ISBN 978-623-198-088-5

DATA MINING

Amna, Wahyuddin S, I Gede Iwan Sudipa, Tri Andi E. Putra,
Ahmad Jurnaidi Wahidin, Wara Alfa Syukrilla,
Anindya Khrisna Wardhani, Nono Heryana,
Tutuk Indriyani, Leo Willyanto Santoso

DATA MINING

Amna
Wahyuddin S
I Gede Iwan Sudipa
Tri Andi E. Putra
Ahmad Jurnaidi Wahidin
Wara Alfa Syukrilla
Anindya Khrisna Wardhani
Nono Heryana
Tutuk Indriyani
Leo Willyanto Santoso

PT GLOBAL EKSEKUTIF TEKNOLOGI

DATA MINING

Penulis:
Amna
Wahyuddin S
I Gede Iwan Sudipa
Tri Andi E. Putra
Ahmad Jurnaidi Wahidin
Wara Alfa Syukrilla
Anindya Khrisna Wardhani
Nono Heryana
Tutuk Indriyani
Leo Willyanto Santoso

ISBN: 978-623-198-088-5

Editor: Dina ediana, S. Kom., M. Kom.
Ari Yanto, M.Pd.
Penyunting: Yuliatri Novita, M. Hum.
Desain Sampul dan Tata Letak: Handri Maika Saputra, S.ST.
Penerbit: PT GLOBAL EKSEKUTIF TEKNOLOGI
Anggota IKAPI No. 033/SBA/2022

Redaksi: Jl. Pasir Sebelah No. 30 RT 002 RW 001
Kelurahan Pasie Nan Tigo Kecamatan Koto Tangah
Padang Sumatera Barat

website: www.globaleksekutifteknologi.co.id
email: globaleksekutifteknologi@gmail.com

Cetakan Pertama, 16 Februari 2023

KATA PENGANTAR

Segala puji syukur penulis panjatkan ke hadirat Allah Ta'ala karena atas limpahan rahmat-Nya sehingga kami dapat menyelesaikan "Buku Data Mining". Data Mining adalah proses penggalian informasi dan pola yang bermanfaat dari suatu data yang sangat besar. Proses data mining terdiri dari pengumpulan data, ekstraksi data, analisa data, dan statistik data. Buku berisi tentang data mining dan knowledge discovery process, data understanding, representation knowledge data mining, data mining roles, cluster analysis, association rules, text mining, ekstraksi fitur, discretization method.

Kami menyadari, bahan Buku ini masih banyak kekurangan dalam penyusunannya. Oleh karena itu, kami sangat mengaharapkan kritik dan saran demi perbaikan dan kesempurnaan Buku ini selanjutnya. Kami mengucapkan terima kasih kepada berbagai pihak yang telah membantu dalam proses penyelesaian Buku ini. Semoga Buku ini dapat bermanfaat bagi pembacanya.

Penulis, 16 Februari 2023

Penulis

DAFTAR ISI

DAFTAR TABEL

BAB 1
DATA MINING DAN KNOWLEDGE DISCOVERY PROCESS

Oleh Amna

1.1 Pengertian Data

Data merupakan fakta-fakta yang menggambarkan suatu kejadian yang sebenarnya pada suatu waktu (Lubis, 2016). Data dapat didefinisikan sebagai nilai yang menggambarkan deskripsi dari suatu objek atau kejadian. Hasil dari pengolahan data dalam suatu bentuk yang lebih berguna dan lebih mudah dipahami penerima dapat disebut sebagai informasi. Informasi mempresentasikan lebih lanjut kejadian-kejadian nyata yang akan diaplikasikan dalam pengambilan keputusan. Keduanya memiliki perbedaan yang signifikan, data bersifat lebih historis, sedangkan informasi memiliki tingkatan yang lebih tinggi, dinamis, juga memiliki nilai yang penting (Pamungkas, 2017). Data adalah deskripsi atau keterangan sebuah objek yang belum memiliki arti sepenuhnya yang dapat berbentuk angka (numerik), karakter (*text*), gambar, suara, ataupun lambang (simbol) (Yusuf dan Daris, 2018). Hasil pengolahan dari data baik dalam bentuk angka (numerik), karakter (*text*), gambar, suara, ataupun lambang (simbol) akan menghasilkan informasi yang lebih diperlukan dalam penyelesaian atau jawapan sesuatu masalah.

1.2 Pengertian Data Mining

Data Mining adalah proses penggalian informasi dan pola yang bermanfaat dari suatu data yang sangat besar. Proses data mining terdiri dari pengumpulan data, ekstraksi data, analisa data, dan statistik data. Ia juga umum dikenal sebagai *knowledge discovery*, *knowledge extraction*, *data/pattern analysis*, *information harvesting*, dan lainnya (Arhami dan Nasir, 2020). Empat proses dalam data mining ini akan menghasilkan model/ pengetahuan yang sangat berguna.

Menurut Muflikhah (2018), data mining dapat didefinisikan sebagai penguraian kompleks dari sekumpulan data menjadi informasi yang memiliki potensi secara implisit (tidak nyata/jelas) yang sebelumnya belum diketahui. Ia juga dapat didefinisikan sebagai penggalian dan analisis dengan menggunakan peralatan otomatis atau semi otomatis, dari sebagian besar data yang memiliki tujuan yaitu menemukan pola yang memiliki arti atau maksud. Data mining termasuk ke dalam *knowledge discovery* di dalam database (KDD).

Gambar 1.1 Bagan Data Mining (Arhami dan Nasir, 2020)

Seperti yang kita ketahui, data mining berfungsi untuk memfasilitasi pekerjaan dengan data yang banyak. Menurut Agrawal dan Agrawal (2017), data mining efektif untuk digunakan tidak hanya di lingkungan bisnis tetapi juga di lingkungan lainnya seperti ramalan cuaca, ilmu kedokteran, perawatan kesehatan, tranportasi, asuransi, dan pemerintahan. Di antara kegunaan di dalam data mining di berbagai industri adalah sebagai berikut:

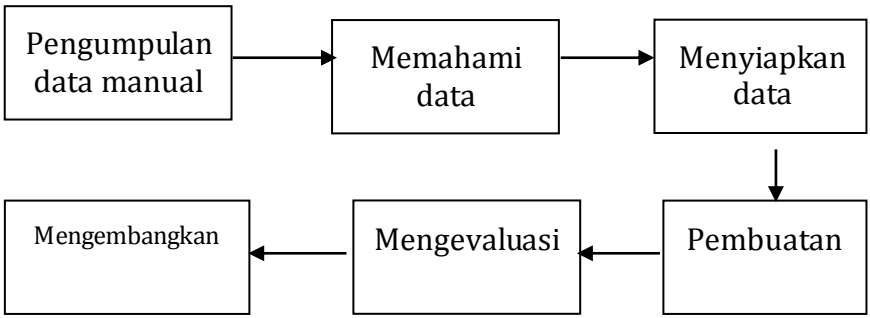
1. Aplikasi data mining dalam pemasaran: data mining dalam pemasaran memudahkan pengelola bisnis untuk mengerti tren tersembunyi di sebalik transaksi pembelian sehingga membantu dalam perencanaan dan peluncuran promosi pemasaran dengan cepat dan rendah biaya. Data mining digunakan untuk analisa keranjang pasar untuk memberikan informasi tentang kombinasi produk yang dibeli, kapan dibeli dan dalam susunan apa oleh pembeli. Informasi ini membantu organisasi bisnis untuk mempromosikan produk yang paling menguntungkan untuk memaksimalkan keuntungan. Sebagai tambahan, ia menggalakkan pembeli untuk membeli produk yang berkaitan yang mungkin terlewatkan. Saat ini, perusahaan retail *online* mengaplikasikan data mining untuk mengidentifikasi pola pembelian pembeli.
2. Aplikasi data mining dalam perbankan/keuangan: beberapa teknik data mining seperti data terdistribusi telah diteliti, dimodelkan dan dikembangkan untuk membantu deteksi penipuan kartu kredit di berbagai bank. Data mining juga digunakan untuk mengidentifikasi loyalitas pelanggan dengan menganalisa data pembelian dari pelanggan seperti frekuensi pembelian dalam periode waktu tertentu, total uang yang digunakan untuk seluruh pembelian dan kapan kali terakhir pembelian dilakukan. Pengaplikasian data mining, pengukuran relatif dapat dihasilkan untuk setiap pelanggan setelah menganalisa dimensi tersebut. Data mining juga digunakan untuk membantu bank mempertahankan pelanggannya. Ia dapat membantu bank untuk memprediksi pelanggan yang memiliki kemungkinan untuk mengganti afiliasi kartu kredit dengan menganalisa data terdahulu, sehingga ia dapat merencanakan dan meluncurkan penawaran spesial untuk mempertahankan pelanggan tersebut. Data mining

digunakan juga untuk menemukan korelasi tersembunyi di antara indikator finansial. Aturan perdagangan saham dapat diidentifikasi dengan menggunakan historis data pemasaran.

3. Aplikasi dalam perawatan kesehatan dan asuransi: Pertumbuhan dari industri asuransi tergantung kepada kemampuan konversi data menjadi pengetahuan, informasi atau inteligensi mengenai pelanggan, pesaing dan pasarnya, yang mana dapat membantu pengaturan dalam waktu yang singkat dan pemilihan keputusan yang baik. Data mining telah membawa kelebihan kompetitif yang yang besar kepada perusahaan yang berhasil mengimplementasikannya. Aplikasi data mining di dalam industri asuransi sangat menonjol. Data mining yang diaplikasikan dalam analisis penyelesaian klaim seperti mengidentifikasi prosedur medis yang diklaim bersamaan. Data mining dapat memprediksi pelanggan yang mana memiliki potensi untuk membeli polisi baru atau mengubah bentuk polisi dari satu perusahaan ke perusahaan yang lain. Data mining juga dapat membolehkan perusahaan asuransi untuk mendeteksi pola pelanggan yang beresiko dalam mendeteksi penipuan.
4. Aplikasi di transportasi: Data mining membantu untuk menentukan jadwal distribusi di antara gudang dan outlet serta menganalisa pola pemuatan. Organisasi tersebut dapat menentukan waktu produksi dan penghantaran dari produk dengan menganalisa pola pasar dengan bantuan data mining.
5. Aplikasi dalam ilmu kedokteran: Data mining membolehkan untuk mengkarakterisasi aktivitas pasien untuk kunjungan ke klinik. Data mining membantu dokter untuk mengidentifikasi pola dari terapi medis yang sukses untuk penyakit yang berbeda. Aplikasi data mining

dikembangkan secara kontinu dalam bidang medis dan industri kesehatan untuk menyediakan pola dan wawasan yang tersembunyi dan menarik untuk membolehkan peningkatan efisiensi bisnis.

6. Aplikasi dalam manufaktur: Data mining dapat digunakan dalam data teknik operasional untuk mendeteksi kesalahan pada peralatan dan menentukan parameter kontrol yang optimal. Data mining juga digunakan untuk mengontrol pembuatan dari produk untuk memaksimalkan keuntungan. Data mining menjadi perangkat penting untuk tetap menyeimbangkan antara permintaan dan penawaran.
7. Aplikasi di pemerintahan: Pada saat ini, skema pemerintah dapat diimplementasikan secara efektif melalui data mining. Data mining membantu divisi pajak penghasilan untuk menggali dan menganalisa rekor transaksi finansial untuk membentuk pola yang dapat mendeteksi pencucian uang atau aktivitas kriminal.



Gambar 1.2 Bagan Proses dari Data Mining (Joseph, 2019)

Proses dari data mining termasuk dari pengumpulan data mentah dari data dasar seperti relasi, gudang data (*data werehouse*), reservasi informasi dan reservasi yang lebih maju,

object-oriented, dan *object-relational*, transaksi dan spasial, heterogen dan legasi, multimedia dan *streaming*, kata, mining kata dan mining web. Proses tersebut melibatkan data mining untuk menghasilkan pemahaman yang menjadikan informasi tersebut lebih dikenal dan dipahami. Prosesnya seperti *knowledge discovery*, pengambilan informasi, analisa pola dan ektraksi pengetahuan yang membolehkan pemahaman terhadap data, menggiring ke pengukuran konstruktif dari area yang terlibat (Joseph, 2019).

1.3 Teknik dalam Data Mining

Berdasarkan Mardi (2017), data mining dapat dikelompokkan menjadi beberapa kelompok, sesuai tugas yang dapat dilakukan yaitu:

1. Deskripsi

   Data mining digunakan dalam mencari metode sederhana untuk penggambaran pola dan kecenderungan yang terdapat pada data. Deskripsi dari pola dan kecenderungan akan memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.
2. Estimasi

   Hampir sama dengan klasifikasi, variabel target pada estimasi lebih cenderung ke arah numerik dibandingkan ke arah kategori. Untuk pembangunan model digunakan rekor lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Kemudian pada peninjauan seterusnya, estimasi nilai dari variabel target dilakukan berdasarkan nilai variabel prediksi.
3. Prediksi

   Memiliki kemiripan dengan klasifikasi dan estimasi, prediksi dapat meramalkan nilai dari hasil yang akan ada di masa mendatang. Terdapat beberapa metode serta

teknik yang digunakan dalam klasifikasi dan estimasi yang dapat digunakan untuk keadaan yang tepat untuk prediksi.

4. Klasifikasi

   Target variabel kategori dijabarkan dalam klasifikasi. Di antara model-model yang telah dikembangkan adalah:

   - Pohon keputusan
   - Pengklasifikasi bayes/*naïve bayes*
   - *Neural network*
   - Analisis statistik
   - Algoritma genetik
   - *Rough sets*
   - Pengklasifikasi *k-nearest neighbour*
   - Metode berbasis aturan
   - *Memory based reasoning*
   - *Support vector machine*

   Di antara klasifikasi yang umum digunakan adalah sebagai berikut:

   - Pohon keputusan

     Teknik pohon keputusan (*decision tree*) dapat diaplikasikan sebagai bagian dari kriteria seleksi. Sebagai tambahan, untuk membantu penggunaan dan seleksi dari data spesifik di antara struktur keseluruhan. Umumnya pada pohon keputusan dimulai dengan pertanyaan sederhana yang memiliki dua atau lebih jawaban. Setiap jawaban akan menggiring ke pertanyaan selanjutnya yang akan membantu dalam klasifikasi dan identifikasi data sehingga dapat dikategorikan atau diprediksi sesuai jawaban (Osman, 2019).

   - *Neural network*

     Neural network merupakan teknik yang umum digunakan pada tahap awal data mining yang berasal

dari *Artificial Inteligence* (AI). Teknologi ini mudah untuk digunakan dikarenakan ia bekerja secara otomatis sehingga pengguna tidak membutuhkan banyak pengetahuan mengenai pekerjaan atau basis data. Untuk menjalankan *neural network* yang efisien, perlu diatur koneksi dari simpulan, nomor unit proses dan kapan pelatihan proses harus dihentikan. *Neural network* sendiri terdiri dari dua bagian utama yaitu simpulan (*node*) dan tautan (Osman, 2019).

- Algoritma genetik
  Algoritma genetik dilakukan dengan memadukan pemikiran mengenai evaluasi alami, seperti pengambilan bagian yang bagus dan menggabungkan dengan bagian yang lain yang bagus untuk memberikan solusi yang lebih baik. Ia menjadi strategi penyelesaian masalah untuk memberikan hasil yang optimal (Jain dan Srivastava, 2013).
- Metode berbasis aturan
  Terdapat tiga kriteria utama dalam metode berbasis aturan untuk evaluasi algoritma: cakupan penggunaan, tipe ketergantungan terhadap kotak hitam dan format dari deskripsi ekstrak. Dimensi pertama berkaitan dengan cakupan penggunaan dari suatu algoritma, sama ada regresi maupun klasifikasi. Dimensi kedua berfokus kepada algoritma ekstraksi pada kotak hitam yang tersembunyi: tidak tergantung terhadap algoritma *dependant*. Dimensi ketiga berfokus kepada aturan yang didapatkan yang mungkin berharga pada waktu sementara: algoritma prediktif vs algoritma deskriptif (Jain da Srivastava, 2013).

5. Pengklusteran

Dalam pengelompokan rekor, pengamatan atau memper-hatikan dan membentuk kelas objek-objek yang

memiliki kemiripan dilakukan secara pengklusteran. Kluster merupakan kumpulan rekor yang memiliki ketidak miripan dengan rekor-rekor dalam kluster lain.

Berbeda dengan klasifikasi, pengklusteran tidak memiliki variabe target. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi atau memprediksi nilai dari variabel target. Namun, algoritma pengklusteran dicoba untuk membagi keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan atau homogen, di mana kemiripan rekor dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan rekor dalam kelompok lain bernilai minimal.

6. Asosiasi

   Asosiasi pada data mining bertugas untuk menemukan atribut yang muncul dalam suatu waktu.

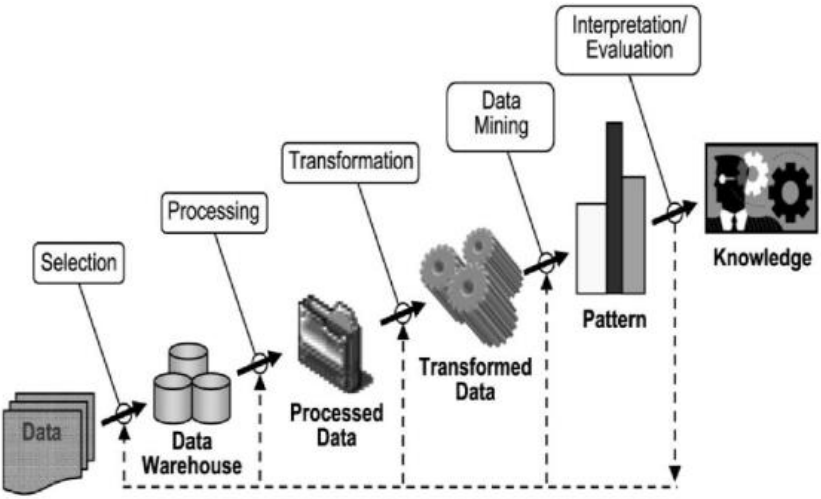
1.4 *Knowledge Discovery*

Knowledge discovery atau *knowledge discovery in database* (KDD) adalah salah satu metode untuk memperoleh pengetahuan dari basis data yang memiliki tabel-tabel yang saling berhubungan atau berelasi. Hasil dari pengetahuan yang didapatkan dari proses tersebut menjadi dasar sebagai basis pengetahuan (*knowledge base*) dalam pengambilan keputusan (Mardi. 2017). Menurut (Siregar & Puspabhuana, tanpa tahun), data mining dan *Knowledge discovery* (KDD) dapat menjelaskan pencarian informasi berguna yang ada pada suatu data yang sangat besar. Namun, kedua istilah ini tetap dua hal yang berbeda walau berkaitan antar satu sama lain. Hal ini karena *data mining* merupakan salah satu proses di dalam KDD. *knowledge discovery in database* memiliki kaitan dengan proses menemukan pengetahuan pada suatu *database*. Pada proses ini dilakukan identifikasi untuk mencari data yang valid, sesuatu yang baru dan memiliki manfaat sehingga dapat ditemukan

suatu pola yang dapat dimengerti yang dapat disebut sebagai suatu proses non-trivial.

Knowledge discovery memiliki kelemahan pada data yang memiliki integrasi atau navigasi yang berbeda.. Pendekatan baru diperlukan jika suatu dimensi yang ada di dalam data juga meningkat. Maka didapatkan kesimpulan pengertian *Knowledge Discovery Data* (KDD) merupakan sejumlah data yang besar yang diekstrak melalui tahapan penggalian dan analisis untuk mendapatkan informasi dan pengetahuan yang berguna (Marisa et al, 2021). Untuk mendapatkan *information discovery* pada analisis data dan bisnis dengan menggunakan *data mining*. *Data mining* akan bekerja dengan mengambil data dari *database* untuk melakukan OLAP (Prasetyowati, 2017).

1.5 Proses *Knowledge Discovery* dan Penyimpanan pengetahuan.



Gambar 1.3 Proses *Knowledge Discovery* In Database (Nofriansyah, 2015)

Menurut (Nofriansyah, 2015), secara detail setiap tahapan proses pada *Knowledge Discovery* dapat dijelaskan sebagai berikut:

1. *Data Selection*

   Tahapan pertama untuk menggali informasi menggunakan KDD yaitu dengan memilih atau menyeleksi beberapa data yang dibutuhkan dari sekumpulan data yang banyak. Data yang telah dipilih akan digunakan pada tahapan *data mining*. Untuk memudahkan penggunaan dan pencarian kembali data yang sudah diseleksi, data tersebut akan disimpan pada suatu berkas yang dibedakan dengan data lainnya.

2. *Pre-processing (Cleaning)*

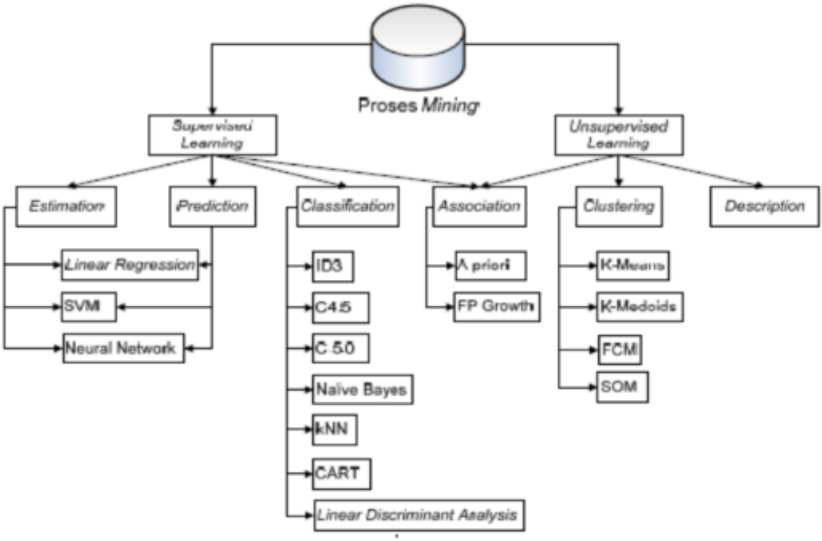
   Untuk memastikan data yang disimpan merupakan data yang bermutu dan akurat, dilakukan tahapan ini yaitu dengan membuang data yang tidak lengkap atau tidak valid. Tahapan ini penting agar proses *data mining* nanti dapat menghasilkan data yang bermutu dan *valid* sesuai dengan data yang dimasukkan.

3. *Transformation*

   Pada tahapan transformasi data, data akan disesuaikan dengan teknik data mining yang akan dipakai. Diterapkan suatu format pada data sehingga nanti data yang tidak sesuai dengan format akan disingkirkan dan menyisakan data yang sesuai untuk digunakan pada tahapan selanjutnya dengan kualitas data yang tepat.

4. Data Mining (Sandi et al, 2020)

   Pada proses data mining, digunakan beberapa teknik atau metode tertentu untuk menganalisis data yang ada sehingga didapatkan hasilnya yaitu data yang berisi pengetahuan penting atau tersembunyi. Pemilihan teknik harus sesuai dengan tujuan KDD dilakukan sehingga hasilnya akan relevan. Beberapa metode pada data mining yaitu:



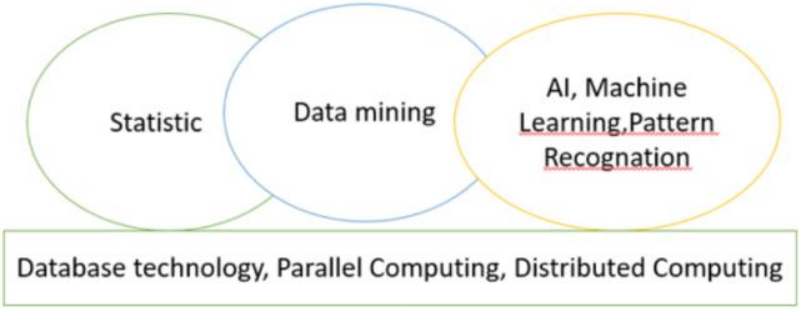
Gambar 1.4 Metode-metode data mining (Sandi et al, 2020)

5. Evaluasi Pola (*Pattern Evaluation*)

   Hasil data yang didapatkan dari proses *data mining* akan diidentifikasi sehingga didapatkan suatu pola yang akan dimasukkan ke dalam *knowledge based* untuk kemudian dilakukan analisis.
6. Presentasi Pengetahuan (*Knowledge Presentation*)

   Pada tahapan ini, data yang ada di dalam *knowledge based* tersebut akan ditampilkan atau diperlihatkan terkait metode yang digunakan untuk memperolehnya kepada pengguna yang melakukan *Knowledge Discovery* tersebut.

1.6 Teknik *Knowledge Discovery*



Gambar 1.5 Pertemuan disiplin ilmu dalam data mining (Ginantra et al, 2021)

Menurut Gorunescu (2011), selain data mining, terdapat suatu istilah yang disebut sebagai *knowledge discovery* yang berasal dari teknik yang dipinjam dan terminologi sebagai berikut:

1. Statistika: statistika merupakan dasar dari data mining. Statistika klasik membawa teknik yang terdefinisi dengan baik yang dapat disimpulkan dalam hal yang umum diketahui sebagai *exploratory data analysis* (EDA), yang digunakan untuk mengidetifikais hubungan sistematis di antara variabel yang berbeda, ketika tidak ada informasi yang mencukupi mengenai hal tersebut. Di antara teknik EDA klasik yang digunakan dalam data mining adalah sebagai berikut:
    - Metode komputasi: statistik deskriptif (distribusi, parameter statistik klasik (mean, median, deviasi standar), korelasi, tabel frekuensi ganda, teknik eksplorasi multivariat (analisis kluster, analisis faktor, prinsip komponen dan analisis klasifikasi, analisis kanon, analisis diskriminan, klasifikasi tree, dan

analisa korespondensi), model linear/non-linear lanjutan (regresi linear/non-linear dan peramalan/seri waktu)

- Visualisasi data bertujuan untuk merpresentasikan informasi dalam bentuk visual, dan ia dapat diketahui sebagai salah satu alat yang sangat berguna dan metode yang menarik mengeksplorasi data. Di antara teknik visualisasi yang paling umum adalah seperti histogram (kolom, silinder, kerucut, piramida, pai, dan batang), plot kotak, plot *scatter*, plot *contour*, plot matriks, dan plot ikon.

2. *Artificial Intelligence* (AI): Tidak seperti statistika, AI dibangun secara heuristik. Sehingga AI berkontribusi dengan teknis memproses informasi, yang berdasarkan model pemikirian manusia ke arah perkembangan data mining. *Machine learing* (ML) juga terkait erat dengan AI. ML mempresentasikan disiplin ilmiah yang penting dalam pengembangan data mining, menggunakan teknik yang membolehkan komputer untuk belajar melalui pelatihan. Dalam konteks ini, dapat dipertimbangkan juga *natural computing* (NC) sebagai dasar tambahan untuk data mining.
3. Sistem basis data (DBS): DBS menjadi akar ketiga dari data mining, dengan menyediakan informasi yang akan ditambang menggunakan metode yang telah disebutkan di atas.

1.7 Implementasi Data Mining dan *Knowledge Discovery*

Terdapat beberapa implementasi pada bidang kehidupan seperti dalam bidang bisnis dan bidang sains dan teknik. Pada bidang bisnis, digunakan suatu metode dikenal dengan nama *Market Basket Analysis* biasa disebut sebagai *association rule*. Metode ini dapat mencari hubungan antara data-data dengan menganalisa data sehingga dapat diketahui suatu pola. Pola ini dapat berupa apa yang dibeli oleh konsumen. Keterkaitan antar data terjadi seperti pada saat konsumen membeli sabun maka kemungkinan besar ia juga akan membeli sampo sehingga diletakkan sampo berdekatan dengan sabun. Pada bidang sains dan teknik, seperti biologi digunakan metode *sequence mining* yang dapat memetakan hubungan darah antar manusia sedangkan pada bidang teknik menggunakan metode *maltifactor dimensionality reduction* dimanfaatkan untuk memprediksi daya per hari kebutuhan dalam bidang kelistrikan (Ginantra, 2021).

DAFTAR PUSTAKA

Agrawal, C.P. dan Agrawal, M. 2017. *Introduction to Data Mining*. Educreation Publishing: New Delhi.

Arhami, M. dan Nasir, M. 2020. *Data Mining: Algoritma dan Implementasi*. Penerbit Andi: Banda Aceh.

Ginantra, N. L. W. S. R., Arifah, F. N., Wijaya, A. H., Septarini, R. S., Ahmad, N., Ardiana, D. P. Y., Effendy, F., Iskandar, A. Hazriani, H., Sari, I. Y., Gustiana, Z., Prianto, C., Gustian, D. dan Negara, E. S. 2021. *Data Mining dan Penerapan Algoritma*. Yayasan Kita Menulis: Medan.

Gorunescu, F. 2011. *Data Mining: Concepts, Models and Techniques*. Springer: Berlin.

Jain, N. dan Srivastava, V. 2013. Data Mining Techniques: A Survey Paper. *International Journal of Research in Engineering and Technology*. 2(11). 116-119.

Joseph, S.I.T. 2019. Survey of Data Mining Algorithms for Intelligent Computing System. *Journal of Trends in Computer Science and Smart Technology* (TCSST). 1(1). 14-23.

Lubis, A. 2016. *Basis Data Dasar*. Penerbit Deepublish: Yogyakarta.

Mardi, Y. 2017. Data Mining: Klarifikasi Menggunakan Algoritma C4.5. *Jurnal Edik Informatika*. 2(2). 213-219.

Marisa, F., Maukar, A. L. dan Akhriza, T. M. 2021. *Data Mining Konsep dan Penerapannya*. Deepublish: Yogyakarta.

