiap kelas satu sama lain termasuk merek dan warna. Pohon keputusan tersebut dapat membantu memahami dampak dari kampanye penjualan tertentu dan merancang kampanye yang lebih efektif di masa mendatang.

4. Analisis Klaster

Tidak seperti klasifikasi dan regresi yang menganalisis kumpulan data pelatihan, klastering menganalisis objek data tanpa berkonsultasi dengan label kelas (pelatihan). Dalam banyak kasus, data pelatihan mungkin tidak ada di awal. Klastering dapat digunakan untuk menghasilkan label kelas untuk sekelompok data. Objek-objek tersebut diklasterkan atau dikelompokkan berdasarkan prinsip memaksimalkan kesamaan *intraclass* dan meminimalkan kesamaan *interclass*. Artinya, klaster-klaster objek dibentuk sedemikian rupa sehingga objek-objek di dalam sebuah klaster memiliki kesamaan yang tinggi dibandingkan satu sama lain, tetapi agak berbeda dengan objek-objek di klaster lain. Setiap klaster yang terbentuk dapat dipandang sebagai kelas objek. Pengelompokan juga dapat memfasilitasi pembentukan taksonomi, yaitu pengorganisasian pengamatan ke dalam hierarki kelas yang mengelompokkan kejadian serupa bersama-sama.

Analisis kluster dapat dilakukan pada data pelanggan "Electronics" untuk mengidentifikasi subpopulasi pelanggan yang homogen. Klaster ini dapat mewakili kelompok sasaran individu untuk pemasaran. Gambar 1.4 menunjukkan plot 2-D pelanggan terhubung dengan lokasi pelanggan di kota.



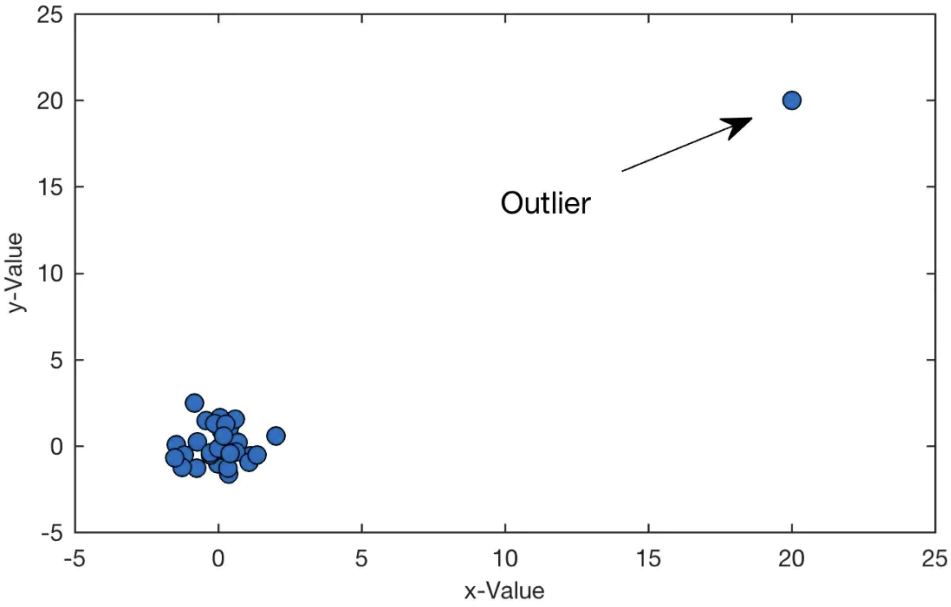
Gambar 4 Plot 2-D dari data pelanggan sehubungan dengan lokasi pelanggan di kota, menampilkan tiga kelompok data.
Sumber: Han, J. et al, 2011

5. Analisis *Outlier*

Sangat memungkinkan kumpulan data berisi objek yang tidak sesuai dengan perilaku pada umumnya atau berbeda dari model data. Objek data tersebut adalah outlier. Beberapa metode penambangand ata dapat digunakan untuk membuang outlier sebagai noise atau pengecualian. Namun dalam beberapa aplikasi, misalnya deteksi penipuan. Peristiwa langka busa lebih menarik daripada kejadian yang umum terjadi. Analisis data outlier sering disebut sebagai analisis outlier atau anomaly mining.

Uji statistik dapat digunakan untuk mendeteksi outlier yang mengasumsikan model distribusi atau probabilitas untuk data. Selain uji statistik, dapat pula menggunakan ukuran jarak di mana objek yang jauh dari klaster lain dianggap outlier atau dalam bahasa Indonesia disebut sempalan. Metode lain berbasis kepadatan juga dapat mengidentifikasi outlier di wilayah lokal, meskipun terlihat normal dari tampilan distribusi statistik global.

Sebagai contoh, analisis *outlier* dapat mengungkap penipuan penggunaan kartu kredit dengan mendeteksi pembelian dalam jumlah besar yang tidak biasa untuk nomor rekening tertentu dibandingkan dengan biaya reguler yang dikeluarkan oleh rekening yang sama. Nilai outlier juga dapat dideteksi sehubungan dengan lokasi dan jenis pembelian, atau frekuensi pembelian.



Gambar 5 Plot 2-D dari data pelanggan sehubungan dengan lokasi pelanggan di kota, menampilkan tiga kelompok data.
Sumber: Han, J. et al, 2011

D. Rangkuman

Penambangan data adalah proses menemukan pola menarik dari sejumlah besar data. Sebagai proses penemuan pengetahuan, biasanya melibatkan pembersihan data, integrasi data, pemilihan data, transformasi data, penemuan pola, evaluasi pola, dan presentasi pengetahuan. Sebuah pola menarik jika valid pada data uji dengan tingkat kepastian tertentu, baru, berpotensi berguna (misalnya, dapat ditindaklanjuti atau memvalidasi firasat yang membuat pengguna penasaran), dan mudah dipahami oleh manusia. Pola yang menarik mewakili pengetahuan. Ukuran ketertarikan pola, baik objektif maupun subjektif dapat digunakan untuk memandu proses penemuan.

E. Evaluasi

Untuk menguji pemahaman Anda terkait materi Pengertian Data Mining, silahkan kerjakan latihan di bawah ini!

1. Jelaskan langkah-langkah yang terlibat dalam penambangan data bila dilihat sebagai proses penemuan pengetahuan.
2. Tentukan masing-masing fungsi penambangan data berikut: karakterisasi, diskriminasi, asosiasi dan analisis korelasi, klasifikasi, regresi, pengelompokan, dan analisis outlier. Berikan contoh masing-masing fungsi penambangan data, menggunakan database kehidupan nyata yang Anda kenal.
3. Jelaskan perbedaan dan kesamaan antara diskriminasi dan klasifikasi, antara karakterisasi dan klasterisasi, dan antara klasifikasi dan regresi.

MATERI POKOK 2:

METODE *DATA MINING*

Indikator Hasil Belajar:
Peserta memahami berbagai metode dala data mining. Peserta dinyatakan lulus materi ini manakala mampu:

1. Mempraktikkan teknik *Classification,*
2. mempraktikkan teknik *Clustering,*
3. mempraktikkan teknik *Regression,*
4. mempraktikkan teknik *Association Rules,*
5. mempraktikkan teknik *Prediction,*
6. mempraktikkan teknik *Sequential Pattern,* dan
7. mempraktikkan teknik *Outlier Detection.*

A. Classification

Analisis ini digunakan untuk mengambil informasi penting dan relevan tentang data, dan metadata. Metode penambangan data ini membantu mengklasifikasikan data dalam kelas yang berbeda.

Teknik *data mining* ini memiliki proses yang sistematis untuk mendapatkan informasi penting dan relevan tentang metadata (data tentang data) dan data; analisis klasifikasi membantu mengidentifikasi berbagai kategori teknik penambangan data. Analisis klasifikasi terkait erat dengan analisis klaster karena mereka secara efektif membuat pilihan yang lebih baik pada alat penambangan data. Email adalah contoh analisis klasifikasi yang terkenal karena menggunakan algoritma untuk melakukan klarifikasi email tergantung pada apakah itu sah atau spam. Ini dilakukan dengan menggunakan perangkat lunak penambangan data pada Email, misalnya, kata-kata dan lampiran yang menunjukkan apakah itu spam atau email yang sah.

B. Clustering

Analisis clustering adalah teknik penambangan data untuk mengidentifikasi data yang mirip satu sama lain. Proses ini membantu untuk memahami perbedaan dan kesamaan antara data. Jenis teknik penambangan data ini mengidentifikasi

penambangan data yang mirip satu sama lain; analisis pengelompokan membantu pemasar memahami persamaan dan perbedaan dalam data. Karena kluster memiliki ciri umum, kluster dapat digunakan untuk meningkatkan algoritme penargetan. Misalnya, jika sekelompok pelanggan tertentu membeli merek produk tertentu, kampanye khusus dapat dibuat untuk membantu menjual produk tersebut.

Memahami hal ini dapat membantu merek secara efektif meningkatkan tingkat konversi penjualan mereka, meningkatkan kekuatan dan keterlibatan merek. Juga, pembuatan persona adalah hasil dari analisis pengelompokan.

Persona adalah karakter fiksi yang mewakili jenis pengguna yang berbeda dalam target demografis, yang mungkin juga menggunakan situs web, merek, atau produk. Seperti ini, aspek penting dari analisis pengelompokan, persona membantu merek membuat pilihan pemasaran yang cerdas dan membuat kampanye yang hebat.

C. Regression

Analisis regresi adalah metode penambangan data untuk mengidentifikasi dan menganalisis hubungan antar variabel. Ini digunakan untuk mengidentifikasi kemungkinan variabel tertentu, mengingat adanya variabel lain. Analisis regresi membantu merek menentukan ketergantungan variabel. Hal ini didasarkan pada asumsi efek kausal satu arah dari satu variabel ke respon variabel lain.

Sementara variabel independen dapat dipengaruhi satu sama lain, ketergantungan umumnya tidak terpengaruh dua arah, seperti halnya analisis korelasi. Analisis regresi dapat menunjukkan bahwa satu variabel tergantung pada yang lain, bukan sebaliknya.

Karena analisis regresi ideal untuk menentukan kepuasan pelanggan, ini dapat membantu merek menemukan wawasan baru dan berbeda tentang loyalitas pelanggan dan bagaimana faktor eksternal dapat memengaruhi tingkat layanan, seperti kondisi cuaca. Contoh analisis regresi yang baik adalah menggunakan teknik penambangan data ini untuk mencocokkan orang di portal kencan. Banyak situs web menggunakan variabel untuk mencocokkan orang menurut kesukaan, minat, dan hobi mereka.

D. Association Rules

Teknik penambangan data ini membantu menemukan hubungan antara dua atau lebih Item. Ini menemukan pola tersembunyi dalam kumpulan data. Teknik penambangan data ini didasarkan pada penemuan hubungan yang menarik antara variabel dalam database besar. Teknik data mining ini digunakan untuk mengungkap pola tersembunyi dalam data.

Mereka dapat digunakan untuk mengidentifikasi variabel dalam data dan kejadian bersama dari variabel berbeda yang muncul dengan frekuensi paling signifikan. Banyak digunakan di toko ritel, teknik penambangan data aturan asosiasi digunakan untuk menemukan pola dalam data titik penjualan.

Alat penambangan data ini dapat merekomendasikan produk baru, terutama untuk mengetahui jenis produk apa yang direkomendasikan orang kepada orang lain atau produk baru untuk direkomendasikan kepada pelanggan

E. Prediction

Prediksi menggunakan kombinasi teknik penambangan data lainnya seperti tren, pengelompokan, klasifikasi, dll. Ini menganalisis peristiwa atau kejadian masa lalu dalam urutan yang tepat untuk memprediksi peristiwa di masa depan. Ini sangat penting untuk proses pengambilan keputusan ketika Anda memiliki banyak faktor untuk dipertimbangkan. Analitik prediktif sangat bagus untuk menghitung hasil dan merencanakan berbagai kampanye iklan atau perubahan dalam proses.

F. Sequential Patterns

Teknik data mining ini membantu untuk menemukan atau mengidentifikasi pola atau tren serupa dalam data transaksi untuk periode tertentu. Tren atau beberapa pola yang konsisten dikenali dalam jenis penambangan data ini. Memahami perilaku pembelian pelanggan dan pola berurutan digunakan oleh toko untuk memajang produk mereka di rak. Misalnya dalam e-commerce dimana ketika Anda membeli barang A, akan terlihat bahwa Barang B sering dibeli dengan Barang A melihat riwayat pembelian sebelumnya.

G. Outlier Detection

Jenis teknik penambangan data ini mengacu pada pengamatan item data dalam kumpulan data yang tidak sesuai dengan pola yang diharapkan atau perilaku yang diharapkan. Teknik ini dapat digunakan di berbagai domain, seperti intrusion, detection, fraud or fault detection, dll. Outlier detection disebut juga dengan *Outlier Analysis* atau *Outlier mining*.

Anomali memberikan informasi penting dan dapat ditindaklanjuti untuk merek dan organisasi karena outlier adalah objek yang menyimpang secara signifikan dari rata-rata umum dalam sekumpulan database atau kombinasi data.

Berbeda dengan data lainnya. Itulah mengapa alat penambangan data outlier memerlukan perhatian dan analisis tambahan, karena memberikan pandangan berbeda tentang masalah tertentu. Jenis teknik penambangan data ini dapat digunakan untuk mendeteksi penipuan dan risiko dalam sistem kritis.

H. Rangkuman

Semua pendekatan *data mining* ini dapat membantu dalam analisis berbagai kumpulan data dari berbagai perspektif. Kita dapat menentukan metode optimal untuk mengubah data menjadi informasi yang berarti. Informasi ini dapat dimanfaatkan untuk memecahkan berbagai masalah bisnis, seperti meningkatkan pendapatan, meningkatkan kebahagiaan pelanggan, memperluas jaringan, atau menurunkan biaya yang tidak perlu.

I. Evaluasi

Untuk menguji pemahaman Anda terkait materi Metode Data Mining, silahkan kerjakan latihan di bawah ini!

1. Dalam sebuah toko online, pemiliknya sangat puas dengan peningkatan penjualan. Hal ini dikarenakan sistem yang dibangun dapat memberikan informasi berupa penawaran produk yang sejenis. Misalnya ketika pelanggan ingin membayar produk sabun mandi, ternyata ada penawaran barang sikat gigi dan pasta gigi. Konsumen tertarik dengan penawaran tersebut sehingga memutuskan untuk membeli tiga produk sekaligus, yaitu: sabun mandi, sikat gigi dan pasta gigi. Keputusan tersebut dipengaruhi oleh informasi yang bermanfaat dari proses *data mining.* Metode data mining apakah yang digunakan di toko tersebut?
2. Metode *data mining* yang cocok digunakan untuk memisahkan objek buah berdasarkan ukuran adalah?

MATERI POKOK 3:

POLA DATA

Indikator Hasil Belajar:
Peserta mampu menganalisis pola data dengan benar. Peserta dinyatakan lulus manakala mampu:

1. menjelaskan pola data pada *data mining,*
2. menjelaskan kelas serta konsep pada data, dan
3. menjelaskan jenis-jenis metode dalam menghasilkan pola data.