Muflikhah, L., Ratnawato, D.E., dan Putri, R.R.M. 2018. *Buku Ajar: Data Mining*. UB Press: Malang.

Nofriansyah, D. 2015. *Konsep Data Mining Vs Sistem Pendukung Keputusan*. Deepublish: Yogyakarta.

Osman, A.S. 2019. Data Mining Techniques: Review. *International Journal of Data Science Research*. 2(1). 1-4.

Pamungkas, C.A. 2017. *Pengantar dan Implementasi Basis Data*. Penerbit Deepublish: Yogyakarta.

Prasetyowati, E. 2017. Data Mining Pengelompokan Data untuk Informasi dan Evaluasi. Data Media Publishing: Pamekasan.

Sandi, K., Habibi, R. dan Fauzan, M. N. 2020. *Tutorial PHP Machine Learning Menggunakan Regresi Linear Berganda pada Aplikasi Bank Sampah Istimewa Versi 2.0 Berbasis Web*. Kreatif Industri Nusantara: Bandung.

Siregar, A. M. dan Puspabhuana, A. Tanpa tahun. *Data Mining*. CV Kekata Group: Sukoharjo.

Yusuf, M. dan Daris, L. 2018. *Analisis Data Penelitian: Teori & Aplikasi dalam Bidang Perikanan*. Penerbit IPB Press: Bogor.

BAB 2
DATA UNDERSTANDING

Oleh Wahyuddin S

2.1 Data Understanding

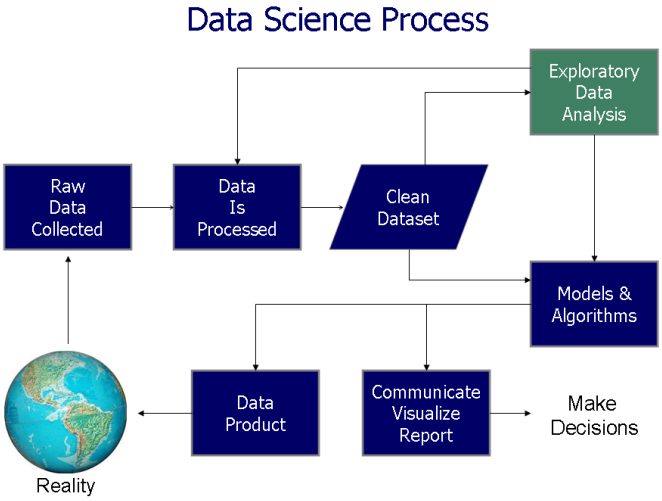
Data understanding merupakan sebuah tahapan dalam metodologi ilmu data dan pengembangan kecerdasan buatan yang bertujuan untuk mendapatkan pemahaman awal tentang data yang diperlukan untuk memecahkan masalah bisnis tertentu. Pengertian data adalah catatan atas kumpulan fakta. Dalam penggunaan sehari-hari, data adalah pernyataan yang terbukti dengan sendirinya. Pernyataan ini merupakan hasil dari pengukuran atau pengamatan terhadap suatu variabel yang dapat berupa angka, kata atau gambar (Wikipedia, 2022b).

2.2 Analisis data

Analisis data merupakan proses pemeriksaan, pembersihan dan pemodelan data dengan tujuan untuk menemukan sebuah informasi yang dapat berguna, menarik kesimpulan dan mendukung pengambilan keputusan. Analisis data memiliki banyak sisi dan pendekatan, mencakup berbagai teknik dan nama yang berbeda serta digunakan di berbagai bidang ekonomi, sains dan ilmu sosial. Dalam dunia bisnis saat ini, analisis data berperan penting dalam pengambilan sebuah keputusan secara ilmiah dan membantu perusahaan bekerja lebih efektif (Schutt & O'Neil, 2013; Wikipedia, 2022a).

2.2.1 Proses Analisis Data

Analisis mengacu pada penguraian keseluruhan menjadi komponen-komponen terpisah untuk dipelajari secara individu. Analisis data merupakan proses untuk memperoleh data mentah kemudian mengubahnya menjadi sebuah informasi yang berguna untuk pengambilan keputusan dari pengguna (Samosir, Wahyuddin, & ..., 2022). Data dikumpulkan dan dianalisis untuk menjawab pertanyaan, menguji hipotesis atau menyangkal teori.



Gambar 2.1 Diagram alur proses ilmu data
Sumber: *Doing Data Science* (Schutt & O'Neil) 2013

2.2.2 Persyaratan Data

Data yang diperlukan sebagai *input* analisis, ditentukan berdasarkan oleh kebutuhan pihak yang melakukan analisis atau pelanggan. Jenis umum entitas di mana data akan dikumpulkan disebut unit eksperimental seperti orang atau populasi. Variabel spesifik populasi dapat ditentukan dan diperoleh seperti usia dan pendapatan dan data dapat berupa numerik atau kategori seperti label teks untuk nomor.

2.2.3 Pengumpulan Data

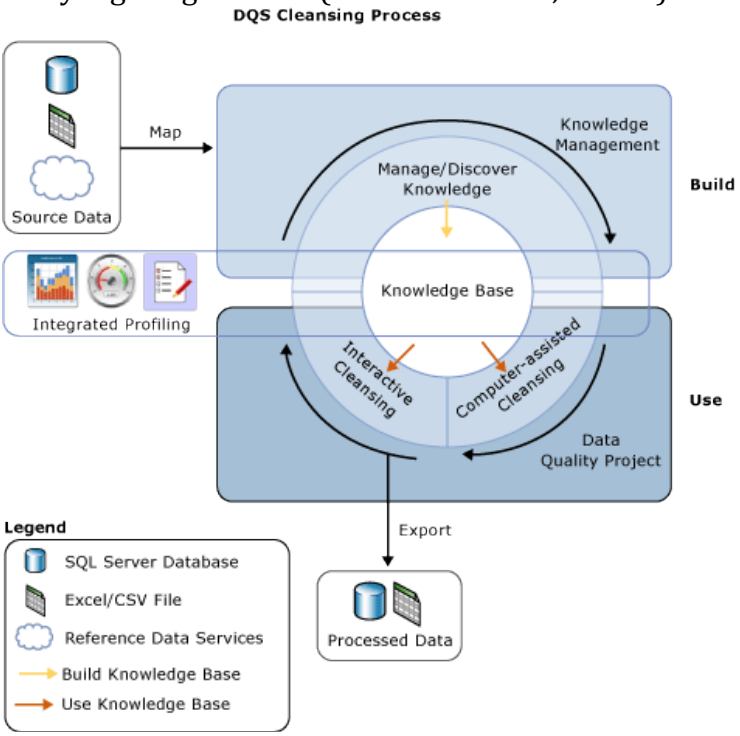
Data dikumpulkan dari berbagi sumber. Analis dapat mengomunikasikan persyaratan kepada pengelola data seperti staf teknologi informasi dalam sebuah organisasi. Data juga dapat dikumpulkan dari sensor di lingkungan, termasuk kamera lalu lintas, satelit, perekam dan lainnya. Data juga dapat dikumpulkan melalui wawancara, mengunduh dari sumber *online* atau membaca dokumentasi.

2.2.4 Pembersihan Data

Pembersihan data merupakan proses menganalisis kualitas data dari sumber data, memungkinkan sistem menerima atau menolak saran secara manual, dan membuat perubahan pada data. Pembersihan data dalam *Data Quality Services* (DQS) melibatkan proses bantuan komputer yang menganalisis bagaimana data mencocokkan pengetahuan dalam basis pengetahuan dan proses interaktif yang memungkinkan pengelola data untuk meninjau dan mengubah hasil proses yang dibantu oleh komputer untuk memastikan bahwa pembersihan data adalah tujuan desainnya (Microsoft Learn, 2022b). Adapun fitur pembersih data pada DQS sebagai berikut:

1. Mengidentifikasi data yang tidak lengkap atau ada kesalahan dari sumber data yang diperoleh (File excel atau database SQL Server), kemudian mengoreksi atau memperingatkan tentang data yang tidak valid.
2. Menyediakan proses dua langkah untuk membersihkan data dibantu dengan komputer dan interaktif. Proses yang dibantu oleh komputer menggunakan pengetahuan dalam pangkalan pengetahuan DQS untuk memproses data secara otomatis dan menyarankan penggantian atau koreksi. Langkah berikutnya adalah interaktif yang dapat memungkinkan pengurus data untuk menyetujui, menolak atau memodifikasi perubahan yang diusulkan oleh DQS selama pembersihan yang dibantu dengan komputer.

3. Menstandarkan serta memperkaya data pelanggan dengan menggunakan nilai domain, aturan domain serta data referensi seperti standarisasi penggunaan istilah dengan mengubah "St." ke "Street", memperkaya data dengan mengisi elemen yang hilang dengan mengubah "*1 Microsoft way Redmond 98006*" menjadi "*1 Microsoft Way, Readmond, WA 98008*".
4. Menyediakan antarmuka seperti *wizard* yang sederhana, intuitif dan konsisten kepada pengguna untuk menavigasi data serta memeriksa kesalahan di antara sekumpulan data yang sangat besar (Microsoft Learn, 2022b).



Gambar 2.2 Proses Pembersihan Data

Sumber: https://learn.microsoft.com/id-id/sql/data-quality-services/data-cleansing?view=sql-server-ver16

2.2.5 Pengolaan Data

Pengolahan data merupakan pengolahan untuk menghasilkan sebuah informasi atau menghasilkan pengetahuan dari data yang belum melalui proses. Setelah terprogram, pengolahan ini dapat dilakukan secara otomatis oleh perangkat komputer. Rangkaian pengolahan data membentuk sebuah sistem informasi. Secara umum pengolahan data meliputi beberapa tahapan seperti:

1. Pengesahan (validasi)
2. Memastikan kebenaran dan relevansi data.
3. Pengurutan
4. Mengatur data agar mengikuti urutan tertentu.
5. Peringkasan
6. Mengurangi perincian data sampai hal-hal utamanya.
7. Penggabungan
8. Menggabungkan beberapa sumber data.
9. Analisis
10. Pengumpulan, pengelolaan, analisis, penafsiran dan penyajian data.
11. Pelaporan
12. Menerangkan ringkasan atau perincian data atau informasi.
13. Pengelompokan (Klasifikasi)
    Pemisahan data berdasarkan sifat tertentu (Wikipedia, 2022c).

2.2.6 Pemodelan Data

Pemodelan data merupakan proses untuk mendefinisikan serta menganalisis persyaratan data yang diperlukan untuk mendukung proses bisnis dalam sistem informasi yang sesuai dalam sebuah organisasi. Proses pemodelan data melibatkan pemodel data professional yang bekerjasama dengan pemangku kepentingan bisnis serta calon pengguna sistem informasi

(Setiawan & Rijanto, n.d.). Terdapat tiga jenis model data yang dibuat selama pembembangan dari persyaratan ke database aktual yang akan digunakan untuk sistem informasi yaitu:

1. *Entity Relationship Model*

   Entity Relationship Model (ERM) merupakan abstrak dan konseptual dari representasi data. ERM adalah salah satu pemodelan pada basis data yang digunakan untuk menghasilkan skema konseptual jenis/model data semantik sistem. Sistem biasanya memiliki basis data relasional dan ketentuanya bersifat *top-down*. Diagram untuk menggambarkan model *Entity-Relationship* disebut dengan *Entity-Relationship Diagram, ER Diagram* atau ERD.

2. *Binary Model*

   Binary Model adalah model data yang memperluas definisi dari sebuah entity, bukan hanya atribute-atributenya, tetapi juga tindakan-tindakan. Relasi memiliki tiga tipe biner yaitu:

   - *One-to-one* (1:1)
     Hubungan yang terjadi apabila instansi entitasnya hanya memiliki satu hubungan dengan instansi entitas lain.
   - *One-to-many* (1-M)
     Relasi ini terjadi apabila setiap instansi dapat memiliki lebih dari satu hubungan terhadap instansi entitas lain tetapi tidak dengan kebalikannya.
   - *Many-to-many* (M-N)
     Hubungan yang saling memiliki lebih dari satu dari setiap instansi entitas terhadap instansi entitas lainnya.

3. Semantik Data Model

   Pada dasarnya, model semantik memiliki arti yang hampir sama dengan model hubungan entitas. Perbedaan hanya tampak pada relasi objek dasar, yang tidak diekspresikan melalui simbol tetapi melalui penggunaan kata-kata (*semantic*).

2.3 Menyebarkan Model Data

Dalam *Master Data Services* (MDS), paket adalah file XML yang berika struktur model yang dapat disebarkan dan secara opsional, data dari model. Untuk dapat bekerja dengan model, terdapat beberapa cara yang dapat digunakan seperti:

1. Alat *MDS Model Deploy*

   Untuk membuat dan menyebarkan objek dan data model, gunakan alat

2. *Wizard* Penyebaran Model

   Untuk membuat dan menyebarkan paket struktur model saja, gunakan wizard di aplikasi web Master Data Manager. Anda tidak dapat menggunakan panduan ini untuk menyebarkan data.

3. Editor Paket Model

   Untuk mengedit paket model, gunakan *ModelPackageEditor.exe* yang meluncurkan wizard Editor Paket Model. Anda menggunakan wizard ini untuk mengedit paket yang dibuat oleh alat *MDSModelDeploy* atau *wizard* Penyebaran Model. Jika Anda memilih jalur default saat menginstal MDS, alat ini terletak di drive:\Program Files\Microsoft SQL Server\130\Master Data Services\ Configuration (Microsoft Learn, 2022a).

2.4 *Exploratory Data Analysis*

Exploratory Data Analysis (EDA) adalah metode eksplorasi data yang menggunakan Teknik aritmatika dan grafis sederhana untuk meringkas data pengamatan. Eksplorasi data merupakan bagian integral dari persepsi kita. Jika tujuan akhir penelitian bukan untuk menarik kesimpulan kausal maka analisis data lebih lanjut tidak diperlukan lagi. Namun, apabila diperlukan, analisis data eksploratif sangat berguna untuk menganalisis dan menemukan sifat-sifat data, yang nantinya dapat berguna saat memilih model statistic yang tepat. Oleh karena itu, dalam analisis data eksplorasi, sifat data yang diamatilah yang dapat menentukan model analisis statistic sehingga sesuai.

Langkah pertama dalam analisis data adalah memeriksa karakteristik data. Ada beberapa alasan penting yang perlu di pertimbangkan dengan hati-hati sebelum melakukan analisis data nyata. Alasan validasi data yang pertama adalah untuk mencari kesalahan yang dapat terjadi pada berbagai tahapan, mulai dari pengumpulan data di lapangan hingga pemasukan data ke dalam komputer. Alasan selanjutnya adalah eksplorasi data agar kita bisa menentukan model analisis yang tepat (BPS (Badan Pusat Statistik), 2020).

2.4.1 Pentingnya *Exploratory Data Analysis*

Saat seseorang melakukan analisis data, salah satu proses yang tidak dapat diabaikan adalah *Exploratory Data Analysis* (EDA). EDA merupakan proses penting dalam analisis data kerena EDA memungkinkan pengguna untuk menghemat lebih banyak waktu dalam proses analisis data dan dapat menemukan beberapa kesalahan dalam data seperti adanya *missing value, outliers,* duplikasi, pengkodean, *noisy data,* data yang todak lengkap dan lainnya. Salah satu hal yang dikhawatirkan ketika gagal melalui proses EDA adalah terjadinya kesalahan yang berulang dalam proses analisis atau hasil analisis menjadi kurang valid dan kurang

relevan dengan tujuan bisnis karena data yang digunakan belum benar0benar siap. Selain itu, melalui EDA dapat membantu pengguna untuk melihat data sebelum membuat asumsi sehingga dapat mengidentifikasi kesalahan dalam data (Dqlab.id, 2020).

2.4.2 Teknik *Exploratory Data Analysis*

Pada proses pengolahan data, dalam melakukan *exploratory data analysis* dapat menggunakan beberapa teknik seperti:

1. Statistik Deskriptif

   Statistik deskriptif adalah menggambarkan atau merangkum data sehingga menghasilkan sebuah informasi secara umum tanpa bertujuan untuk menarik kesimpulan. Statistik deskriptif dapat menunjukkan informasi kunci seperti rata-rata median, modus, standar deviasi, varians dan kecekungan. Statistik deskriptif ini dapat ditampilkan dalam berbagai bentuk tabel, diagram, grafik dan lainnya.

2. *Univariate Analysis*

   Analisis univariat yaitu menganalisis kolom secara terpisah dan melihat distribusi data. Analisis univariat secara umum dibagi menjadi dua, yaitu analisis numerik dan kategorikal. Analisis ini juga dapat digunakan dengan tujuan untuk menarik kesimpulan dengan menggunakan berbagai analisis inrefensial yang dapat digunakan.

3. *Multivariate Analysis*

   Analisis multivariat menggabungkan beberapa kolom untuk menemukan hubungan antara satu kolom dengan kolom lainnya. Analisis multivariat ini mencakup variabel dalam jumlah lebih besar dari atau sama dengan tiga variabel.

DAFTAR PUSTAKA

BPS (Badan Pusat Statistik). (2020). SP2020 - Analisis Data Eksploratif. Retrieved November 28, 2022, from https://qasp2020.bps.go.id/posts/dda93a4b648c406f9bd8db4488e3a4e0/data-exploration/analisis-data-eksploratif

Dqlab.id. (2020). Exploratory Data Analysis : Pahami Lebih Dalam untuk Siap Ha... Retrieved November 28, 2022, from https://www.dqlab.id/data-analisis-machine-learning-untuk-proses-pengolahan-data

Microsoft Learn. (2022a). Menyebarkan Model - SQL Server Master Data Services | Microsoft Learn. Retrieved November 28, 2022, from https://learn.microsoft.com/id-id/sql/master-data-services/deploying-models-master-data-services?view=sql-server-ver16

Microsoft Learn. (2022b). Pembersihan data - Data Quality Services (DQS) | Microsoft Learn. Retrieved November 28, 2022, from https://learn.microsoft.com/id-id/sql/data-quality-services/data-cleansing?view=sql-server-ver16

Samosir, K., Wahyuddin, S., & ... (2022). *Sistem Basis Data*. Retrieved from https://books.google.com/books?hl=en&lr=&id=m-KWEAAAQBAJ&oi=fnd&pg=PA18&dq=%22wahyuddin+s%22&ots=guOhBDOwEM&sig=ITEuW6N3w8Wz-u9Lky2Vgt1xF-E

Schutt, R., & O'Neil, C. (2013). Doing Data Science. In *Foreign Affairs* (Vol. 91). Retrieved from https://archive.org/details/doingdatascience0000schu

Setiawan, A., & Rijanto, E. (n.d.). *An ICT Platform Design for Traceability and Big Data Analytics of Sugarcane Harvesting Operation*.

Wikipedia. (2022a). Analisis data - Wikipedia bahasa Indonesia, ensiklopedia bebas. Retrieved November 28, 2022, from https://id.wikipedia.org/wiki/Analisis_data

Wikipedia. (2022b). Data - Wikipedia bahasa Indonesia, ensiklopedia bebas. Retrieved November 28, 2022, from Wikipedia website: https://id.wikipedia.org/wiki/Data
Wikipedia. (2022c). Pengolahan data - Wikipedia bahasa Indonesia, ensiklopedia bebas. Retrieved November 28, 2022, from https://id.wikipedia.org/wiki/Pengolahan_data

BAB 3
REPRESENTATION KNOWLEDGE DATA MINING

Oleh I Gede Iwan Sudipa

3.1 Definisi Data Mining

Penambangan data adalah proses pengumpulan data dan informasi penting pada kumpulan data besar. Alat seperti statistik, kecerdasan buata, dan matematika sering digunakan dalam proses ini (Joko Suntoro 2019). Penambangan data adalah proses mengekstraksi konsep inti dari data mentah untuk membangun kerangka kerja yang dikenali (Dr.Suyanto,2019).

Penambangan data juga dikenal sebagai pengenalan pola dan penemuan pengetahuan. Ini adalah nama yang akurat untuk praktik penambangan data. Penambangan data adalah proses menemukan informasi tersembunyi dala sekumpulan data. Oleh karena itu, istilah penemuan pengetahuan digunakan untuk tujuan ini. Atau, pengenalan pola digunakan untuk menemukan pola yang akan dieksplorasi dalam satu set data.

Penambangan data adalah istilah yang memiliki arti berbeda bagi banyak orang. Tidak ada satu definisi istilah yang disepakati. Namun, itu mengacu pada proses yang berfokus pada penggalian pengetahuan dari data yang dikumpulkan oleh upaya ilmiah masa lalu(Portisch, Heist, and Paulheim 2022).

3.2 Fungsi Data Mining

Penambangan data adalah proses otomatis yang digunakan untuk menemukan informasi yang berguna dari kumpulan data yang besar. Ini memiliki banyak efek samping positif, termasuk menemukan data dan informasi prediktif tentang subjek tertentu. Data mining juga memiliki fungsi lain seperti klasifikasi, regresi, pengurutan, asosiasi, peramalan, dan pengelompokan (Suyanto 2017).

1. Deskriftif (*description*)
   Penambangan data adalah proses mempelajari data untuk menemukan pola dan karakteristik yang terkandung di dalamnya. Melakukan proses ini lebih dalam membantu mencapai tujuan penambangan data. Menentukan pola dan sifat berbeda yang terkandung dalam data. Penambangan data menggunakan pola yang tidak dicari dalam sekumpulan data untuk menemukan informasi baru.
2. Peramalan (*forecasting*)
   Data dapat digunakan untuk meramalkan masa depan, selama sejumlah besar informasi dikumpulkan. Beginilah cara kerja peramalan; mengumpulkan data sebanyak mungkin, membantu menciptakan visi masa depan. Salah satu contohnya adalah data permintaan konsumen untuk satu produk setelah dirilis ke publik.
3. Regresi (*Regression*)
   Fungsi regresi memperoleh pola numerik alih-alih klasifikasi seperti Klasifikasi. Ini mencari fungsi regresi dengan permainan akhir tertentu: menemukan rumus numerik alih-alih skema klasifikasi. Fungsi regresi menghargai data input karena menghasilkan fungsi yang menentukan hasil berdasarkan nilai.
4. Klasifikasi (*classification*)
   Fungsi klasifikasi digunakan untuk menemukan kesimpulan tentang karakteristik sekelompok data. Misalnya, konsumen

yang berhenti membeli barang dari suatu perusahaan karena ketidakpuasannya terhadap barang perusahaan tersebut. Atas konsumen, yang beralih menggunakan barang pesaing karena memberikan nilai lebih.

5. Pengelompokan (*clustering*)
   Pakar produk menggunakan fitur khusus untuk mengidentifikasi kelompok produk yang dimaksud. Ini disebut sebagai pengelompokan dan merupakan salah satu dari banyak fungsi yang termasuk dalam proses indentifikasi kelompok.
6. Asosiasi (*association*)
   Menganalisis asosiaso antar *record* adalah fungsi asosiasi dalam data mining. Proses ini melihat hubungan yang ada disetiap data yang ada. Data yang digunakan dalam proses ini dapat bersifat terkini atau historis.
7. Pengurutan (*sequencing*)
   Penambangan data bergantung pada fungsi yang disebut *Sequencing* untuk menentukan nilai titik data yang diberikan. Proses ini memperhitungkan data sebelumnya saat membuat prediksi.

3.2.1 Tahapan Data Mining

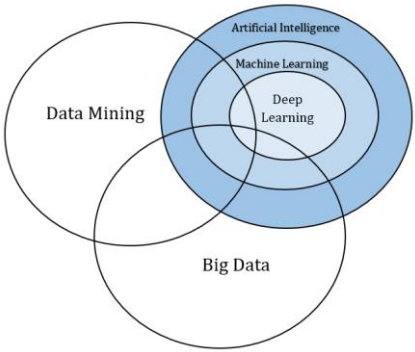
Penambang data perlu menyelesaikan beberapa fase untuk menemukan data terbanyak. Fase-fase tersebut antara lain sebagai berikut (Sinaga and Husein 2019):

1. *Data Selection*
   Sebelum menemukan fakta dari database, langkah pertama adalah memilih data dari data operasional. Ini dilakukan melalui penambangan data dan berdasarkan hasil.
2. *Pre-processing*
   Pembersihan data memerlukan penghasupan informasi duplikat, mengidentifikasi ketidakkonsistenan data, dan memperbaiki kesalahan data. Proses ini diperlukan sebelum penambangan data dapat dilakukan.

3. *Transformation*
   Transformasi adalah tidakan mengubah data mentah menjadi bentuk yang cocok untuk penambangan data. Proses kreatif ini bergantung pada informasi yang dikumpulkan untuk menemukan pola data tertentu.
4. *Data Mining*
   Penambangan data bergantung pada banyak teknik dan metode untuk menemukan pola atau informasi yang tersembunyi di dalam kumpulan data. Teknik-teknik ini seringkali bervariasi dari satu algoritma ke algoritma lainnya, tergantung pada jenis data yang ditambang. Penambangan data juga membutuhkan alasan dan tujuan yang berbeda.
5. *Evalution*
   Evalution merupakan sebelum mengumpulkan data, seorang peneliti harus terlebih dahulu melakukan interpretasi terhadap fakta dan hipotesis yang melingkupi proyeknya. Tahap KDD yang disebut evaluasi ini diperlukan untuk menghindari kontradiksi informasi atau temuan sebelumnya.

3.3 Teknik dalam Data Mining

Penambangan data dan pembelajaran mesin memiliki tingkat tumpang tindih yang tinggi penambang data mempelajari berbagai teknik yang ditemukan dalam pembelajaran mesin yang dapat mengekstrak kumpulan data besar. Banyak orang kesulitan membedakan *data* mining dari pembelajaran mesin, karena banyak ahli menganggap keduanya sebagai ilmu. Istilah“pembelajaran mesin“ dan “kecerdasan buatan“ sering dikaitkan dengan penambangan data. Itu juga istilah yang digunakan untuk menggambarkan teknik yang digunakan dalam penambangan data yang mendukungnya. Penambangan data, data besar, dan kecerdasan buatan semuanya memiliki hubungan yang erat. Gambar 2.1 menunjukkan hubungan mereka secara alami.



Gambar 3.1 Data mining, big data, artificial intelligence, machine learning, deep learning

Istilah alternatif untuk kecerdasan buatan adalah kecerdasan mesin. Ini adalah subbagian dari kecerdasan buatan, yang juga mencakup teknik pencarian, penalaran, perencanaan, dan pembelajaran. Karena istilah-istilah tersebut terkait, orang sering menggunakan istilah kecerdasan buatan secara bergantian. Ada empat teknik pembelajaran. Yang paling berkembang disebut sebagai pembelajaran mesin, karena dapat dengan mudah diterapkan pada kumpulan data yang lebih besar daripada tiga teknik pembelajaran lainnya. Mereka juga berevolusi dari sejumlah besar data yang tersedia secara online melalui *internet of things* dan media sosial.

3.4 Himpunan Data dan Jenis-Jenis Atribut

Kumpulan data terdiri dari objek data, yang mewakili entitas dunia yata. Misalnya, perguruan tinggi biasanya menggunakan objek data untuk merepresentasikan mahasiswa, kelas, dan dosen (Sani Susanto and Dedy Suryadi, S.T. 2010). Data dapat direpresentasikan menggunakan berbagai atribut. Sebagai contoh, *database* menyimpan tupel, yaitu objek data dengan atribut berupa baris dan kolom. Demikian pula, data dapat direpresentasikan

menggunakan objek yang memiliki banyak atribut lainnya(Ginantra et al. 2020). Berbagai bidang menggunakan istilah dimensi dan fitur secara bergantian saat mengacu pada karakteristik suatu objek atau individu. Misalnya, dimensi atau ciri siswa sekolah menengah diberi label dilengan baju dengan label SMA. Atribut adalah simbol yang mewakili identitas atau karakteristik individu atau objek.

3.4.1 Atribut Nominal

Contoh atribut kategorical adalah kode area, yang tidak memiliki urutan tertentu. Itu memenuhi syarat sebagai nominal karena menggambarkan nilai tanpa hierarki apa pun. Orang terkadang menyebut atribut nominal sebagai 'kategori' atau 'kode'.

3.4.2 Atribut Biner

Data mining khusus, atribut biner karena karakteristik uniknya. Ini dibagai dengan atribut nominal lainnya tetapi memiliki dua kategori nilai. Ini biasanya adalah atribut *boolean*, yang hanya memiliki nilai 0 atau 1

a. Atribut Biner Simetris, Dalam atribut binae simetris, signifikansi setiap nilai dianggap sama.
b. Atribut Biner Asimetris, Satu sisi atribut biner memiliki efek yang berbeda secara signifikan dari yang lain.

3.4.3 Atribut Numerik

Atribut numerik dihitung dengan cara yang dapat dengan diukur. Mereka memiliki nilai berdasarkan jumlah bilangan bulat atau bilangan *real*. Banyak atribut dapat diskalakan dengan membagi skala dengan interval atau rasio. Misalnya, suhu udara dapat diskalakan dengan interval dengan membagi suhu Celcius dengan rasio.

3.4.4 Atribut Ordinal

Atribut ordinal dapat digunakan dalam survei untuk mengukur data kualitatif, seperti opini dan perasaan. Atribut ini menggambarkan nilai yang menunjukkan urutan atau peringkat subjek (Sato et al. 2019). Namun, tidak jelas seberapa perbedaan antara dua nilai berurutan. Survei kepuasan konsumen biasanya menggunakan angka urut : 0 untuk sangat tidak puas, 1 untuk tidak puas, 2 untuk cukup puas, 3 untuk puas dan 4 untuk sangat puas. Bilangan urut sama seperti bilangan biner dan nominal ; semuanya kuantitatif dan berlaku untuk rentang nilai.