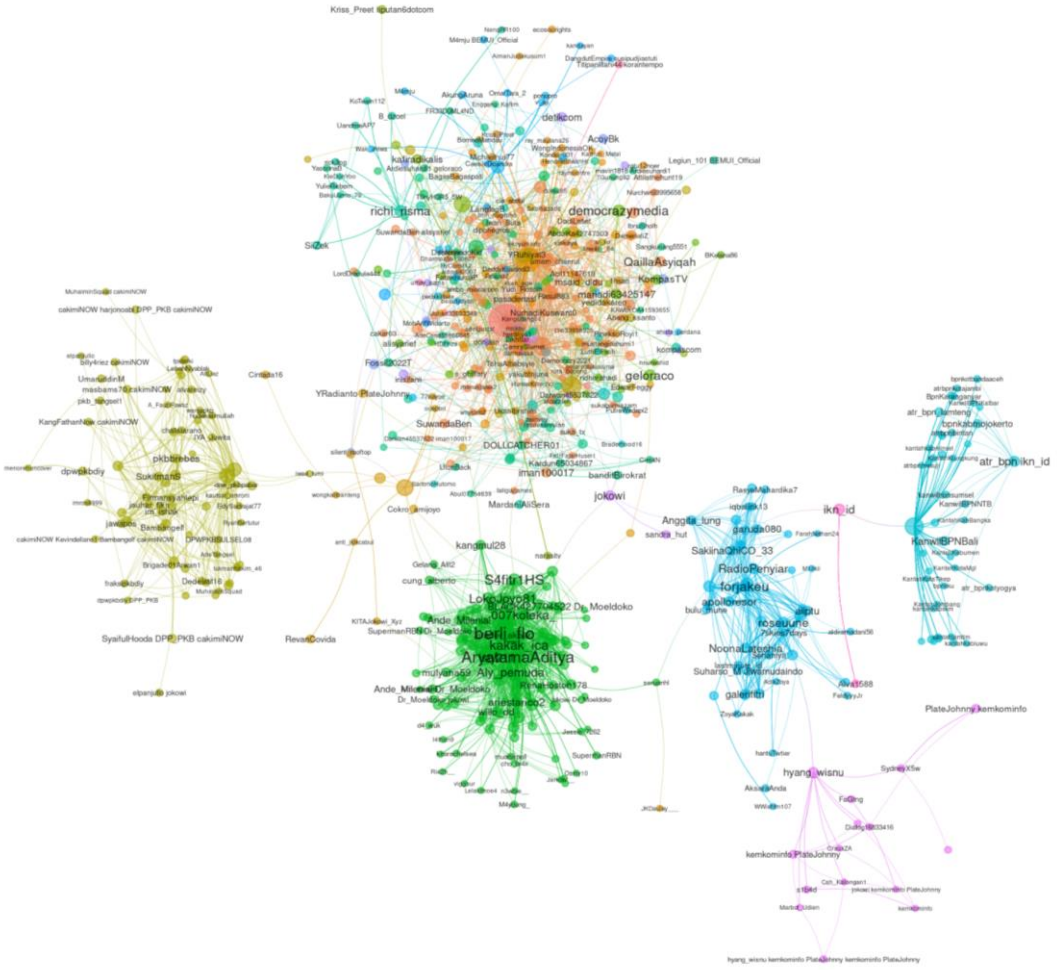
A. Definisi Pola Data

Pola data dalam penambangan data atau *data mining* secara sederhana diartikan sebagai pola-pola khusus yang biasanya dideskripsikan sebagai pola menarik serta tersembunyi. Pola ini selanjutnya berusaha dimunculkan atau diungkap melalui serangkaian proses ekstraksi data dengan menggunakan sejumlah metode tertentu. Sebuah pola dikatakan menarik jika mudah dipahami oleh manusia, valid pada data baru atau data uji dengan tingkat kepastian tertentu, berpotensi berguna, dan mengandung kebaruan. Sebuah pola juga menarik jika memvalidasi hipotesis yang ingin dikonfirmasi oleh pengguna. Pola yang menarik mewakili pengetahuan yang dapat membantu pengguna data untuk dapat lebih memahami data serta melakukan pengambilan keputusan secara lebih objektif berdasarkan data tersebut.

Pada contoh pola data di atas, tergambar polarisasi jaringan pengguna yang terbentuk berdasarkan pola *retweet* pada data. Polarisasi ini selanjutnya membentuk kluster-kluster berdasarkan kesamaan topik pembicaraan. kluster serta topik yang berhasil ditampilkan inilah yang selanjutnya dapat memberikan pengetahuan yang sebelumnya tidak terlihat kepada pengguna.

Terdapat beberapa jenis data yang dapat digunakan dalam penambangan data antara lain data berupa basis data atau *database*, *data warehouse* serta data transaksional. Pola data juga dapat dimunculkan melalui bentuk-bentuk data lain seperti data berupa aliran atau *data streams,* data berurut atau *ordered/sequence data,* data berupa grafik/jaringan atau *graph or networked data,* data spasial, data

text, data multimedia ataupun bentuk lainnya.



Gambar 6 Contoh pola data yang dihasilkan melalui Social Network Analysis (SNA) pengguna Twitter
Sumber: Dokumentasi Penulis

Jenis pola data yang dapat dimunculkan melalui penambangan data dibedakan berdasarkan metode serta fungsi dari penambangan data yang dilakukan. Fungsi dari penambangan data tersebut antara lain, karakterisasi dan diskriminasi, penambangan pola yang sering muncul atau *frequent pattern*, asosiasi dan korelasi, klasifikasi dan regresi, analisa kluster serta analisa *outlier.* Secara garis besar fungsi-fungsi penambangan data di atas dapat dibedakan menjadi dua kategori utama yaitu deskriptif dan prediktif. Penambangan deskriptif berusaha mencirikan properti dalam suatu kumpulan data. sementara penambangan prediktif berusaha melakukan induksi pada data terkini untuk membuat prediksi pada masukkan data selanjutnya.

B. Kelas/Konsep (Karakterisasi/Diskriminasi)

Data dapat dideskripsikan sebagai kelas serta konsep. Kelas secara sederhana merupakan karakteristik dasar dari suatu objek data yang secara jelas dapat membedakan objek data tersebut dengan objek lainnya pada data. Sedangkan konsep merupakan karakteristik lainnya pada suatu objek data yang biasanya dapat ditemukan melalui pengamatan terhadap fungsi atau proses yang dilakukan oleh objek data tersebut.

Tabel 2 Titanic dataset

Passengerid	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Survived
1	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	0
2	1	Cumings, Mrs. John Bradley (Fiorence Briggs)	female	38.0	1	0	PC 17599	71.2833	C85	C	1
3	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2.3 101282	7.9250	NaN	S	1
4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	1
5	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	0

Sumber: Dokumentasi Penulis

Tabel 2 *Titanic dataset*, berisikan data penumpang saat terjadinya peristiwa tenggelamnya *RMS Titanic.* Kolom Pclass atau kelas tiket merupakan contoh atribut kelas sementara *Survived* atau label penumpang yang berhasil selamat merupakan contoh atribut konsep.

Dengan mendeskripsikan serta menyimpulkan data menjadi kelas atau konsep akan lebih memudahkan pengguna data dalam memahami informasi serta pengetahuan yang terkandung pada data. Pendeskripsian data ini dapat dilakukan melalui berbagai cara antara lain karakterisasi data atau *data characterization,* diskriminasi data atau *data discrimination* atau penggunaan kedua cara tersebut secara bersamaan.

Karakterisasi data adalah ringkasan dari karakteristik atau fitur umum dari suatu kelas. Data terkait pengguna biasanya dikumpulkan melalui pengumpulan data dengan proses *query*. Bentuk keluaran karakterisasi data biasanya disajikan dalam berbagai bentuk seperti diagram lingkaran, diagram batang, kurva, kubus data multidimensi, dan tabel multidimensi, termasuk tab silang atau *crosstab*. Deskripsi yang dihasilkan juga dapat disajikan sebagai hubungan umum atau

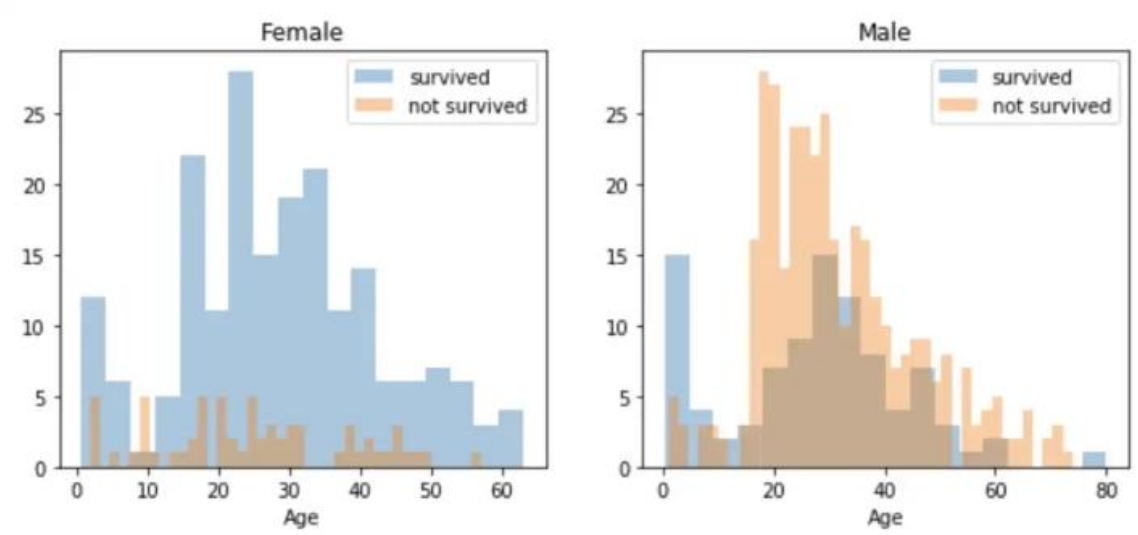
generalized relations atau dalam bentuk aturan atau *characteristic rules*.

Diskriminasi data adalah perbandingan fitur umum dari suatu kelas data terhadap fitur umum dari satu atau lebih kelas data lainnya. diskriminasi dilakukan dengan memunculkan fitur yang dinilai kontras antar kelas data sehingga dapat dibedakan untuk selanjutnya dibuat pengelompokan. Karakteristik perbedaan antar kelas data biasanya ditentukan oleh pengguna dan objek data yang diperlukan dapat diambil melalui proses *query* pada basis data.

C. *Frequent Pattern,* Asosiasi dan Korelasi

Frequent pattern secara sederhana adalah pola tertentu yang memiliki frekuensi kemunculan yang tinggi atau tergolong sering dalam suatu data. Terdapat beberapa jenis *frequent pattern*, termasuk *frequent itemset*, *frequent subsequences* (juga dikenal sebagai *sequential pattern*), dan *frequent substructures*. *Frequent itemset* biasanya mengacu pada sekumpulan item yang sering muncul bersamaan dalam suatu kumpulan data transaksional. *Frequent subsequences* adalah kecenderungan kemunculan data secara berurutan. *Frequent substructures* merujuk ke berbagai bentuk struktural (misalnya grafik, pohon, atau kisi-kisi) yang dapat digabungkan dengan *frequent itemset* atau *frequent subsequences*. Penambangan data melalui *frequent pattern* seringkali mengarah pada penemuan asosiasi dan korelasi menarik yang terdapat pada data.

Analisis asosiasi selanjutnya dapat dibedakan menjadi dua bagian berdasarkan jumlah dimensi atau fitur yang terlibat dalam keterkaitannya. Pembagian tersebut meliputi aturan asosiasi berdimensi tunggal atau *single-dimensional association rules* dimana hanya melibatkan sebuah dimensi atau fitur serta aturan asosiasi berdimensi banyak atau *multidimensional association rule* dimana terdapat beberapa dimensi atau fitur yang digunakan.



Gambar 7 Contoh Pola Frequent Itemset pada Titanic Dataset
Sumber: Dokumentasi Penulis

Gambar 7 berisi pola *frequent itemset* pada *titanic dataset* dimana secara umum atribut *sex - female* memiliki frekuensi berpasangan dengan atribut *survived* atau selamat yang lebih tinggi dibandingkan atribut *sex - male.*

D. Klasifikasi dan Regresi untuk Analisis Prediktif

Klasifikasi adalah serangkaian proses untuk menemukan model (atau fungsi) yang menggambarkan dan membedakan kelas atau konsep dari data. Model diturunkan berdasarkan analisis sekumpulan data pelatihan atau *data training* yaitu objek data yang label kelasnya telah diketahui sebelumnya. Model ini selanjutnya digunakan untuk memprediksi label yang belum diketahui dari kelas objek lainnya.

Beberapa metode dalam menentukan model klasifikasi antara lain aturan klasifikasi atau *classification rules, decision tree,* formula matematika, hingga *neural network. Decision tree* secara sederhana merupakan suatu *flowchart* berbentuk pohon dimana tiap *node* yang ada merupakan proses pengecekan terhadap nilai dari suatu atribut dari data. Tiap cabang dari diagram ini merupakan keluaran dari proses pengecekan tersebut sedangkan daunnya merupakan pendistribusian kelas atau konsep yang ada. *Neural network* merupakan contoh lain dari metode yang dapat digunakan dalam pengklasifikasian. Pada *neural network* terdapat kumpulan *neuron* yang berfungsi sebagai unit pemrosesan dengan koneksi beban atau *weighted connection* antar unit. Koneksi beban ini selanjutnya akan memberikan

karakteristik tertentu berdasarkan nilai atribut dari objek pada data.

Berbeda dengan klasifikasi yang digunakan untuk memprediksi label kategorikal yang bersifat diskrit ataupun tidak berurutan, model yang dihasilkan proses regresi digunakan untuk memprediksi label berupa nilai numerik yang bersifat berkelanjutan atau *continuous* yang hilang ataupun tidak/belum tersedia pada data. Analisa regresi juga sering digunakan dalam mengidentifikasi tren distribusi pada data yang tersedia.

Proses analisis yang relevan perlu dilakukan guna menentukan tipe proses prediksi yang dilakukan apakah akan menggunakan klasifikasi atau regresi. Suatu kumpulan data dapat mengandung beberapa atribut yang berbeda yang selanjutnya akan diseleksi guna mengetahui tipe proses prediksi yang diperlukan.

E. Analisa Kluster

Sering kali data sebagai objek pengamatan tidak memiliki label atau kelas tertentu. Analisa kluster selanjutnya dapat digunakan untuk menghasilkan label kelas untuk sekumpulan data. Analisa kluster menganalisis objek data tanpa mempertimbangkan label kelas dari data tersebut. Objek-objek data tersebut dikelompokkan atau berdasarkan pada prinsip memaksimalkan kesamaan sesama anggota dalam suatu kelas serta meminimalkan kesamaan anggota suatu kelas dengan kelas lainnya. kluster-kluster objek dibentuk sedemikian rupa sehingga objek-objek di dalam sebuah kluster memiliki kesamaan yang tinggi antara satu sama lain, tetapi memiliki perbedaan dengan objek-objek di pada kluster lain. Setiap kluster yang terbentuk dapat dipandang sebagai kelas objek, dimana aturan dapat diturunkan.

F. Analisa *Outlier*

Outlier pada data adalah objek atau kumpulan objek yang tidak memiliki kesesuaian dengan perilaku umum atau model data. Pada beberapa kasus penambangan data, objek data yang dianggap sebagai *outlier* akan dieliminasi sebagai *noise* atau pengecualian. Namun, dalam beberapa kasus penambangan data lainnya, objek data yang memiliki kelainan atau penyimpangan nilai atribut tersebut justru menjadi sasaran utama pengamatan. Uji statistik dapat digunakan

dalam mendeteksi *outlier.* Metode ini mengasumsikan model distribusi atau probabilitas terkait suatu data, atau menggunakan ukuran jarak di mana objek yang jauh dari kluster lain dapat diasumsikan sebagai *outlier*. Metode berbasis kepadatan selanjutnya dapat mengidentifikasi *outlier* di wilayah lokal, meskipun terlihat normal jika dilihat melalui distribusi statistik global.

G. Rangkuman

Pola data merupakan pola pada data yang memiliki kekhususan sehingga dikatakan “menarik” dan merupakan representasi dari pengetahuan. Pola ini dapat dimunculkan melalui serangkaian metode penambangan data atau *data mining.* Metode atau fungsi tersebut antara lain, karakterisasi dan diskriminasi, penambangan pola yang sering muncul atau *frequent pattern*, asosiasi dan korelasi, klasifikasi dan regresi, analisa kluster serta analisa *outlier.* Secara garis besar fungsi-fungsi penambangan data di atas dapat dibedakan menjadi dua kategori utama yaitu deskriptif dan prediktif. Penambangan deskriptif berusaha mencirikan properti dalam suatu kumpulan data. sementara penambangan prediktif berusaha melakukan induksi pada data terkini untuk membuat prediksi pada masukkan data selanjutnya.

H. Evaluasi

Untuk menguji pemahaman Anda terkait materi Pola Data, silahkan kerjakan latihan di bawah ini!

1. Sebutkan dan jelaskan kriteria pola data yamg dikatakan menarik!
2. Jelaskan perbedaan antara klasifikasi dengan regresi!
3. Jelaskan metode yang dapat dilakukan terhadap data yang belum memiliki label atau kelas!

MATERI POKOK 4:

DATA SUMMARIZATION

Indikator Hasil Belajar:
Peserta mampu merangkum data (*data summarization*). Peserta dinyatakan lulus manakala mampu :

1. menjelaskan tujuan dari *data summarization* dengan tepat,
2. menjelaskan pengertian dari *univariate analysis* dan *multivariate analysis* dengan tepat,
3. menjelaskan contoh-contoh dari visualisasi data dengan tepat.

Data Summarization atau peringkasan data dapat didefinisikan sebagai penyajian ringkasan/laporan dari suatu data yang dihasilkan secara komprehensif dan informatif. Ringkasan diperoleh melalui serangkaian proses eksplorasi pada data atau *data exploration.* Proses ini dilakukan untuk menggali informasi yang terkandung dalam dataset. Eksplorasi hingga peringkasan data dilakukan untuk menyampaikan tren dan pola dari kumpulan data dengan cara yang sederhana sehingga lebih mudah untuk dipahami.

Dalam proses *data mining*, eksplorasi data dilakukan dalam berbagai tingkatan dan tahapan seperti pra pemrosesan atau persiapan data, membangun model hingga memberikan intepretasi terhadap hasil sebagai keluaran dari model. Selanjutnya eksplorasi data memiliki memiliki beberapa tujuan sebagai berikut:

1. Pemahaman data atau *data understanding*, eksplorasi data menghasilkan sudut pandang tingkat tinggi dari setiap atribut dalam kumpulan data serta dapat memberikan gambaran terkait interaksi antar atribut.
2. Persiapan data atau *data preparation,* melalui eksplorasi data dapat dilakukan identifikasi terhadap anomali yang mungkin terdapat pada data untuk selanjutnya dilakukan penanganan. Eliminasi terhadap anomali merupakan salah satau tahap persiapan data yang penting untuk dilakukan sebelum data tersebut dimasukan ke dalam algoritma *data mining.*
3. Eksplorasi data pada beberapa kasus dapat memenuhi fungsi dari keseluruhan proses *data mining.* Salah satu contohnya adalah *scatterplot*

yang dapat secara jelas menggambarkan terbentuknya kluster-kluster berdimensi rendah.

4. Eksplorasi data digunakan dalam memahami prediksi, klasifikasi, dan pengklusteran dari proses data mining. Dalam pengelompokan berdimensi rendah, *scatterplot* adalah cara yang efisien untuk memvisualisasikan kelompok atau kluster. Histogram memungkinkan untuk memahami distribusi atribut dan juga berguna untuk memvisualisasikan prediksi numerik, estimasi tingkat kesalahan, dan lainnya.

Pada proses *data mining,* data sebagai masukkan dapat datang dalam berbagai format serta tipe yang berbeda. Jenis dari operasi *data mining* yang cocok digunakan pada suatu jenis data dapat diketahui dengan memahami properti dari tiap atribut atau fitur yang memberikan informasi. Terdapat dua pembagian umum terkait tipe data yaitu numerik atau kontinu serta kategorikal atau nominal. Tipe data numerik merupakan tipe data yang menggunakan bilangan atau angka dimana nilainya bersifat berurutan dan dapat dilakukan operasi matematika pada nilai tersebut. Sedangkan tipe data kategorikal merupakan variabel-variable berisikan simbol tertentu dimana tidak ada hubungan langsung nilai data sehingga tidak dapat dilakukan operasi matematika pada data tersebut.

A. Univariate Analysis

Univariate Analysis atau Analisis univariat merupakan salah satu metode pengukuran dalam mempelajari karakteristik data. *Univariate Analysis* digunakan untuk mempelajari perilaku dari tiap atribut yang ada pada suatu dataset, atribut ini dianggap sebagai entitas yang independen dari variabel lain pada dataset.

Univariate analysis pada atribut bertipe numerik dapat dilakukan melalui metode statistik deskriptif. Statistik deskriptif merupakan metode statistik yang dapat digunakan untuk mendeskripsikan karakteristik dari suatu variabel numerik. Terdapat dua jenis pengukuran dalam statistik deskriptif yaitu:

1. Pengukuran Tendensi Sentral atau *Measure of Central Tendency*

   Pengukuran ini bertujuan untuk menemukan lokasi pusat dari suatu variabel. Hal ini dilakukan untuk mengukur kumpulan data dengan satu nilai pusat atau paling umum. Terdapat beberapa teknik perhitungan yang dapat digunakan untuk mengetahui lokasi pusat dari suatu data.

a. *Mean* adalah nilai rata-rata dari seluruh nilai pengamatan dalam suatu dataset.
b. *Median* adalah nilai tengah dari distribusi nilai pengamatan pada suatu dataset.
c. *Mode* adalah nilai pengamatan yang paling banyak muncul dalam suatu dataset.

2. Pengukuran sebaran atau *Measure of Spread*

Selain pengukuran terhadap lokasi pusat pada data, pengukuran terhadap sebaran data juga perlu dilakukan untuk mengetahui sejauh apa nilai data tersebar dari nilai lokasi pusatnya. Terdapat 2 jenis metrik yang dapat digunakan untuk mengetahui sebaran data.

a. *Range* adalah perbedaan nilai antara nilai maksimum dengan nilai minimum pada suatu variabel. *Range* memiliki kelemahan karena nilainya akan sangat terpengaruh dengan munculnya *outlier*.
b. *Deviation* yang meliputi varian dan standar deviasi adalah dua metrik yang mempertimbangkan nilai dari semua nilai data pada suatu atribut. *Deviation* secara sederhana mengukur perbedaan dari setiap nilai data dengan nilai *mean* nya. Varian adalah jumlah dari kuadrat penyimpangan semua titik data dari titik data rata-rata dibagi dengan jumlah titik data. Standar deviasi adalah akar kuadrat dari varian untuk kumpulan data dengan N pengamatan.

$$\text{Variance} = s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Univariate analysis untuk atribut data bertipe kategorikal berbeda jika dibandingkan dengan pengukuran pada atribut data bertipe numerik. Pada atribut bertipe kategorikal, pengukuran terhadap kecenderungan sentral, serta lokasi relatif dari data tidak dapat dilakukan. Untuk itu pengukuran untuk atribut kategorikal dilakukan dengan menggambarkan pengaturan dari tiap-tiap nilai data. Pengukuran ini dapat dilakukan melalui pengukuran tingkat heterogenitas f_h dari setiap titik data. Pengukuran ini biasanya dilakukan menggunakan *gini index* dan *entropy index.*

1. *Gini index* bernilai sama dengan 0 dalam kasus heterogenitas terendah, yaitu ketika sebuah kelas diasumsikan pada frekuensi sama dengan 1 dan semua kelas lainnya tidak pernah diasumsikan. Sebaliknya, ketika semua kelas memiliki frekuensi relatif yang sama atau pada kasus heterogenitas tertinggi terjadi, *gini index* mencapai nilai maksimumnya. *Gini index* diformulasikan sebagai berikut

$$G = 1 - \sum_{h=1}^{H} f_h^2.$$

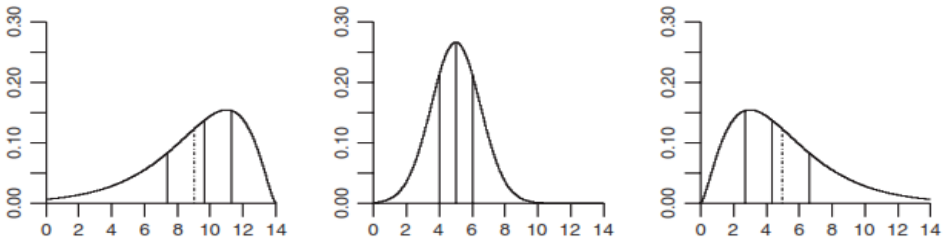
dimana f_h merupakan nilai heterogenitas.

2. *Entropy index* bernilai sama dengan 0 dalam kasus heterogenitas terendah, yaitu ketika sebuah kelas diasumsikan pada frekuensi sama dengan 1 dan semua kelas lainnya tidak pernah diasumsikan. Sebaliknya, ketika semua kelas memiliki frekuensi relatif yang sama dan heterogenitas tertinggi, indeks entropi mencapai nilai maksimumnya. *Entropy index* diformulasikan sebagai berikut

$$E = -\sum_{h=1}^{H} f_h \log_2 f_h.$$

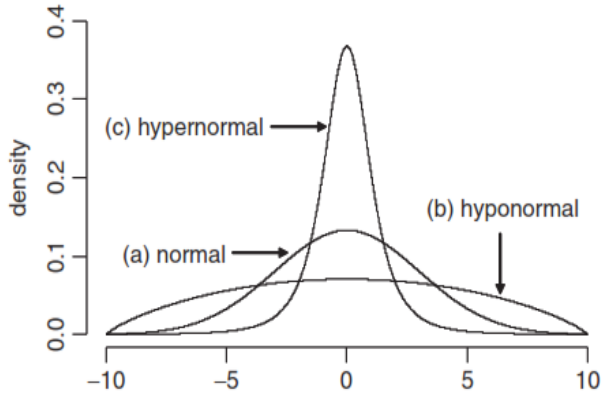
Analisis terhadap frekuensi kepadatan data dapat juga dilakukan melalui pengamatan secara visual terhadap *histogram* yang menggambarkan bentuk kurva dari kepadatan data. Terdapat 2 jenis pengamatan yang dapat dilakukan yaitu *skewness* dan *kurtosis.*

Skewness merupakan pengamatan secara visual terhadap bentuk asimetris dari kurva kepadatan distribusi data. Bentuk asimetri ini dapat terjadi jika nilai *mean* dan *median* persis sama. Jika nilai *mean* lebih besar dari *median* maka kurva dikatakan bergeser ke kanan sebaliknya jika nilai *mean* lebih kecil dari *median* maka kurva dikatakan bergeser ke kiri.



Gambar 8 Pergeseran / skewness puncak pada kurva kepadatan distribusi data
Sumber: Carlo

Kurtosis merupakan pengamatan secara visual terhadap bentuk kurva kepadatan distribusi data yang dibandingkan dengan bentuk kurva distribusi normal. Jika bentuk kurva yang mewakili frekuensi empiris bertepatan bentuknya dengan bentuk kurva kerapatan normal (a), kurtosis dikatakan bernilai 0. Jika bentuk kurva lebih rendah dari bentuk kurva normal maka dikatakan kurtosis bernilai hiponormal, karena menunjukkan dispersi yang lebih besar daripada kerapatan normal (frekuensi yang lebih rendah ke nilai yang mendekati *mean*, seperti pada kurva (b)). Jika bentuk kurva lebih tinggi daripada kurva normal maka kurtosis dikatakan hipernormal, karena menunjukkan dispersi yang lebih rendah dari kerapatan normal dan oleh karena itu memiliki frekuensi kepadatan data yang lebih tinggi ke nilai yang mendekati *mean*, seperti pada kurva (c).



Gambar 9 Bentuk perbandingan kurva kepadatan distribusi data
Sumber: Carlo

B. Multivariate Analysis

Multivariate analysis atau analisis multivariat adalah analisis yang dilakukan pada lebih dari satu atribut secara bersamaan. Teknik ini dilakukan untuk memahami hubungan antar atribut dalam suatu data.

Central data point atau titik pusat data adalah titik data yang merepresentasikan nilai *mean* dari tiap atribut untuk selanjutnya dituangkan ke dalam koordinat kartesian. Penggambaran secara visual hanya mungkin untuk dilakukan pada penggunaan atribut sebanyak 3 dimensi.

Correlation atau korelasi adalah pengukuran yang dilakukan untuk mengetahui hubungan antara dua variabel atau keterkaitan antara satu variabel dengan variabel lainnya. Pada saat dua variabel memeliki keterkaitan yang tinggi maka kedua variabel tersebut akan memiliki variasi pola nilai yang mirip sehingga kedua variabel tersebut dapat dijadikan fitur yang baik untuk memprediksi variabel lainnya. Namun korelasi antar dua variabel bukanlah suatu hubungan sebab akibat karena keterkaitan mungkin saja terjadi dikarenakan adanya pengaruh dari variabel lain yaitu variabel ketiga.

Korelasi antara 2 variabel biasanya diukur menggunakan *pearson correlation coefficient* (r) yang menghitung besarnya keterkaitan linear antara 2 variabel. Nilai dari koefisien korelasi ini memilii rentang dari -1 <= r >= 1 dimana nilai mendekati -1 atau 1 menandakan bahwa kedua variabel memiliki keterkaitan yang tinggi (contoh harga barang dan total penjualan). sebaliknya, nilai koefisien korelasi 0 menandakan bahwa kedua variabel tidak memiliki keterkaitan sama sekali.

Pearson correlation coefficient dirumuskan sebagai berikut

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

$$= \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N * S_x * S_y}$$

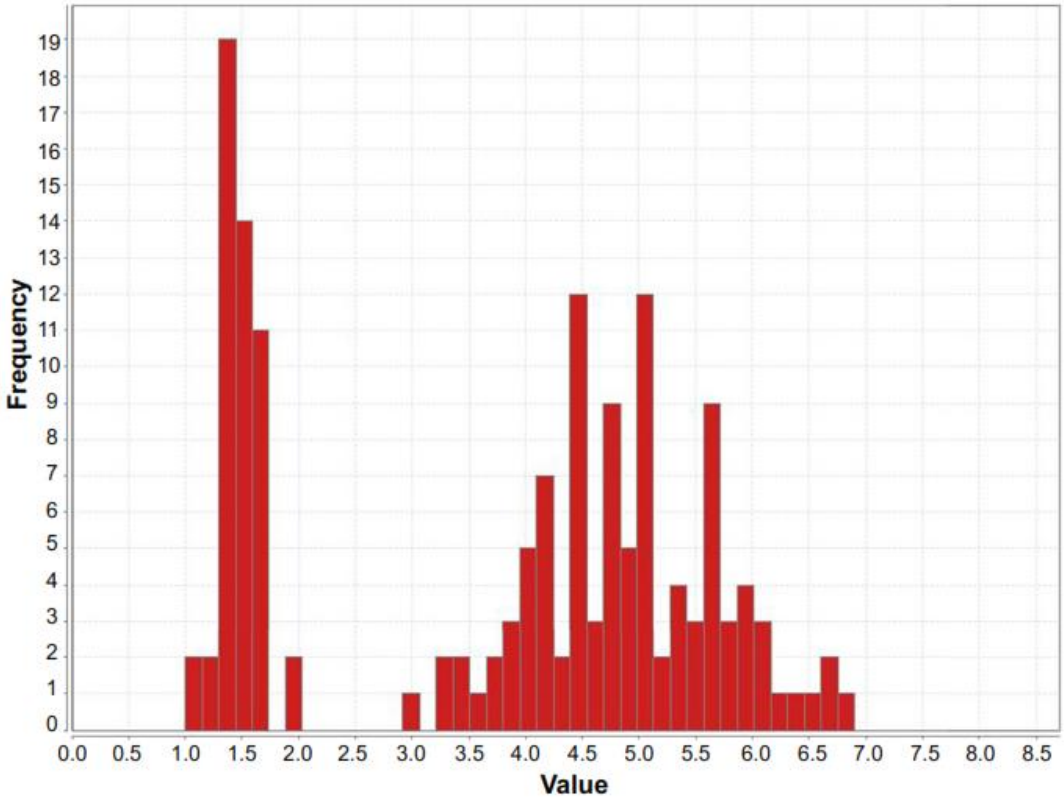
Dimana S_x dan S_y merupakan standar deviasi dari variabel x dan y