3.4.5 Atribut Diskrit dan Kontinu

Atribut dapat dibagi menjadi beberapa kategori berdasarkan atributnya yang berubah dan yang tidak. Atribut kontinu memiliki jumlah yang tidak terbatas ; atribut diskrit memiliki sejumlah nilai tertentu. Misalnya, indeks suhu Celcius akan memiliki kisaran antara 0 dan 150. Dalam kode komputer bilangan *real* atau *floating-point* mewakili atribut dengan nilai pecahan. Misalnya harga sepeda motor seharga Rp. 17.750.000,99 diwakili oleh bilangan *real.*

3.5 Deskripsi dan Pengetahaun yang Dihasilkan

Ada banyak cara untuk meringkas kumpulan data besar, seperti membuat metode untuk mendeskripsikan kumpulan data besar.

3.5.1 Deskripsi Grafis

Deskripsi grafis menggunakan gambar untuk mewakili data, bukan kata atau angka. Lebih mudah memahami gambar daripada kata-kata atau angka karena dapat menjelaskan ribuan kata. Gambar umum yang digunakan untuk data grafis adalah histogram dan diagram titik.

3.5.2 Deskripsi Lokasi

Sebelum terlalu kasar dan tidak praktis untuk digunakan, representasi grafis dari data tidak lengkap. Proposal memerlukan lokasi geografis karena mewakili data dari tempat tertentu.

1. **Rata – rata (*Mean*)**

   Nilai rata-rata dari sekumpulan data dianggap sebagai pusat atau titik tengah dari data tersebut. Ini juga disebut sebagai rata-rata.

   Persamaan Rata-rata

   $\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$

   Tinggi badan seseorang dihitung dengan menjumlahkan semua datanya, lalu membaginya dengan jumlah data. Misalnya, seseorang dengan data berikut akan menggunakan persamaan ini :

   Data Kelas Premium : 168, 164, 167, 164, 171, 169, 172, 166, 166

   Menjumlahkan semua data menghasilkan total 1507, yang dikonversi menjadi rata-rata 1507/9 = 167,4. Hal ini menunjukkan bahwa siswa dikelas premium umumnya mempersepdikan informasi mereka dengan tinggi rata-rata 167,4 cm.

2. **Nilai Tengah (*Median*)**

   Data di tengah harus menjadi fokus perhatian. Pertama, semua data harus diatur berdasarkan nilai. Meskipun ini tampak jelas, perlu dicatat bahwa pengurutan data dari terkecil hingga terbesar diperlukan. Ini karena datanya seperti ini :

   Data asli: 168, 164, 167, 164, 171, 166, 169,172, 166, 166

   Data urut : 164, 164, 166, 166, 166, 167, 168, 169, 171, 172

   Saat berhadapan dengan data genap, lokasi tengah berada diantara data ke-2 dan ke-3. Saat berhadapan dengan data ganjil, seperti lima atau enam, nilai tengahnya adalah tiga. Karena ada sepuluh data dan lokasi tengah ini berada di antara

data pertama dan kedua, semua fakta ini digabungkan untuk membuat kebenaran lengkap tentang data.

Ketika n titik data dipertimbangkan, median dihitung dengan mengambil [(n+1)/2] dari titik data ke-n dan menambahkan satu lagi. Ini diikuti dengan membagi hasil penjumlahan dengan 2. Jika n genap, maka perhitungan median dilakukan dengan menambahkan satu lagi dari n/2 titik data. Agar perhitungan ini dilakukan, tambahkan 1 ke masing-masing angka ini hingga Anda mencapai n/2. Kemudian hitung [(n/2)+1] dan bagi dengan 2 :

Median = [(Data ke-5 + Data ke-6)/2] = [(166+167)/2] = 166.5

3. Modus

Modus atau nilai yang sering muncul dapat digunakan untuk mengukur pusat himpunan n nilai dalam suatu atribut x. Misalkan

Data urut : 164, 164, 166, 166, 166, 167, 168, 169, 171, 172

Siswa sering datang dengan modus 166 data 3 kali berturut-turut. Hal ini menunjukkan bahwa banyak siswa memiliki tinggi badan 166 cm.

4. Kuartil

Untuk mencari median daru tengah nilai data, pisahkan data menjadi empat bagian yang sama. Selanjutnya, cari nilai disetiap bagian, atau kuartil, untuk menentukan mediannya. Contoh urutan angka ini terlihat seperti ini :

Data urut : 164 , 164, 166, 166, 166, 167, 168, 169, 171, 172

q1 q2 q3

Kuartil Pertama = 166

Kuartil kedua = [(166+167)/2] = 166.5 (sama dengan media)

Kuartil ketiga = 169

5. Persentil

Bagian tengah kumpulan data disebut sebagai p0.50 dalam hal persentase. Dan membagi data menjadi 100 bagian menghasilkan persentil atas dan bawah. Menemukan persentil ke-83, presentil ke-46, dan persentil ke-10 semuanya adalah kasus persentil khusus. Persentil ke-83 adalah 0,75 ; yang ke-46 adalah 0,25 ; dan tanggal 10 adalah 0. Untuk mempelajari lebih lanjut, kunjungi halaman ini.

Data urut : 164 , 164, 166, 166, 166, 167, 168, 169, 171, 172

q1 q2 q3

Persentil-10 = [(164+164)/2] = 164 (diantari data ke-1 & 2)
Persentil-46 = 166
Persentil-83 = 171

3.5.3 Deskripsi Keragaman

Saat ini, deskripsi lokasi memberikan gambaran umum tentang pusat data dengan mengukur rata-rata, median, dan mode lokasi. Keanekaragaman data tambahan diperlukan untuk melengkapi gambaran ; ini dicapai dengan ukuran keragaman yang disebut standar deviasi, rentang dan varians.

1. *Range*

Data yang menjangkau rentang jarak besar menonjol karena berisi berbagai macam nilai yang berbeda. Misalnya, ini mungkin muncul sebagai berikut :

Data I : 6, 6, 7, 7, 7, 8, 8
Range data I = 8-6 =2
Data II : 0, 1, 3, 7, 7, 12, 19
Range data II = 19-0 = 19

Data II memiliki data yang lebih beragam dengan *range* yang jauh lebih besar daripada Data I.

2. *Varians* dan Standar Deviasi

Nilai rentang data terbukti tidak cukup untuk menilai keragaman data secara akurat. Mencapai ini membutuhkan menemukan jarak antara setiap data dan pusatnya melalui variabilitas. Ini dilakukan melalui penggunaan persamaan yang secara konsisten merata-ratakan nilai rentang setiap data dengan nilai rentang data terdekat berikutnya. Persamaan yang dihasilkan terlihat seperti ini :

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}$$

Untuk menentukan rata-rata semua data, digunakan rumus yang mengurangi setiap pengamatan. Selanjutnya, data ini dikuadratkan, dan semuanya dijumlahkan untuk membentuk jumlah akhir. Kemudian, hasil penjumlahan tersebut dibagi dengan bilangan bulat yang kurang dari n, yaitu banyaknya pengamatan.

DAFTAR PUSTAKA

Ginantra, Ni Luh Wiwik Sri Rahayu et al. 2020. *Basis Data: Teori Dan Perancangan*. Yayasan Kita Menulis. https://kitamenulis.id/2020/10/08/basis-data-teori-dan-perancangan/.

Joko Suntoro. 2019. *Data Mining Algoritma Dan Implementasi Dengan Pemograman PHP*. Jakarta: PT Elex Media Komputindo Kompas Gramedia.

Portisch, Jan, Nicolas Heist, and Heiko Paulheim. 2022. "Knowledge Graph Embedding for Data Mining vs. Knowledge Graph Embedding for Link Prediction–Two Sides of the Same Coin?" *Semantic Web* (Preprint): 1–24.

Sani Susanto, Ph.D, and M.S Dedy Suryadi, S.T. 2010. *Pengantar Data Mining Menggali Pengetahuan Dari Bongkahan Data*. Bandung: ANDI Yogyakarta.

Sato, Yuki, Kazuhiro Izui, Takayuki Yamada, and Shinji Nishiwaki. 2019. "Data Mining Based on Clustering and Association Rule Analysis for Knowledge Discovery in Multiobjective Topology Optimization." *Expert Systems with Applications* 119: 247–61.

Sinaga, Sriyuni, and Amir Mahmud Husein. 2019. "Penerapan Algoritma Apriori Dalam Data Mining Untuk Memprediksi Pola Pengunjung Pada Objek Wisata Kabupaten Karo." *Jurnal Teknologi dan Ilmu Komputer Prima (JUTIKOMP)* 2(1): 49–54.

Suyanto, Dr. 2017. "Data Mining Untuk Klasifikasi Dan Klasterisasi Data." *Bandung: Informatika Bandung*.

BAB 4
DATA MINING ROLES

Oleh Tri Andi E. Putra

4.1 Pendahuluan

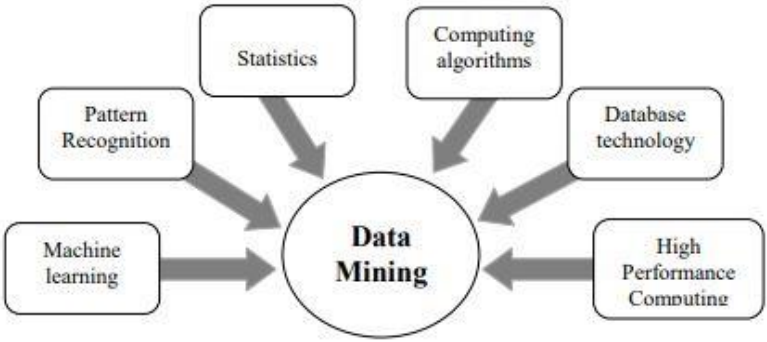
Kebutuhan informasi yang tinggi kadang tidak sebanding dengan penyajian informasi yang memadai. Informasi yang disajikan sering kali masih harus digali dari data dalam jumlah besar. Salah satu contoh yaitu data yang tumbuh dalam bidang kesehatan. Data kesehatan menyimpan banyak sekali data-data yang terkait dalam lingkungan kesehatan seperti data pasien, data obat, data penyakit, yang sangat penting untuk dapat diolah supaya lebih bermanfaat.

Metode tradisional yang biasa digunakan untuk menganalisis data, tidak dapat menangani data dalam jumlah besar. Oleh karena itu data tersebut dapat diolah menjadi pengetahuan menggunakan teknik yang disebut data mining. Sebagai bidang ilmu yang relatif baru, data mining menjadi pusat perhatian para akademisi maupun praktisi. Beragam penelitian dan pengembangan data mining banyak diaplikasikan pada bidang kesehatan. Bab ini memberikan pandangan secara singkat mengenai definisi data mining, dataset, jenis dataset, dan jenis atribut (Blikstein and Worsley, 2016)

Data mining dikenal sejak tahun 1990-an, ketika adanya suatu pekerjaan yang memanfaatkan data menjadi suatu hal yang lebih penting dalam berbagai bidang, seperti marketing dan bisnis, sains dan teknik, serta seni dan hiburan. Sebagian ahli menyatakan bahwa data mining merupakan suatu langkah untuk menganalisis pengetahuan dalam basis data atau biasa disebut Knowledge Discovery in Database (KDD). Data mining merupakan proses

untuk menemukan pola data dan pengetahuan yang menarik dari kumpulan data yang sangat besar. Sumber data dapat mencakup database, data warehouse, web, repository, atau data yang dialirkan ke dalam sistem dinamis.

Data mining, secara sederhana merupakan suatu langkah ekstraksi untuk mendapatkan informasi penting yang sifatnya implisit dan belum diketahui. Selain itu, data mining mempunyai hubungan dengan berbagai bidang diantaranya statistik, machine learning (pembelajaran mesin), pattern recognition, computing algorithms, database technology, dan high performance computing. Diagram hubungan data mining disajikan pada Gambar 5.1.



Gambar 4.1. Diagram hubungan data mining
(**Sumber:** (Muslim *et al.*, 2019))

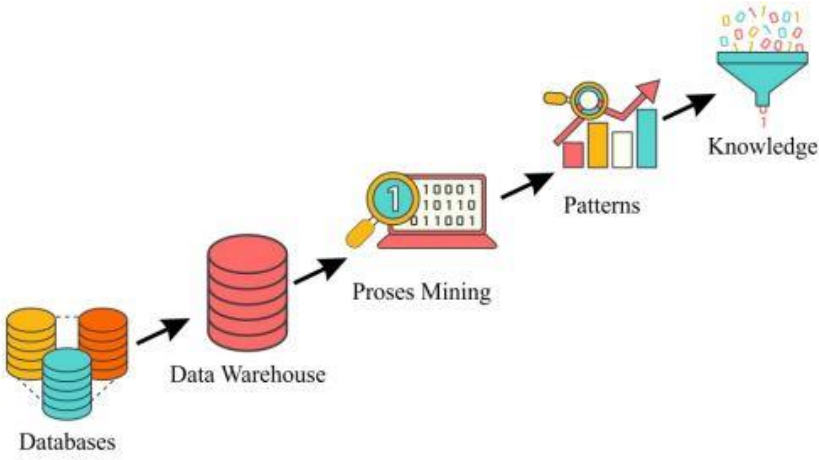
Secara sistematis, langkah utama untuk melakukan data mining terdiri dari tiga tahap, yaitu sebagai berikut:

1. Eksplorasi atau pemrosesan awal data Eksplorasi atau pemrosesan awal data terdiri dari pembersihan data, normalisasi data, transformasi data, penanganan missing value, reduksi dimensi, pemilihan subset fitur, dan sebagainya.
2. Membangun model dan validasi Membangun model dan validasi, yaitu melakukan analisis dari berbagai model dan

memilih model sehingga menghasilkan kinerja yang terbaik.

3. Pembangunan model dilakukan menggunakan metode-metode seperti klasifikasi, regresi, analisis cluster, dan asosiasi.
4. Penerapan Penerapan dilakukan dengan menerapkan model yang dipilih pada data yang baru untuk menghasilkan kinerja yang baik pada masalah yang diinvestigasi.

Tahapan proses data mining ada beberapa yang sesuai dengan proses KDD sebagaimana seperti yang digambarkan pada Gambar 5.2 yaitu:



Gambar 4.2. Proses KKD
(**Sumber:** (Silwattananusarn, 2012)

1. Cleaning and Integration

a. *Data Cleaning* (Pembersihan Data)

Data cleaning (Pembersihan data) adalah proses yang dilakukan untuk menghilangkan noise pada data yang tidak konsisten atau bisa disebut tidak relevan. Data yang diperoleh dari database suatu perusahaan

maupun hasil eksperimen yang sudah ada, tidak semuanya memiliki isian yang sempurna misalnya data yang hilang, data yang tidak valid, atau bisa juga hanya sekedar salah ketik. Data yang tidak relevan itu dapat ditangani dengan cara dibuang atau sering disebut dengan proses cleaning. Proses cleaning dapat berpengaruh terhadap performa dari teknik data mining.

b. Data Integration (Integrasi data)

Integrasi data merupakan proses penggabungan data dari berbagai database sehingga menjadi satu database baru. Data yang diperlukan pada proses data mining tidak hanya berasal dari satu database tetapi juga dapat berasal dari beberapa database.

2. Selection and Transformation

a. Data Selection (Seleksi Data)

Tidak semua data yang terdapat dalam database akan dipakai, karena hanya data yang sesuai saja yang akan dianalisis dan diambil dari database. Misalnya pada sebuah kasus market basket analysis yang akan meneliti faktor kecenderungan pelanggan, maka tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan saja.

b. Data Transformation (Transformasi Data)

Transformasi data merupakan proses pengubahan data dan penggabungan data ke dalam format tertentu. Data mining membutuhkan format data khusus sebelum diaplikasikan. Misalnya metode standar seperti analisis asosiasi dan clustering hanya bisa menerima input data yang bersifat kategorikal. Karenanya data yang berupa angka numerik apabila mempunyai sifat kontinyu perlu

dibagi-bagi menjadi beberapa interval. Proses ini sering disebut dengan transformasi data.

3. **Proses Mining**

   Proses mining dapat disebut juga sebagai proses penambangan data. Proses mining merupakan proses utama yang menggunakan metode untuk menemukan pengetahuan berharga yang tersembunyi dari data.

4. **Evaluation and Precentation**

   a. Evaluasi Pola (*Pattern Evaluation*)

      Evaluasi pola bertugas untuk mengidentifikasi pola-pola yang menarik ke dalam knowledge based yang ditemukan. Pada tahap ini dihasilkan polapola yang khas dari model klasifikasi yang dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai. Bila ternyata hasil yang diperoleh tidak sesuai dengan hipotesa, terdapat beberapa alternatif yang bisa diambil seperti menjadikannya umpan balik untuk memperbaiki proses data mining, atau mencoba metode data mining lain yang lebih sesuai.

   b. Presentasi Pengetahuan (*Knowledge Presentation*)

      Knowledge presentation merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan atau informasi yang telah digali oleh pengguna. Tahap terakhir dari proses data mining adalah memformulasikan keputusan dari hasil analisis yang didapat.

4.2 Tahapan Proses Data Mining

Kinerja yang baik tidak lepas dari proses data mining yang baik pula. Oleh karena itu, pada bab ini akan membahas mengenai proses data mining, diantaranya ada himpunan data, metode data mining, pengetahuan (*knowledge*), dan *evaluation*. Pemilihan

algoritma dan teknik yang tepat akan bergantung pada proses Knowledge Discovery in Database (KDD) secara keseluruhan. Tahapan untuk menemukan informasi atau pola dari sekumpulan data dengan menggunakan algoritma dan teknik tertentu merupakan pengertian dari data mining.(Mustika *et al.*, 2021)

Himpunan Data (Pemahaman dan pengolahan data)

Bukan data mining namanya jika tidak ada dataset yang diolah di dalamnya. Sedangkan saat ini jumlah data sangat meningkat dari berbagai media seperti ecommerce, e-government, media elektronik, dan sebagainya. Dari berbagai media tersebut akan menghasilkan data yang sangat besar. Selain itu dataset juga dapat diambil dari berbagai organisasi yang akan dijadikan objek penelitian, misalnya Bank, Rumah sakit, Industri, Pabrik, dan sebagainya yang biasa disebut dengan private dataset.

Untuk mendapatkan hasil dari proses data mining yang optimal, maka perlu diadakannya pre-processing data. Tujuan dari preprocessing, yaitu untuk memudahkan peneliti dalam memahami data yang belum dilakukan proses data mining, selain itu bisa juga digunakan untuk meningkatkan kualitas data sehingga hasil data mining akan menjadi lebih baik, dan pre-processing juga dapat meningkatkan efisiensi serta memudahkan proses penambangan data. Pengolahan data sebelum dilakukan proses data mining atau yang sering disebut dengan pre-processing data ada empat metode, yaitu pembersihan data, integrasi data, reduksi data, dan transformasi data. Selanjutnya, akan dibahas setiap teknik dari pre-processing data secara detail pada subbab berikut ini.(Atika and Priatna, 2020)

1. *Data Cleaning* (Pembersihan data)

Data Cleaning merupakan suatu pemrosesan terhadap data untuk penanganan terhadap data yang mempunyai missing value pada suatu record dan menghilangkan noise. Cleaning

data merupakan langkah awal yang perlu dilakukan, karena dalam record data yang digunakan terdapat data yang bernilai karakter "?" atau dapat disebut juga dengan data yang salah (missing value) sebagaimana seperti yang ditunjukkan pada Tabel 4.3 merupakan contoh record dataset crhonic kidney disease yang mempunyai nilai missing value. Oleh karena itu, diberi perlakuan atau penanganan untuk menangani data yang salah (missing value). Jika pada data masih ada yang mempunyai nilai kosong dapat dibersihkan dengan cara sebagai berikut:

a. Abaikan tuple Mengabaikan tuple biasanya dilakukan pada data yang tidak mempunyai label kelas. Cara ini lebih efekif apabila tuple tersebut memiliki banyak atribut kosong. Namun, metode ini kurang efektif untuk data yang mempunyai banyak tuple dengan sedikit missing value.
b. Isi atribut kosong secara manual Cara ini dapat digunakan untuk mengatasi kelemahan metode pertama. Namun cara ini tentu saja memerlukan banyak waktu dan seringkali tidak layak untuk himpunan data besar yang mengandung banyak atribut kosong.
c. Menggunakan nilai tendensi sentral rata-rata (average) Atribut kosong dapat diisi dengan menggunakan model average di mana dapat menggantikan missing value tersebut dengan nilai rataan berdasarkan nilai yang tersedia pada fitur tersebut.(Siguenza-guzman, Saquicela and Vandewalle, 2019).

2. *Data Integration* (Integrasi Data)

Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru. Integrasi data yang baik akan menghasilkan data gabungan dengan sedikit redundansi/atau inkonsistensi sehingga meningkatkan

akurasi dan kecepatan proses data mining. Permasalahan utama dalam integrasi data adalah heterogenitas semantik dan struktur dari semua data yang diintegrasikan. (Computing, 2013)

3. *Data Reduction* (Reduksi Data)

Tahap seleksi data ini terjadi pengurangan dimensi/ atribut pada dataset guna mengoptimalisasi atribut yang akan berpengaruh pada akurasi algoritma dalam me-mining dataset. Pengurangan dimensi atau seleksi atribut dapat dilakukan dengan menggunakan beberapa metode, diantaranya sebagai berikut:

1. Information Gain

   Information Gain merupakan ekspektasi dari pengurangan entropi yang dihasilkan dari partisi objek dataset berdasarkan fitur tertentu. Terdapat dua kasus berbeda pada saat perhitungan information gain, pertama untuk kasus perhitungan atribut tanpa missing value dan kedua, perhitungan atribut dengan missing value.

   - Perhitungan information gain tanpa Missing value
     Menghitung information gain tanpa missing value digunakan rumus seperti pada persamaan berikut.

$$IG(S,A) = Entropy(S) - \sum_{c \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Dimana:
S : himpunan kasus
A : atribut
Si: jumlah kasus pada partisi ke-i
S: jumlah kasus dalam S
Sementara itu, untuk menghitung nilai entropy dari koleksi label benda S dan A didefinisikan pada persamaan berikut.

$$Entropy\ (S) = \sum_{i=1}^{c} -p_i log_2 p_i$$

Dimana:
S : himpunan kasus
c : jumlah partisi S
Pi : proporsi dari Si, terhadap S.
Penghapusan atribut dilakukan satu persatu dari atribut yang memiliki nilai information gain yang paling kecil lalu akan di-mining. Proses pembuangan dan mining ini akan berhenti saat hasil akurasi masingmasing algoritma mengalami penurunan.(Rao, Govardhan and Rao, 2012)

DAFTAR PUSTAKA

Atika, P. D. and Priatna, W. 2020. Modul Perkuliahan Data Mining', *Modul Data Mining*, p. 106. Available at: http://repository.ubharajaya.ac.id/6318/1/modul fix %285%29.pdf.

Blikstein, P. and Worsley, M. 2016. Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, 3(2), pp. 220–238. doi: 10.18608/jla.2016.32.11.

Computing, M. 2013. Role of Data Mining in. 2(April), pp. 374–383. Available at: http://www.enggjournals.com/ijcse/doc/IJCSE13-05-01-051.pdf.

Muslim, M. A. *et al.* 2019. Data Mining Algoritma C4.5. in *buku data mining*. Available at: https://www.ptonline.com/articles/how-to-get-better-mfi-results.

Rao, K. V., Govardhan, A. and Rao, K. V. C. 2012. K.Venkateswara Rao 1 , A.Govardhan 2 and K.V.Chalapati Rao 1 1. 3(1), pp. 39–52.

Siguenza-guzman, L., Saquicela, V. and Vandewalle, J. 2019. Affiliations. 32(0).

Silwattananusarn, T. 2012. Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012. *International Journal of Data Mining & Knowledge Management Process*, 2(5), pp. 13–24. doi: 10.5121/ijdkp.2012.2502.

BAB 5
CLASSIFICATION AND PREDICTION

Oleh Ahmad Jurnaidi Wahidin

5.1 Pendahuluan

Di Industri 4.0, berbagai teknologi berkembang sangat pesat, keempat teknologi ini saling terkait. termasuk Internet of Things (IoT), data science, kecerdasan buatan (AI), dan big data. Dengan berkembangnya internet, data yang disimpan baik berupa teks, gambar, suara maupun video juga berkembang dengan signifikan dan sangat pesat. Dengan penggunaan *Internet of Things (*IoT) di masyarakat, banyak data dihasilkan. Teknologi big data memungkinkan kemampuan untuk memperoleh, memproses dan menyimpan data.

Data adalah aset berharga lembaga, dan dengan menggunakan data yang diekstraksi dan kemudian diproses, lembaga dapat memprediksi keputusan tentang pertumbuhan lembaga. Proses penggalian atau penambangan informasi dari suatu data disebut data mining.

Data mining yang merupakan tahapan penggalian dengan data sangat besar yang diubah menjadi basis data besar yang memfasilitasi pengambilan keputusan tentang masalah serta prediksi masa depan. Data mining menggunakan beberapa teknik guna mengetahui pola yang sebelumnya tidak diketahui. Sumber untuk data mining berasal dari gudang data yang sebelumnya dikonsolidasikan.

Data mining merupakan suatu bidang dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistic, database, dan visualisasi untuk

penanganan permasalahan pengambilan informasi dari database yang besar menurut Larose dalam (Dasril Aldo et al., 2021)

Data mining yaitu studi mengenai pengumpulan, pembersihan, pemrosesan, analisis serta penggalian wawasan dari data. Dan menjadi istilah yang sering digunakan untuk menjelaskan berbagai unsur pemrosesan data (Purwati and Kurniawan, 2021).

Fungsi data mining adalah sebagai berikut: clustering, predictive, descriptive, classification, association, characterization, outlier and trend analysis, discrimination, dan lainnya. Berbagai teknik dapat digunakan dalam proses data mining, contohnya pemodelan prediktif yang terdapat teknik klasifikasi dan prediksi. Pada data mining terdapat tiga metode yaitu Prediction, Segmentation dan Association. Tipe Prediction terbagi menjadi tiga yaitu *Classification, Time Series* dan *Regression.*

5.2 *Classification & Prediction*

Klasifikasi (*Classification*) data mining merupakan proses mendapatkan penjelasan kesamaan karakteristik pada suatu kelas atau kelompok dengan tujuan untuk memperkirakan kelas dari suatu objek yang belum diketahui labelnya. Dan metode klasifikasi menjadi salah satu yang paling sering digunakan pada data mining.

Prediksi (*Prediction*) mirip klasifikasi dan estimasi, kecuali dalam prediksi nilai dari hasil akan ada di masa mendatang. Beberapa teknik dan metode pada prediksi bisa digunakan pada klasifikasi dan estimasi.

Klasifikasi pada data mining melakukan prosesnya dengan cara mempelajari data yang ada sebelumnya, selanjutnya mengklasifikasikan data baru, metode ini menghasilkan kategorikal (ordinal ataupun nominal). Untuk mengetahui apakan perkiraan akurasi yang dihasilkan benar, maka dapat diketahui melalui confusion matrix.

		True/Observed Class	
		Positive	Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
	Negative	False Negative Count (FN)	True Negative Count (TN)

Gambar 5.2 Confusion Matrix

Menggunakan matrix pada gambar 1 orang yang melakukan data mining dapat mengetahui perkiraan akurasi dari proses yang sudah dijalankan.

Klasifikasi memakai data uji guna menentukan keakuratan model. Umumnya kumpulan data yang digunakan selanjutnya dibagi menjadi dua bagian, bagian pertama adalah data latih dan bagian kedua adalah data uji. Model yang diharapkan dibentuk menggunakan data latih kemudian proses pengujian menggunakan data uji.

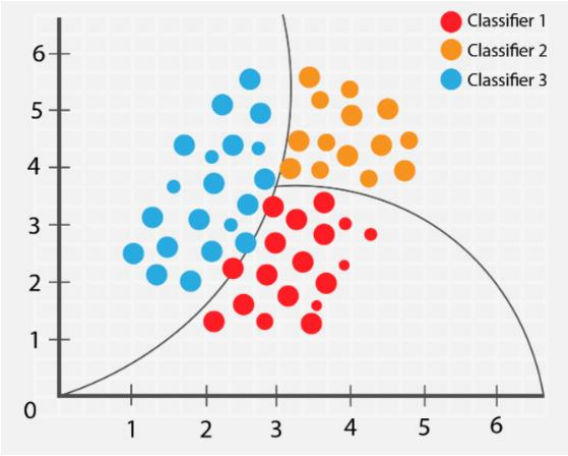
Proses pembersihan data dapat mempengaruhi kinerja metode klasifikasi. Tentunya jika data yang akan digunakan tidak bersih, memiliki banyak anomali, missing value dan masalah lainnya, maka hasil klasifikasi yang didapatkan tidak akan optimal. Confusion matrix juga dapat digunakan untuk menghitung kualitas metode klasifikasi yang digunakan dengan menghitung akurasi, spesifisitas, F-score dan sensitivitas.

Beberapa metode klasifikasi yang sering digunakan pada data mining, yaitu:

1. **Naive Bayes**

   Naive Bayes merupakan metode klasifikasi yang berdasarkan teorema Bayes, yang konsep dasarnya adalah probabilitas bersyarat, yang memprediksi kemungkinan masa depan berdasarkan pengalaman masa lalu. Naive Bayes adalah jenis algoritma supervised learning yang tidak dapat belajar sendiri, tetapi harus menerima contoh terlebih

dahulu dengan memberi label pada kumpulan data yang digunakan. Metode ini dianggap sederhana dan efektif untuk digunakan dalam analisis bisnis.



Gambar 5.3 Naive Bayes Classifier
(Sumber: kdagiit.medium.com)

Metode ini cocok untuk klasifikasi biner dan multikelas, dikenal juga sebagai Naive Bayes Classifier, metode ini memakai teknik supervised klasifikasi objek di masa depan dengan menetapkan pengidentifikasi kelas ke kasus/catatan menerapkan probabilitas bersyarat. Probabilitas bersyarat yaitu ukuran probabilitas suatu peristiwa berdasarkan peristiwa lain yang telah diasumsikan terjadi. Meskipun asumsi independensi ini sering dilanggar dalam praktiknya, Naive Bayes sering memberikan akurasi klasifikasi yang kompetitif. Ditambah dengan efisiensi komputasinya dan banyak fitur lain yang diinginkan, membuat penggunaan Naive Bayes secara luas dalam praktiknya (Webb, Keogh and Miikkulainen, 2010).

Berdasarkan fungsinya Metode Naive Bayes digolongkan menjadi tiga:

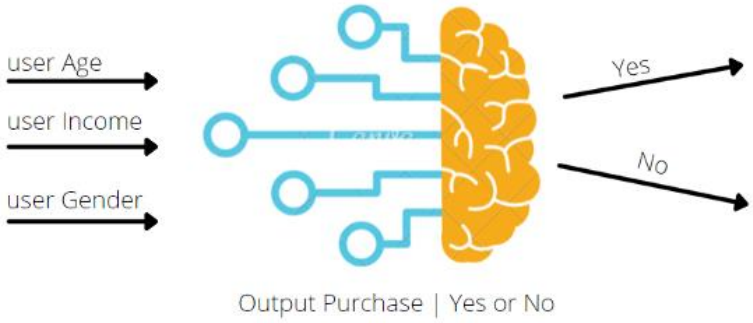
a. Multinomial Naive Bayes
   Multinomial digunakan untuk mengklasifikasikan kelas dokumen. Sebuah dokumen dapat diklasifikasikan sebagai topik olahraga, politik, teknis atau lainnya tergantung pada seberapa sering kata-kata tersebut muncul dalam dokumen.
b) Bernoulli Naive Bayes
   Bernoulli mirip dengan Multinomial, tetapi klasifikasinya lebih berfokus pada hasil ya atau tidak. Prediktor yang dimasukan yaitu variabel boolean. Misalnya, untuk memprediksi apakah suatu kata muncul dalam teks atau tidak.
c) Gaussian Naive Bayes
   Distribusi Gaussian merupakan asumsi distribusi nilai kontinu yang dikaitkan dengan setiap fitur yang berisi nilai numerik. Saat diplot, kurva berbentuk lonceng simetris muncul di sekitar rata-rata nilai fitur.

2. Logistic Regression

Dalam klasifikasi data mining, Logistic Regression merupakan algoritma yang mempunyai performance tinggi. Pada penerapan data mining, algoritma ini mempunyai perfomance yang lebih baik dibandingkan dengan algoritma lain seperti Support Vector Mechine (SVM), K-Nearest Neighbor (KNN) dan Naive Bayes (Mandiri, 2015). Hasil akurasi Logistic Regression akan rendah jika pada dataset kelasnya tidak seimbang.

Logistic Regression adalah teknik statistik yang umum dimanfaatkan untuk menganalisis data yang menggambarkan variabel respon dengan variabel prediksi satu atau lebih. Variabel respon pada dasarnya memiliki

sifat dikotomis dengan memiliki nilai 1 (ya) dan 0 (tidak), sehingga mengikuti distribusi Bernoulli untuk variabel respon yang dihasilkan(Hosmer Jr, Lemeshow and Sturdivant, 2013).



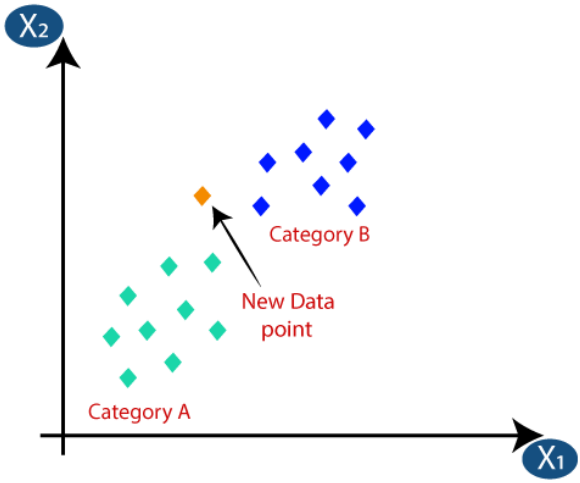
Gambar 5.4 Ilustrasi Logistic Regression
(Sumber: blog.devgenius.io)

Metode ini termasuk dalam kategori suppervised learning, yang dapat dimanfaatkan guna menyelesaikan berbagai masalah binary classification. Untuk data mengikuti persyaratan dataset untuk supervised learning. Dataset ini berpasangan (*input/output*) yang disebut dengan dataset berlabel (*labeled dataset*).

3. K-Nearest Neighbour

K-Nearest Neighbor (KNN) yaitu metode pengklasifikasian objek berdasarkan data training yang paling dekat dengan objek tersebut. Memiliki fungsi untuk mengklasifikasikan data berdasarkan data pembelajaran (*training data sets*), yang diambil dari k tetangga terdekatnya (*nearest neighbors*). Dengan k merupakan banyaknya tetangga terdekat. Teknik KNN sangat sederhana dan mudah diterapkan terutama klasifikasi, namun bisa juga digunakan untuk prediksi ataupun estimasi. Mirip dengan

metode clustering, yang melakukan penglompokan data yang baru berdasarkan jarak data baru tersebut terhadap beberapa data atau tetangga yang paling dekat.



Gambar 5.5 Ilustrasi K-Nearest Neighbor
(Sumber: javatpoint.com)

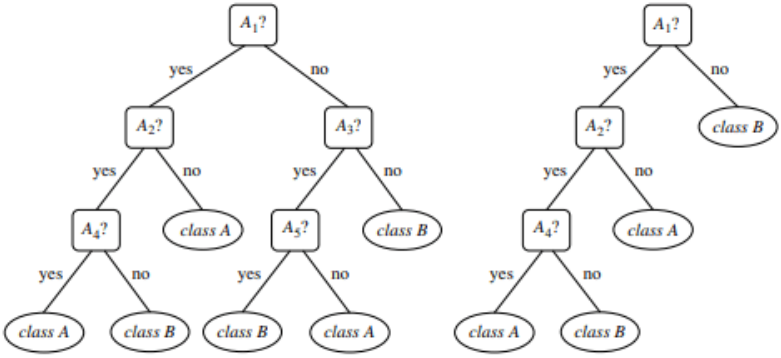
Proses metode K-Nearest Neighbor melakukan pencarian data uji untuk kelompok k objek yang paling dekat dengan objek pada data baru atau data uji. KNN merupakan algoritma supervised learning, artinya algoritma KNN memanfaatkan data yang ada sebelumnya dan sudah diketahui hasilnya (Wahidin and Maulana, 2021). KNN merupakan contoh basis pembelajaran yang menyimpan data training sehingga klasifikasi untuk data yang belum terklasifikasi dapat ditemukan dengan cara membandingkannya dengan data training.

4. Decision Tree

Decision Tree adalah salah satu teknik data mining yang terkenal dan salah satu metode paling populer untuk menentukan keputusan suatu kasus. Metode ini tidak

memerlukan proses pengolahan pengetahuan sebelumnya dan dapat digunakan untuk menyelesaikan kasus besar sekalipun. Metode ini adalah cara pengolahan data untuk memprediksi masa mendatang dengan membuat model regresi atau klasifikasi menggunakan bentuk struktur pohon. Model Decision Tree yang menggunakan struktur hierarki atau struktur pohon konsepnya adalah dengan mengubah data dan dijadikan aturan keputusan serta pohon keputusan. Dilakukan dengan membaginya lebih lanjut menjadi himpunan bagian yang lebih kecil dan mengembangkan secara bertahap pohon keputusan. Pada tahapan tersebut hasil akhirnya yaitu pohon yang memiliki node keputusan dan node daun. Contoh dari note keputusan adalah cuaca dan memiliki cabang hujan, mendung dan cerah.

Decision Tree digunakan mengeksplorasi data dan menemukan kaitan beberapa kandidat variabel input dengan variabel target. Dalam proses pemodelan data mining dan decision tree adalah langkah awal yang sangat baik.

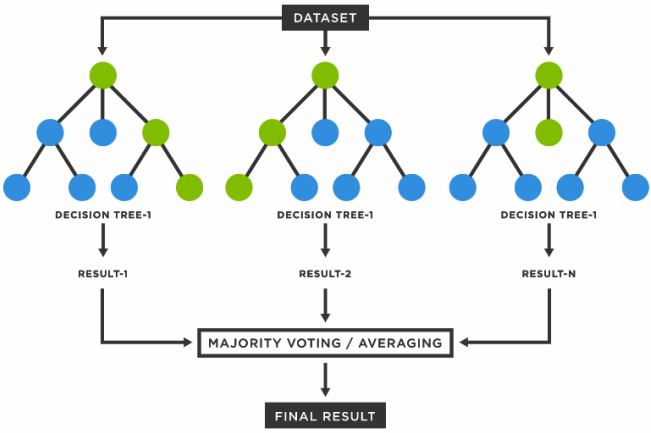


Gambar 5.6 Ilustrasi Decision Tree
(Sumber: softwaretestinghelp.com)

Keuntungan dari Decision Tree yaitu data atau perhitungan yang tidak perlu dapat dihapus. Karena biasanya sampel yang ada hanya diperiksa menurut kategori tertentu. Selain keuntungan tersebut, metode ini mempunyai kekurangan. Decision Tree ini dapat tumpang tindih, apalagi ketika kriteria dan kelas sangat sering digunakan, yang dapat meningkatkan waktu dalam pengambilan keputusan bergantung pada jumlah memori yang dibutuhkan.

5. Random Forest

Random Forest adalah salah satu metode dalam Decision Tree. Merupakan perpaduan dari masing–masing tree yang baik yang selanjutnya digabungkan menjadi sebuah model. Metode ini bergantung kepada nilai vector acak yang memiliki distribusi yang sama di semua pohon, di mana setiap decision tree mempunyai kedalaman yang maksimal (Breiman, 2001).



Gambar 5.7 Diagram Random Forest
(Sumber: tibco.com)

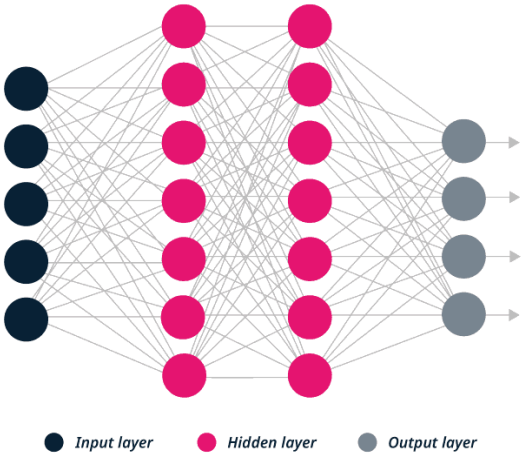
Kelebihan dari random forest adalah bisa menaikan hasil akurasi jika ada data yang hilang, serta sebagai resisting outliers, dan efisien untuk menyimpan data.

Random Forest juga memiliki proses seleksi fitur yang mana dapat mengambil fitur terbaik untuk meningkatkan performa terhadap model klasifikasi.

6. Artificial Neural Network

Artificial Neural Networks (ANN) yang dalam bahasa Indonesia adalah Jaringan Saraf Tiruan adalah model klasifikasi yang meniru cara kerja dari sistem jaringan saraf biologi otak manusia. dan merupakan metode yang sekarang ini dikembangkan oleh ahli data mining. Metode ini awalnya terinspirasi oleh jaringan saraf makhluk hidup yang diharapkan bisa meniru kinerja otak manusia dan muncul sebagai alternatif pendekatan konvensional, yang umumnya kurang fleksibel dalam menanggapi perubahan struktur masalah. Kelebihan metode ini seperti kemampuan prediksi yang memiliki pola nonlinear, kuat terhadap missing value dan waktu penyelesaian yang cepat.

Artificial Neural Networks (ANN) sering digunakan untuk data mining yang efektif, mengubah data mentah menjadi informasi yang layak. Mencari pola dalam kumpulan big data, memungkinkan bisnis untuk mempelajari lebih lanjut tentang pelanggan, yang dapat menginformasikan strategi pemasaran, meningkatkan penjualan, dan menurunkan biaya.



Gambar 5.8 Model Artificial Neural Network
(Sumber: Getsmarter, 2022)

Metode ini memiliki faktor-faktor yang sangat berperan penting dalam implementasinya pada data mining, maksudnya yaitu kombinasi yang kuat dari Neural Network itu sendiri dan teknologi data mining yang biasanya digunakan, sangat membutuhkan penelitian agar mendapatkan hasil dari inovasi data mining yang berfungsi untuk memecahkan masalah dengan tingkat akurasi yang tinggi.

DAFTAR PUSTAKA

Breiman, L. 2001. Random Forests. *Machine Learning*, 45(1), pp. 5–32. doi: 10.1023/A:1010933404324.

Dasril Aldo, S. K. M. K. *et al.* 2021. *DATA MINING*. Insan Cendekia Mandiri. Available at: https://books.google.co.id/books?id=zWgtEAAAQBAJ.

Getsmarter. 2022. *How Artificial Neural Networks Can Be Used for Data Mining*, *getsmarter.com*. Available at: https://www.getsmarter.com/blog/career-advice/how-artificial-neural-networks-can-be-used-for-data-mining/.

Hosmer Jr, D. W., Lemeshow, S. and Sturdivant, R. X. 2013. *Applied logistic regression*. John Wiley & Sons.

Mandiri, K. N. 2015. Resampling Logistic Regression untuk Penanganan Ketidakseimbangan Class pada Prediksi Cacat Software. *Journal of Software Engineering*, 1(1).

Purwati, N. and Kurniawan, H. 2021. *Data Mining*. Zahira Media Publisher (data mining). Available at: https://books.google.co.id/books?id=Q3NHEAAAQBAJ.

Wahidin, A. J. and Maulana, R. 2021. Classification of Super Air Jet Initial Cabin Crew Candidates Using K-Nearest Neighbor (KNN) Method: Klasifikasi Calon Awak Kabin Awal Super Air Jet Menggunakan Metode K-Nearest Neighbor (KNN)', *SYSTEMATICS*, 3(2), pp. 249–262.

Webb, G. I., Keogh, E. and Miikkulainen, R. 2010. Naïve Bayes. *Encyclopedia of machine learning*, 15, pp. 713–714.

BAB 6
CLUSTER *ANALYSIS*

Oleh Wara Alfa Syukrilla

6.1 Pendahuluan

Cluster analysis adalah sebuah teknik eksplorasi data yang bertujuan untuk membagi data ke dalam kelompok-kelompok dimana data dalam satu kelompok yang sama memiliki variasi sekecil mungkin (homogen) dan antar kelompok memiliki variasi sebesar mungkin (Wierzchoń and Kłopotek 2017). Penentuan banyaknya kelompok dapat dilakukan dengan dua cara, yaitu ditentukan oleh peneliti atau ditentukan oleh data (*data driven*), kemudian pengelompokan dilakukan berdasarkan statistik tertentu misalnya *Euclidean distance*. Contoh saat banyaknya kelompok ditentukan oleh peneliti adalah ketika peneliti memiliki dana terbatas dan ingin mengetahui segmentasi pasar bagi tokonya yang bejumlah 3 toko cabang. Pada kasus seperti ini, peneliti akan menetapkan banyaknya kelompok adalah 3. Sedangkan peneliti yang tidak memiliki informasi dan preferensi khusus pada data dapat membiarkan analisis klaster menemukan sendiri jumlah kelompok yang paling optimum bagi data berdasarkan kriteria statistik tertentu. Cara kedua lebih fleksibel tetapi memungkinkan antar peneliti menghasilkan jumlah kelompok dan keanggotaan yang berbeda untuk data yang sama.

Cluster analysis berguna untuk mengungkapkan karakteristik dari setiap struktur atau pola yang terkandung dalam data (Landau et al. 2011). Analisis klaster dapat menyederhanakan data berukuran besar sehingga data dapat dipahami dengan mudah dan informasi dapat diambil secara lebih efisien (Everitt, Landau, and

Leese 2001). *Cluster analysis* bukanlah sebuah algoritma. Ada beberapa algoritma yang bekerja berdasarkan prinsip *cluster analysis* (Makajić-Nikolić 2018), diantaranya dua algoritma yang banyak dipakai adalah *k-means* dan *hierarchical clustering* yang akan dibahas pada chapter ini.

6.2 K-means *Clustering*

Cara kerja k-means adalah dengan menetapkan titik tengah setiap kelompok dan mengelompokkan data berdasarkan kedekatannya terhadap titik tengah tersebut. Misal terdapat data pengamatan sebanyak n yaitu $x_1, x_2, \dots, x_n$. Setiap amatan akan masuk ke satu dari k kelompok dimana besaran k umumnya jauh lebih kecil dari n. Setiap kelompok memiliki nilai tengah yang dinamakan *centroids* dan disimbolkan dengan $c_1, c_2, \dots, c_k$, yaitu centroid dari kelompok ke-1, 2, ..., k. Setiap amatan akan dihitung jarak kedekatannya terhadap *centroids* dan suatu amatan akan masuk menjadi bagian dari kelompok ke-j jika jaraknya paling dekat dengan centroids c_j. Setelah masing-masing kelompok memiliki anggota, akan dihitung *update* rata-rata kelompok dan rata-rata tersebut menjadi centroids baru. Kemudian dilakukan perhitungan kedekatan setiap amatan kepada centroids secara berulang-ulang sampai tidak ada perubahan lagi pada posisi centroids.

6.2.1 Algoritma K-means

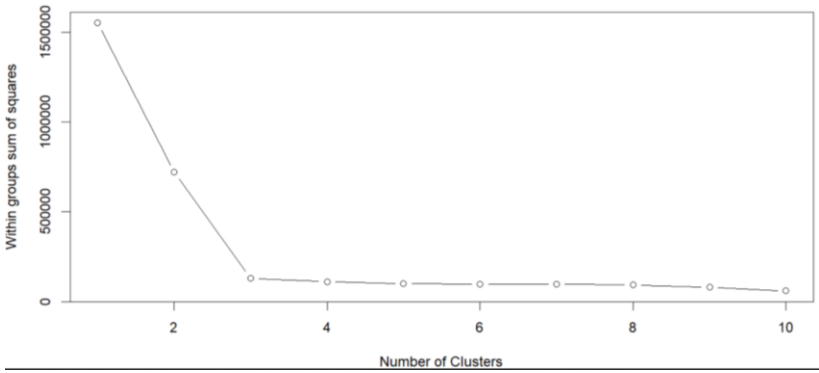
Algoritma k-means menurut (James et al. 2013) adalah:

1. Tentukan banyaknya kelompok, misal k
2. Bubuhkan angka secara acak (*random*), dari 1 hingga k, kepada masing-masing amatan. Angka-angka ini mengilustrasikan inisialisasi keanggotaan awal dari data amatan pada kelompok 1 hingga kelompok k.
3. Lakukan iterasi berikut hingga keanggotaan kelompok tidak berubah lagi:

a. Pada setiap kelompok *k*, hitung nilai *centroids*. *Centroid* dari kelompok ke-*k* adalah vektor nilai rata-rata *p* variabel dari seluruh amatan di dalam kelompok ke-*k*.
b. Masukkan amatan pada kelompok yang memiliki centroids paling dekat dengan amatan tersebut. Definisi "paling dektat" adalah berdasarkan jarak Euclidean.

Hasil *clustering* berdasarkan k-means sangat bergantung pada pengelompokan awal yang dilakukan secara *random* karena algoritma k-means cenderung akan menemukan titik *local optimum* daripada *global optimum* (James et al. 2013). Jika pengelompokan awal menggunakan angka random yang berbeda, maka hasil akhir pengelompokan k-means juga berbeda. Oleh karena itu, dalam *clustering* menggunakan k-means sangat direkomendasikan bagi peneliti untuk menjalankan algoritma k-means berulang kali dengan angka inisial *random* yang beragam. Kemudian pilih satu hasil pengelompokan yang memiliki hasil terbaik, yaitu pengelompokan yang paling meminimumkan variasi di dalam *cluster* setelah dijumlah dengan seluruh *cluster*. Variasi di dalam *cluster* ini diindikasikan dengan nilai *within cluster variance sum of square*. Hasil pengelompokan yang baik adalah yang memiliki nilai terkecil pada *within cluster variance sum of square.* Dalam menerapkan k-means, kita harus menentukan banyaknya kelompok *k* di awal. Penentuan banyaknya kelompok dapat dilakukan dengan meminta bantuan *scree plot* yang menampilkan grafik hubungan antara besarnya variasi di dalam grup (*within cluster variation sum of square*) yang dijumlahkan untuk seluruh cluster yang terbentuk. Dalam menentukan banyaknya kelompok yang optimal dalam menggambarkan data, kita berharap untuk mendapatkan jumlah klaster yang sedikit namun memiliki nilai jumlah kuadrat variasi di dalam kelompok (*within cluster variation*

sum of square) terkecil. Bentuk grafik yang memenuhi kriteria ini biasanya berbentuk siku dimana bagian sikunya menunjukkan banyaknya klaster paling kecil dari sekian pilihan banyak klaster yang memiliki *within cluster variation* rendah. Berikut ini adalah ilustrasi penentuan banyaknya kelompok berdasarkan *scree plot.*



Gambar 6.1 Scree plot antara banyaknya klaster terhadap skor *within groups sum of square*
(Sumber : data pribadi)

Berdasarkan plot di atas, banyaknya kelompok yang dianggap optimum dalam menggambarkan data adalah 3. Hal ini dikarenakan k=3 memiliki variasi dalam grup (*within group variation*) yang kecil dibandingkan k=1 dan k=2, serta k=4 dan selebihnya memiliki variasi dalam grup yang mirip dengan k=3. Oleh karena itu tidak perlu memilih jumlah klaster yang lebih banyak jika nilai variasi dalam grup terbilang mirip. Posisi k=3 berada di siku garis diagram. Adanya bentuk siku pada garis grafik memudahkan pengambilan keputusan tentang berapa jumlah klaster yang dibentuk.

6.2.2 Praktik K-means dengan Software R

Pada bagian ini akan diilustrasikan penerapan *clustering analysis* dengan algoritma k-means menggunakan software R pada data bangkitan. Kita awali dengan membangkitkan data terlebih dahulu menggunakan fungsi `rnorm()` berupa data dengan 2 variabel yang rata-ratanya telah digeser sedemikian hingga agar terbentuk tiga kelompok yang berbeda.

```
set.seed (2)
x=matrix (rnorm (60*2) , ncol =2)
x[1:20 , ]=x[1:20 ,]+2
x[21:40 , ]=x[21:40 ,] -4
x[41:60 , ]=x[41:60 ,] +6
```

Kemudian kita jalankan algoritma k-means pada data x di atas dengan banyaknya klaster adalah tiga `(k=3)` dan `nstart=20`.

```
km.out <- kmeans(x,3, nstart=20)
```

Hasil keanggotaan k-means *clustering* untuk setiap amatan dapat dilihat melalui syntax berikut.

```
> km.out$cluster
 [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3
[23] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1
[44] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

K-means *Clustering* dengan `k=3` dan `nstart=20` menghasilkan *within clusters variance sum of squares* sebesar 147.1611.

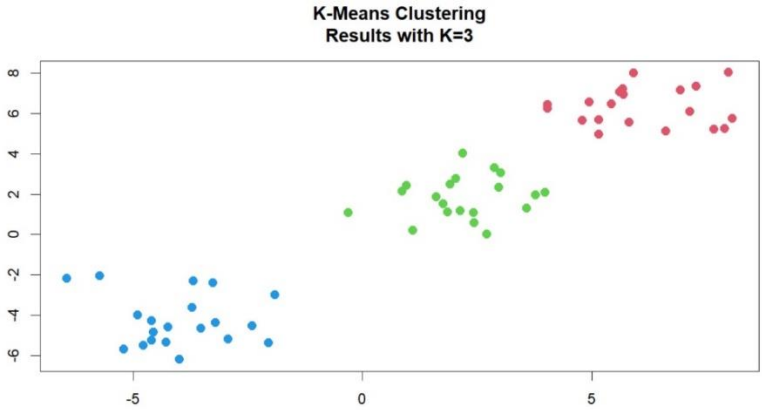
```
> km.out$tot.withinss
[1] 147.1611
```

Pada analisis k-means *clustering* di atas, digunakan `nstart=20` yang bermakna kita menggunakan 20 set nilai inisiasi awal yang berbeda dan algoritma k-means akan memilih satu hasil *clustering* terbaik yaitu yang memiliki nilai *within cluster variation sum of squares* terkecil. Mengingat hasil k-means sangat bergantung pada inisiasi *means* pengelompokan awal, maka disarankan untuk menggunakan nstart yang besar, setidaknya 20-50.

Berikut ini adalah ilustrasi bahwa jika menggunakan `nstart =1` akan mendapatkan hasil *clustering* dengan *total within cluster variation sum of squares* lebih besar daripada saat menggunakan `nstart=20`. Saat `nstart=1` nilai *total within cluster sum of square* adalah 109.369, sedangkan nilai *total within cluster sum of square* ketika menggunakan nstart=20 hanya 98.2478.

```
> km.out4 = kmeans (x,4, nstart =1)
> km.out4$tot.withinss
[1] 109.369
> km.out4 = kmeans (x,4, nstart =20)
> km.out4$tot.withinss
[1] 98.24779
```

Hasil k-means *clustering* dengan `k=3` disajikan pada Gambar 6.2. Penggunaan jumlah klaster sebanyak tiga `(k=3)` menunjukkan hasil berupa data amatan terkelompokkan dengan optimal.



Gambar 6.2 Hasil *clustering* 3 kelompok dengan k-means
(Sumber : data pribadi)

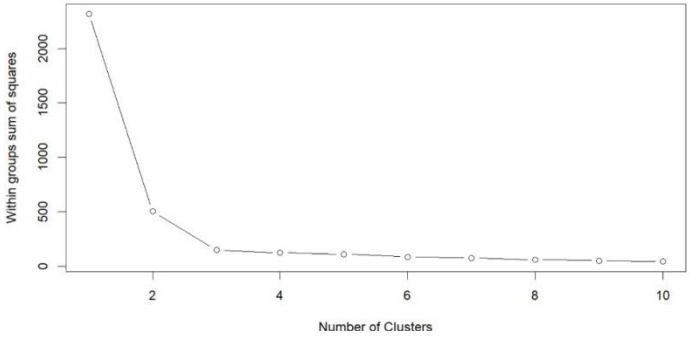
Pada contoh kasus di atas, data telah diatur agar menjadi 3 kelompok. Pada data riil, mungkin kita tidak tahu berapa jumlah klaster dalam data. Untuk itu, penentuan jumlah klaster saat *clustering* menggunakan k-means dapat dilakukan dengan membaca scree plot yang dibuat dari syntax fungsi berikut.

```
wssplot<-function(data, nc=15, seed=1234){
  wss<-(nrow(data)-1)*sum(apply(data,2,var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] <-sum(kmeans(data,
centers=i)$withinss)}
  plot(1:nc, wss, type="b", xlab="Number of
Clusters",ylab="Within groups sum of squares")}
```

Dengan menggunakan fungsi di atas, dapat kita panggil scree plot untuk data x dengan cara sebagai berikut.