C. Graphical Summarization

Visualisasikan data adalah salah satu aspek penting dalam penemuan dan eksplorasi data. Visualisasi data meliputi berbagai metode merepresentasikan data dalam bentuk visual yang abstrak. Representasi visual data memberikan pemahaman yang mudah tentang suatu data yang kompleks dengan banyak variabel dan hubungan yang mendasarinya. Terdapat beberapa diagram yang biasa digunakan dalam memvisualisasikan data antara lain:

1. *Histogram*

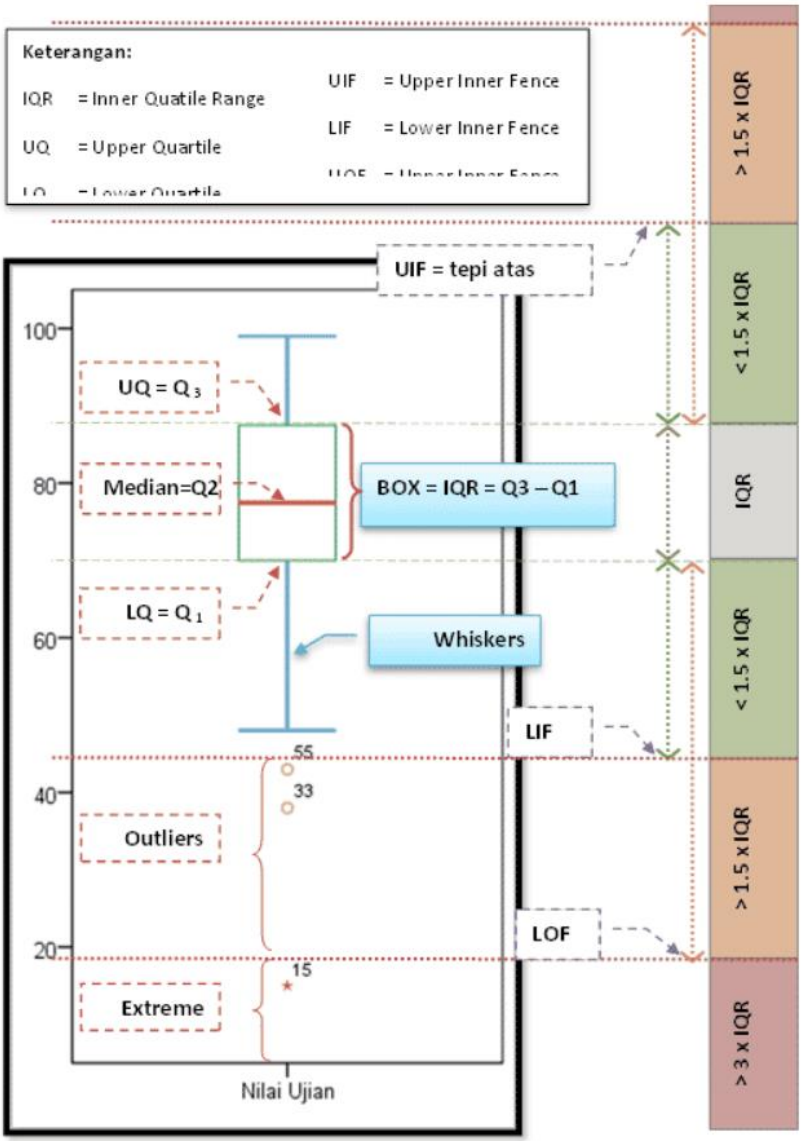
   Histogram adalah salah satu cara visual yang paling mendasar untuk memahami frekuensi kemunculan rentang nilai untuk suatu variabel. *Histogram* menentukan distribusi data dengan memplot frekuensi kejadian dalam suatu rentang. Dalam *histogram*, variabel kontinu yang diselidiki ditempatkan pada sumbu horizontal dan frekuensi kejadian ditempatkan pada sumbu vertikal.



Gambar 10 Contoh histogram yang menampilkan penempatan nilai dan frekuensi
Sumber: Kotu

2. *Box-plot*

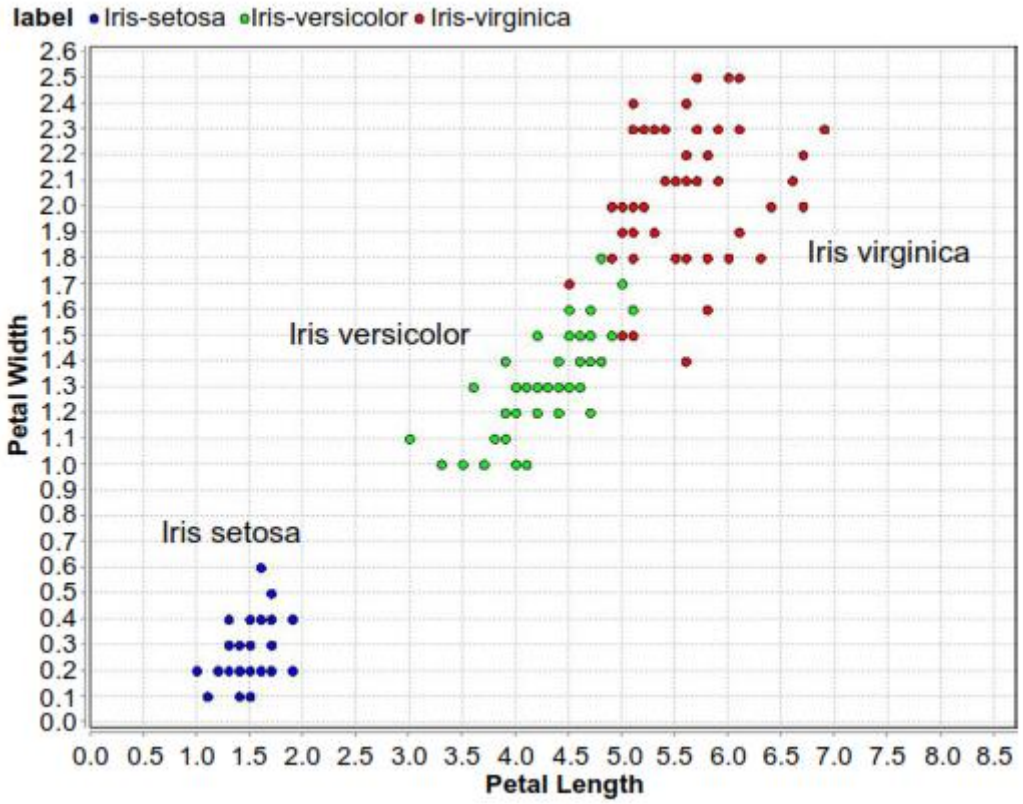
Box-plot merupakan ringkasan dari sebaran nilai data yang disajikan secara grafis yang dapat menggambarkan bentuk sebaran data (skewness), ukuran tendensi sentral dan ukuran sebaran (keanekaragaman) data yang diamati. *Box-plot* menggambarkan beberapa ukuran statistik pada atribut data antara lain nilai terkecil, kuartil terendah atau kuartil pertama (Q1), nilai tengah atau *median* (Q2), kuartil tertinggi atau kuartil ketiga (Q3), nilai maksimum serta nilai-nilai yang menjadi nilai *outlier* atau nilai ekstrim.



Gambar 11 Contoh box-plot beserta keterangannya
Sumber: smartstat.info

3. *Scatterplot*

Pada suatu *scatterplot,* titik-titik data ditempatkan ke dalam suatu koordinat kartesian berdasarkan besaran nilai dari variabel yang digambarkan dimana variabel yang digambarkan biasanya bertipe kontinu. *Scatterplot* secara mudah dapat menggambarkan hubungan dari 2 variabel yang berbeda. Hubungan yang erat antara 2 variabel biasanya tergambar melalui titik-titik data yang bersusun mendekati suatu garis lurus imajiner. Sebaliknya, jika tidak terdapat hubungan maka titik-titik data hanya akan tersebar secara acak. *Scatterplot* juga dapat digunakan dalam menentukan kumpulan kluster yang dibentuk dari titik data.



Gambar 12 Contoh *scatterplot* yang menggambarkan sebaran titik data pada *iris dataset.*
Sumber: Kotu

D. Rangkuman

Data summarization merupakan bentuk penyajian ringkasan/laporan dari suatu proses eksplorasi data yang menghasilkan kesimpuilan atau visualisasi secara komprehensif dan informatif. Eksplorasi data dilakukan dengan mendeskrisikan masing-masing atribut yang terdapat pada data. Eksplorasi atribut data dilakukan dengan melakukan pengamatan pada tiap-tiap atribut (*univariate analysis)* ataupun dengan pengamatan terhadap hubungan antar atribut (*multivariate analysis).*

E. Evaluasi

1. Sebutkan tujuan-tujuan dari *data summarization* melalui eksplorasi data!
2. Jelaskan perbedaan antara *univariate analysis* dengan *multivariate analysis!*
3. Jelaskan ukuran-ukuran yang dapat digambarkan melalui penggunaan *box-pot*

MATERI POKOK 5:

KATEGORISASI DATA

Indikator Hasil Belajar:

Peserta mampu mengkategorikan data: dengan tepat. Peserta dinyatakan lulus manakala mampu :

1. mengidentifikasi data terstruktur dan tidak terstruktur,
2. menjelaskan tipe-tipe atribut beserta dengan pengoperasiannya.

Proses penambangan data memerlukan suatu input, yang biasanya berbentuk tabel dua dimensi, yang dinamakan dengan *dataset*. Baris-baris di dalam *dataset* biasanya disebut juga dengan *objects*, *samples*, *instances*, atau *records*. Sementara itu, kolom-kolom pada *dataset* merepresentasikan informasi yang terkandung dalam masing-masing obyek, yang biasanya disebut juga dengan *attributes*, *dimensions*, atau *features*. Distribusi data yang melibatkan satu atribut dinamakan *univariate dataset*. Sementara itu, distribusi *bivariate* melibatkan dua atribut dan *multivariate* melibatkan banyak atribut.

A. Data Terstruktur dan Tidak Terstruktur

Data yang terstruktur mempunyai format yang standar, struktur yang terdefinisi dengan baik sesuai dengan model, mengikuti urutan yang tetap, dan mudah diakses. Data terstruktur biasanya tersimpan dalam bentuk tabular pada basis data relasional. Tabel 5.1 merupakan contoh kumpulan data yang terstruktur. Sementara itu, data yang tidak terstruktur tidak mempunyai format yang standar dan struktur serta urutan yang baku. Contoh data yang tidak terstruktur adalah dokumen teks, video, gambar, file audio, dan lain-lain. Tabel 5.2 merupakan contoh kumpulan data yang tidak terstruktur yang bersumber dari tulisan di media sosial. Atribut dari *dataset* ini adalah kata-kata. Biasanya teknik temu kembali informasi atau pemrosesan bahasa alami diperlukan untuk membuat *dataset* ini menjadi lebih terstruktur seperti pada Tabel 5.3.

Tabel 3 *Dataset* Terstruktur

ID	Kota	Waktu	Temperatur
1	Bandung	2 Desember 2022 19:45	17^{o}C
2	Surabaya	2 Desember 2022 19:45	34^{o}C

Tabel 4 *Dataset* tidak Terstruktur

ID	Teks
139492	Malam ini, Bandung dingin banget.
154239	Buset, panas sekali Surabaya malam ini.

Tabel 5 Hasil Konversi Dataset tidak Terstruktur menjadi Terstruktur

ID	Kota	Waktu	Temperatur
139492	Bandung	2 Desember 2022 19:45	Dingin
154239	Surabaya	2 Desember 2022 19:45	Panas

B. Tipe Atribut Data Beserta Pengoperasiannya

Memahami tipe-tipe atribut penting untuk mengetahui jenis operasi apa yang dapat dilakukan pada atribut tersebut. Contohnya, kedua *dataset* pada Tabel 5.1 dan 5.3 mempunyai nama atribut atau fitur yang sama, yaitu ID, kota, waktu, dan temperatur. Namun, tipe atribut temperatur berbeda pada kedua *dataset*, sehingga operasi yang dapat dilakukan pada *dataset* pertama mungkin saja tidak dapat dilakukan pada *dataset* kedua dan begitu pula sebaliknya. Materi pada bagian ini menjelaskan kategori atribut yang biasanya muncul dalam suatu *dataset*. Kategori tersebut yaitu nominal, binari, numerikal, dan ordinal atau diskrit dan kontinu.

B.1 Nominal

Nilai-nilai dari atribut nominal adalah nama-nama kategori dengan jumlah tertentu. Misalnya atribut kota (pada *dataset* kependudukan di ibukota provinsi di pulau Jawa) termasuk tipe nominal karena nilainya hanya punya lima kategori yaitu, “Jakarta”, “Bandung”, “Semarang”, “Yogyakarta”, dan

“Surabaya”. Contoh lainnya adalah status perkawinan penduduk, dengan nilai “belum menikah”, “menikah”, “cerai hidup”, dan “cerai mati”. Pekerjaan penduduk juga termasuk dalam atribut nominal, dengan kemungkinan nilai seperti “guru”, “dosen”, “peneliti”, “perekayasa”, “analis data ilmiah”, dan lain-lain.

Nilai dari atribut nominal juga dapat berupa angka yang mereprensentasikan satu kategori. Misalnya, “Jakarta” dapat diganti dengan kode (kodifikasi/ *encoding*) dengan nilai 0, kemudian nilai 1 untuk “Bandung”, dan seterusnya. Namun, operasi aritmetik tidak dapat dilakukan untuk tipe ini walaupun nilai-nilai dari atribut tersebut numerik. Misalnya, nilai *mean* (rata-rata) atau median untuk atribut nominal tidak berarti apapun. Hanya operator *logical* yang dapat dilakukan untuk atribut dengan tipe ini. Pada literatur tipe atribut ini juga disebut dengan multinominal atau polynominal.

B.2 Biner

Nilai-nilai dari atribut biner (binominal) hanya ada dua kategori, misalnya 0 atau 1. Terkadang atribut biner juga merepresentasikan *boolean*, *true* atau *false*. Pada contoh *dataset* di dalam Tabel 5.3, atribut temperatur termasuk dalam tipe biner, dengan nilai “dingin” atau “panas”. Contoh lainnya, jenis kelamin, dengan nilai “laki-laki” atau “perempuan”. Kodifikasi juga dapat dilakukan untuk tipe data ini, misalnya kode “0” untuk “laki-laki” dan “1” untuk perempuan, ataupun sebaliknya. Sama halnya dengan tipe nominal, operator matematika, kecuali operator *logical*, tidak dapat dilakukan untuk atribut biner.

B.3 Numerikal

Nilai-nilai dari atribut numerikal adalah angka-angka kuantitatif yang dapat diukur. Misalnya pada contoh dataset dalam Tabel 5.1, atribut temperatur termasuk dalam tipe numerikal. Contoh lainnya, atribut usia penduduk termasuk juga dalam tipe ini. Operasi-operasi aritmetik seperti tambah, kurang, bagi, kali, menghitung *mean* serta median, dan lain-lain dapat dilakukan untuk atribut numerikal.

Atribut numerikal terbagi menjadi dua kategori, yaitu *interval-scaled* dan *ratio-scaled*. Jika, *zero-point* dapat didefinisikan, maka atribut tersebut termasuk dalam kategori *ratio-scaled*. Atribut usia penduduk termasuk dalam *ratio-scaled* karena 0 adalah *zero-point* untuk usia manusia sehingga operasi perbandingan dapat dilakukan untuk atribut ini. Angka 0 dalam konteks ini berarti kelahiran dan belum mempunyai usia. Andaikan saat ini usia Pak Slamet 60 tahun dan usia Pak Zaenal 30 tahun, maka pernyataan berikut valid: "Usia Pak Slamet dua kali lebih tua dari pada Pak Zaenal". Sementara itu, atribut temperatur dalam derajat Celcius termasuk dalam kategori *interval-scaled* karena *zero-point* tidak dapat ditentukan. Temperatur 0^{o}C bukanlah *zero-point*, karena angka 0 dalam hal ini bukan berarti tidak punya temperatur melainkan titik beku air. Sehingga, pernyataan berikut tidak valid: "Temperatur kota Surabaya dua kali lebih hangat daripada kota Bandung".

B.4 Ordinal

Nilai-nilai dari atribut ordinal pada dasarnya adalah nilai nominal yang mempunyai urutan (*ranking*). Misalnya level pendidikan, dengan kemungkinan nilai "SD", "SMP", "SMA", "S-1", "S-2", dan "S-3". Contoh lainnya yaitu jabatan, dengan nilai "ahli pertama", "ahli muda", "ahli madya", dan "ahli utama". Nilai-nilai ini mempunyai urutan yang relevan namun perbedaan ataupun rasio antara nilai-nilai ini tidak dapat dihitung dengan pasti.

Atribut ordinal juga kerap digunakan pada penilaian subyektif terhadap sesuatu yang tidak dapat diukur secara obyektif. Misalnya dalam survei tingkat kepuasan konsumen, nilai ordinal dengan *range* 1-5 digunakan untuk 1: sangat tidak puas, 2: tidak puas, 3: netral, 4: puas, dan 5: sangat puas. Nilai ordinal juga bisa didapatkan dengan melakukan kategorisasi pada nilai-nilai numerik. Misalnya, gaji kurang dari 5 juta dikategorikan menjadi penduduk dengan pendapatan rendah, dan gaji sama dengan atau lebih dari 5 juta dikategorikan menjadi penduduk dengan pendapatan tinggi. Namun, berbeda dengan kodifikasi yang tidak menghilangkan informasi, konversi dengan cara seperti ini dapat mengakibatkan hilangnya sebagian informasi.

B.5 Diskrit dan Kontinyu

Selain dibagi menjadi empat kategori, tipe atribut juga dapat dibagi menjadi dua kategori lainnya yaitu diskrit dan kontinyu. Atribut diskrit mempunyai himpunan nilai yang terbatas atau himpunan nilai tidak terbatas yang dapat dihitung. Atribut-atribut nominal, biner, dan ordinal seperti kota, status perkawinan, pekerjaan, jenis kelamin, level pendidikan, dan jabatan mempunyai himpunan nilai yang terbatas sehingga atribut-atribut ini termasuk dalam kategori diskrit. Kode pos juga merupakan salah satu contoh atribut diskrit karena kode pos merupakan nilai tak hingga yang dapat dihitung. Kode pos bisa bertambah hingga tak terbatas, namun himpunan nilai sebenarnya dapat dihitung. Contoh lainnya yaitu, NIK (Nomor Induk Kependudukan). Sementara itu, atribut kontinyu adalah atribut numerikal yang diasumsikan mempunyai nilai tak terhingga yang tidak dapat dihitung. Pada contoh atribut numerikal di atas, temperatur dalam derajat Celsius termasuk dalam kategori atribut kontinyu. Namun, usia termasuk dalam kategori atribut diskrit.

C. Rangkuman

Data dapat dibagi menjadi dua, yaitu data terstruktur dan tidak terstruktur. Suatu *dataset* terdiri dari objek-objek dan atribut-atribut yang mendeskripsikan masing-masing objek. Tipe atribut dapat dibagi menjadi empat kategori, yaitu nominal, biner, numerikal, dan ordinal. Selain itu, tipe atribut juga dapat dibagi menjadi dua kategori yang lain, yaitu diskrit dan kontinyu.

D. Evaluasi

Untuk menguji pemahaman Anda terkait materi Kategorisasi Data, silahkan kerjakan latihan di bawah ini!

1. Operasi *logical* apa saja yang dapat dilakukan pada atribut nominal?
2. Pada kasus *dataset* diabetes terkandung atribut apakah seseorang positif mengidap diabetes atau tidak. Bagaimana sebaiknya kodifikasi biner diberikan pada atribut ini? Apakah kode 0 diberikan untuk nilai positif atau negatif?
3. Apakah temperatur dalam derajat Fahrenheit dan Kelvin termasuk ke dalam kategori *interval-scaled* atau *ratio-scaled*?

MATERI POKOK 6:

DATA *RESCALLING*

Indikator Hasil Belajar:
Peserta mampu melakukan *data rescalling* dengan benar. Peserta dinyatakan lulus apabila mampu:

1. melakukan praktik reduksi data dengan benar.
2. melakukan praktik diskritisasi data dengan benar.

Rescalling merupakan salah satu teknik dalam *preprocessing* data untuk menstandarisasi data agar mempunyai skala yang sama atau bahkan mereduksi data sehingga dimensinya menjadi lebih kecil.

A. Reduksi Data

Dataset yang sangat besar memerlukan analisis dan penambangan yang sangat kompleks. Hal ini pada banyak kasus tidak praktis dan tidak bisa dilakukan. Teknik reduksi data dapat diaplikasikan pada kasus ini untuk mendapatkan representasi dataset dengan volume yang jauh lebih kecil, namun tetap mempertahankan integritas data aslinya. Sehingga, penambangan pada data yang tereduksi dapat dilakukan lebih efisien dan dapat memproduksi hasil analisis yang (hampir) sama. Materi pada bagian ini menjelaskan teknik-teknik reduksi data seperti reduksi dimensi, reduksi numerik, dan kompresi data.

A.1 Reduksi dimensi

Reduksi dimensi adalah proses mengurangi jumlah atribut atau fitur suatu *dataset*, sehingga dapat mereduksi volume data asli. Ada dua tipe teknik dalam mereduksi dimensi, yang pertama adalah dengan *features extraction*, yaitu mengekstrak fitur-fitur asli dan membuat fitur-fitur baru yang merupakan transformasi atau kombinasi dari atribut-atribut aslinya dengan jumlah yang sama atau bahkan lebih kecil. Contoh-contoh teknik yang dapat digunakan pada metode ini adalah *wavelet transform* dan *principal component analysis* (PCA). Sedangkan teknik yang kedua adalah dengan *attribute subset selection*, yaitu

memilih fitur-fitur yang penting saja dan menghapus atribut yang tidak relevan dan redundan.

Transformasi wavelet mengubah vektor A menjadi vektor A' yang berbeda secara numerik. Walaupun, vektor A dan vektor A' memiliki panjang yang sama, data asli dapat direduksi karena data yang diperoleh dari hasil transformasi wavelet dapat terpotong. Data terkompresi diperoleh dengan mempertahankan fragmen terkecil dari koefisien wavelet terkuat. Sementara itu, PCA dapat mereduksi dimensi dengan “menggabungkan” esensi atribut-atribut asli dan membuat alternatif suatu kumpulan atribut dengan jumlah yang lebih kecil. Misalkan data yang ingin direduksi terdiri dari vektor dengan jumlah atribut sebanyak n. PCA mencari k vektor orthogonal/independen berdimensi n yang paling baik digunakan untuk merepresentasikan data, dimana $k \leq n$. Data asli dengan demikian diproyeksikan ke ruang yang jauh lebih kecil, menghasilkan reduksi dimensi.

Tujuan pemilihan sub himpunan atribut adalah untuk menemukan sekumpulan atribut minimum sedemikian rupa sehingga distribusi probabilitas kelas data sedekat mungkin dengan distribusi *dataset* asli yang menggunakan semua atribut. Secara umum ada tiga metode pemilihan fitur, yaitu *filter*, *wrapper*, dan *embedded*. Metode *filter* tidak bergantung pada algoritma pembelajaran mesin, tetapi fitur dipilih dengan menggunakan berbagai analisis statistik.

Metode *wrapper* memerlukan percobaan-percobaan pelatihan dan pengujia model pembelajaran mesin dengan menggunakan sub himpunan fitur terpilih. Metode ini biasanya memerlukan komputasi yang tinggi karena pada umumnya merupakan algoritma *greedy* dalam melakukan pencarian yang menyeluruh. Ada tiga kategori metode-metode dengan pendekatan ini, yaitu *stepwise forward selection*, *stepwise backward elimination*, serta kombinasi dari *forward selection* dan *backward elimination*. Prosedur *stepwise forward selection* dimulai dari himpuan atribut yang kosong. Kemudian, atribut asli yang terbaik pada tiap iterasi ditentukan dan ditambahkan ke himpunan tersebut sampai memenuhi kriteria henti. Sementara itu prosedur *stepwise backward elimination* dimulai dengan himpunan atribut asli, kemudian atribut yang terburuk yang pada tiap iterasi dibuang sampai memenuhi kriteria henti.

Kombinasi dari *forward selection* dan *backward elimination* dapat menentukan serta menambahkan atribut asli yang terbaik sekaligus membuang atribut asli yang terburuk pada tiap iterasi sampai memenuhi kriteria henti.

Sementara itu, metode *embedded* merupakan kombinasi dari metode *filter* dan *wrapper*. Salah satu teknik reduksi dimensi yang termasuk dalam metode ini adalah induksi pohon keputusan. Algoritma pohon keputusan, seperti ID3, C4.5, dan CART, digunakan untuk membuat pohon yang terdiri dari atribut-atribut terbaik yang dikonstruksi dari *dataset*. Atribut-atribut yang tidak muncul dalam pohon keputusan diasumsikan tidak relevan sehingga dapat dibuang.

Materi pada bagian ini menjelaskan salah satu teknik yang sederhana dalam memiilih fitur, yang termasuk dalam metode *filter*, yaitu Pearson *correlation*. Korelasi dalam istilah statistik mengacu pada seberapa dekat dua variabel memiliki hubungan linier satu sama lain. Misalnya, dua variabel yang berkorelasi secara linear ($x = 2y$) memiliki korelasi yang lebih tinggi daripada dua variabel yang tidak bergantung secara linier ($u = v^2$). Fitur dengan korelasi tinggi memiliki efek yang hampir sama pada variabel dependen. Jadi, ketika dua fitur memiliki korelasi yang tinggi, salah satu dari dua fitur tersebut dapat dibuang. Koefisien Person *correlation* (r) antara variable x dan y dengan jumlah data n, dapat dihitung dengan formula berikut:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\, n\sum x^2 - (\sum x)^2 \,][\, n\sum y^2 - (\sum y)^2 \,]}}$$

Tabel 6 Dataset Diabetes

ID	Tinggi badan (cm)	Berat badan (kg)	Usia (tahun)	Diabetes
1	155	84	50	Ya
2	165	73	40	Tidak
3	167	88	54	Ya
4	150	76	45	Tidak

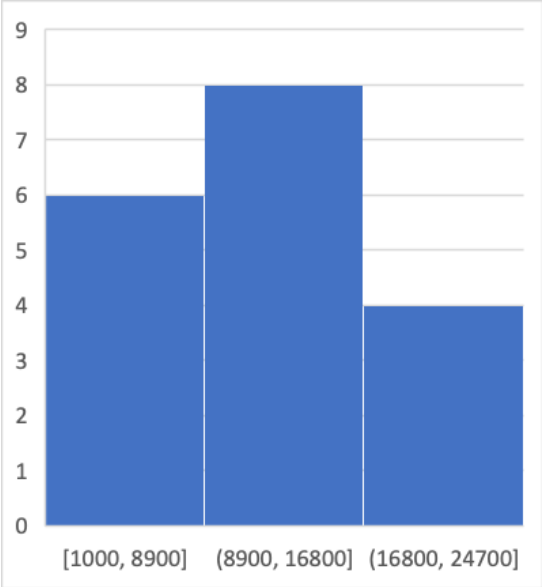
Pada contoh *dataset* dalam Tabel 6.1, koefisien Pearson *correlation* antara attribut tinggi badan dan berat badan adalah 0.25, antara atribut tinggi badan dan usia adalah 0.13, dan antara atribut berat badan dan usia adalah 0.99. Karena berat badan dan usia mempunyai korelasi yang tinggi, maka salah

satu dari kedua atribut ini dapat dibuang untuk mereduksi dimensi data.

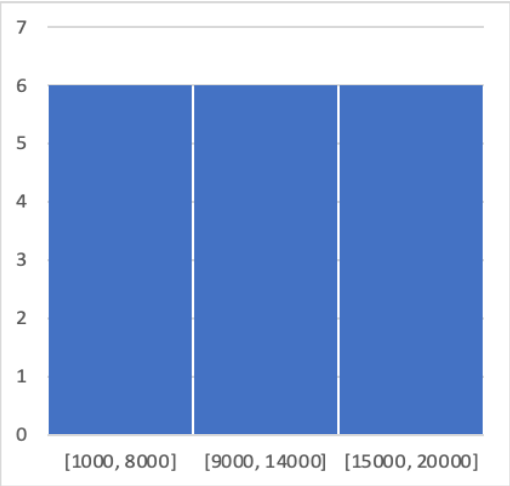
A.2 Reduksi numerik

Reduksi numerik merupakan proses menggantikan nilai asli dengan representasi yang lebih kecil. Ada dua tipe teknik dalam mereduksi data dengan attribut numerik, yaitu parametrik dan non-parametrik. Pada pendetakatan parametrik, sebuah model digunakan untuk mengestimasi data, sehingga pada umumnya hanya data parameter saja yang butuh disimpan, alih-alih data yang sebenarnya. Model regresi dan log-linear adalah contoh teknik dengan pendekatan ini. Sementara itu, contoh pendekatan non-parametrik antara lain *histogram*, *clustering*, *sampling*, dan *data cube aggregation*.

Materi pada bagian ini menjelaskan salah satu teknik yang sederhana dalam mereduksi data dengan atribut numerik, yaitu *binning*. Histogram menggunakan *binning* untuk memperkirakan distribusi data. Histogram untuk suatu atribut mempartisi distribusi data menjadi himpunan bagian yang terpisah-pisah, yang disebut dengan *bin*. Misalkan data harga barang-barang yang terjual di suatu toko adalah 1000, 1000, 5000, 5000, 5000, 8000, 9000, 10000, 10000, 12000, 14000, 14000, 15000, 15000, 18000, 18000, 18000, 20000. Histogram dapat digunakan untuk mereduksi data ini. Seperti terlihat pada Gambar 6.1, ada tiga *bin* yang dibuat dengan *range* yang seragam, dengan nilai 8900, untuk mengelompokkan data. Sehingga, atribut dengan nilai 1000-8900, dapat digantikan dengan nilai 1, atribut dengan nilai 8900-16800 dapat digantikan dengan nilai 2, dan atribut dengan nilai 16800-24700 dapat digantikan dengan nilai 3. Aturan partisi ini dinamakan dengan *equal-width binning*. Selain *equal-width*, histogram juga dapat dibuat dengan aturan *equal-frequency* (atau *equal-depth*), dimana frekuensi tiap *bin* bernilai (kurang lebih) sama, seperti terlihat pada Gambar 6.2. Sehingga, atribut dengan nilai 1000-8000, dapat digantikan dengan nilai 1, atribut dengan nilai 9000-14000, dapat digantikan dengan nilai 2, dan atribut dengan nilai 15000-20000 dapat digantikan dengan nilai 3. Dengan *binning*, nilai atribut harga yang tadinya mempunyai *range* antara 1000 sampai 20000 dapat direduksi menjadi nilai 1-3.



Gambar 13 Histogram equal-width



Gambar 14 Histogram equal-frequency

A.3 Kompresi Data

Dalam kompresi data, transformasi diterapkan untuk mendapatkan representasi data asli yang direduksi. Jika data asli dapat direkonstruksi dari data terkompresi tanpa kehilangan informasi, teknik mereduksi data ini disebut *lossless*. Sebaliknya, jika hanya perkiraan data asli yang dapat direkonstruksi, maka teknik mereduksi data ini disebut *lossy*. Reduksi dimensi dan reduksi numerik juga dapat dianggap sebagai bentuk kompresi data.

B. Diskritisasi

Diskritisasi adalah proses mengubah nilai pada atribut numerik dengan memetakan nilai-nilai pada atribut tersebut ke atribut diskrit, dapat berupa label interval ataupun label konseptual. Histogram dan *binning* yang sudah dijelaskan sebelumnya termasuk teknik diskritisasi. *Binning* adalah teknik diskritisasi *top-down* karena jumlah *bin* dispesifikasikan oleh *user*. Seperti *binning*, histogram juga adalah teknik diskritisasi *unsupervised* karena keduanya tidak menggunakan informasi kelas. Selain itu, diskritisasi juga dapat dilakukan dengan klasterisasi, pohon keputusan, atau analisis korelasi.