```
wssplot(x, nc=10)
```

Hasil scree plot dari syntax di atas disajikan pada Gambar 6.3. Pada Gambar 6.3 terdapat bentuk siku pada posisi *number of clusters* = 3. Pada k=3 variasi di dalam klaster terbilang kecil, lebih kecil dibandingkan variasi saat k=2 dan k=1. Klaster sebanyak 4 dan seterusnya tidak dipilih karena nilai variasinya mirip dengan variasi pada saat k=3.



Gambar 6.3 Scree plot *clustering* data x.
(Sumber : data pribadi)

Seringkali data yang kita miliki terdiri dari variabel-variabel yang berbeda satuan dan besaran angka. Mengingat algoritma k-means berjalan dengan menetapkan centroids yang dihitung dari rata-rata tiap klaster, sebaiknya dilakukan standarisasi data terlebih dahulu dengan mengubah data mentah menjadi Z-score. Berikut ini adalah ilustrasi dampak dari transformasi data ke Z-score dalam k-means *clustering*. Ketika data tidak ditransformasi ke Z-score, nilai *total within clusters variation sum of squares* adalah 147.1611, sedangkan setelah distandarisasi total variasinya hanya 7.517.

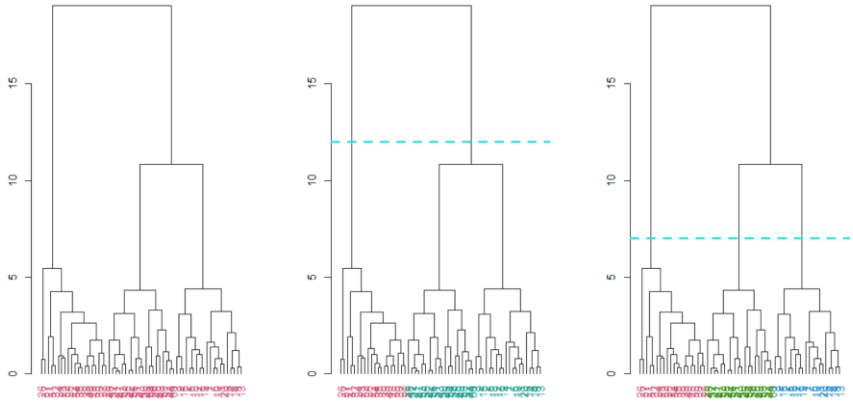
```
> x.std <- scale(x)
> km.out.std =kmeans (x.std,3, nstart =30)
> km.out.std$tot.withinss
[1] 7.517658
```

6.3 *Hierarchical Clustering*

Pada *hierarchichal clustering*, peneliti tidak diwajibkan untuk menentukan banyaknya kelompok di awal sebelum menjalankan algoritma *clustering* sebagaimana pada k-means. Hierarchichal *clustering* akan menghasilkan *dendogram* yaitu sebuah ilustrasi berbentuk seperti cabang-cabang pohon yang merupakan representasi dari data amatan. Banyaknya kelompok yang terbentuk akan ditentukan dari membaca dendogram tersebut. Oleh karena itu peneliti perlu memahami cara menginterpretasikan dendogram dengan benar.

Gambar 6.4 menyajikan contoh dendogram yang dibedakan dalam 3 tipe bagian kiri, tengah, dan kanan. Pada gambar dendogram di sebelah kiri, di bagian paling bawah dari pohon dendogram terdapat 60 ranting dimana setiap ranting merupakan representasi dari setiap amatan dalam data pada Gambar 6.2. Label di bawah daun dendogram yang berada di sumbu X adalah label untuk setiap amatan. Kemudian semakin ke atas pohon, amatan yang berdekatan akan melebur menjadi satu dan menjadi cabang. Demikian seterusnya hingga seluruh klaster melebur menjadi satu

pada puncak pohon tertinggi. Metode pembentukan dendogram analisis klaster dengan cara seperti ini disebut dengan *agglomerative method*. Amatan yang berdekatan dan saling melebur menandakan kemiripan antar amatan. Semakin cepat dua amatan melebur, dengan kata lain peleburan antar amatan terjadi di posisi pohon yang rendah, maka semakin mirip amatan tersebut. Namun semakin lama dua amatan bertemu dan melebur, atau peleburan terjadi di posisi tinggi dari sebuah pohon dendogram, maka semakin berbeda dua amatan tersebut. Sehingga dapat dikatakan bahwa dengan melihat ketinggian letak titik peleburan dua amatan pada sumbu vertikal akan diperoleh pandangan tentang seberapa mirip atau berbeda dua amatan.



Gambar 6.4 Dendogram hasil *hierarchical clustering* data x.
(Sumber : data pribadi)

Melihat gambar dendogram pada Gambar 6.4, dendogram di sebelah kiri menampilkan bahwa seluruh amatan saling bersambungan dan melebur hingga pada puncak pohon dendogram menjadi 1 klaster. Jika pohon dendogram dipotong pada posisi ketinggian 12 (ditandai dengan garis horizontal putus-putus) maka akan dihasilkan 2 klaster yang diberi warna merah dan hijau. Sedangkan jika dipotong pada ketinggian 7 akan

menghasilkan 3 klaster yang diindikasikan dengan warna berbeda untuk klaster yang berbeda. Jika pemotongan dilakukan pada posisi yang lebih rendah, misal pada ketinggian 0, maka artinya masing-masing amatan adalah satu klaster tersendiri.

Pengambilan keputusan tentang banyaknya klaster dengan memotong pohon dendogram merupakan kelebihan yang dimiliki *hierarchical clustering* dibanding kmeans. Sebab kmeans menuntut kita untuk menentukan jumlah klaster di awal, sedangkan *hierarchical clustering* tidak perlu dan kita bisa memilih ingin memiliki jumlah klaster berapapun. Namun, terdapat kekurangan pada *hierarchical clustering* yaitu tidak ada standar baku mengenai dimana letak pemotongan garis dendogram yang optimal untuk menentukan banyak klaster dalam data kita. Seringkali peneliti melihat dendogram lalu memilih letak pemotongan dendogram berdasarkan ketinggian leburan, banyak klaster yang realistis, serta banyak klaster yang disesuaikan dengan harapan.

Kekurangan lain dari *hierarchical clustering* adalah tidak berlaku untuk sembarang data karena klaster pada *hierarchical clustering* tersarang pada klaster lainnya (M.Kom et al. 2020). Sebagai contoh, kita memiliki data dimana banyaknya laki-laki dan perempuan adalah sama dan banyaknya subjek yang beragama Islam, Kristen, Katolik, Hindu, Budha, Konghucu juga sama rata di dalam data kita. Jika dilakukan *clustering* dengan 2 kelompok, skenario pembagian kelompok paling baik adalah data dikelompokkan berdasarkan jenis kelamin menjadi 2 kelompok. Sedangkan jika dilakukan *clustering* dengan k=6, pembagian kelompok paling baik adalah *clustering* berdasarkan agama. Namun realitanya, pengelompokan ini tidak bersarang, dalam arti pembagian kelompok menjadi 6 bukan diperoleh dari pembagian kelompok menjadi 2 lalu masing-masing kelompok dipecah lagi menjadi 3 sehingga didapat 6 kelompok. Untuk kasus seperti ini, penggunaan *hierarchical clustering* mungkin kurang akurat dibandingkan k-means.

6.3.1 Algoritma *Hierarchical Clustering*

Cara kerja hierarchichal *clustering* adalah dengan menentukan kriteria kemiripan sebagai tahapan awal. Ada beberapa kriteria kemiripan yang dapat digunakan, tetapi yang paling sering digunakan adalah jarak Euclidean. Kemudian, jadikan setiap amatan sebagai satu klaster sehingga kita memiliki n buah klaster yang diilustrasikan dengan bagian paling bawah pada dendogram. Selanjutnya, amatan yang paling mirip akan melebur menjadi satu sehingga diperoleh $n-1$ klaster. Selanjutnya, amatan yang paling mirip akan melebur juga sehingga kita memiliki $n-2$ klaster. Proses peleburan tersebut dilakukan terus menerus secara iteratif hingga seluruh amatan tergabung menjadi satu klaster. Metode ini disebut *agglomerative method.*

Menurut James et al. (2013), algoritma *hierarchical clustering* adalah:

1. Mulai dengan n amatan dan sebuah ukuran kemiripan antar pasangan amatan (contohnya jarak Euclidean) sehingga diperoleh $C_2^n = \frac{n(n-1)}{2}$ pasangan ukuran kemiripan antar amatan. Setiap amatan diperlakukan sebagai klaster sendiri.
2. Untuk $i = n, n = 1, 2, \ldots, n$, lakukan:
   a. Pengecekan kemiripan pada seluruh pasangan antar-klaster pada i klaster. Identifikasi mana pasangan klaster yang paling mirip. Lebur dua klaster ini. Derajat kemiripan dua klaster ini menentukan ketinggian dendogram terkait dimana posisi peleburan dua klaster ini seharusnya diletakkan.
   b. Hitung skor kemiripan antar pasangan klaster pada $n-1$ klaster yang tersisa.

6.3.2 Praktik *Hierarchical Clustering* dengan Software R

Pada software R, akan digunakan fungsi `hclust()` untuk menjalankan perintah hirearchical *clustering*. Selain itu, akan digunakan fungsi `dist()` untuk menghitung $n \times n$ matriks jarak Euclidean bagi seluruh pasangan antar amatan/klaster.

Pertama kita bangkitkan data terlebih dahulu. Data yang digunakan adalah data yang sama dengan data bangkitan untuk ilustrasi k-means.

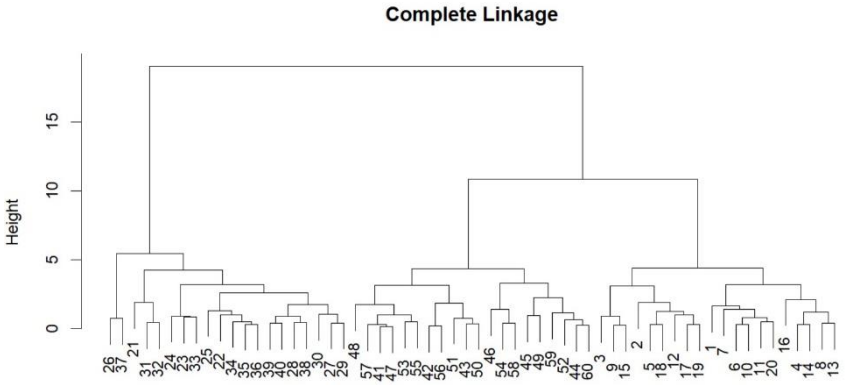
```
set.seed (2)
x=matrix (rnorm (60*2) , ncol =2)
x[1:20 , ]=x[1:20 ,]+2
x[21:40 , ]=x[21:40 ,] -4
x[41:60 , ]=x[41:60 ,] +6
```

Kemudian *hierarchical clustering* dengan complete linkage diterapkan.

```
hc.complete=hclust(dist(x),method="complete")
```

Dendogram hasil *clustering* diperoleh menggonakan fungsi plot().

```
plot(hc.complete, main = "Complete Linkage",
xlab="", sub ="", cex =.9)
```



Gambar 6.5 Dendogram Hasil *Hierarchical Clustering* Complete Linkage

(Sumber : Dokumentasi Pribadi)

Hasil keanggotaan setiap amatan pada klaster yang terbentuk dapat dipanggil menggunakan perintah `cutree()`, dimana elemen pertama dari fungsi tersebut diisi objek hasil *clustering* dengan `hclust()`, dan elemen kedua diisi banyaknya kelompok yang diharapkan.

```
> cutree (hc.complete , 3)
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2
[24] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3
[47] 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

DAFTAR PUSTAKA

Everitt, Brian S., Sabine Landau, and Morven Leese. 2001. *Cluster Analysis*. Taylor & Francis.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 103. New York, NY: Springer New York.

Landau, Sabine, Morven Leese, Daniel Stahl, and Brian S. Everitt. 2011. *Cluster Analysis*. John Wiley & Sons.

Makajić-Nikolić, Dragana. 2018. *XIII Balkan Conference on Operational Research Proceedings*. FON.

M.Kom, Dwi Ely Kurniawan, Afdhol Dzikri M.T, Sudra Irawan M.Sc, and Oktavianto Gustin IPM M. T. 2020. *Sistem Informasi Geografis: Praktikum Dan Penerapan Dalam Pengambilan Keputusan*. PolibatamPress.

Wierzchoń, Slawomir, and Mieczyslaw Kłopotek. 2017. *Modern Algorithms of Cluster Analysis*. Springer.

BAB 7
ASSOCIATION RULES

Oleh Anindya Khrisna Wardhani

7.1 Pendahuluan

Association rules mining (ARM) pertama kali dikemukakan oleh Agrawal dkk pada tahun 1993, yang kemudian diteliti oleh banyak peneliti lainnya. Mereka mengoptimalkan algoritma asli, seperti memasukkan sampel acak, pemikiran paralel, menambahkan titik referensi, menolak aturan, mengubah kerangka penyimpanan, dan lain-lain (Yunita Ardilla & Anindya Khrisna Wardhani, 2021). Pekerjaan tersebut bertujuan untuk meningkatkan efisiensi aturan algoritma, menyebarkan penerapan association rules dari bisnis awal arah ke bidang lain, seperti pendidikan, penelitian ilmiah, kedokteran, dll.

Penambangan menggunakan association rules adalah untuk menemukan asosiasi dan hubungan di antara set item data besar. Penambangan association rules adalah cabang penting dari penelitian data mining, dan association rules adalah gaya data mining yang paling khas(Khrisna Wardhani et al., 2020). Saat ini, masalah penambangan Association rules sangat dihargai oleh para peneliti dalam database, kecerdasan buatan, statistik, pencarian informasi, terlihat, ilmu informasi, dan banyak bidang lainnya(Ghafari & Tjortjis, 2019). Banyak hasil luar biasa telah ditemukan. Apa yang secara efisien dapat menangkap hubungan penting antar data adalah bentuk Association rules yang sederhana dan mudah dijelaskan dan dipahami. Masalah Association rules penambangan dari database besar telah menjadi konten penelitian yang paling matang, penting, dan aktif(Estrela et al., 2011).

7.2 Kelemahan Association rules

Aturan asosiasi adalah salah satu jenis pengetahuan yang paling umum, yang memiliki aplikasi luas. Namun, terdapat tiga bug dalam algoritma aturan asosiasi saat ini. Pertama, sebagian besar algoritma didasarkan pada Apriori, yang membutuhkan biaya tinggi untuk menggali aturan asosiasi multi-layer dan multi-dimensi. Kedua, lapisan konsep digali dengan menggunakan mode top-down, tidak mendukung penggalian lintas lapisan. Ketiga, aturan asosiasi yang diperoleh memiliki redundansi yang lebih besar(Khrisna Wardhani et al., 2022). Selain itu, aturan klasifikasi dapat dianggap sebagai semacam aturan asosiasi khusus, karena distribusi data yang tidak merata dalam aplikasi dunia nyata, algoritma umum sering menggunakan metode penambangan dua tahap, yang hanya dapat menggunakan tingkat dukungan tunggal, efisiensi penambangan dan skalabilitas sistem tidak tinggi, perlu menggunakan beberapa tingkat dukungan (Rekik et al., 2018).

ARM adalah area penelitian yang dieksplorasi dengan baik, seperti yang dinyatakan sebelumnya, submasalah kedua ARM sangat mudah, sebagian besar pendekatan tersebut berfokus pada submasalah pertama. Submasalah pertama dapat dibagi lagi menjadi dua submasalah: calon proses pembuatan itemset besar dan proses pembuatan frequent itemsets. Kumpulan item yang dukungannya melebihi ambang dukungan sebagai kumpulan item besar atau sering, kumpulan item yang diharapkan atau memiliki harapan untuk menjadi besar atau sering disebut kumpulan item kandidat(Wardhani et al., 2018).

Sebagian besar algoritma ARM sangat mirip, perbedaannya adalah sejauh mana peningkatan tertentu telah dilakukan, jadi hanya beberapa tonggak pencapaian asosiasi algoritma penambangan aturan akan diperkenalkan. Pertama kami akan memperkenalkan beberapa algoritma naive dan dasar untuk penambangan aturan asosiasi, pendekatan seri Apriori (Wang et al., 2018).

7.3 Implementasi Perhitungan Asosiation Rules

Untuk membuat aturan asosiasi, setidaknya diperlukan dua langkah untuk menggunakan nilai dukungan dan kepercayaan di atas:

Frequent itemset

Langkah pertama adalah menentukan kumpulan objek yang sering, sering kali berarti kombinasi objek yang sering muncul (data peristiwa), dalam kumpulan data ini. sehingga aturan yang dibuat dapat menghasilkan tingkat kepercayaan yang tinggi. Kembali ke studi kasus, mari kita mulai dengan mencari barang-barang umum. Membuat kemungkinan kombinasi item berdasarkan set k-item (seringkali merupakan generasi item), di sini k berarti jumlah item yang akan digabungkan.

k-itemset (k=1)

Kita mulai dengan membangkitkan subset k = 1, maka subset yang dapat digenerate beserta banyaknya event pada semua event adalah sebagai berikut:

Tepung = 6
Gula=4
Minyak=6
Telur=6
Mentega=3

Berdasarkan delapan sampel pembelian, terdapat 6 produk terigu, 4 produk gula, 6 produk minyak, 6 telur, dan 3 produk mentega.

k-itemset (k=2)
pada langkah iterasi kedua kita lanjutkan dengan k = 2, yaitu membentuk kombinasi dari 2 set elemen sebagai berikut:
{Tepung,Gula} = 2
{Tepung,Minyak}=4
{Tepung,Telur}=5
{Tepung,Mentega}=2
{Gula,Minyak}=3
{Gula,Telur}=2
{Gula,Mentega}=1
{ Telur,Minyak}=5
{Minyak,Mentega}=2
{Telur,Mentega}=2
Dan seterusnya……

Dari delapan transaksi pembelian pada contoh, minimal 2 dari delapan transaksi pembelian produk tepung dan gula secara bersama-sama, 4 transaksi pembelian produk tepung dan minyak, 5 transaksi pembelian produk tepung dan telur, dst.

k-itemset (k=3)
maka pada iterasi ketiga dengan k = 3 terbentuk gabungan dari 3 himpunan elemen sebagai berikut :
{Telur,Minyak,Tepung}=4
{Telur,Minyak,Gula}=2
{Telur,Minyak,Mentega}=2
{Telur,Tepung,Gula}=1
{Telur,Tepung,Mentega}=2
{Tepung,Gula,Mentega}=0
{Tepung,Mentega,Minyak}=2
{Tepung,Gula,Minyak,}=1
{Gula,Minyak,Mentega}=1
Dan seterusnya…….

Dari 8 transaksi pada contoh, 4 transaksi membeli produk {telur, minyak dan tepung} secara bersamaan, 2 transaksi membeli barang {telur, minyak dan gula}, 2 transaksi membeli produk {telur, minyak dan mentega} dan sebagainya.

k-itemset (k=4)
terakhir pada iterasi keempat dengan nilai k=4, akan dibentuk kombinasi dari 4 buah itemset sebagai berikut:

{Telur,Minyak,Tepung,Gula}=1
{Telur,Minyak,Tepung,Mentega}=2
{Telur,Minyak,Gula,Mentega}=0
Dan seterusnya…..

Dari delapan transaksi pada contoh, 1 transaksi membeli produk {telur, minyak, tepung dan gula} secara bersamaan, 1 transaksi membeli produk {telur, minyak, tepung dan mentega} secara bersamaan, dan seterusnya. Anda dapat membayangkan bahwa hanya 8 transaksi dengan maksimal 4 produk yang dibeli per transaksi dapat menghasilkan 31 kombinasi asosiasi.

Pada fase audiens umum, kami membuat setidaknya 31 kemungkinan kombinasi audiens. Selain itu, kita dapat menghitung skor kepercayaan untuk setiap rangkaian item umum yang digunakan sebagai aturan asosiasi saat skor kepercayaan tinggi.

Ekstraksi Association rules

Setelah mendapatkan kombinasi subset yang frequent, langkah selanjutnya adalah mengekstrak aturan asosiasi dari kombinasi subset dengan nilai kepercayaan tinggi. Untuk menghitung nilai confidence juga perlu menghitung nilai support agar dapat diketahui korelasi antara nilai support

dengan nilai safe. Misalnya, nilai dukungan dan kepercayaan dari set target k (k=2) dihitung sebagai berikut:

{Tepung,Gula} = 2
s(Tepung->Gula) = 2/8 = 0.25 = 25%
c(Tepung->Gula)=2/6 = 0.33 = 33%
{Tepung,Telur}=5
s(Tepung->Telur) = 5/8 = 0.625= 62.5%
c(Tepung->Telur)=5/6 = 0.83 = 83%

Seperti yang dapat dilihat dari contoh perhitungan di atas, kami melanjutkan ini untuk menentukan nilai dukungan dan kepercayaan untuk 31 set target lain yang umum digunakan. Butuh banyak waktu, tapi coba lihat contoh perhitungan di atas. Ada korelasi yang dengannya kita dapat membuat hipotesis yang memfasilitasi penentuan aturan asosiasi yang sesuai dengan konsep algoritma apriori.

7.4 Algoritma Apriori

Algoritma Apriori merupakan algoritma yang efisien untuk menentukan jumlah himpunan elementer yang berulang. Prinsip dasar dari algoritma ini adalah jika suatu himpunan objek merupakan himpunan objek yang berulang, maka semua subset (bagian) dari himpunan objek tersebut juga berulang dan sebaliknya. jika mis. Misalnya, jika grup objek A tidak berulang (tidak sering terjadi dalam peristiwa), maka tidak ada item yang digabungkan dari item A yang membuat objek A sering terjadi (sering terjadi dalam peristiwa).

Nah, itulah yang dimanfaatkan oleh algoritma Apriori untuk mengurangi/membatasi ruang pencarian judul-judul umum. Hal ini tentu saja ditandai dengan batasan nilai support threshold (minSupport). Nilai support yang rendah

mempengaruhi nilai kepercayaan. Misalnya nilai support produk set (tepung->gula) adalah 25%, nilai support minimal justru mempengaruhi nilai confidence yaitu hanya 33%. Tentu saja kelompok sasaran tidak boleh dijadikan salah satu aturan asosiasi, karena nilai pastinya hanya 33%. Tentu saja, algoritma apriori memotong aturan dengan nilai kepercayaan minimum sehingga nantinya aturan asosiasi dihasilkan dengan nilai kepercayaan yang sesuai.

Pada Algortima Apriori langkah pertama yang harus dilakukan adalah:

Menentukan nilai minimum Support

Misalnya, nilai dukungan minimum yang kami gunakan adalah minSupport=4 (setara dengan 4/8 = 0,5 atau 50%). Kemudian pada iterasi pertama dari set target k (k = 1) terbentuk aturan sebagai berikut:

Tepung = 6
Gula=4
Minyak=6
Telur=6
Mentega=3

Dari lima item set, produk mentega (3/8 = 0,375 atau 37,5%) tidak mencapai level support minimum = 50%. Dengan demikian, pada iterasi kedua, set target k (k=2) menjadi. Semua barang yang mengandung mentega juga dihilangkan sesuai dengan prinsip algoritma apriori.

{Tepung,Gula} = 2
{Tepung,Minyak}=4
{Tepung,Telur}=5
{Tepung,Mentega}=2

{Gula,Minyak}=3
{Gula,Telur}=2
{Gula,Mentega}=1
{ Telur,Minyak}=5
{Minyak,Mentega}=3
{Telur,Mentega}=3
Dan seterusnya……

Pada k-itemset (k=2) di atas, itemset {Tepung,Gula} (2/8=0.25 atau **25%**), {Gula,Minyak} (3/8=0.375 atau **37.5%**) dan {Gula,Telur} (2/8=**25%**) tidak memenuhi nilai minimum support, sehingga itemset tersebut juga dieliminasi. Pada iterasi ketiga k-itemset (k=3) hanya tersisa 1 itemset yang memenuhi minimum nilai support yaitu itemset {Telur,Minyak,Tepung} (4/8 =0.5 atau **50%**)

{Telur,Minyak,Tepung}=4
{Telur,Minyak,Gula}=2
{Telur,Minyak,Mentega}=2
{Telur,Tepung,Gula}=1
{Telur,Tepung,Mentega}=2
{Tepung,Gula,Mentega}=0
{Tepung,Mentega,Minyak}=2
{Tepung,Gula,Minyak,}=1
{Gula,Minyak,Mentega}=1
Dan seterusnya…….

Berdasarkan algoritma Apriori, maka Association rules yang berhasil didapatkan adalah sebagai berikut :

1. *{Tepung,Minyak}*
   Nilai confident, c(Tepung->Minyak) = 4/6 = 0.67 = ***67%***
2. *{Tepung,Telur}*
   Nilai confident, c(Tepung->Telur) = 5/6 = 0.83 = ***83%***

3. *{Minyak,Telur}*
   Nilai confident, c(Minyak->Telur) = 5/6 = 0.83 = ***83%***
4. *{Telur,Minyak,Tepung}*
   Nilai confident, c(Telur,Minyak->Tepung) = 4/5 = 0.67 = ***80%***

Association rules :

1. *If Tepung, Maka Minyak*
2. *If Tepung, Maka Telur*
3. *If Minyak, Maka telur*
4. *If Telur dan Minyak, Maka Tepung*

DAFTAR PUSTAKA

Estrela, C., Guedes, O. A., Silva, J. A., Leles, C. R., Estrela, C. R. de A., & Pécora, J. D. 2011. Diagnostic and clinical factors associated with pulpal and periapical pain. *Brazilian Dental Journal*, *22*(4), 306–311. https://doi.org/10.1590/S0103-64402011000400008

Ghafari, S. M., & Tjortjis, C. 2019. A survey on association rules mining using heuristics. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (Vol. 9, Issue 4). Wiley-Blackwell. https://doi.org/10.1002/widm.1307

Khrisna Wardhani, A., Novita Putri, A., Fathi Salim Ashour, S., Medis dan Informasi Kesehatan, R., Informatika, T., Science, C., & Rukun Abdi Luhur, P. 2020. *Telematika An Improved K-NN Algorithm and Bagging for Liver Disease Classification*. *15*(2), 100–107. https://doi.org/10.35671/telematika.v15i2.1247

Khrisna Wardhani, A., Nugraha, E., Ulfiana, Q., Medis, R., Kesehatan, I., Rukun, P., & Luhur, A. 2022. Optimization of the Decision Tree Method using Pruning on Liver Disease Classification. In *Journal of Applied Informatics and Computing (JAIC)* (Vol. 6, Issue 2). http://jurnal.polibatam.ac.id/index.php/JAIC

Rekik, R., Kallel, I., Casillas, J., & Alimi, A. M. 2018. Assessing web sites quality: A systematic literature review by text and association rules mining. *International Journal of Information Management*, *38*(1), 201–216. https://doi.org/10.1016/j.ijinfomgt.2017.06.007

Wang, F., Li, K., Duić, N., Mi, Z., Hodge, B. M., Shafie-khah, M., & Catalão, J. P. S. 2018. Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns. *Energy Conversion and Management*, *171*, 839–854. https://doi.org/10.1016/j.enconman.2018.06.017

Wardhani, A. K., Widodo, C. E., & Suseno, J. E. 2018. *Information System for Culinary Product Selection Using Clustering K-Means and Weighted Product Method*. *165*(ICCSR), 18–22. https://doi.org/10.2991/iccsr-18.2018.5

Yunita Ardilla, & Anindya Khrisna Wardhani. 2021. *DATA MINING DAN APLIKASINYA*. https://repository.penerbitwidina.com/media/publications/351768-data-mining-dan-aplikasinya-7b2a8129.pdf

BAB 8
TEXT MINING

Oleh Nono Heryana

8.1 Apa itu Text Mining?

Text mining adalah suatu teknik komputasi dan pembelajaran mesin yang digunakan untuk mengolah dan menganalisis teks secara otomatis. Text mining bertujuan untuk mengekstrak informasi yang bermanfaat dari teks, seperti menentukan tema atau topik, mengidentifikasi entitas atau kata kunci, atau mengukur sentiment atau emosi dari teks.

Text mining menggunakan berbagai macam teknik dan algoritma, seperti pemrosesan natural language, analisis korpus, atau pembelajaran mesin, untuk mengolah dan menganalisis teks secara otomatis. Text mining juga menggunakan konsep-konsep dari linguistik komputasi, seperti tokenisasi, stemming, atau N-gram, untuk memperoleh informasi yang bermanfaat dari teks.

Text mining memiliki banyak aplikasi dan manfaat dalam berbagai bidang, seperti pemasaran, penelitian, atau pemerintahan. Dengan text mining, dapat dilakukan berbagai macam analisis teks, seperti klasifikasi teks, cluster atau pengelompokan teks, atau sumber daya linguistik yang terkait dengan teks. Dengan demikian, text mining memiliki peran yang penting dalam mengelola dan menganalisis teks secara otomatis.

8.1.1 Terminologi

Beberapa terminologi yang sering digunakan dalam text mining adalah:

1. *Corpus*: Ini adalah kumpulan teks yang akan dianalisis menggunakan text mining. Corpus dapat berupa teks-teks yang terkait dengan satu topik atau tema, atau teks-teks yang dikumpulkan dari sumber-sumber yang berbeda.
2. *Token*: Ini adalah unit dasar dari teks yang akan dianalisis menggunakan text mining. Token biasanya merupakan kata-kata atau frasa-frasa yang terdapat dalam teks, yang akan diolah dan dianalisis secara terpisah.
3. *Term*: Ini adalah kata atau frasa yang terdapat dalam teks, yang akan dianalisis menggunakan text mining. Term biasanya merupakan kata-kata yang memiliki makna atau nilai informatif dalam teks, yang akan digunakan untuk mengidentifikasi tema atau topik dari teks.
4. *Term frequency (tf)*: Ini adalah jumlah kemunculan suatu term dalam teks. Term frequency digunakan untuk mengukur seberapa sering suatu term muncul dalam teks, sehingga dapat menunjukkan kepentingan atau relevansi suatu term dalam teks.
5. *Document frequency (df)*: Ini adalah jumlah dokumen yang mengandung suatu term dalam corpus. Document frequency digunakan untuk mengukur seberapa banyak dokumen yang mengandung suatu term, sehingga dapat menunjukkan kemunculan atau distribusi suatu term dalam corpus.
6. *Inverse document frequency (idf)*: Ini adalah nilai logaritmik yang mengukur seberapa jarang suatu term muncul dalam corpus. Inverse document frequency digunakan untuk mengimbangi term frequency, sehingga dapat menghitung nilai informatif suatu term dalam teks.
7. N-gram: Ini adalah gabungan dari N token yang terdapat dalam teks, yang akan dianalisis menggunakan text mining. N-

gram biasanya digunakan untuk mengidentifikasi frasa-frasa yang sering muncul dalam teks, sehingga dapat membantu mengidentifikasi pola-pola yang terdapat dalam teks.

8. *Stemming*: Ini adalah proses mengubah kata-kata menjadi kata dasar, sehingga dapat digunakan untuk mengelompokkan kata-kata yang memiliki akar yang sama dalam teks. Stemming biasanya digunakan untuk mengurangi variasi kata-kata yang muncul.

8.1.2 Tujuan Text Mining

Tujuan utama dari text mining adalah untuk mengolah dan menganalisis teks secara otomatis menggunakan teknik-teknik komputasi dan pembelajaran mesin. Text mining bertujuan untuk mengekstrak informasi yang bermanfaat dari teks, seperti menentukan tema atau topik, mengidentifikasi entitas atau kata kunci, atau mengukur sentiment atau emosi dari teks.

Dengan text mining, dapat dilakukan berbagai macam analisis teks, seperti klasifikasi teks, cluster atau pengelompokan teks, atau sumber daya linguistik yang terkait dengan teks. Text mining juga dapat membantu mengidentifikasi pola-pola yang tidak terlihat secara manual dari teks, sehingga dapat membantu mengambil keputusan atau menemukan pola-pola yang bermanfaat dari teks.

Selain itu, text mining juga dapat membantu mengoptimalkan pencarian dan navigasi informasi di dalam teks, sehingga dapat memudahkan pengguna untuk menemukan informasi yang diinginkan dengan cepat dan tepat. Dengan demikian, text mining memiliki banyak manfaat dan tujuan yang bermanfaat dalam mengelola dan menganalisis teks secara otomatis.

8.1.3 Manfaat Text Mining

Menurut (Kwartler, 2017) ada banyak manfaat yang bisa didapatkan dari *text mining* yaitu:

1. Kepercayaan yang timbul antar pemangku kepentingan (*stakeholders*);
2. Metodologi dapat diterapkan dengan cepat;
3. Memungkinkan untuk diaudit dan diulang;
4. Text mining mengidentifikasi wawasan baru atau memperkuat persepsi yang ada berdasarkan semua informasi yang relevan.

Sedangkan Manfaat text mining secara umum meliputi:

1. Mempermudah akses terhadap informasi: Text mining membantu mengkonversi teks menjadi bentuk yang lebih mudah dianalisis dan diolah oleh mesin, sehingga mempermudah akses terhadap informasi yang terkandung dalam teks.
2. Mempercepat proses analisis: Text mining menggunakan algoritma dan teknik-teknik machine learning untuk mengekstrak informasi dari teks dengan cepat dan akurat, sehingga mempercepat proses analisis dan menghasilkan keputusan yang lebih cepat.
3. Menghasilkan analisis yang lebih akurat: Text mining menggunakan teknik-teknik *natural language processing* yang canggih untuk mengekstrak informasi dari teks, sehingga dapat menghasilkan analisis yang lebih akurat dan terpercaya.
4. Membantu mengambil keputusan yang lebih baik: Text mining mengekstrak informasi penting dari teks yang mungkin tidak terlihat secara visual, sehingga dapat membantu individu atau organisasi mengambil keputusan yang lebih baik berdasarkan data yang lebih lengkap dan akurat.

5. Menghasilkan pengetahuan baru: Text mining dapat mengidentifikasi pola-pola dan hubungan-hubungan yang mungkin tidak terlihat secara visual dalam teks, sehingga dapat menghasilkan pengetahuan baru yang bermanfaat bagi berbagai bidang.

8.1.4 Penggunaan Text Mining

Text mining dapat digunakan dalam berbagai bidang, termasuk :

1. Bisnis : Text mining dapat digunakan untuk mengekstrak informasi penting dari ulasan pelanggan, laporan keuangan, dan lainnya untuk membantu perusahaan mengambil keputusan yang lebih baik.
2. Kesehatan : Text mining dapat digunakan untuk mengekstrak informasi dari catatan medis, laporan klinis, dan sumber-sumber lainnya untuk membantu dokter mengambil keputusan diagnosis dan pengobatan yang lebih baik.
3. Penelitian : Text mining dapat digunakan untuk mengekstrak informasi dari artikel-artikel ilmiah, laporan penelitian, dan sumber-sumber lainnya untuk membantu peneliti menemukan pola dan hubungan yang mungkin tidak terlihat secara visual.
4. Media : Text mining dapat digunakan untuk mengekstrak informasi dari berita-berita, artikel-artikel, dan sumber-sumber lainnya untuk membantu jurnalis mengumpulkan fakta dan menulis artikel yang lebih baik.
5. Pemerintahan : Text mining dapat digunakan untuk mengekstrak informasi dari dokumen-dokumen pemerintah, laporan-laporan, dan sumber-sumber lainnya untuk membantu pemerintah membuat kebijakan yang lebih baik.

8.2 Proses Text Mining

Untuk menemukan pengetahuan atau *knowledge* dalam sekumpulan teks melibatkan beberapa langkah yang berbeda. Menurut (Žižka, Dařena and Svoboda, 2019) proses *text mining* secara umum sebagai berikut:

1. Mendefinisikan masalah, Langkah ini sebenarnya tidak tergantung pada tindakan apa pun yang mungkin diambil selanjutnya. Di sini, domain masalah perlu dipahami dan pertanyaan yang harus dijawab, didefinisikan.
2. Mengumpulkan data yang diperlukan, Sumber teks yang berisi informasi yang diinginkan perlu diidentifikasi dan dokumen dikumpulkan. Teks dapat berasal dari dalam perusahaan (database atau arsip internal) atau dari sumber eksternal, misalnya dari web. Dalam hal ini, untuk mengambil konten halaman web secara langsung. Atau, API dari beberapa sistem berbasis web dapat digunakan untuk mengambil data. Setelah pengambilan, teks disimpan sehingga siap untuk analisis lebih lanjut.
3. Mendefinisikan fitur, Fitur yang mencirikan teks dengan baik dan cocok untuk tugas yang diberikan perlu didefinisikan. Fitur biasanya didasarkan pada isi dokumen. Pendekatan yang sangat sederhana, *bag-of-words* dengan bobot atribut biner, mengambil setiap kata sebagai fitur boolean. Nilainya menunjukkan apakah kata tersebut ada dalam dokumen atau tidak. Metode lain mungkin menggunakan skema pembobotan yang lebih rumit atau fitur yang diturunkan dari kata (kata yang dimodifikasi, kombinasi kata, dll.).
4. Menganalisis data, Ini adalah proses menemukan pola dalam data menurut jenis tugas yang harus diselesaikan (misalnya, klasifikasi), model atau algoritma tertentu dipilih dan properti serta parameternya ditentukan. Kemudian, model tersebut dapat diterapkan pada data

sehingga dapat ditemukan solusi dari masalah yang dipecahkan. Untuk memecahkan masalah tertentu, biasanya tersedia lebih banyak model. Beberapa model memiliki kompleksitas komputasi yang lebih tinggi daripada yang lain. Menurut penggunaan model, pembuatan cepat dapat lebih disukai daripada aplikasi cepat atau sebaliknya. Kesesuaian model seringkali sangat bergantung pada data. Model yang sama dapat memberikan hasil yang sangat baik untuk satu kumpulan data sementara itu dapat sepenuhnya gagal untuk yang lain. Jadi, memilih model yang tepat, menemukan struktur yang tepat untuknya, dan menyetel parameter seringkali membutuhkan banyak upaya eksperimental.

5. Menginterpretasi hasil, Di sini, beberapa hasil diperoleh dari analisis. Kita perlu hati-hati melihatnya dan menghubungkannya dengan masalah yang ingin kita pecahkan. Fase ini mungkin mencakup langkah-langkah verifikasi dan validasi untuk meningkatkan keandalan hasil.

Proses text mining melibatkan beberapa tahapan, seperti preprocessing data teks, ekstraksi fitur, dan aplikasi algoritma pembelajaran mesin. Pada tahap preprocessing, data teks diperoleh dan dibersihkan dari noise atau karakter yang tidak relevan. Setelah itu, fitur-fitur yang dianggap penting dari data teks tersebut diekstrak, seperti kata-kata yang sering muncul atau entitas yang disebutkan. Kemudian, algoritma pembelajaran mesin seperti klasifikasi atau clustering diterapkan untuk memproses data teks dan menemukan pola atau hubungan di dalamnya. Hasil dari proses text mining kemudian dapat digunakan untuk mengambil keputusan atau menyusun laporan.

Proses text mining biasanya meliputi tahapan-tahapan berikut:

1. Pembersihan teks: Tahap pertama dalam proses text mining adalah membersihkan teks dari *noise*, seperti tanda baca, simbol-simbol, dan angka, sehingga data teks yang akan dianalisis menjadi lebih bersih dan mudah diolah oleh mesin.
2. Tokenisasi: Tahap selanjutnya adalah memecah teks menjadi token-token yang lebih kecil, seperti kata-kata atau frase-frase, sehingga dapat digunakan untuk analisis lebih lanjut.
3. *Stemming* atau lemmatisasi: Tahap selanjutnya adalah mengubah kata-kata menjadi bentuk dasar atau kata dasar, sehingga kata-kata yang berbeda tetapi memiliki makna yang sama dapat dianggap sebagai satu kata.
4. Ekstraksi fitur: Tahap selanjutnya adalah mengidentifikasi fitur atau ciri-ciri penting dalam teks yang dapat digunakan untuk mengklasifikasikan atau mengelompokkannya. Fitur-fitur ini dapat berupa kata-kata yang sering muncul, frase-frase yang spesifik, atau pola-pola kata yang umum.
5. Pengkategorian: Tahap selanjutnya adalah mengelompokkan teks ke dalam kategori atau label yang sesuai, seperti topik yang dibicarakan, sentimen yang tersirat, atau kelas yang diprediksi.
6. Klasifikasi: Tahap selanjutnya adalah menggunakan model machine learning untuk memprediksi kelas atau label suatu teks berdasarkan data yang telah diberikan. Model ini dapat dibuat dengan menggunakan algoritma-algoritma seperti Naive Bayes, SVM, atau RNN.
7. Analisis sentimen: Tahap selanjutnya adalah mengidentifikasi dan mengekstrak informasi tentang sentimen atau perasaan yang tersirat dalam teks. Ini dapat dilakukan dengan menggunakan teknik-teknik natural language processing atau analisis kata-kata kunci yang sering muncul dalam teks.
8. Analisis tema: Tahap terakhir adalah mengidentifikasi dan mengekstrak informasi tentang tema utama atau topik yang

dibicarakan dalam suatu teks. Ini dapat dilakukan dengan menggunakan algoritma-algoritma seperti LDA atau Word2Vec.

8.3 Metode Text Mining

Metode text mining adalah suatu proses ekstraksi informasi bermanfaat dan bermakna dari data teks. Ini biasanya dilakukan dengan menggunakan teknik pengolahan bahasa alami (NLP), yang melibatkan menggunakan algoritma dan perangkat lunak untuk menganalisis dan menafsirkan teks bahasa alami. Text mining dapat digunakan untuk mendapatkan wawasan dari data teks tidak terstruktur besar, seperti ulasan pelanggan, posting media sosial, dan artikel berita. Hal ini juga dapat digunakan untuk membantu mengotomatisasi tugas-tugas seperti analisis sentimen dan model topik.

Beberapa metode text mining yang umum digunakan meliputi:

1. Analisis sentimen: Mengidentifikasi dan mengevaluasi emosi dan perasaan dari teks, seperti kepuasan atau kekecewaan pelanggan.
2. Klasifikasi dokumen: Membagi dokumen teks ke dalam kelompok-kelompok yang berbeda berdasarkan tema atau kategori yang relevan.
3. Ekstraksi entitas: Mengekstrak informasi yang berkaitan dengan entitas spesifik dari teks, seperti nama perusahaan, lokasi, atau produk.
4. *Topic modeling*: Mengidentifikasi dan mengelompokkan dokumen teks berdasarkan tema yang terkait.
5. *Text summarization*: Membuat ringkasan teks yang menyajikan informasi penting dari dokumen teks asli dengan jumlah kata yang lebih sedikit.

8.3.1 Analisis Sentimen

Analisis sentimen adalah suatu teknik text mining yang digunakan untuk mengidentifikasi dan mengevaluasi emosi dan perasaan dari teks. Ini sering digunakan untuk menentukan apakah suatu teks bersifat positif, negatif, atau netral. Analisis sentimen dapat digunakan dalam berbagai situasi, seperti menganalisis ulasan pelanggan untuk menentukan tingkat kepuasan mereka terhadap suatu produk atau layanan, atau menganalisis posting di media sosial untuk menentukan apakah suatu topik memiliki sentimen positif atau negatif. Teknik ini biasanya dilakukan dengan menggunakan algoritma klasifikasi yang telah dilatih dengan menggunakan data teks yang telah diberi label secara manual untuk mengklasifikasikan teks baru sebagai positif, negatif, atau netral.

Beberapa tahapan umum dalam melakukan analisis sentimen meliputi:

1. Pembersihan data: Mengikis karakter yang tidak diinginkan dari teks, seperti tanda baca, angka, dan simbol.
2. Pemotongan kata: Memisahkan teks menjadi kata-kata individu atau frase-frase yang lebih kecil, yang disebut "token".
3. Pemilihan fitur: Memilih fitur teks yang akan digunakan untuk mengklasifikasikan teks sebagai positif, negatif, atau netral. Fitur ini bisa berupa kata-kata atau frase-frase yang sering muncul dalam teks positif atau negatif.
4. Klasifikasi: Menggunakan algoritma klasifikasi yang telah dilatih dengan data teks yang telah diberi label untuk mengklasifikasikan teks baru sebagai positif, negatif, atau netral.
5. Evaluasi: Mengevaluasi hasil klasifikasi untuk menentukan seberapa akurat algoritma dalam mengklasifikasikan teks baru.

8.3.2 Klasifikasi Dokumen

Klasifikasi dokumen adalah suatu teknik text mining yang digunakan untuk membagi dokumen teks ke dalam kelompok-kelompok yang berbeda berdasarkan tema atau kategori yang relevan. Ini sering digunakan untuk mengelompokkan dokumen teks menjadi kategori yang sudah ditentukan sebelumnya, seperti topik, genre, atau tipe dokumen. Klasifikasi dokumen dapat digunakan untuk membantu mengatur dan mengelompokkan dokumen teks yang besar sehingga lebih mudah dicari dan dianalisis. Teknik ini biasanya dilakukan dengan menggunakan algoritma klasifikasi yang telah dilatih dengan menggunakan data teks yang telah diberi label secara manual untuk mengklasifikasikan dokumen baru ke dalam kelompok yang relevan.

8.3.3 Ekstraksi Entitas

Ekstraksi entitas adalah suatu teknik text mining yang digunakan untuk mengekstrak informasi yang berkaitan dengan entitas spesifik dari teks. Entitas ini bisa berupa nama perusahaan, nama orang, lokasi, produk, atau hal-hal lain yang terkait dengan teks tersebut. Ekstraksi entitas sering digunakan untuk mengumpulkan informasi yang berkaitan dengan suatu topik tertentu dari teks yang besar. Teknik ini biasanya dilakukan dengan menggunakan algoritma yang telah dilatih dengan data teks yang telah diberi label secara manual untuk mengekstrak entitas yang relevan dari teks baru.

8.3.4 Topic Modeling

Topic modeling adalah suatu teknik text mining yang digunakan untuk mengidentifikasi dan mengelompokkan dokumen teks berdasarkan tema yang terkait. Ini biasanya dilakukan dengan menggunakan algoritma yang dapat mengelompokkan dokumen teks yang memiliki kata-kata atau

frase-frase yang sama atau serupa menjadi topik yang terkait. Topic modeling dapat membantu mengelompokkan dokumen teks yang besar menjadi tema yang lebih kecil dan lebih mudah dianalisis. Teknik ini biasanya dilakukan dengan menggunakan algoritma yang telah dilatih dengan data teks yang telah diberi label secara manual untuk mengelompokkan dokumen baru ke dalam topik yang relevan.

Topic modeling memiliki beberapa kegunaan, di antaranya:

1. Membantu mengelompokkan dokumen teks yang besar menjadi tema yang lebih kecil dan lebih mudah dianalisis.
2. Memungkinkan pengguna untuk menemukan dokumen teks yang relevan dengan topik yang dicari dengan lebih cepat dan mudah.
3. Membantu mengidentifikasi tema yang mungkin tidak diketahui sebelumnya dari dokumen teks yang besar.
4. Membantu menyajikan informasi dari dokumen teks dalam bentuk yang lebih mudah dipahami.
5. Membantu mengelompokkan dokumen teks berdasarkan kesamaan tema, sehingga memungkinkan analisis lebih lanjut tentang perbedaan dan persamaan antara tema yang berbeda.

8.3.5 Text Summarization

Text summarization adalah suatu teknik text mining yang digunakan untuk membuat ringkasan teks yang menyajikan informasi penting dari dokumen teks asli dengan jumlah kata yang lebih sedikit. Ini sering digunakan untuk membuat ringkasan dari dokumen teks yang panjang atau berulang-ulang sehingga lebih mudah dibaca dan dipahami. *Text summarization* dapat membantu menyajikan informasi penting dari teks dengan lebih cepat dan efisien. Teknik ini biasanya dilakukan dengan menggunakan algoritma yang telah dilatih dengan data teks yang telah diberi label secara manual untuk mengekstrak informasi penting dari teks baru.

8.4 Algoritma Text Mining

Beberapa algoritma yang dapat digunakan dalam text mining meliputi:

1. *Naive Bayes*: Ini adalah algoritma klasifikasi yang menggunakan teori probabilitas untuk memprediksi kelas suatu teks berdasarkan fitur-fitur yang terkandung di dalamnya.
2. *Support Vector Machine (SVM)*: Ini adalah algoritma klasifikasi yang menggunakan teknik-teknik pembelajaran mesin untuk membuat garis batas yang memisahkan teks-teks berdasarkan kelasnya.
3. *Latent Dirichlet Allocation (LDA)*: Ini adalah algoritma pengkategorian yang menggunakan teori probabilitas untuk mengelompokkan teks-teks berdasarkan tema atau topik yang dibicarakan.
4. Word2Vec: Ini adalah algoritma representasi vektor yang menggunakan teknik-teknik deep learning untuk mengkonversi kata-kata menjadi vektor-vektor yang dapat digunakan untuk analisis lebih lanjut.
5. *Recurrent Neural Network (RNN)*: Ini adalah algoritma pembelajaran mesin yang menggunakan teknik-teknik deep learning untuk mengekstrak informasi secara terus menerus dari teks yang panjang.

DAFTAR PUSTAKA

Almuzaini, H.A. and Azmi, A.M., 2020. Impact of stemming and word embedding on deep learning-based Arabic text categorization. IEEE Access, 8, pp.127913-127928.

Chen, Z., Teng, S., Zhang, W., Tang, H., Zhang, Z., He, J., Fang, X. and Fei, L., 2018, August. LSTM sentiment polarity analysis based on LDA clustering. In CCF Conference on Computer Supported Cooperative Work and Social Computing (pp. 342-355). Springer, Singapore.

de Oliveira Lima, T., Colaço, M., Prado, K.H.D.J. and de Oliveira, F.R., 2021, December. A Big Data Experiment to Evaluate the Effectiveness of Traditional Machine Learning Techniques Against LSTM Neural Networks in the Hotels Clients Opinion Mining. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 5199-5208). IEEE.

Dieng, A.B., Ruiz, F.J. and Blei, D.M., 2020. Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics, 8, pp.439-453.

El-Kassas, W.S., Salama, C.R., Rafea, A.A. and Mohamed, H.K., 2021. Automatic text summarization: A comprehensive survey. Expert Systems with Applications, 165, p.113679.

Justicia De La Torre, C., Sánchez, D., Blanco, I. and Martín-Bautista, M.J., 2018. Text mining: techniques, applications, and challenges. International journal of uncertainty, fuzziness and knowledge-based systems, 26(04), pp.553-582.

Kim, D., Seo, D., Cho, S. and Kang, P., 2019. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. Information Sciences, 477, pp.15-29.

Kwartler, T. (2017) Text Mining in Practice with R. 1st edn. Wiley. Available at: https://doi.org/10.1002/9781119282105.

Li, C., Lu, Y., Wu, J., Zhang, Y., Xia, Z., Wang, T., Yu, D., Chen, X., Liu, P. and Guo, J., 2018, April. LDA meets Word2Vec: a novel model for academic abstract clustering. In Companion proceedings of the the web conference 2018 (pp. 1699-1706).

Li, Q., Li, S., Zhang, S., Hu, J. and Hu, J., 2019. A review of text corpus-based tourism big data mining. Applied Sciences, 9(16), p.3300.

Yadav, A. and Vishwakarma, D.K., 2020. Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review, 53(6), pp.4335-4385.

Žižka, J., Dařena, F. and Svoboda, A. (2019) Text mining with machine learning: principles and techniques. First. Boca Raton: CRC Press.

BAB 9
EKSTRAKSI FITUR

Oleh Tutuk Indriyani

9.1 Pengertian dan Manfaat

Fitur adalah atribut yang berasal dari data yang telah diproses sebelumnya untuk mencirikan sifat-sifat data. Ekstraksi fitur adalah proses pengurangan dimensi di mana sekumpulan awal data mentah direduksi menjadi grup yang lebih mudah dikelola untuk diproses. Karakteristik kumpulan data besar ini adalah sejumlah besar variabel yang membutuhkan banyak sumber daya komputasi untuk diproses. Ekstraksi fitur juga merupakan metode yang memilih atau menggabungkan variabel ke dalam fitur, secara efektif mengurangi jumlah data yang harus diproses, sambil tetap mendeskripsikan kumpulan data asli secara akurat dan lengkap. Teknik lain yang umum digunakan untuk mengurangi jumlah fitur dalam sebuah dataset adalah seleksiatau pemilihan fitur (Nixon and Aguado, 2018). Perbedaan antara pemilihan fitur dan ekstraksi fitur adalah bahwa pemilihan fitur bertujuan untuk menentukan peringkat pentingnya fitur yang ada dalam kumpulan data dan ekstraksi fitur membuang fitur yang kurang penting (tidak ada fitur baru yang dibuat). Ciri khas dari kumpulan data besar ini adalah bahwa mereka mengandung sejumlah besar variabel dan selain itu, variabel-variabel ini membutuhkan banyak sumber daya komputasi untuk memprosesnya. Oleh karena itu Ekstraksi Fitur dapat berguna dalam hal ini dalam memilih variabel tertentu dan juga menggabungkan beberapa variabel terkait yang akan mengurangi jumlah data. Hasil yang diperoleh akan dievaluasi dengan bantuan

langkah-langkah presisi dan penarikan kembali. Salah satu teknik reduksi dimensi linier yang paling banyak digunakan juga adalah ***Principal Component Analysis*** **(**PCA).

Maanfaat dari ekstraksi fitur adalah berguna untuk mengurangi jumlah sumber daya yang diperlukan untuk pemrosesan tanpa kehilangan informasi penting atau relevan. Ekstraksi Fitur bertujuan untuk mengurangi jumlah fitur dalam dataset dengan membuat fitur baru dari yang sudah ada dan kemudian membuang fitur aslinya yangdikurangi tersebut. Serangkaian fitur baru yang dikurangi ini kemudian harus dapat meringkas sebagian besar informasi yang terkandung dalam rangkaian fitur asli. Dengan cara ini, versi ringkasan dari fitur asli dapat dibuat dari kombinasi set aslinya. Ekstraksi fitur juga dapat mengurangi jumlah data yang berlebihan untuk analisis tertentu. Juga, pengurangan data dan upaya mesin dalam membangun kombinasi variabel (fitur) memfasilitasi kecepatan pembelajaran dan langkah-langkah generalisasi dalam proses pembelajaran mesin (Nixon and Aguado, 2018). Menggunakan regularisasi tentu dapat membantu mengurangi risiko overfitting, tetapi sebaliknya menggunakan teknik ekstraksi fitur juga dapat menghasilkan jenis keuntungan lain seperti: Peningkatan akurasi, pengurangan risiko overfitting. percepatan dalam latihan dan peningkatan visualisasi data.

Generasi fitur adalah proses menciptakan fitur baru dari fitur yang sudah ada. Karena ukuran kumpulan data sangat bervariasi, menjadi tidak mungkin untuk mengelola kumpulan data yang lebih besar. Dengan demikian proses pembuatan fitur ini dapat memainkan peran penting untuk memudahkan tugas. Untuk menghindari pembuatan fitur yang tidak berarti, kami menggunakan beberapa rumus matematika dan model statistik untuk meningkatkan kejelasan dan akurasi. Proses ini biasanya menambahkan lebih banyak informasi ke model agar lebih akurat (Nixon and Aguado, 2018). Jadi meningkatkan akurasi model

adalah sesuatu yang dapat dicapai melalui proses ini. Proses ini dengan cara mengabaikan interaksi yang tidak bermakna dengan mendeteksi interaksi yang bermakna. Evaluasi fitur sangatlah penting untuk memprioritaskan fitur terlebih dahulu untuk menyelesaikan pekerjaan kami dengan cara yang terorganisir dengan baik dan dengan demikian evaluasi fitur dapat menjadi alat untuk ini. Di sini setiap fitur dievaluasi untuk menilai mereka secara objektif dan selanjutnya memanfaatkannya berdasarkan kebutuhan saat ini. Yang tidak penting bisa diabaikan. Jadi evaluasi fitur adalah tugas penting yang harus dilakukan untuk mendapatkan hasil akhir yang tepat dari model dengan mengurangi bias dan inkonsistensi dalam data.

Saat ini, *Machine Learning* digunakan untuk menganalisis data yang semakin banyak dan data yang tersedia menjadi semakin kompleks. Dalam dekade terakhir, munculnya *Deep Learning* membantu menciptakan model pembelajaran yang lebih efisien (Devulapalli S. *et al*. 2021). Banyak tugas *Machine Learning* menargetkan masalah klasifikasi. Sistem seperti itu bekerja dengan cara tertentu. Langkah pertama yang dilakukan adalah fitur diekstrak dari data masukan. Hal ini dapat dilihat sebagai pembuatan representasi baru dari data khusus untuk tugas saat ini. Sistem klasifikasi kemudian dipelajari di atas fitur-fitur ini untuk mencapai terselesaikannya tugas. Setelah dilatih, sistem sekarang harus dapat digunakan pada data yang belum terlihat selama fase pelatihan dan memprediksi responsnya secara akurat, dalam hal ini label kelas. Seringkali, dan khususnya hingga beberapa tahun terakhir, fitur yang diekstraksi dari input merupakan fitur buatan tangan. Kualifikasi ini berarti bahwa fitur tersebut dirancang khusus untuk input data dan tugas yang ada. Mereka umumnya terikat tidak hanya pada jenis data, misalnya gambar, kata-kata tulisan tangan, tetapi juga pada subset tertentu seperti gambar dan kata-kata tulisan tangan yang ditulis dengan tinta. Sebagian besar fitur ini umumnya tidak kuat untuk diubah.

Pendekatan lain untuk mengekstrak fitur dari data adalah mempelajari ekstraksi fitur menggunakan *Machine Learning*. Alih-alih membangun sistem untuk mengklasifikasikan beberapa gambar, sistem pembelajaran dibangun untuk mengekstraksi fitur dari input. Dalam kasus gambar, ini berarti bahwa jaringan mempelajari fitur tingkat tinggi langsung dari piksel masukan, bahwa pendekatan ini lebih baik daripada menggunakan fitur buatan tangan, karena beberapa alasan. Dengan melatih model pada setiap kumpulan data, model yang dilatih dapat diadaptasi ke berbagai jenis input sedangkan fitur buatan tangan mungkin memerlukan penyetelan tangan untuk setiap kumpulan data. Selain itu, pendekatan ini seharusnya tidak membutuhkan seorang ahli. pengetahuan tentang gambar yang dianalisis, menganalisis mendalam tentang teknik ekstraksi fitur dari data. Seperti yang ditunjukkan pada bagian berikut yaitu fokus khusus diberikan pada ekstraksi fitur dari gambar dan lebih khusus lagi dari tulisan tangan. Kelas metode pembelajaran fitur yang dianalisis mencakup karya *Geoffrey Hinton* sebelumnya, yang disebut *"Hinton Approach"* (Huang P. *et al*, 2021).

9.2 Penggunaan Praktis Ekstraksi Fitur

9.2.1 Auto-Encoder

Tujuan autoencoder adalah pembelajaran tanpa pengawasan dari pengkodean data yang efisien. Ekstraksi fitur digunakan di sini untuk mengidentifikasi fitur kunci dalam data untuk pengkodean dengan belajar dari pengkodean kumpulan data asli untuk mendapatkan yang baru. Autoencoder adalah teknik pembelajaran tanpa pengawasan untuk jaringan saraf yang mempelajari representasi data (pengkodean) yang efisien dengan melatih jaringan untuk mengabaikan sinyal *noise.* Jaringan autoencoder memiliki tiga lapisan: input, lapisan tersembunyi untuk pengkodean, dan lapisan decoding keluaran. Menggunakan propagasi balik, algoritma tanpa pengawasan terus menerus

melatih dirinya sendiri dengan menggunakan propagasi balik, algoritma tanpa pengawasan terus menerus melatih dirinya sendiri dengan menyesuaikan nilai output target agar sama dengan input. Hal ini memaksa lapisan penyandian tersembunyi yang lebih kecil untuk menggunakan pengurangan dimensi untuk menghilangkan *noise* dan merekonstruksi inputan. Cara kerja Autoencoder dijelaskan sebagai berikut: Jaringan autoencoder belajar sendiri cara mengompresi data dari lapisan input menjadi kode yang lebih pendek, lalu membuka kompresi kode tersebut ke dalam format apa pun yang paling cocok dengan input aslinya. Proses ini terkadang melibatkan banyak autoencoder, seperti lapisan autoencoder ditumpuk yang digunakan dalam pemrosesan gambar. Sebagai contoh, proses autoencoder pertama akan belajar mengkodekan fitur yang mudah seperti sudut atap, sementara yang kedua menganalisis keluaran lapisan pertama untuk mengkodekan fitur yang kurang jelas seperti kenop pintu. Kemudian yang ketiga mengkodekan seluruh pintu dan seterusnya hingga autoencoder terakhir mengkodekan seluruh gambar menjadi kode yang cocok dengan konsep rumah. Hal ini juga dapat digunakan untuk pemodelan generatif. Misalnya, jika sebuah sistem secara manual diberi kode yang dipelajarinya untuk rumah dan terbang, sistem tersebut dapat menghasilkan gambar kucing terbang, meskipun sistem tersebut tidak pernah memproses gambar tersebut.

Jenis Auto-Encoder:

1. Denoifikasi Auto-Encoder, menggunakan input yang rusak sebagian untuk mempelajari cara memulihkan input asli yang tidak terdistorsi.
2. Auto-Encoder jarang, ini menggunakan lebih banyak lapisan pengkodean tersembunyi daripada input, dan beberapa menggunakan output dari autoencoder terakhir sebagai inputnya.

3. Variational Auto-Encoder (VAE), dalam pembelajaran representasi laten, komponen kerugian tambahan digunakan untuk mendekati distribusi posterior.
4. Auto-Encoder kontraktif (CAE) ini menggunakan *regularizer eksplisit* yang memaksa model untuk mempelajari fungsi yang kuat terhadap berbagai variasi nilai input.