Algoritma-algoritma klasterisasi dapat dilakukan untuk mendiskritiisasi atribut-atribut numerik. Algoritma k-means, misalnya, dapat mempartisi dataset menjadi beberapa klaster sejumlah *k*. Sementara itu, pohon keputusan merupakan teknis diskritisasi supervised. Sedangkan, analisis korelasi, misalnya dengan menggunakan ChiMerged, merupakan pendekatan *bottom-up* berbasis x^2. Awalnya, setiap nilai yang berbeda dari suatu atribut numerik dianggap sebagai satu interval. Uji x^2 dilakukan untuk setiap pasangan interval yang berdekatan. Interval yang berdekatan dengan nilai x^2 terkecil digabungkan menjadi satu, karena nilai x^2 yang rendah menunjukkan distribusi kelas yang mirip. Proses penggabungan ini berlangsung secara rekursif sampai kriteria henti yang telah ditetapkan terpenuhi.

C. Rangkuman

Rescalling merupakan salah satu teknik dalam *preprocessing* data untuk menstandarisasi data agar mempunyai skala yang sama atau bahkan mereduksi data sehingga dimensinya menjadi lebih kecil. *Rescalling* dilakukan dengan cara melakukan reduksi dan diskritisasi data.

Teknik reduksi diaplikasikan untuk mendapatkan representasi dataset dengan volume yang jauh lebih kecil, namun tetap mempertahankan integritas data aslinya. Reduksi dimensi adalah proses mengurangi jumlah atribut atau fitur suatu *dataset*, sehingga dapat mereduksi volume data asli. Reduksi numerik merupakan proses menggantikan nilai asli dengan representasi yang lebih kecil.

Diskritisasi adalah proses mengubah nilai pada atribut numerik dengan memetakan nilai-nilai pada atribut tersebut ke atribut diskrit.

D. Evaluasi

Untuk mengetahui tingkat pemahaman Anda terkait materi Data *Rescaling,* lakukanlah reduksi dan kompresi untuk data di bawah ini!

1. Lakukan data rescalling untuk data di bawah ini. Praktikkan teknik reduksi dan diskritisasi data!

MATERI POKOK 8:

MISSING VALUE

Indikator Hasil Belajar:
Peserta mampu mendeteksi dan mengatasi missing value. Peserta dinyatakan lulus apabila mampu:

1. mengidentifikasi data dengan *missing value,*
2. melakukan penanganan *missing value* dengan metode yang tepat.

A. Definisi Missing Value

Missing value adalah adanya sebagian data yang hilang nilainya. Penanganan Missing Value menjadi bagian dari penyiapan data (data preprocessing) yang penting, yaitu *data cleaning* (pembersihan data. Data yang mengandung missing value sering juga disebut sebagai data tidak lengkap (incomplete data). *Missing value*, *noise* dan inkonsistensi berkontribusi kepada data yang tidak akurat. Data Mining melakukan pemodelan berbasiskan dataset input, sehingga bila dataset input tidak akurat, maka model yang dihasilkan juga akan memiliki akurasi output yang rendah.

B. Teknik Penanganan Missing Value

Untuk mengkoreksi missing value, kita dapat mengadopsi beberapa teknik. Berikut adalah teknik yang bisa digunakan untuk memperbaiki data yang mengalami *missing value.*

1. Teknik Eliminasi

   Teknik eliminasi, yaitu dengan membuang seluruh record yang satu atau lebih attributnya hilang atau kosong. Teknik ini cocok bila dalam satu record yang memiliki beberapa atribut yang kosong. Dalam kasus analisis Data Mining bertipe *supervised* seperti klasifikasi dan regresi, sangat penting untuk membuang record yang atribut targetnya *missing* atau kosong. Teknik ini juga cocok untuk kasus di mana distribusi *missing value* sangat besar variasinya dan irreguler tersebar di atribut-atribut yang berbeda. Dalam hal, kolom/atribut yang sebagian besarnya kosong, maka lebih baik kita membuang atribut kolom tersebut.

2. Teknik Inspeksi

   Teknik inspeksi**,** yaitu pakar melakukan inspeksi secara manual kepada setiap *missing value* dalam rangka mendapatkan rekomendasi nilai-nilai yang mungkin dapat diisikan manual sebagai pengganti *missing value* tersebut. Teknik ini memerlukan waktu yang banyak dan tidak cocok untuk dataset yang besar. Teknik ini juga lemah untuk data dengan ketidakaturan yang tinggi dan subyektifitas pakar yang tinggi. Tetapi bila dilakukan oleh pakar dengan kemampuan tinggi, pengalaman menunjukkan bahwa teknik ini menghasilkan tingkat keakuratan yang tinggi.

3. Teknik Identifikasi

   Teknik identifikasi**,** yaitu menggantikan *missing value* dengan nilai konstan tertentu atau label sebagai kode, misal “NA”, “unknown” atau lainnya. Meskipun demikian, bila *missing value* digantikan oleh misal “unknown” maka program penambang (data mining) bisa jadi secara salah menganggap sebagai bentuk konsep tertentu, karena mereka memiliki bentuk yang sama, yaitu “unknown”. Oleh karena itu, meski teknik ini sederhana, tetapi tidak selalu benar.

4. Teknik Substitusi

   Teknik substitusi*,* yaitu penggantian *missing value* secara otomatis, misalkan menggunakan ukuran tengah (rata-rata atau median) dari atribut. Untuk data dengan distribusi normal (simetris) dapat digunakan nilai rata-rata atribut. Untuk data dengan distribusi tidak normal (miring, skewed), hendaknya digunakan nilai median.

5. Variasi dari Teknik Substitusi

   Variasi dari teknik substitusi**,** khususnya untuk teknik data mining supervised, adalah menggunakan nilai rata-rata atau median untuk seluruh sample/record dalam kelas yang sama, atau kita sebut dengan **subsitusi per class**. Misal untuk dataset *credit_risk* (https://www.kaggle.com/datasets/laotse/credit-risk-dataset), mungkin kita akan mengganti *missing value* dengan nilai rata-rata *income* (pemasukan) untuk seluruh

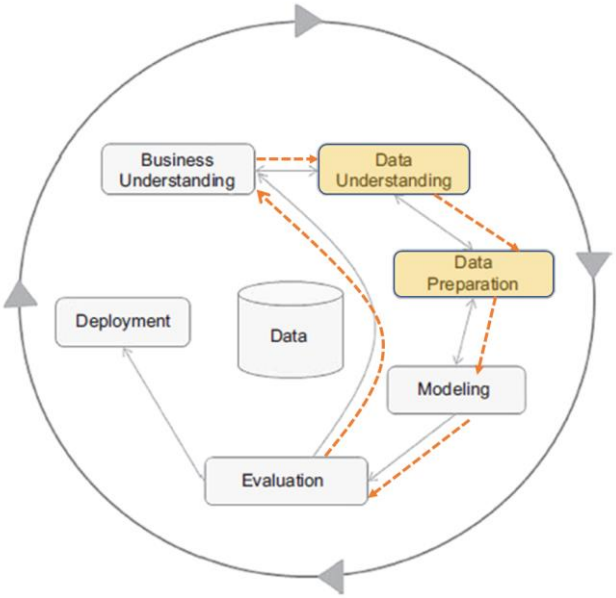
customer dengan kategori *credit_risk* yang sama denga record/tuple tersebut. Bila distribusi data untuk *class* tersebut miring/skewed, maka penggantian dengan nilai median lebih baik.

6. Variasi Lain dari Teknik Substitusi

   Variasi lain dari teknik **substitusi,** selain dengan nilai rata-rata atau median, adalah dengan penggantian yang ditentukan oleh algoritma data mining seperti regresi, tool berbasis bayesian atau induksi *decision-tree*. Kita sebut ini dengan **substitusi dengan algoritma.**

   Metode identifikasi dan substitusi (teknik 3 - 6), sangat bias terhadap data, dengan kata lain, nilai yang dimasukkan bisa jadi tidak benar. Namun demikian teknik 6 merupakan strategi yang populer. Dibandingkan dengan teknik lainnya, teknik 6 ini lebih cepat dengan otomasi, dan juga teknik 6 ini memanfaatkan informasi yang ada dalam data secara maksimal, untuk memprediksi *missing value*.

C. Praktik Penanganan Missing Value



Gambar 15 Proses CRISP Data Mining

Ruang lingkup materi *Missing Value* ini ada di *Data Preparation*

sebagaimana dalam Gambar XX. Proses CRISP Data Mining. Fokus utama kita di praktik ini adalah di 2 tahapan: *Data Understanding* dan *Data Preparation* (khususnya dengan *Missing Value*). Tetapi praktik yang akan kita lakukan dalam tahapan *Business Understanding → Data Understanding → Data Preparation → Modeling → Evaluation → Business Understanding → …*dan seterusnya.

Tahapan *Business Understanding.* Di tahapan ini kita mengetahui bahwa kita akan mencoba membangun model untuk prediksi tingkat keselamatan penumpang dari kecelakaan sebagaimana kecelakaan kapal Titanic.

Tahapan *Data Understanding.* Di tahapan ini kita mencoba memahami data. Informasi data secara lengkap didapatkan dari link url di atas. Bagaimana karakteristik data, khususnya kondisi *missing value (incomplete data)*.

Tahapan *Data Preparation*. Di tahapan ini kita mencoba memperbaiki data khususnya *missing value*, dengan memilih salah satu teknik penangangannya.

Tahapan *Modeling*. Di tahapan ini kita akan mencoba menggunakan teknik data mining kategori klasifikasi untuk latihan ini.

Tahapan *Evaluasi.* Di tahapan ini kita mengukur akurasi model yang telah dihasilkan.

Kemudian kita ulangi lagi proses dengan teknik penanganan *missing value* yang berbeda, dan kita bandingkan akurasinya.

Praktik penanganan Missing Value akan menggunakan dataset publik dari Kaggle, yaitu dataset *Titanic*, dataset *credit risk* dan dataset *Melbourne Housing*. Di sesi ini, contoh akan menggunakan dataset *Titanic* dan latihan akan menggunakan dataset *credit risk* dan *Melbourne Housing*.

Sumber dataset:
https://www.kaggle.com/competitions/titanic/data
https://www.kaggle.com/datasets/laotse/credit-risk-dataset
https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot

Dataset *Titanic* adalah data penumpang kapal Titanic yang mengalami kecelakaan. Dataset *Titanic* ini merupakan dataset untuk prediksi dan sudah dipisahkan antara dataset *training* dan dataset *test*. Pengembangan model prediktif (learning) dilakukan menggunakan dataset *training*, sedangkan dataset *test* untuk mengukur keakuratan model prediktif yang dihasilkan. Penanganan *missing value* yang lebih

baik, akan menghasilkan model prediktif yang lebih akurat.

Dalam praktik ini, kita akan melakukan semua teknik penanganan *missing value*, kecuali teknik kedua inspeksi. Kemudian kita akan mencoba membandingkan efektifitas semua teknik tersebut dengan cara melihat perbedaan akurasi algoritma tertentu, misalkan *Decision Tree*.

1. Praktik dengan Python

Dalam praktik dengan Python, kita dapat menggunakan contoh *source code* yang dapat dilihat atau didownload di :

- https://github.com/jfadibrin/modul2022/tree/main/M06_DataMining/M06S08_MissingValue

Lingkungan Python yang digunakan dalam praktik ini adalah Jupyter Notebook, berbasis Docker Container jupyter/datascience-notebook. Informasi lebih lanjut tentang lingkungan docker tersebut dapat dilihat di

- https://hub.docker.com/r/jupyter/datascience-notebook

a. Teknik Eliminasi

Tahap pertama adalah memahami data (*Data Understanding*), yang dapat kita ikuti sesuai *source code* titanic00_EDA.ipynb. Pertama menggunakan library standard Pandas, untuk melakukan Analisis Eksplorasi Data (*Exploratory Data Analysis)*. Berdasarkan analisis eksplorasi data tersebut, diketahui jenis, karakter masing-masing atribut data dan juga atribut yang memiliki *missing value*.

Berdasarkan hasil EDA sebelumnya, dapat diketahui bahwa:

- field/kolom/atribut **PassengerId** karena hanya sebagai nomor urut atau unique variabel untuk setiap record, tidak kita perlukan.
- field/kolom/atribut **Name** memiliki jumlah unique sebanyak jumlah record total, yaitu 891, sehingga tidak signifikan untuk pemodelan prediksi.
- field/kolom/atribut **Ticket** memiliki jumlah unique hampir sebanyak jumlah record total, yaitu 681 (85%), sehingga tidak signifikan untuk pemodelan prediksi.
- ada 3 field yang memiliki missing value yaitu: **age, cabin** dan

embarked. Dengan teknik eliminasi ini, kita membuang ketiga atribut tersebut.

Sekarang kita masuk ke tahapan *Data Preparation* (Penyiapan Data), dan berdasarkan hasil analisis EDA di atas, maka kita akan membuang enam atribut: PassengerId, Name, Ticket, Age, Cabin dan Embarked. Berikut urutan program python: load library, load data, eliminasi dan menggunakan data hasil olahan (eliminasi) untuk algoritma *Decision Tree.*

import library Python yang diperlukan untuk analisi data: Numpy dan Panda

[1]:
```
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.tree import DecisionTreeClassifier # Import Decision Tree Classifier
from sklearn.model_selection import train_test_split # Import train_test_split function
from sklearn import metrics #Import scikit-learn metrics module for accuracy calculation
```

[2]:
```
df = pd.read_csv('../dataset/titanic/train.csv')
dft = df.copy()
```

Eliminasi attribut dengan perintah **drop()** dari library Pandas

[3]:
```
dft.drop(["PassengerId","Name","Ticket","Cabin","Age","Embarked"], axis=1, inplace=True)
dft
```

[3]:

	Survived	Pclass	Sex	SibSp	Parch	Fare
0	0	3	male	1	0	7.2500
1	1	1	female	1	0	71.2833
2	1	3	female	0	0	7.9250
3	1	1	female	1	0	53.1000
4	0	3	male	0	0	8.0500
...	...	...	...	...	...	...
886	0	2	male	0	0	13.0000
887	1	1	female	0	0	30.0000
888	0	3	female	1	2	23.4500
889	1	1	male	0	0	30.0000
890	0	3	male	0	0	7.7500

891 rows × 6 columns

```
[5]: dft.isnull().sum()
```

```
[5]: Survived    0
     Pclass      0
     Sex         0
     SibSp       0
     Parch       0
     Fare        0
     dtype: int64
```

```
[6]: dft.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Survived  891 non-null    int64
 1   Pclass    891 non-null    int64
 2   Sex       891 non-null    object
 3   SibSp     891 non-null    int64
 4   Parch     891 non-null    int64
 5   Fare      891 non-null    float64
dtypes: float64(1), int64(4), object(1)
memory usage: 41.9+ KB
```

Dari perintah **isnull()**, dan **info()** terlihat bahwa sudah tidak ada lagi *missing value (null)* di dalam data. Sekarang berada pada tahap *modeling* (pemodelan) dengan *Decision Tree*. Pertama-tama kita melakukan **Seleksi Fitur *(Feature Selection).*** Sebagaimana algoritma *supervised* lainnya, kita perlu membagi atribut menjadi 2 bagian. Satu bagian terdiri dari 1 atribut, sebagai **variabel *target***, dan sisanya adalah sebagai **variabel *independent***.

Dalam hal ini, atribute **survived** adalah variabel *target*, dan sisanya adalah variabel *independent*. Kemudian, kita membagi data menjadi data training (learning) dan data testing. Kali ini kita memilih cara pembagian data yang sederhana yaitu data testing adalah 30% dari total dataset.

Selanjutnya, untuk pemodelan dengan algoritma Decision Tree, jenis data harus numerik. Sedangkan kondisi dataset sekarang masih ada beberapa yang non-numerik, yaitu Sex:male,female. Oleh karena itu atribute Sex harus ditransformasi menjadi nilai numerik. Dengan menggunakan tool LabelEncoder dari sklearn.preprocessing, kita dapat merubah secara otomatis atribute Sex menjadi male = 1, dan female = 0.