Dalam pembelajaran dalam Auto-Encoder terdapat tiga pembelajaran yaitu pembelajaran tanpa pengawasan, pembelajaran dengan pengawasan dan pembelajaran penguatan. Pembelajaran yang diawasi merupakan jenis algoritma pembelajaran mesin yang paling sederhana adalah algoritma pembelajaran yang diawasi. Dalam pembelajaran terawasi, sebuah model dilatih dengan data dari kumpulan data berlabel, yang terdiri dari sekumpulan fitur, dan sebuah label. Ini biasanya tabel dengan beberapa kolom yang mewakili fitur, dan kolom terakhir untuk label. Model kemudian belajar memprediksi label untuk contoh yang tidak terlihat.

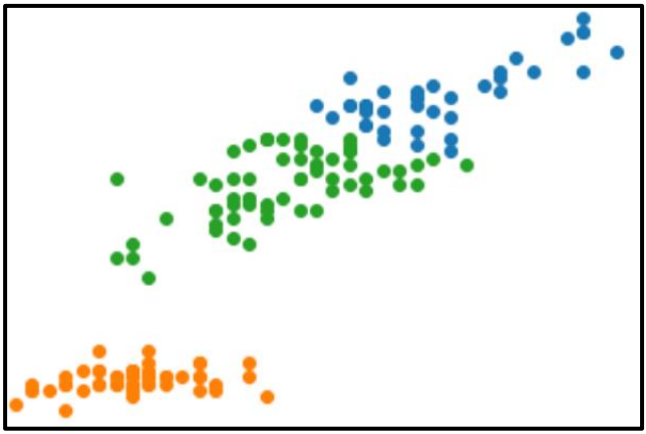
Pembelajaran tanpa pengawasan adalah sejenis pembelajaran mesin di mana model harus mencari pola dalam kumpulan data tanpa label dan dengan pengawasan manusia yang minimal. Hal ini berbeda dengan teknik pembelajaran terawasi, seperti klasifikasi atau regresi, di mana sebuah model diberikan serangkaian input pelatihan dan serangkaian observasi, dan harus mempelajari pemetaan dari input hingga observasi. Dalam pembelajaran tanpa pengawasan, hanya masukan yang tersedia, dan model harus mencari pola yang menarik dalam data. Nama lain untuk pembelajaran tanpa pengawasan adalah penemuan pengetahuan. Teknik pembelajaran tanpa pengawasan yang umum termasuk pengelompokan, dan pengurangan dimensi. Dalam pembelajaran tanpa pengawasan, kumpulan data disediakan tanpa label, dan model mempelajari properti yang berguna dari struktur

kumpulan data. Kita tidak memberi tahu model apa yang harus dipelajari, tetapi membiarkannya menemukan pola dan menarik kesimpulan dari data yang tidak berlabel. Algoritme dalam pembelajaran tanpa pengawasan lebih sulit daripada pembelajaran dengan pengawasan, karena kita memiliki sedikit atau tidak ada informasi tentang data. Tugas pembelajaran tanpa pengawasan biasanya melibatkan pengelompokan contoh serupa bersama-sama, pengurangan dimensi, dan estimasi kepadatan.

Selain pembelajaran tanpa pengawasan dan pengawasan, ada pembelajaran mesin jenis ketiga yang disebut pembelajaran penguatan. Dalam pembelajaran penguatan, seperti pembelajaran tanpa pengawasan, tidak ada data berlabel. Sebaliknya, model belajar dari waktu ke waktu dengan berinteraksi dengan lingkungannya. Misalnya, jika robot sedang belajar berjalan, ia dapat mencoba berbagai strategi untuk mengambil langkah dalam urutan yang berbeda. Jika robot berhasil berjalan lebih lama, maka hadiah diberikan ke strategi yang menghasilkan hasil tersebut. Seiring waktu, model pembelajaran penguatan belajar seperti seorang anak, dengan menyeimbangkan eksplorasi atau mencoba strategi baru dan eksploitasi atau memanfaatkan Teknik yang diketahui. Dalam pembelajaran terbimbing, model mengamati beberapa contoh variabel x, masing-masing dipasangkan dengan vektor y, dan belajar memprediksi y dari x. Garis antara pembelajaran yang diawasi dan tidak diawasi tidak selalu jelas. Dengan kata lain, mereka bukan konsep yang didefinisikan secara formal, dan banyak algoritma dapat digunakan untuk melakukan kedua tugas tersebut.

Kadang-kadang dimungkinkan untuk mengungkapkan kembali masalah pembelajaran yang diawasi sebagai masalah pembelajaran yang tidak diawasi, dan sebaliknya. Misalnya, masalah pembelajaran yang diawasi dapat diekspresikan kembali melalui teorema *Bayes* sebagai masalah pembelajaran distribusi bersama yang tidak diawasi. Meskipun demikian, konsep

pembelajaran terbimbing dan tidak terbimbing merupakan pembagian yang sangat berguna untuk dipraktikkan. Secara tradisional, masalah regresi dan klasifikasi dikategorikan dalam pembelajaran yang diawasi, sementara estimasi kepadatan, pengelompokan, dan pengurangan dimensi dikelompokkan dalam pembelajaran yang tidak diawasi. Salah satu contoh teknik pembelajaran tanpa pengawasan adalah *analisis clustering* memiliki tugas mengelompokkan satu set item sehingga setiap item ditugaskan ke grup yang sama dengan item lain yang mirip dengannya. Clustering umumnya digunakan untuk eksplorasi data dan penambangan data (Jianwu W. *et al*., 2021). Tidak ada satu algoritma pengelompokan tunggal, tetapi algoritma umum termasuk pengelompokan k-means, pengelompokan hierarkis, dan model campuran, hal ini dapat ditunjukkan pada Gambar 9.1.



Gambar 9.1. Pengelompokan data sesuai dengan grup yang sama

9.2.2 Bag-of-Words

Teknik pemrosesan bahasa alami yang mengekstraksi kata atau fitur yang digunakan dalam kalimat, dokumen, situs web, dll. dengan mengklasifikasikannya berdasarkan frekuensi penggunaan. Teknik ini juga dapat diterapkan pada pengolahan citra. *Bag of words* adalah salah satu teknik *Natural Language Processing* (NLP) dari pemodelan teks, Kita akan memahami konsepnya dengan bantuan sebuah contoh, belajar lebih banyak tentang implementasinya di Python, dan banyak lagi. Menggunakan pemrosesan bahasa alami, dengan memanfaatkan data teks yang tersedia di internet untuk menghasilkan wawasan bagi bisnis. Untuk memahami jumlah data yang sangat besar ini dan membuat wawasan darinya, kita perlu membuatnya dapat digunakan pemrosesan bahasa alami membantu kita melakukannya.

Apa itu *bag* kata-kata di NLP atau *Bag of words* adalah salah satu teknik *natural language processing* dari pemodelan teks. Dalam istilah teknis, kita dapat mengatakan bahwa ini adalah metode ekstraksi fitur dengan data teks. Pendekatan ini adalah cara yang sederhana dan fleksibel untuk mengekstraksi fitur dari dokumen. Sebuah tas kata-kata adalah representasi dari teks yang menggambarkan terjadinya kata-kata dalam sebuah dokumen. Kita hanya melacak jumlah kata dan mengabaikan detail tata bahasa dan urutan kata yang disebut *Bag of words* karena setiap informasi tentang urutan atau struktur kata dalam dokumen akan dibuang. Model ini hanya memperhatikan apakah kata-kata yang dikenal muncul di dalam dokumen atau bukan di dalam dokumen. Mengapa algoritma *Bag-of-Words* digunakan karena salah satu masalah terbesar dengan teks adalah berantakan dan tidak terstruktur, dan algoritma pembelajaran mesin lebih memilih input panjang tetap yang terstruktur dan terdefinisi dengan baik. Dengan menggunakan teknik *Bag-of-Words* dapat mengonversi teks panjang variabel menjadi panjang tetap vector, dengan tingkat

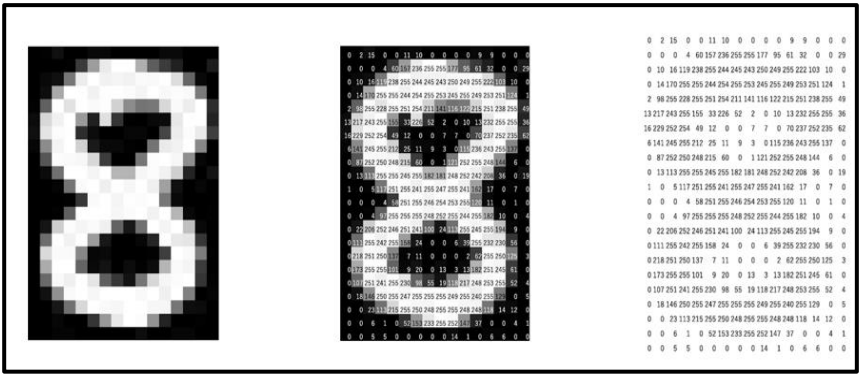
yang lebih terperinci, model pembelajaran mesin bekerja dengan data numerik daripada data tekstual. Jadi untuk lebih spesifik, dengan menggunakan teknik *Bag-of-Words* (BoW), dapat mengubah teks menjadi vektor angka yang setara.

9.2.3 Pemrosesan Gambar

Algoritma digunakan untuk mendeteksi fitur seperti bentuk, tepi, gerakan dalam gambar atau video digital. Saat ini menjadi sangat umum untuk bekerja dengan kumpulan data dari ratusan atau bahkan ribuan fitur. Jika jumlah fitur menjadi serupa atau bahkan lebih besar daripada jumlah pengamatan yang disimpan dalam kumpulan data, kemungkinan besar hal ini dapat menyebabkan model *machine learning* mengalami *overfitting*. Untuk menghindari jenis masalah ini, perlu diterapkan teknik regularisasi atau reduksi dimensi ekstraksi fitur. Dalam *machine learning*, dimensi suatu dataset sama dengan jumlah variabel yang digunakan untuk mewakilinya. Jika objek yang digunakan adalah citra maka kemungkinan bekerja dengan gambar menggunakan teknik visi komputer tidak terbatas. Tetapi satelah melihat tren dikalangan ilmuwan data baru-baru ini, ada keyakinan kuat bahwa dalam hal bekerja dengan data tidak terstruktur, terutama data gambar, model *deep learning* adalah solusinya. Teknik pembelajaran mendalam tidak diragukan lagi bekerja dengan sangat baik, tetapi apakah itu satu-satunya cara untuk bekerja dengan gambar, tidak semua memiliki sumber daya tak terbatas seperti raksasa teknologi besar seperti *Google* dan *Facebook*.

Jadi bagaimana kita bisa bekerja dengan data gambar jika tidak melalui lensa *deep learning*? image_data_*machine learning* kita dapat memanfaatkan kekuatan *machine_learning*, Benar – kita dapat menggunakan model *machine_learning* sederhana seperti pohon keputusan atau *Support Vector Machines* (SVM). Jika disediakan data dan fitur yang tepat, model *machine_learning* ini dapat bekerja dengan baik dan bahkan dapat digunakan sebagai

solusi tolak ukur. Bagaimana mesin menyimpan gambar? Mari kita mulai dengan dasar-dasarnya. Penting untuk memahami bagaimana kita dapat membaca dan menyimpan gambar di mesin kita sebelum kita melihat yang lain. contoh sederhana lihat Gambar 2. yaitu gambar angka 8. Perhatikan baik-baik gambar tersebut yang terdiri dari kotak persegi kecil yang disebut piksel. Gambar dapat dilihat sebagaimana adanya dalam bentuk visualnya sehingga dapat dengan mudah membedakan tepi dan warna untuk mengidentifikasi apa yang ada di dalam gambar tersebut. Mesin di sisi lain, berjuang untuk melakukan hal ini untuk menyimpan gambar dalam bentuk angka. Mesin menyimpan gambar dalam bentuk matriks angka (Indriyani *et al*, 2019). Ukuran matriks ini bergantung pada jumlah piksel yang kita miliki dalam gambar tertentu. Katakanlah dimensi sebuah gambar adalah 180 x 200 atau n x m. Dimensi ini pada dasarnya adalah jumlah piksel pada gambar yaitu tinggi x lebar. Angka-angka ini, atau nilai piksel, menunjukkan intensitas atau kecerahan piksel. Angka yang lebih kecil mendekati nol menunjukkan warna hitam, dan angka yang lebih besar mendekati 255 menunjukkan warna putih pada gambar 2 merupakan gambar gray dengan dimensi gambar adalah 22 x 16, yang dapat Anda verifikasi dengan menghitung jumlah piksel. Dari citra tersebut (citra warna, gray, dan biner) kita dapat menganalisanya misal dapat di segmentasi, dicari tepi objeknya, dikelompokkan nilai pikselnya dan diekstraksi fiturnya dll.



Gambar 9.2. Contoh citra gray angka delapan

9.3 Ekstraksi Fitur Lokal

Dua area utama tercakup di sini. Pendekatan tradisional bertujuan untuk mendapatkan fitur lokal dengan mengukur properti gambar tertentu. Target utamanya adalah memperkirakan kelengkungan: puncak kelengkungan lokal adalah sudut, dan menganalisis gambar berdasarkan sudutnya sangat cocok untuk gambar objek buatan. Area kedua mencakup pendekatan yang lebih modern yang meningkatkan kinerja dengan menggunakan analisis berbasis wilayah atau area. Lokasi ekstraksi pertama adalah operator berbasis kelengkungan yang lebih mapan, sebelum beralih ke analisis berbasis area atau kawasan.

9.3.1 Mendeteksi Kelengkungan Gambar (Ekstraksi Sudut)

Tepi objek merupakan fitur gambar tingkat rendah yang paling jelas terlihat oleh penglihatan manusia. Mereka mempertahankan fitur yang signifikan, jadi biasanya dapat mengenali isi gambar dari versi deteksi tepinya. Namun, ada fitur tingkat rendah lainnya yang dapat digunakan dalam visi komputer. Salah satu fitur penting adalah kelengkungan. Secara intuitif, kita dapat menganggap kelengkungan sebagai laju perubahan arah tepi.

Laju perubahan ini mencirikan titik-titik dalam kurva, titik di mana arah tepi berubah dengan cepat adalah sudut, sedangkan titik di mana ada sedikit perubahan arah tepi sesuai dengan garis lurus (Indriyani *et al*, 2020). Titik ekstrem seperti itu sangat berguna untuk deskripsi dan pencocokan bentuk, karena mewakili informasi yang signifikan dengan data yang direduksi. Kelengkungan biasanya didefinisikan dengan mempertimbangkan bentuk parametrik dari kurva planar. Itu kontur parametrik $v(t) = x(t)U_x + y(t)U_y$ menggambarkan titik-titik dalam kurva kontinu sebagai titik akhir dari vektor posisi. Di sini, nilai t menentukan parameterisasi arbitrer, vektor satuan lagi $U_x = [1,0]$ dan $U_y = [0,1]$. Perubahan vektor posisi diberikan oleh fungsi vektor tangen dari kurva $v(t)$ yaitu, $\dot{v}(t) = \dot{x}(t)U_x + \dot{y}(t)U_y$. Ekspresi vektor ini memiliki makna intuitif yang sederhana. Jika kita menganggap jejak kurva sebagai gerak titik dan t terkait dengan waktu, vektor singgung mendefinisikan gerak sesaat. Setiap saat, titik bergerak dengan kecepatan yang diberikan oleh $|\dot{v}(t)| = \sqrt{\dot{x}^2(t) + \dot{y}^2(t)}$ searah $\varphi(t) = \tan^{-1}(\dot{y}(t)/\dot{x}(t))$ Kelengkungan pada titik $v(t)$ menggambarkan perubahan arah $\varphi(t)$ sehubungan dengan perubahan panjang busur. Itu adalah,

$$k(t) = \frac{d\varphi(t)}{ds} \qquad (1)$$

di mana s adalah panjang busur, sepanjang tepi itu sendiri. Berikut adalah sudut garis singgung kurva. Yaitu, $\varphi = \theta \pm 90^o$, dimana θ arah gradien didefinisikan tanda Mx dan My dapat digunakan untuk menentukan kuadran yang sesuai untuk tepi arah, dapat ditunjukkan dalam persamaan 2.

$$\theta(x, y) = \tan^{-1}\left(\frac{My(x, y)}{Mx(x, y)}\right) \qquad (2)$$

Yaitu, jika kita menerapkan operator pendeteksi tepi ke gambar, kita memiliki nilai arah gradien untuk setiap piksel yang mewakili arah normal ke setiap titik dalam kurva. Garis singgung kurva diberikan oleh vektor ortogonal. Kelengkungan diberikan sehubungan dengan panjang busur karena kurva yang diparameterisasi oleh panjang busur mempertahankan kecepatan gerak yang konstan. Dengan demikian, kelengkungan mewakili perubahan arah untuk perpindahan konstan sepanjang kurva. Dengan mempertimbangkan aturan rantai, kita dapatkan pada persamaan 3.

$$k(t) = \frac{d\varphi(t)}{ds}\frac{dt}{ds} \tag{3}$$

Diferensial ds/dt mendefinisikan perubahan panjang busur sehubungan dengan parameter t. Jika kita kembali menganggap kurva sebagai gerakan suatu titik, diferensial ini mendefinisikan perubahan jarak seketika terhadap waktu. Artinya, kecepatan sesaat. Dengan demikian, dapat ditunjukkan pada persamaan

$$ds / dt = \left|\dot{v}(t)\right| = \sqrt{\dot{x}^2(t) + \dot{y}^2(t)} \tag{4}$$

dan

$$dt / ds = 1 \Big/ \sqrt{\dot{x}^2(t) + \dot{y}^2(t)} \tag{5}$$

engan mempertimbangkan bahwa $\varphi(t) = \tan^{-1}(\dot{y}(t) / \dot{x}(t))$ maka kelengkungan pada titik $v(t)$ pada Persamaan 2 diberikan oleh

$$k(t) = \frac{\dot{x}(t)\ddot{y}(t) - \dot{y}(t)\ddot{x}(t)}{[\dot{x}^2(t) + \dot{y}^2(t)]^{3/2}} \tag{6}$$

Hubungan ini disebut fungsi kelengkungan dan merupakan ukuran standar kelengkungan untuk kurva planar (Yang, Liu *et al*., 2021). Fitur penting dari kelengkungan adalah menghubungkan turunan dari vektor tangensial ke vektor normal.

9.3.2 Menghitung Perbedaan Arah Tepi

Mungkin cara yang lebih mudah untuk menghitung kelengkungan pada gambar digital adalah dengan mengukur perubahan sudut sepanjang jalur kurva. Pendekatan ini dipertimbangkan dalam teknik deteksi sudut awal (Bennett dan MacDonald, 1975; Groan dan Verbeek, 1978; Kitchen dan Rosenfeld, 1982) dan hanya menghitung perbedaan arah tepi antara piksel yang terhubung membentuk kurva diskrit. Artinya, ini mengaproksimasi turunan dalam Persamaan 1 sebagai selisih antara piksel tetangga. Dengan demikian, kelengkungan hanya diberikan oleh

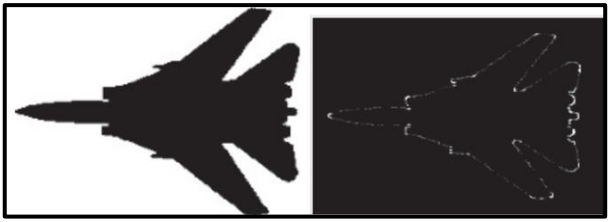
$$k(t) = \varphi_{t+1} - \varphi_{t-1} \tag{8}$$

di mana urutan $\varphi_{t-1}, \varphi_t, \varphi_{t+1}, \varphi_{t+2}$ mewakili arah gradien dari urutan piksel yang mendefinisikan segmen kurva. Arah gradien dapat diperoleh sebagai sudut yang diberikan oleh operator detektor tepi. Atau, dapat dihitung dengan mempertimbangkan posisi piksel dalam urutan. Artinya, dengan mendefinisikan $\varphi_t = (y_{t-1} - y_{t+1}) / (x_{t-1} - x_{t+1})$, di mana (x_t, y_t) menunjukkan piksel t dalam urutan. Karena titik tepi hanya ditentukan pada titik diskrit, sudut ini hanya dapat mengambil delapan nilai, sehingga kelengkungan yang dihitung sangat kasar. Hal ini dapat dihaluskan dengan mempertimbangkan perbedaan arah sudut rata-rata n piksel pada segmen kurva terdepan dan tertinggal. Itu adalah,

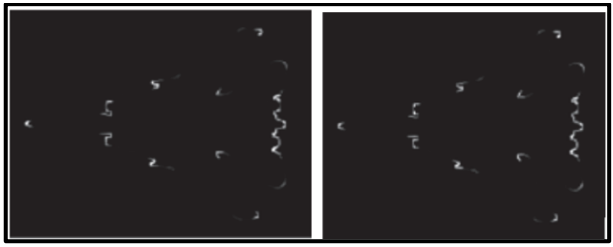
$$k_n(t) = \frac{1}{n}\sum_{i=1}^{n} \varphi_{1+i} - \frac{1}{n}\sum_{i=-n}^{-1} \varphi_{1+i} \tag{9}$$

Rata-rata juga memberikan kekebalan terhadap *noise* dan dapat diganti dengan rata-*rata threshold* jika *Gaussian smoothing* diperlukan. Jumlah piksel yang dipertimbangkan, nilai n, menentukan kompromi antara akurasi dan sensitivitas noise. Perhatikan bahwa teknik pemfilteran juga dapat digunakan untuk mengurangi efek kuantisasi ketika sudut diperoleh oleh operator

deteksi tepi. Untuk menghitung perbedaan sudut, kita perlu menentukan sisi yang terhubung. Ini bisa dengan mudah diimplementasikan dengan kode yang sudah dikembangkan untuk ambang histeresis di operator Canny edge. Untuk menghitung perbedaan titik dalam kurva, hanya perlu diatur untuk menyimpan perbedaan arah tepi antara titik-titik yang terhubung. *Gaussian operator* menunjukkan implementasi untuk deteksi kelengkungan. Pertama, tepi dan besaran ditentukan. Kelengkungan hanya terdeteksi pada titik tepi (Karthickmanoj *et al*. 2021). Karena itu, kami menerapkan penekanan maksimal. Fungsi Cont mengembalikan matriks yang berisi piksel tetangga yang terhubung dari setiap sisi. Setiap piksel tepi terhubung ke satu atau dua tetangga. Matriks Next hanya menyimpan arah piksel yang berurutan di sebuah tepi. Kami menggunakan nilai −1 untuk menunjukkan bahwa tidak ada tetangga yang terhubung. Fungsi NextPixel memperoleh posisi piksel tetangga dengan mengambil posisi piksel dan arah tetangganya. Kelengkungan dihitung sebagai perbedaan arah gradien piksel tetangga yang terhubung (Indriyani *et al*, 2021). Hasil penerapan bentuk pendeteksian kelengkungan ini pada sebuah citra ditunjukkan pada Gambar 3 (a) memuat siluet suatu objek, Gambar 4 adalah kelengkungan yang diperoleh dengan menghitung laju perubahan arah tepi. Pada gambar ini, kelengkungan hanya ditentukan pada titik-titik tepi. Di sini, dengan perumusannya, pengukuran kelengkungan hanya memberikan garis tipis perbedaan arah tepi yang dapat dilihat untuk melacak titik-titik perimeter bentuk (pada titik-titik di mana ada kelengkungan yang diukur). Titik paling terang adalah mereka dengan kelengkungan terbesar. Untuk menunjukkan hasil, kami telah menskalakan nilai kelengkungan menggunakan 256 nilai intensitas. Estimasi titik sudut dapat diperoleh dengan versi ambang yang seragamdapat dilihat pada gambari Gambar 3 (b).



Gambar 9.3.a. Citra Asli b. Hasil deteksi kelengkungan citra



Gambar 9.4. Kelengkungan perubahan arah tepi

Dapat dilihat pada Gambar 9.4, pendekatan ini tidak memberikan hasil yang dapat diandalkan. Ini pada dasarnya adalah formulasi ulang dari proses deteksi tepi orde pertama dan mengandaikan bahwa informasi sudut terletak di dalam data ambang batas (dan tidak menggunakan struktur sudut dalam deteksi). Salah satu kesulitan utama dengan pendekatan ini adalah bahwa pengukuran sudut dapat sangat dipengaruhi oleh kesalahan kuantisasi dan akurasi terbatas (Alzubaidi, *et al* 2021), sebuah faktor yang akan kembali mengganggu kita nanti ketika kita mempelajari metode untuk mendeskripsikan bentuk.

9.4 Aplikasi Ekstraksi Fitur

Kita dapat mengotomatiskan identifikasi dan ekstraksi fitur dari volume besar data teks dan membuat ringkasan semua fitur unik atau kombinasi fitur dengan solusi ekstraksi teks. Merek dan pemasaran perlu memahami kebutuhan audien target mereka. Gunakan ekstraksi teks untuk mendapatkan wawasan berharga

dan membuat produk dengan serangkaian fitur yang menarik. Sistem rekomendasi dapat mendeteksi produk lain yang berisi entitas serupa dengan mengidentifikasi entitas secara otomatis dalam spesifikasi produk. Kita dapat membuat grup fitur yang berbeda dengan mengelompokkan fitur serupa atau terkait. Mengidentifikasi fitur produk baru atau segmen pelanggan baru dan mengembangkan produk yang lebih personal dan menarik. Dengan membuatnya mudah untuk menemukan, mengatur, dan mengakses hal-hal yang relevan, dan dapat mempelajari hal-hal baru (Zheng X. *et al.* 2021). Alat untuk mengekstraksi teks meliputi:

1. Bytesview

   Solusi ekstraksi fitur canggih BytesView dapat menganalisis volume besar data teks dan mendeteksi fitur bersama dengan wawasan terkait. Tentukan sebelumnya tag untuk mengidentifikasi konten topikal, intelijen bisnis, opini pelanggan, dan tiket berulang.

2. IBM Watson

   IBM Watson adalah platform AI pilihan perusahaan. API Pemahaman Bahasa Alami Watson memberi pengembang alat dan fitur canggih untuk membuat model analisis teks pembelajaran mendalam. Watson Natural Language Classifier (untuk klasifikasi teks), Watson Tone Analyzer (untuk analisis emosi), dan Watson Personality Insights semuanya menggunakan API dalam lingkungan Watson atau untuk segmentasi pelanggan.

3. Monkeylearn

   MonkeyLearn adalah program analisis teks yang dikenal dengan kemampuan beradaptasinya. Cukup buat tag, lalu sorot bagian teks yang berbeda secara manual untuk menunjukkan konten mana yang termasuk dalam tag tersebut. Seiring waktu, perangkat lunak belajar sendiri dan dapat memproses banyak file pada saat yang sama Ini berisi

kumpulan model terlatih untuk tugas-tugas seperti analisis sentimen, ekstraksi kata kunci, deteksi urgensi, dan banyak lagi

a. API Bahasa alami untuk Google Cloud

   Menggunakan pembelajaran mesin Google, Google Cloud Natural Language API membantu bisnis memahami dan membantu memajukan informasi dalam teks. Ini pada dasarnya menawarkan dua jenis opsi: sekumpulan model terlatih untuk menganalisis sentimen, menemukan entitas, dan mengkategorikan konten, dan Cloud Auto ML, rangkaian untuk membuat model pembelajaran mesin kustom. Membuat model Anda sendiri sangatlah mudah, dan ada banyak panduan yang tersedia untuk membantu Anda menavigasi API.

b. Leksalitik

   Lexalytics adalah solusi analitik teks yang menganalisis berbagai jenis teks. Lexalytics dapat menganalisis komentar, survei, dan ulasan media sosial, serta jenis dokumen teks lainnya. Selain analisis sentimen, alat ini juga melakukan klasifikasi, ekstraksi tema, dan deteksi niat, yang memungkinkan pengguna melihat konteks lengkap.

9.5 Ekstraksi Fitur Linear dan Non-Linear

Ekstraksi Fitur dapat dibagi menjadi dua kategori besar yaitu linier dan non-linier. Salah satu contoh ekstraksi ciri linier adalah PCA (*Principal Component Analysis*). Komponen utama adalah kombinasi linier yang dinormalisasi dari fitur asli dalam kumpulan data. PCA pada dasarnya adalah metode untuk mendapatkan variabel yang diperlukan atau yang penting dari sekumpulan besar variabel yang tersedia dalam kumpulan data. PCA cenderung menggunakan transformasi ortogonal untuk mengubah data menjadi ruang berdimensi lebih rendah yang pada gilirannya memaksimalkan varians data.

PCA dapat digunakan untuk deteksi anomali dan outlier karena ini dianggap sebagai data derau atau tidak relevan di seluruh kumpulan data.
Langkah-langkah yang diikuti dalam membangun PCA dari awal adalah:

a. Pertama, standarisasi data
b. Setelah itu, hitung matriks Kovarian
c. Kemudian, hitung Eigenvector & Eigenvalues untuk Covariance-matrix.
d. Susun semua Nilai Eigen dalam urutan menurun.
e. Normalisasikan nilai Eigen yang diurutkan.
f. Tumpuk nilai Normalized_Eigen secara horizontal

PCA gagal saat datanya non-linier yang dapat dianggap sebagai salah satu kelemahan terbesar PCA. Di sinilah Kernel-PCA memainkan perannya. Kernel-PCA mirip dengan SVM (*support vector machine)* karena keduanya mengimplementasikan Kernel–Trick untuk mengonversi data non-linier menjadi data dimensi yang lebih tinggi hingga data dapat dipisahkan. Pendekatan Non-Linear dapat digunakan dalam kasus pengenalan wajah untuk mengekstraksi fitur pada kumpulan data besar. Ekstraksi Fitur juga memberi kita visualisasi yang jelas dan improvisasi dari data yang ada dalam kumpulan data karena hanya data penting dan diperlukan yang telah diekstraksi Ekstraksi Fitur membantu dalam melatih model dengan cara yang lebih efisien yang pada dasarnya mempercepat keseluruhan proses. PCA gagal saat datanya non-linier yang dapat dianggap sebagai salah satu kelemahan terbesar PCA. Di sinilah Kernel-PCA memainkan perannya. Kernel-PCA mirip dengan SVM karena keduanya mengimplementasikan Kernel–Trick untuk mengonversi data non-linear menjadi data berdimensi lebih tinggi hingga saat data dapat dipisahkan. Pendekatan Non-Linear dapat digunakan dalam kasus pengenalan wajah untuk mengekstraksi fitur pada kumpulan data besar. Jadi

Ekstraksi Fitur memiliki penggunaan yang beragam di sebagian besar domain. Ini muncul pada tahap awal proyek apa pun yang menggunakan kumpulan data besar (Tamililakkiya *et al.*, 2011). Jadi seluruh prosedur ekstraksi fitur harus dijalankan dan dievaluasi dengan hati-hati untuk mendapatkan hasil yang dioptimalkan dengan akurasi yang lebih tinggi yang pada gilirannya akan membantu kita untuk mendapatkan wawasan yang lebih baik tentang hubungan antara variabel yang ada dalam dataset dan untuk selanjutnya merencanakan tahap berikutnya.