[7]:
```
le = LabelEncoder()
dft['Sex'] =  le.fit_transform(dft['Sex'])
```

[8]:
```
dft.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Survived  891 non-null    int64
 1   Pclass    891 non-null    int64
 2   Sex       891 non-null    int64
 3   SibSp     891 non-null    int64
 4   Parch     891 non-null    int64
 5   Fare      891 non-null    float64
dtypes: float64(1), int64(5)
memory usage: 41.9 KB
```

[11]:
```
#split dataset in features and target variable
feature_cols = ['Pclass','Sex','SibSp','Parch','Fare']
X = dft[feature_cols] # Features
y = dft.Survived # Target variable
```

[15]:
```
# Split dataset into training set and test set
# 70% training and 30% test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
```

Selanjutnya dalam tahapan pemodelan ini kita melakukan proses *training,* yaitu menggunakan data sebagai input untuk algoritma *Decision Tree*.

[13]:
```
# Create Decision Tree classifer object
clf = DecisionTreeClassifier()

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)

#Predict the response for test dataset
y_pred = clf.predict(X_test)
```

Tahap akhir adalah mengukur akurasi dari model berbasis *Decision Tree* dan data hasil penanganan *missing value* dengan teknik **eliminasi.**

[14]:
```
# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7611940298507462
```

b. Teknik Identifikasi

Teknik identifikasi dilakukan dengan mengganti *missing value*, misal dengan NaN. Karena kita menggunakan Pandas saat meload data ke python, secara otomatis missing value digantikan dengan nilai NaN (not a number). Tetapi algoritma *Decision Tree*, yang memanfaatkan perhitungan nilai numerik, sehingga kita perlu mengubah nilai non-numerik variable Age, Cabin dan Embarked menjadi numerik dengan memanfaatkan tool **LabelEncoder** sebagaimana perlakuan kita pada variable Sex.

Berikut contoh kode:

import library Python yang diperlukan untuk analisi data: Numpy dan Panda

```
[1]: import pandas as pd
     from sklearn.preprocessing import LabelEncoder
     from sklearn.tree import DecisionTreeClassifier # Import Decision Tree Classifier
     from sklearn.model_selection import train_test_split # Import train_test_split function
     from sklearn import metrics #Import scikit-learn metrics module for accuracy calculation
```

Load data dan memahami data

Untuk secara detil analisis eksplorasi data secara singkat, dapat melihat ke Notebook titanic00_EDA

```
[2]: df = pd.read_csv('../dataset/titanic/train.csv')
     dft = df.copy()
```

Sama saat kita menggunakan teknik Eliminasi di atas, kita membuang variabel PassengerId, Name dan Ticket. Namun berbeda dengan teknik Eliminasi, di teknik subsitusi ini kita mempertahankan variabel Age, Cabin dan Embarked.

```
[3]: dft.drop(["PassengerId","Name","Ticket"], axis=1, inplace=True)
```

```
[4]: dft.describe(include = "all")
```

[4]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
count	891.000000	891.000000	891	714.000000	891.000000	891.000000	891.000000	204	889
unique	NaN	NaN	2	NaN	NaN	NaN	NaN	147	3
top	NaN	NaN	male	NaN	NaN	NaN	NaN	B96 B98	S
freq	NaN	NaN	577	NaN	NaN	NaN	NaN	4	644
mean	0.383838	2.308642	NaN	29.699118	0.523008	0.381594	32.204208	NaN	NaN
std	0.486592	0.836071	NaN	14.526497	1.102743	0.806057	49.693429	NaN	NaN
min	0.000000	1.000000	NaN	0.420000	0.000000	0.000000	0.000000	NaN	NaN
25%	0.000000	2.000000	NaN	20.125000	0.000000	0.000000	7.910400	NaN	NaN
50%	0.000000	3.000000	NaN	28.000000	0.000000	0.000000	14.454200	NaN	NaN
75%	1.000000	3.000000	NaN	38.000000	1.000000	0.000000	31.000000	NaN	NaN
max	1.000000	3.000000	NaN	80.000000	8.000000	6.000000	512.329200	NaN	NaN

Sebagaimana terlihat di gambar di atas, masih ada beberapa variabel yang memiliki missing value. Ditunjukkan juga oleh perintah isnull() dan info() seperti gambar di bawah ini. Variabel Sex, Cabin dan Embarked memiliki tipe data Object, artinya non numerik. Sedangkan variabel Age, meskipun bertipe data numerik (float64), tetapi kita tahu ada 177 record yang *missing value*, artinya juga bernilai NaN.

```
[5]: dft.isnull().sum()
```

```
[5]: Survived      0
     Pclass        0
     Sex           0
     Age         177
     SibSp         0
     Parch         0
     Fare          0
     Cabin       687
     Embarked      2
     dtype: int64
```

```
[6]: dft.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Survived  891 non-null    int64
 1   Pclass    891 non-null    int64
 2   Sex       891 non-null    object
 3   Age       714 non-null    float64
 4   SibSp     891 non-null    int64
 5   Parch     891 non-null    int64
 6   Fare      891 non-null    float64
 7   Cabin     204 non-null    object
 8   Embarked  889 non-null    object
dtypes: float64(2), int64(4), object(3)
memory usage: 62.8+ KB
```

Untuk dapat digunakan sebagai data input di algoritma *Decision Tree* dari library sklearn, maka variabel Sex, Age, Cabin dan Embarked harus dirubah menjadi kategorikal dengan kode numerik secara otomatis dengan memanfaatkan tool LabelEncoder.

[7]:
```
le = LabelEncoder()
dft['Sex'] =  le.fit_transform(dft['Sex'])
dft['Age'] = le.fit_transform(dft['Age'])
dft['Cabin'] =  le.fit_transform(dft['Cabin'])
dft['Embarked'] = le.fit_transform(dft['Embarked'])
```

[8]:
```
dft.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Survived  891 non-null    int64
 1   Pclass    891 non-null    int64
 2   Sex       891 non-null    int64
 3   Age       891 non-null    int64
 4   SibSp     891 non-null    int64
 5   Parch     891 non-null    int64
 6   Fare      891 non-null    float64
 7   Cabin     891 non-null    int64
 8   Embarked  891 non-null    int64
dtypes: float64(1), int64(8)
memory usage: 62.8 KB
```

Sekarang data kita telah siap untuk diaplikasikan sebagai data input kepada pemodelan dengan algoritma *Decision Tree*.

[9]:
```
#split dataset in features and target variable
feature_cols = ['Pclass','Sex','SibSp','Parch','Fare','Age','Cabin','Embarked']
X = dft[feature_cols] # Features
y = dft.Survived # Target variable
```

[10]:
```
# Split dataset into training set and test set
# 70% training and 30% test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
```

Training: membangun model prediksi dengan Decision Tree

[11]:
```
# Create Decision Tree classifer object
clf = DecisionTreeClassifier()

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)

#Predict the response for test dataset
y_pred = clf.predict(X_test)
```

Evaluasi akurasi dari model yang dihasilkan:

[12]:
```
# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7425373134328358
```

Akurasi dengan teknik identifikasi seperti di atas, menghasilkan akurasi yang lebih rendah dibandingkan dengan teknik eliminasi

sebagaimana penjelasan sebelumnya.

c. Teknik Substitusi

Untuk mengganti *missing value* di dalam Python, selain menggunakan fungsi fillna() atau tools LabelEncoder sebagaimana dijelaskan di teknik sebelumnya, teknik lain adalah menggunakan sklearn.impute.SimpleImputer. Dengan library sklearn.impute.SimpleImputer, kita dengan mudah melakukan penggantian variabel *missing value* dengan substitusi. Ada 4 kategori substitusi, yaitu substitusi dengan:

a) Rata-rata (mean)
b) Median
c) Nilai terbanyak (most frequent)
d) Nilai Constant

Contoh kode program untuk subsitusi *missing value* dapat dilihat pada SimpleImputer.ipynb

```
import pandas as pd
import numpy as np
from sklearn.impute import SimpleImputer

# Importing the SimpleImputer class
from sklearn.impute import SimpleImputer
```

```
dataSiswa = [[85, 'L', 'sangatbagus'],
             [95, 'P', 'sempurna'],
             [75, None,'bagus'],
             [np.NaN, 'L', 'cukup'],
             [70, 'L', 'bagus'],
             [np.NaN, None, 'sangatbagus'],
             [92, 'P', 'sangatbagus'],
             [98, 'L', 'sempurna']]
```

[46]:
```
dfsiswa = pd.DataFrame(dataSiswa)
dfsiswa.columns = ['nilai', 'gender', 'hasil']
dfsiswa
```

[46]:

	nilai	gender	hasil
0	85.0	L	sangatbagus
1	95.0	P	sempurna
2	75.0	None	bagus
3	NaN	L	cukup
4	70.0	L	bagus
5	NaN	None	sangatbagus
6	92.0	P	sangatbagus
7	98.0	L	sempurna

[40]:
```
dfsiswa.isnull().sum()
```

[40]:
```
nilai     2
gender    2
hasil     0
dtype: int64
```

Field nilai dan gender, masing-masing memiliki 2 *missing value*

Subsitusi dengan mean.

[11]:
```
dfsiswa2 = dfsiswa.copy()
#
# Missing values direpresentatikan dengan NaN sehingga terspesifikasi.
# Jika dia adalah data kosong, maka missing value terspesifikasi
# sebagai ''
imputer = SimpleImputer(missing_values=np.NaN, strategy='mean')

dfsiswa2.nilai = imputer.fit_transform(dfsiswa2['nilai'].values.reshape(-1,1))[:,0]

dfsiswa2
```

[11]:

	nilai	gender	hasil
0	85.000000	L	sangatbagus
1	95.000000	P	sempurna
2	75.000000	None	bagus
3	85.833333	L	cukup
4	70.000000	L	bagus
5	85.833333	None	sangatbagus
6	92.000000	P	sangatbagus
7	98.000000	L	sempurna

Di contoh ini, terlihat bahwa *missing value* pada atribut/variabel nilai tergantikan dengan nilai rata-rata yaitu: 85.833333

Sebagaimana dijelaskan di sesi A, bahwa data dengan tipe numerik dapat menggunakan 4 strategi: mean, median, most_frequent dan constant. Tetapi untuk data kategorikal seperti variabel/atribute gender, tidak dapat menggunakan strategi mean dan median.

Strategi dengan median

[12]:
```
dfsiswa3 = dfsiswa.copy()
#
# Missing values direpresentatikan dengan NaN sehingga terspesifikasi.
# Jika dia adalah data kosong, maka missing value terspesifikasi
# sebagai ''
imputer = SimpleImputer(missing_values=np.NaN, strategy='median')

dfsiswa3.nilai = imputer.fit_transform(dfsiswa3['nilai'].values.reshape(-1,1))[:,0]

dfsiswa3
```

[12]:

	nilai	gender	hasil
0	85.0	L	sangatbagus
1	95.0	P	sempurna
2	75.0	None	bagus
3	88.5	L	cukup
4	70.0	L	bagus
5	88.5	None	sangatbagus
6	92.0	P	sangatbagus
7	98.0	L	sempurna

Untuk data kategorikal hanya dapat dilakukan dengan strategi nilai terbanyak muncul (most_frequent) atau digantikan dengan nilai tertentu (constant). Penggantian dengan nilai tertentu (constant) adalah teknik identifikasi, sebagaimana dijelaskan sebelumnya. Conoh di bawah ini adalah teknik substitusi dengan SimpleImputer

[23]:
```
dfsiswa4 = dfsiswa.copy()

imputer = SimpleImputer(missing_values=None, strategy='most_frequent')

dfsiswa4.gender = imputer.fit_transform(dfsiswa4['gender'].values.reshape(-1,1))[:,0]
dfsiswa4
```

[23]:

	nilai	gender	hasil
0	85.0	L	sangatbagus
1	95.0	P	sempurna
2	75.0	L	bagus
3	NaN	L	cukup
4	70.0	L	bagus
5	NaN	L	sangatbagus
6	92.0	P	sangatbagus
7	98.0	L	sempurna

Berikut kita akan mempratekkan SimpleImputer untuk data Titani. Contoh kode program dapat dilhat pada titanic03_Substitusi.ipynb

Kondisi data sebelum SimpleImputer, bahwa variabel Age, Cabin dan Embarked mengandung *missing value*. Age bertipe numerik, sehingga kita ganti dengan nilai mean, serta variabel Cabin dan Embarked kita ganti dengan nilai most_frequent.

[33]:
```
dft.columns = ['Survived','Pclass',
               'Sex', 'Age',
               'SibSp','Parch',
               'Fare', 'Cabin',
              'Embarked']
```

[45]:
```
#
# Missing values direpresentatikan dengan NaN sehingga terspesifikasi.
# Jika dia adalah data kosong, maka missing value terspesifikasi
# sebagai ''
imputer = SimpleImputer(missing_values=np.NaN, strategy='mean')

dft.Age = imputer.fit_transform(dft['Age'].values.reshape(-1,1))[:,0]
```

[46]:
```
#
# Missing values direpresentatikan dengan NaN sehingga terspesifikasi.
# Jika dia adalah data kosong, maka missing value terspesifikasi
# sebagai ''
imputer = SimpleImputer(missing_values=np.NaN, strategy='most_frequent')

dft.Cabin = imputer.fit_transform(dft['Cabin'].values.reshape(-1,1))[:,0]
dft.Embarked = imputer.fit_transform(dft['Embarked'].values.reshape(-1,1))[:,0]
```

[47]:
```
dft.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Survived  891 non-null    int64
 1   Pclass    891 non-null    int64
 2   Sex       891 non-null    int64
 3   Age       891 non-null    float64
 4   SibSp     891 non-null    int64
 5   Parch     891 non-null    int64
 6   Fare      891 non-null    float64
 7   Cabin     891 non-null    int64
 8   Embarked  891 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 62.8 KB
```

[31]:
```
dft.isnull().sum()
```

[31]:
```
Survived      0
Pclass        0
Sex           0
Age         177
SibSp         0
Parch         0
Fare          0
Cabin       687
Embarked      2
dtype: int64
```

[32]:
```
dft.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Survived  891 non-null    int64
 1   Pclass    891 non-null    int64
 2   Sex       891 non-null    object
 3   Age       714 non-null    float64
 4   SibSp     891 non-null    int64
 5   Parch     891 non-null    int64
 6   Fare      891 non-null    float64
 7   Cabin     204 non-null    object
 8   Embarked  889 non-null    object
dtypes: float64(2), int64(4), object(3)
memory usage: 62.8+ KB
```

Sekarang seluruh *missing value* sudah tergantikan. Kemudian kita jalankan kode program pemodelan dan perhitungan akurasi sebagaimana teknik sebelumnya (teknik identifikasi).

[13]:
```
#split dataset in features and target variable
feature_cols = ['Pclass','Sex','SibSp','Parch','Fare','Age','Cabin','Embarked']
X = dft[feature_cols] # Features
y = dft.Survived # Target variable
```

[14]:
```
# Split dataset into training set and test set
# 70% training and 30% test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
```

Training: membangun model prediksi dengan Decision Tree

[15]:
```
# Create Decision Tree classifer object
clf = DecisionTreeClassifier()

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)

#Predict the response for test dataset
y_pred = clf.predict(X_test)
```

Evaluasi akurasi dari model yang dihasilkan:

[16]:
```
# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7611940298507462
```

Sedangkan bila nilai Age kita ganti dengan median (titanic04_Substitusi.ipynb), nilai akurasi adalah: 0.7649253731343284

2. **Teknik Substitusi berbasis kelas**
3. **Teknik Substitusi dengan algoritma**

D. Komparasi

Berikut adalah rekapitulasi akurasi teknik penanganan *missing value* pada variabel Age, Cabin dan Embarked, serta akurasi hasil pemodelan dengan *Decision Tree* berbasis contoh Python.

Tabel 7 Contoh Rekapitulasi Akurasi Teknik Penanganan Missing Value

No	Teknik penangangan missing value	Akurasi
C.1.1.	Eliminasi	0.7611940298507462
C.1.2.	Identifikasi (dengan NaN)	0.7425373134328358
C.1.3	Substitusi (Age=mean, Cabin=most_frequent Embarked=most_frequent)	0.7649253731343284
C.1.3.	Substitusi (Age=median, Cabin=most_frequent Embarked=most_frequent)	0.7611940298507462
C.1.4	Substitusi per kelas	
C.1.5	Substitusi otomatis dg algoritma KNN	

1. Praktik dengan R

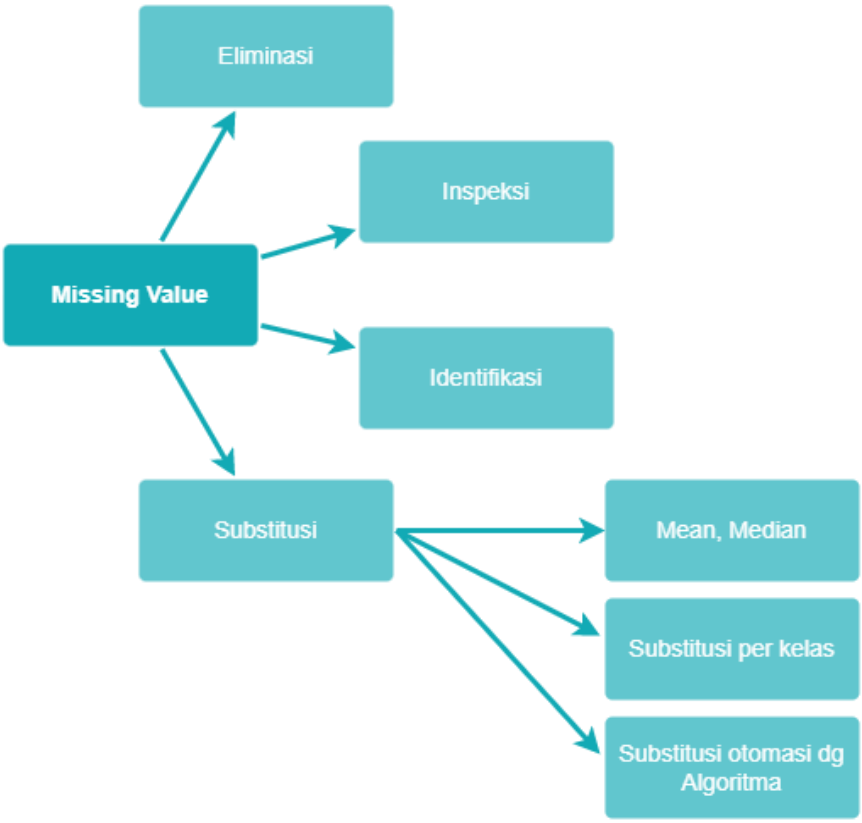
2. Praktik dengan RapidMiner

E. Rangkuman

Berisi uraian teknik penanganan *missing value* (data tidak lengkap) yaitu teknik eliminasi, inspeksi, identifikasi, substitusi, substitusi per kelas dan substitusi otomatis. Praktik teknik ini dilakukan untuk seluruh teknik, kecuali inspeksi. Praktik dengan contoh python, R dan RapidMiner.

Missing value adalah adanya sebagian data yang hilang nilainya. Untuk mengkoreksi missing value, kita dapat mengadopsi beberapa teknik antara lain eliminasi, inspeksi, identiikasi, substitusi (dan teknik modiikasinya)

Penanganan missing value bisa dilakukan dengan beberapa aplikasi seperti *phyton R*, dan *RapidMinder*



Gambar *16 Mind Map* Materi *Missing Value*

F. Evaluasi

Untuk menguji pemahaman Anda yerkait dengan materi Missing Value, silahkan kerjakan latihan di bawah ini!

1. Apa perbedaan antara teknik eliminasi, identifikasi dan substitusi?
2. Gunakan dataset Melbourne Housing Snapshot Dataset untuk mempraktekan minimal 2 teknik penanganan data tidak lengkap (*missing value*).

MATERI POKOK 9:

DATA SMOOTHING

Indikator Hasil Belajar:
Peserta mampu mengoreksi *noise* yang ada dalam data. Peserta dinyatakan lulus manakala mampu :

1. menjelaskan konsep teknik smoothing dengan benar,
2. mempraktikkan teknik *smoothing* data set dengan benar.

A. Teknik Smoothing

Suatu dataset dapat memiliki semacam *noise*. Noise adalah kesalahan acak atau variasi dalam suatu variabel. Teknik-teknik berikut ini dapat digunakan untuk kita bisa "memuluskan" data untuk menghilangkan *noise*.

1. Binning

Metode Binning memuluskan nilai data yang diurutkan dengan berkonsultasi melihat "lingkungannya", yaitu, nilai-nilai di sekitarnya. Nilai-nilai dalam variabel diurutkan dan didistribusikan ke dalam beberapa `bin ("ember") atau keranjang,j kemudian angka-angka dalam satu bin, dibuat sama. Metode binning ini adalah perataan lokal. Contoh di bawah ini menggambarkan proses *smoothing* lokal dengan binning.

Data asli. Harga dalam dalam dolar yang telah diurutkan: 4, 8, 15, 21, 21, 24, 25, 28, 34
Partisi dengan bin ***equal-frequency***: Bin 1: 4, 8, 15 Bin 2: 21, 21, 24 Bin 3: 25, 28, 34
Smoothing bin dengan **mean**: Bin 1: 9, 9, 9 Bin 2: 22, 22, 22

Bin 3: 29, 29, 29
Smoothing bin dengan ***boundary/*batas**: Bin 1: 4, 4, 15 Bin 2: 21, 21, 24 Bin 3: 25, 25, 34

Alternatif lainnya, bin mungkin memiliki lebar yang sama, di mana rentang interval nilai di setiap bin adalah konstan. Binning juga digunakan sebagai salah satu teknik diskritisasi.

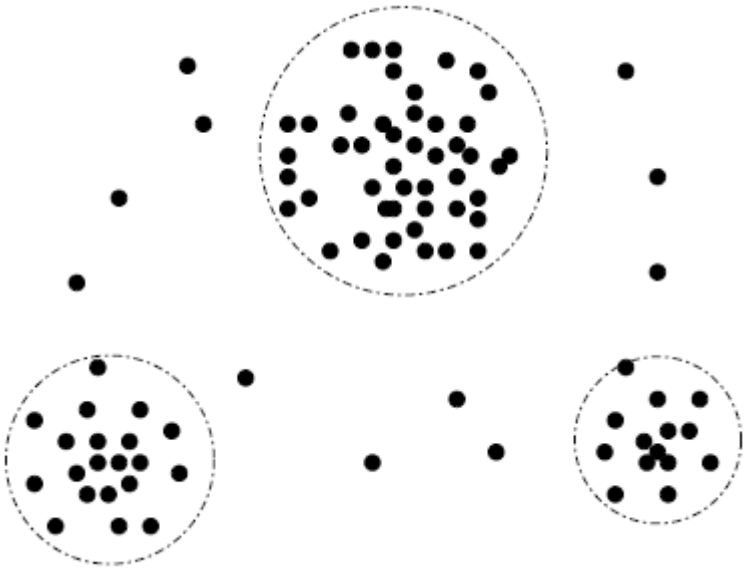
2. Regresi

Smoothing data juga dapat dilakukan dengan regresi, suatu teknik yang menyesuaikan nilai data ke suatu fungsi. Regresi linier melibatkan pencarian garis lurus "terbaik" yang ***fit*** (pas) menghubungkan dua data dalam suatu atribut/variabel, sehingga satu data dapat digunakan untuk memprediksi data lainnya.

Regresi linier berganda merupakan perluasan dari regresi linier, dimana lebih dari dua\ atribut terlibat dan data sesuai dengan permukaan multidimensi.

3. Deteksi Outlier

Outlier dapat ditemukan dengan teknik clustering. Sebagai contoh ketika nilai-nilai yang sama terorganisir dalam beberapa grup atau cluster. Secara intuitif nilai-nilai yang terletak jauh dari kumpulan cluster, dapat dianggap sebagai outliner.



Gambar 17 Deteksi *Outliner*

B. Rangkuman

Teknik smoothing digunakan untuk menghilangkan noise. Noise adalah kesalahan acak atau variasi dalam suatu variabel. Terdapat beberapa metode yang bisa digunakan untuk smoothing data antara lain:

- Binning
- Regresi
- Deteksi outliner

Metode Binning memuluskan nilai data yang diurutkan dengan berkonsultasi melihat "lingkungannya". Metode regresi dilakukan dengan cara menyesuaikan nilai data ke suatu fungsi. Sedangkan deteksi outliner dilakuan dengan cara mengklaster data berdasarkan kriteria tertentu.

C. Evaluasi

Untuk menguji pemahaman Anda terkkait materi *smoothing data,* kerjakanlah latihan di bawah ini!

1. Lakukan smoothing untuk dataset Adult!
   Link: (https://archive-beta.ics.uci.edu/dataset/2/adult)
2. Lakukan analisis untuk menentukan teknik yang akan digunakan!
3. Tulis/gambar hasil smoothing dari dataset tersebut!

MATERI POKOK 10:

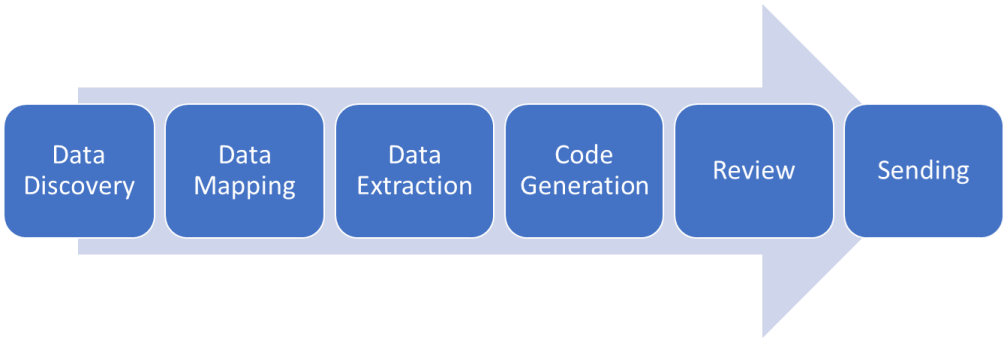
TRANSFORMASI DATA

Indikator Hasil Belajar:
Peserta mampu melakukan transformasi data dengan tepat. Peserta dinyatakan lulus manakala mampu :

1. menjelaskan urgensi transformasi data dengan tepat,
2. menjelaskan ragam transformasi data dengan tepat, dan
3. mempraktikan transformasi data dengan tepat.

A. Transformasi Data dalam Data Mining

Bab ini menyajikan metode transformasi data. Pada tahap preprocessing ini, data ditransformasikan atau dikonsolidasikan sehingga proses penambangan yang dihasilkan bisa lebih banyak efisien, dan pola yang ditemukan mungkin lebih mudah dipahami. Transformasi data juga digunakan untuk menggabungkan data terstruktur dan tidak terstruktur untuk analisis selanjutnya. Transformasi data ini juga penting saat mentransfer data ke data center atau data cloud yang baru. Akan lebih mudah untuk menganalisis dan mencari pola ketika datanya homogen dan terstruktur dengan baik. Proses transformasi data dilakukan dalam 6 langkah:



Gambar 18 Proses Tranformasi Data

1. **Data Discovery, penemuan data (EDA, Exploratory Data Analysis):**
   Analis bekerja pada tahap pertama untuk memahami dan mengidentifikasi data dalam format aslinya. Mereka akan menggunakan alat pemrofilan data untuk melakukannya. Langkah ini membantu analis dalam menentukan

tindakan apa yang diperlukan untuk mengubah data menjadi format yang diinginkan.

2. **Data Mapping, Pemetaan Data:** Analis melakukan pemetaan data selama fase ini untuk menentukan bagaimana masing-masing bidang dimodifikasi, dipetakan, difilter, digabungkan, dan digabungkan. Pemetaan data sangat penting untuk banyak proses data, dan satu kesalahan dapat menyebabkan analisis yang salah dan efek riak di seluruh organisasi.
3. **Ekstraksi Data**: Analis mengekstrak data dari sumber aslinya selama fase ini. Sumber terstruktur seperti database dapat digunakan, demikian pula sumber streaming seperti file log pelanggan dari aplikasi web.
4. **Pembuatan dan Eksekusi Kode**: Setelah mengekstraksi data, analis harus menulis kode untuk menyelesaikan transformasi. Analis sering membuat kode dengan bantuan platform atau alat Transformasi Data
5. **Review**: Setelah mengubah data, analis harus memeriksa ulang untuk memastikan bahwa semuanya telah diformat dengan benar.
6. **Pengiriman**: Langkah terakhir adalah mengirim data ke tujuan akhirnya. Gudang data atau basis data yang dapat menangani data terstruktur dan tidak terstruktur bisa menjadi tujuannya.

B. Ragam Teknik Transformasi Data

Menurut definisi, Transformasi Data adalah proses mengambil data dari sumber dan mengubahnya menjadi format tujuan yang dapat digunakan untuk berbagai keperluan.

Itu terjadi selama proses ETL (Extract, Load, Transform), di mana data harus dikenali dan diekstraksi dari tempat penyimpanannya dan dipindahkan ke dalam satu tujuan atau repositori. Data mentah dari Proses Transformasi Data dalam Data Mining ini harus dibersihkan dan disiapkan untuk transformasi dengan mengatasi masalah seperti *missing value* (lihat Bab 8) dan ketidakkonsistenan. Teknik Transformasi Data dalam Penambangan Data kemudian digunakan:

- Data Smoothing
- Data Aggregation
- Discretization

- Generalization
- Construction of Attributes
- Normalization

1. **Data Smoothing**

   Teknik ini, sebagaimana di bahas bab 9, digunakan untuk menghilangkan noise dari dataset. Data yang terdistorsi dan tidak berarti dalam kumpulan data disebut sebagai noise. Algoritma smoothing digunakan untuk menyoroti fitur khusus data. Setelah menghilangkan noise, proses dapat mendeteksi setiap perubahan kecil pada data untuk mendeteksi pola khusus.

   Metode ini dapat mendeteksi perubahan atau tren data apa pun dan dapat membantu dalam Proses Transformasi Data dalam Data Mining

2. **Data Aggregation**
3. **Discretization**
4. **Generalization**
5. **Konstruksi Atribut**
6. **Normalisasi**

C. Rangkuman

Transformasi data juga digunakan untuk menggabungkan data terstruktur dan tidak terstruktur. Tranformasi data dilakukan melalui enam tahapan antara lain :

- data discoveri
- data mapping
- data extraction
- code generation
- review
- sending

D. Evaluasi

Untuk menguji pemahaman Anda terkait materi transformasi data, silahkan kerjakan latihan di bawah ini!

1. Apa manfaat dari transformasi data?
2. Apa perbedaan antara smoothing.dan discretization?
3. Lakukan data transformasi data dengan salah satu teknik di atas,untuk dataset Adult (https://archive-beta.ics.uci.edu/dataset/2/adult)
4. Jelaskan kenapa anda memilih teknik tersebut.

DAFTAR PUSTAKA

Arhatin, R. E. & Wahyuningrum, P. I. (2013). Algoritma indeks vegetasi mangrove menggunakan satelit landsat ETM+. *Buletin PSP*, *21*(2): 215-227.

Chavez Jr., P. S. (1988). An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. *Remote sensing of environment*, *24*(3): 459-479.

Danoedoro, P. (2012). *Pengantar penginderaan jauh digital*. Yogyakarta: Andi Publisher.

Du, Y., Zhang, Y., Ling, F., Wang, Q., Li, W. & Li, X. (2016). Water bodies' mapping from Sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution produced by sharpening the SWIR band. *Remote Sensing*, *8*(4): 354-362.

Haboudane, D., Miller, J. R., Tremblay, N., Zarco-Tejada, P. J. & Dextraze, L. (2002). Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote sensing of environment*, *81*(2): 416-426.

https://github.com/jfadibrin/modul2022/tree/main/M06_DataMining/M06S08_MissingValue

https://github.com/jfadibrin/modul2022/tree/main/M06_DataMining/dataset

https://www.kaggle.com/competitions/titanic/data

https://www.kaggle.com/datasets/laotse/credit-risk-dataset

https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot

https://archive-beta.ics.uci.edu/dataset/2/adult