DAFTAR PUSTAKA

Alzubaidi M., *et al.,* 2021. Comprehensive and Comparative Global and Local Feature Extraction Framework for Lung Cancer Detection Using CT Scan Images. *IEEE Access.* Digital Object Identifier 10.1109/ACCESS.2021.3129597.

Devulapalli S., *et al*., 2021. Experimental evaluation of unsupervised image retrieval application using hybrid feature extraction by integrating deep learning and handcrafted techniques. *Materials Today: Proceedings Available online* June 2021. https://doi.org/10.1016/j.matpr.2021.04.326.

Huang P. *et al.,* 2021. Double L2, p-norm based PCA for feature extraction. *Information Sciences* v. 573, September 2021, P. 345-359, https://doi.org/10.1016/j.ins.2021.05.079.

Indriyani T., Utoyo M.I, Rulaningtyas R. 2019. Comparison of Image Smoothing Methods on Potholes Road Images. *Journal of Physics*: Conference Series. ICComSET IOP Publishin. v. 1477.

Indriyani T., Utoyo I. Rulaningtyas R. 2020. Comparison of Image Edge Detection Methods on Potholes Road Images. *Journal of Physics*: Conference Series, 1613(1).

Indriyani T., Utoyo I., Rulaningtyas R. 2021. A New watershed Algorithm for Pothole Image Segmentation, *Studies in Informatics and Control*, 30(3) 131-139.

Jianwu W., *et al*. 2021 Joint feature extraction and classification in a unified framework for cost-sensitive face recognition', *Pattern Recognition* 115 (2021) 107927.

Karthickmanoj R., *et al*. 2021 A novel pixel replacement-based segmentation and double feature extraction techniques for efficient classification of plant leaf diseases Elsevier Science'. *Materials Today: Proceedings* v. 47, part 9, 2021, p. 2048-2052.

Nixon M. and Aguado A. 2018. Feature Extraction and Image Processing', *Academic Press is an imprint of Elsevier*. Second edition.

Tamililakkiya V., *et al*. 2011.Linear and Non-Linear Feature Extraction Algorithms For Lunar Images'. *Signal & Image Processing: An International Journal (SIPIJ)* v.2, No.4. doi: 10.5121/sipij.2011.2414.

Yang Y. , Liu Y., *et al*. 2021. Improvement of feature extraction and intelligent identification method for the edge coherent mode in EAST. *Fusion Engineering and Design,* v. 171, October 2021, 112552.
https://doi.org/10.1016/j.fusengdes.2021.112552 .

Zheng X., *et al.* 2021.A Text-Image Matching Network Integrating Multi-Stage Feature Extraction with Multi-Scale Metrics, *Neurocomputing* v. 465, 20 November 2021, p. 540-548.

BAB 10
DISCRETIZATION METHOD

Oleh Leo Willyanto Santoso

10.1 Pendahuluan

Discretization/diskritisasi adalah prosedur pemrosesan data yang mengubah data kuantitatif menjadi data kualitatif. Aplikasi Data Mining seringkali melibatkan data kuantitatif. Namun, terdapat banyak algoritma pembelajaran yang terutama berorientasi untuk menangani data kualitatif (Kerber, 1992, Dougherty et al., 1995, Kohavi dan Sahami, 1996). Bahkan untuk algoritma yang secara langsung dapat menangani data kuantitatif, pembelajaran seringkali kurang efisien dan kurang efektif (Catlett, 1991, Santoso, 2018, Richeldi dan Rossotto, 1995, Frank dan Witten, 1999). Karenanya diskritisasi telah lama menjadi topik aktif dalam *Data Mining and Knowledge Discovery*. Banyak algoritma diskritisasi telah diusulkan. Evaluasi algoritma ini sering menunjukkan bahwa diskritisasi membantu meningkatkan kinerja pembelajaran dan membantu memahami hasil pembelajaran.

10.2 Kuantitatif vs. Kualitatif

Diskritisasi mengubah satu jenis data ke jenis lainnya. Dalam sejumlah besar literatur yang membahas tentang diskritisasi, ada banyak variasi dalam terminologi yang digunakan untuk mendeskripsikan kedua jenis data ini, termasuk 'kuantitatif' vs. 'kualitatif', 'kontinu' vs. 'diskrit', 'ordinal' vs. 'nominal', dan 'numerik' vs. 'kategorikal'. Penting untuk memperjelas perbedaan diantara berbagai istilah dan karenanya memilih terminologi yang paling cocok untuk diskritisasi. Kami mengadopsi terminologi

statistik (Bluman, 1992, Samuels dan Witmer, 1999), yang menyediakan dua cara paralel untuk mengklasifikasikan data ke dalam jenis yang berbeda. Data dapat diklasifikasikan menjadi kualitatif atau kuantitatif. Data juga dapat diklasifikasikan ke dalam berbagai tingkat skala pengukuran.

Data kualitatif, juga sering disebut sebagai data kategorikal, adalah data yang dapat ditempatkan ke dalam kategori yang berbeda. Data kualitatif terkadang dapat disusun dalam urutan yang bermakna, tetapi tidak ada operasi aritmatika yang dapat diterapkan padanya. Contoh data kualitatif adalah: golongan darah seseorang: A, B, AB, O; dan penilaian tugas: gagal, lulus, baik, sangat baik. Data kuantitatif bersifat numerik, sehingga dapat diberi peringkat secara berurutan. Operasi aritmatika dapat kita lakukan pada data jenis ini. Data kuantitatif dapat diklasifikasikan lebih lanjut menjadi dua kelompok, yaitu: diskrit atau kontinu.

Data diskrit diasumsikan sebagai nilai yang dapat dihitung. Data diskrit tidak dapat diasumsikan pada semua nilai pada garis bilangan dalam rentang nilainya. Contohnya adalah: jumlah anak dalam sebuah keluarga. Contoh lainnya adalah dadu. Saat kita melempar dadu tidak akan muncul angka pecahan, seperti 1.5, 2.5, dan lain sebagainya. Angka yang muncul di dalam dadu hanya bilangan bulat saja, yaitu: 1, 2, 3, 4, 5 dan 6. Data kontinu dapat diasumsikan dengan semua nilai pada garis bilangan dalam rentang nilainya. Nilai diperoleh dengan mengukur. Contohnya adalah: suhu.

Selain diklasifikasikan menjadi kualitatif atau kuantitatif, data juga dapat diklasifikasikan berdasarkan bagaimana mereka dikategorikan, dihitung atau diukur. Jenis klasifikasi ini menggunakan skala pengukuran, dan empat tingkatan skala yang umum adalah: nominal, ordinal, interval, dan rasio.

Tingkat nominal skala pengukuran yang mengklasifikasikan data ke dalam kategori yang saling eksklusif (tidak tumpang tindih), lengkap di mana tidak ada urutan atau peringkat yang

berarti yang dapat dikenakan pada data. Contohnya adalah: golongan darah seseorang: A, B, AB, O

Tingkat ordinal skala pengukuran yang mengklasifikasikan data ke dalam kategori yang dapat diberi peringkat. Namun, perbedaan antara peringkat tidak dapat dihitung dengan aritmatika. Contohnya adalah: penilaian kuliah: gagal, lulus, baik, sangat baik. Penting untuk mengatakan bahwa evaluasi penilaian kuliah peringkat lulus lebih tinggi daripada yang gagal. Sebaliknya, kita tidak bisa mengatakan bahwa golongan darah A lebih tinggi daripada golongan darah B.

Tingkat interval data menandakan peringkat pada skala pengukuran, dan perbedaan antara satuan ukuran dapat dihitung dengan aritmatika. Contohnya adalah: suhu Celcius, dimana memiliki perbedaan yang berarti satu derajat antara setiap unit. Tapi 10^0 Celcius bukan berarti tidak ada panas. Penting untuk mengatakan bahwa 74 derajat adalah dua derajat lebih tinggi dari 72 derajat.

Tingkat rasio skala pengukuran memiliki semua karakteristik pengukuran interval, dan terdapat nol yang sama dengan nol aritmatika, berarti 'nihil' atau 'tidak ada'. Contoh nya adalah: jumlah anak dalam sebuah keluarga. Penting untuk mengatakan bahwa keluarga X memiliki anak dua kali lebih banyak daripada keluarga Y. Sebaliknya, dengan cara yang sama tidak bisa dikatakan bahwa 100^0 Celcius dua kali lebih panas dari 50^0 Celcius.

Level nominal adalah level skala pengukuran yang paling rendah atau paling tidak kuat, karena memberikan informasi paling sedikit tentang data. Tingkat ordinal lebih tinggi, diikuti oleh tingkat interval. Tingkat rasio adalah yang tertinggi. Setiap konversi data dari tingkat skala pengukuran yang lebih tinggi ke tingkat skala pengukuran yang lebih rendah akan kehilangan informasi.

11.3 Metode Diskritisasi

Dari studi literatur, ada beragam pengelompokan untuk mengklasifikasikan metode diskritisasi. Pengelompokan yang berbeda menekankan aspek yang berbeda dari perbedaan antara metode diskritisasi. Biasanya, metode diskritisasi dapat berupa primer atau komposit. Metode primer menyelesaikan diskritisasi tanpa mengacu pada metode diskritisasi lainnya. Metode komposit dibangun di atas beberapa metode primer.

Metode primer dapat diklasifikasikan sebagai berikut ini.

1. **Supervised** vs. **Unsupervised** (Dougherty et al., 1995). Metode yang menggunakan informasi *class* dari instance pelatihan untuk memilih titik potong diskritisasi disebut sebagai metode *supervised*. Sebaliknya metode yang tidak menggunakan informasi *class*, disebut metode *unsupervised*.

   Selanjutnya, diskritisasi *supervised* dapat digolongkan menjadi: berbasis kesalahan (*error-based*), berbasis entropi (*enthropy-based*), atau berbasis statistik (*statistics-based*) berdasarkan apakah interval dipilih menggunakan metrik berdasarkan kesalahan pada data pelatihan, entropi interval, atau beberapa ukuran statistik.
2. **Parametrik** vs. **Non-parametrik**. Diskritisasi parametrik memerlukan input dari pengguna, seperti jumlah maksimum interval diskrit. Diskritisasi non parametrik hanya menggunakan informasi dari data dan tidak memerlukan input dari pengguna.
3. **Univariat** vs. **Multivariat** (Bay, 2000). Metode yang mendiskritisasi setiap atribut dalam isolasi bersifat univariat. Metode yang mempertimbangkan hubungan antar atribut selama diskritisasi bersifat multivariat.
4. **Disjoint** vs **Non-disjoint** (Yang dan Webb, 2002). Metode disjoint mendiskritisasi rentang nilai atribut di bawah diskritisasi ke dalam interval disjoin. Tidak ada interval

yang tumpang tindih. Metode non-disjoint mendiskritkan rentang nilai ke dalam interval yang dapat tumpang tindih.

5. ***Eager*** vs ***Lazy*** (Hsu et al., 2000, Hsu et al., 2003). Metode *eager* melakukan diskritisasi sebelum waktu klasifikasi. Metode *lazy* melakukan diskritisasi selama waktu klasifikasi.
6. **Ordinal** vs **Nominal**. Diskritisasi ordinal mengubah data kuantitatif menjadi data kualitatif ordinal. Hal ini bertujuan untuk memanfaatkan informasi pengurutan yang tersirat dalam atribut kuantitatif, agar tidak membuat nilai 1 dan 2 berbeda dengan nilai 1 dan 10. Diskritisasi nominal mengubah data kuantitatif menjadi data kualitatif nominal.
7. **Fuzzy** vs **Non-fuzzy** (Wu, 1995, Wu, 1999, Ishibuchi et al., 2001). Metode diskritisasi fuzzy pertama-tama akan mendiskritisasi nilai atribut kuantitatif ke dalam interval, kemudian menempatkan semacam fungsi keanggotaan (*membership function*) pada setiap titik potong sebagai batas fuzzy.

Fungsi keanggotaan mengukur derajat setiap nilai yang termasuk dalam setiap interval. Dengan batas fuzzy ini, sebuah nilai dapat didiskritisasi menjadi beberapa interval yang berbeda pada waktu yang sama, dengan derajat yang berbeda-beda. Sebaliknya pada metode diskritisasi non-fuzzy tanpa digunakan fungsi keanggotaan apa pun. Metode komposit pertama-tama memilih beberapa metode diskritisasi primer untuk membentuk titik potong awal. Kemudian metode tersebut fokus pada bagaimana menyesuaikan titik potong awal ini untuk mencapai tujuan tertentu.

10.4 Diskritisasi dengan lebar sama (*equal-width*), frekuensi sama (*equal-frequency*), dan frekuensi tetap (*fixed-frequency*)

Saat melakukan diskritisasi atribut kuantitatif, diskritisasi lebar sama (EWD) (Catlett, 1991, Kerber, 1992, Dougherty et al., 1995) menentukan k dan jumlah interval. Kemudian membagi garis bilangan antara v_{min} dan v_{max} menjadi interval k dengan lebar yang sama, di mana v_{min} adalah nilai observasi minimum, v_{max} adalah nilai observasi maksimum. Jadi interval memiliki lebar $w = (v_{max} - v_{min})/k$ dan titik potong berada di $v_{min} + w, v_{min} + 2w, \cdots, v_{min} + (k-1)w$.

Ketika mendiskritisasi atribut kuantitatif, diskritisasi frekuensi yang sama (EFD) menentukan k, jumlah interval. Kemudian membagi nilai yang diurutkan menjadi interval k sehingga setiap interval berisi kira-kira jumlah contoh pelatihan yang sama. Misalkan ada n instance pelatihan, setiap interval kemudian berisi n/k instance pelatihan dengan nilai yang berdekatan (mungkin identik). Perhatikan bahwa instance pelatihan dengan nilai identik harus ditempatkan dalam interval yang sama. Akibatnya tidak selalu mungkin untuk menghasilkan k interval frekuensi yang sama.

Saat mendiskritisasi atribut kuantitatif, diskritisasi frekuensi tetap (FFD) (Yang dan Webb, 2004) menentukan k sebagai frekuensi interval. Kemudian mendiskritisasi nilai yang diurutkan ke dalam interval sehingga setiap interval memiliki kira-kira jumlah k yang sama dari instance pelatihan dengan nilai yang berdekatan (mungkin identik). Ada baiknya membandingkan EFD dan FFD, keduanya membentuk interval dengan frekuensi yang sama.

10.5 Diskritisasi Iterative Dichotomiser 3 (ID3)

ID3 merupakan contoh diskritisasi lokal. ID3 (Quinlan, 1986) merupakan program pembelajaran induktif yang menyusun aturan-aturan klasifikasi dalam bentuk pohon keputusan. ID3 menggunakan diskritisasi lokal untuk menangani atribut kuantitatif. Untuk setiap atribut kuantitatif, ID3 membagi nilai yang diurutkan menjadi dua interval dengan semua cara yang memungkinkan. Untuk setiap divisi, *information gain* yang dihasilkan dari data dihitung. Atribut yang memperoleh *information gain* maksimum dipilih menjadi simpul pohon saat ini. Kemudian, data dibagi menjadi himpunan bagian yang sesuai dengan dua interval nilainya. Di setiap subset, proses yang sama dilakukan secara rekursif untuk menumbuhkan pohon keputusan. Atribut yang sama dapat didiskritisasi secara berbeda jika muncul di cabang pohon keputusan yang berbeda.

10.6 Diskritisasi Fuzzy

Diskritisasi fuzzy (FD) (Ishibuchi et al., 2001) digunakan untuk menghasilkan aturan asosiasi linguistik, di mana banyak istilah linguistik, seperti 'pendek' dan 'tinggi', tidak dapat secara tepat diwakili oleh interval dengan titik potong tertentu. Oleh karena itu digunakan fungsi keanggotaan (*membership function*), seperti pada contoh berikut, sehingga tinggi 150 milimeter adalah 0 derajat untuk menunjukkan 'tinggi'; tinggi 175 milimeter adalah 0,5 derajat untuk menunjukkan 'tinggi' dan tinggi 190 milimeter adalah 1,0 derajat untuk menunjukkan 'tinggi'.

$$Mem_{tall}(x) = \begin{cases} 0, & \text{if x <= 170;} \\ (x-170)/10 & \text{if 170<x<180;} \\ 1, & \text{if x>=180.} \end{cases}$$

FD menggunakan *domain knowledge* untuk mendefinisikan fungsi keanggotaan linguistiknya. Ketika berurusan dengan data tanpa pengetahuan domain seperti itu, batas fuzzy masih dapat diatur dengan fungsi yang biasa digunakan seperti linier, polinomial dan arctan, untuk mengaburkan batas yang tajam (Wu, 1999).

DAFTAR PUSTAKA

Bay, S. D. 2000. Multivariate discretization of continuous variables for set mining. In Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 315–319.

Bluman, A. G. 1992. Elementary Statistics, A Step By Step Approach. Wm.C.Brown Publishers. page5-8.

Catlett, J. (1991). On changing continuous attributes into ordered discrete attributes. In Proceedings of the European Working Session on Learning, pages 164–178.

Dougherty, J., Kohavi, R., and Sahami, M. 1995. Supervised and unsupervised discretization of continuous features. In Proceedings of the 12th International Conference on Machine Learning, pages 194–202.

Frank, E. and Witten, I. H. 1999. Making better use of global discretization. In Proceedings of the 16th International Conference on Machine Learning, pages 115–123. Morgan Kaufmann Publishers.

Hsu, C.-N., Huang, H.-J., and Wong, T.-T. 2000. Why discretization works for naïve Bayesian classifiers. In Proceedings of the 17th International Conference on Machine Learning, pages 309–406.

Hsu, C.-N., Huang, H.-J., and Wong, T.-T. 2003. Implications of the Dirichlet assumption for discretization of continuous variables in naive Bayesian classifiers. Machine Learning, 53(3):235–263.

Ishibuchi, H., Yamamoto, T., and Nakashima, T. 2001. Fuzzy Data Mining: Effect of fuzzy discretization. In The 2001 IEEE International Conference on Data Mining.

Kerber, R. 1992. Chimerge: Discretization for numeric attributes. In National Conference on Artificial Intelligence, pages 123–128. AAAI Press.

Kohavi, R. and Sahami, M. 1996. Error-based and entropy-based discretization of continuous features. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pages 114–119.

Quinlan, J. R. 1986. Induction of decision trees. Machine Learning, 1:81–106.

Richeldi, M. and Rossotto, M. 1995. Class-driven statistical discretization of continuous attributes (extended abstract). In European Conference on Machine Learning, 335-338. Springer.

Samuels, M. L. and Witmer, J. A. 1999. Statistics For The Life Sciences, Second Edition. Prentice-Hall. page10-11.

Santoso, Leo Willyanto and Yulia. 2018 Predicting Student Performance Using Data Mining. In: 5th International Conference on Communication and Computer Engineering, 19-07-2018 - 19-07-2018, Malacca - Malaysia.

Wu, X. 1995. Knowledge Acquisition from Databases. Ablex Publishing Corp. Chapter 6.

Wu, X. 1999. Fuzzy interpretation of discretized intervals. IEEE Transactions on Fuzzy Systems, 7(6):753–759.

Yang, Y. and Webb, G. I. 2002. Non-disjoint discretization for naive-Bayes classifiers. In Proceedings of the 19th International Conference on Machine Learning, pages 666–673.

Yang, Y. and Webb, G. I. 2004. Discretization for naive-Bayes learning: Managing discretization bias and variance. Submitted for publication.

BIODATA PENULIS

Drs. Amna, M.IT.
Dosen Teknik Informatika
Universitas Gajah Putih

Penulis merupakan anak dari Bapak Abdurrahman AS dan ibu Fatimah dan menikah dengan Dr. Anna Permatasari Kamarudin, S.Tp, MBA sejak tahun 2000 dan dikaruniakan dengan 3 putri diberi nama Adiba Imani Salsabila, Najla Raihana Kamila dan Khayla Rahadatul Fakhira. Penulis memulai pendidikan dasar dan menengah di kota kelahirannya Takengon Aceh Tengah dan kemudian melanjutkan pendidikan tinggi di bidang matematika di USK Banda Aceh pada tahun 1988 dan memperoleh gelar sarjana pada 1992. Tahun 1993 melanjutkan program Pra S2 Matematika di ITB Bandung. Program S2 pula di tempuh pada Mathematics Department Fakulti Sains matematik dan komputer Universiti Kebangsaan Malaysia (UKM). Akhir tahun 1994 menukar jurusan ke Sains Komputer seiring dengan terbentuknya Fakulti Teknologi dan Sains Maklumat UKM. Selama melanjutkan studi S2 sangat aktif sebagai Tutor sambilan untuk mata kuliah matematik, statistik dan ilmu komputer, juga sebagai Asisten Research untuk bidang ilmu komputer dan rekayasa perangkat lunak. Setelah menamatkan S2 tahun 1997 dunia Dosenpun digeluti dengan

memulai sebagai Dosen program degree sains sistem di kolej Politek MARA Bangi untuk franchise program UKM-KYPM. Kemudian sebagai dosen D3 sains komputer prancise dengan UiTM Shah Alam, UPM Serdang juga program degree software engineering dengan MMU Cyberjaya di KYPM Kuala Lumpur. Tahun 2001 kembali ke Kampus UKM Bangi sebagai dosen sains sistem. Tahun 2003 bergabung dengan Kolej SAL Kuala Lumpur sebagai dosen program prancise degree in software engineering dan degree in networking dengan Edith Cowan University Western Australia. Tahun 2004 bergabung dengan Universiti Tenaga Nasional (UNITEN) sebagai asisten research bidang parallel processing dan dosen matematik di college Of engineering sambil melanjutkan studi S3. Tahun 2008 bergabung dengan College of IT UNITEN sebagai dosen system and network departement. Tahun 2010 hingga sekarang bergabung dengan Universitas Gajah Putih sebagai dosen teknik informatika. Banyak jurnal dan proseding baik internasional dan nasional yang telah di hasilkan dalam jangka waktu tersebut. Juga mengisi berbagai macam konferensi, simposium, dan webinar. Sejak 2017 hingga sekarang sebagai dosen tamu prodi teknik informatika di STTP Paya kumbuh. Empat book chapter yang telah di hasilkan: *The art of digital marketing* strategi pemasaran generasi milanial, *Fintech innovation essence, posisition & strategy*, Sistem operasi dan Tahapan rekayasa perangkat lunak. Hal terbaru dari riset pada 27 Desember 2021 mendapat tempat harapan 1 lomba Innovasi LISIK 2021menerusi judul riset Kamus Bahasa Gayo Berbasis Android.

Email Penulis: amnaa98@hotmail.com

BIODATA PENULIS

Wahyuddin S, S.Kom., M.Kom
Dosen STMIK Amika Soppeng

Wahyuddin S. was born at Malaka-Bone-Sulawesi Selatan in 1992. In 2011 he attended STMIK Dipanegara Makassar and was completed in 2015. He was completed after attending 7 semesters and active on an XPcom (Extreme Programmer Computer) campus organization. He was also active as a lecturer assistant for three semesters and taught several courses on programming. He continued his Master of Information systems at UNIKOM Bandung in 2016 and was completed in April 2019. He worked as a lecturer at a campus (STMIK AMIKA Soppeng) 2019 to present and also a Freelance Web Programmer. Has competence in the field of software engineer, application developer, multimedia, web developer, network security, and data analyst.

BIODATA PENULIS

I Gede Iwan Sudipa, S.Kom., M.Cs.
Dosen Program Studi Teknik Informatika
Fakultas Teknologi dan Informasi
Institut Bisnis dan Teknologi Indonesia (INSTIKI)

Penulis lahir di Singaraja, Bali. Penulis adalah dosen tetap pada Program Studi Dosen Program Studi Teknik Informatika, Fakultas Teknologi dan Informasi, Institut Bisnis dan Teknologi Indonesia (INSTIKI. Penulis menyelesaikan pendidikan S1 pada Jurusan Teknik Informatika, STMIK AKAKOM Yogyakarta dan melanjutkan S2 pada Jurusan Ilmu Komputer, Universitas Gadjah Mada. Penulis mengampu mata kuliah Algoritma dan Pemrograman, Sistem Pendukung Keputusan, Analisa dan Desain Sistem Informasi, Obyek Oriented Analysis Design, Basis Data, dan lainnya Penulis juga aktif dalam menulis karya tulis ilmiah pada scope data mining, sistem pendukung keputusan, data analisis. Email penulis : iwansudipa@instiki.ac.id

BIODATA PENULIS

Tri Andi E. Putra, S.Kom, M.Kom, CPS, C.DMS.
Dosen Program Studi Bisnis Digital Fakultas Ekonomi dan Bisnis
Universitas Fort De Kock Bukittinggi

Penulis lahir di Desa Pasar Bantal, Mukomuko, Bengkulu, pada tanggal 13 Juni 1994, Penulis adalah dosen tetap di Universitas Fort De Kock Bukittinggi pada Program Studi Bisnis Digital Fakultas Ekonomi dan Bisnis. Menyelesaikan Pendidikan S1 Sistem Informasi (2017), Magister Komputer (2019) di Universitas Putra Indonesia YPTK Padang.

Minat riset dan keahlian penulis ialah dibidang Sistem Pakar, Image Processing, Digital Marketing, Sistem Analis, penulis, telah menerbitkan 3 buku yaitu Buku Sistem Informasi Manajemen Bisnis, Pengantar Teknologi Informasi, Pengantar Bisnis, Manajemen Pemasaran Jasa. Penulis juga mempunyai Certified Public Speaking (CPS) dan Certified Digital Marketing Strategy (C.DMS) yang resmi dari Lembaga Kursus dan Pelatihan (LKP).

BIODATA PENULIS

Ahmad Jurnaidi Wahidin, M.Kom.
Dosen Prodi Teknologi Informasi
Universitas Bina Sarana Informatika Jakarta

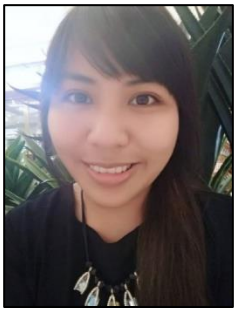
Penulis merupakan dosen dari kampus ranking ke-66 dari 100 kampus terbaik di Indonesia versi UniRank 2022, menyelesaikan pendidikan S2 di Universitas Budi Luhur pada 2018, mulai aktif mengajar pada usia 23 tahun di kampus STMIK Mahakarya (sekarang: Universitas Mahakarya Asia). Dan pada tahun 2020 mendapatkan Nomor Induk Dosen Nasional (NIDN) dari kampus Universitas Bina Sarana Informatika, Jakarta pada prodi Teknologi Informasi dan aktif mengajar sampai sekarang. Tergabung dalam Perkumpulan Dosen Peneliti Indonesia dan Asosiasi Kolaborasi Dosen Lintas Negara.

BIODATA PENULIS

Wara Alfa Syukrilla, M.Sc.
Dosen Statistika
Fakultas Psikologi UIN Syarif Hidayatullah Jakarta

Penulis adalah dosen tetap statistika pada Fakultas Psikologi UIN Syarif Hidayatullah Jakarta. Penulis menyelesaikan pendidikan S1 pada Jurusan Statistika di Universitas Brawijaya dan melanjutkan S2 Master of Statistics di Universiteit Hasselt, Belgia dengan beasiswa penuh dari pemerintah Belgia yaitu beasiswa VLIR UOS. Penulis menekuni bidang statisika, machine learning, spatial statistics, psychometrics, dan biostatistics. Saran yang membangun terkait tulisan ini dapat dikomunikasikan dengan penulis di email wara.alfa@uinjkt.ac.id .

BIODATA PENULIS



Anindya Khrisna Wardhani, S.Kom., M.Kom.,
Dosen Program Studi Rekam Medis dan Informasi Kesehatan
Politeknik Rukun Abdi Luhur, Kudus, Jawa Tengah

Penulis lahir di Kudus, Jawa Tengah pada tanggal 15 Agustus 1994. Saat ini penulis merupakan pengajar dan peneliti pada Program Studi Rekam Medis dan Informasi Kesehatan pada Politeknik Rukun Abdi Luhur, Kudus, Jawa Tengah. Penulis menyelesaikan Program Sarjana Teknik Informatika di Fakultas Ilmu Komputer Universitas Dian Nuswantoro (Udinus) Semarang pada tahun 2016. Semangat untuk terus belajar membuat penulis terpilih sebagai penerima Beasiswa Unggulan Masyarakat Berprestasi oleh Kementerian Pendidikan dan kebudayaan RI pada Program Magister Sistem Informasi di Universitas Diponegoro (Undip) Semarang dan selesai pada tahun 2018.

BIODATA PENULIS

Nono Heryana, M. Kom.
Dosen Program Studi Sistem Informasi
Fakultas Ilmu Komputer Universitas Singaperbangsa Karawang

Penulis lahir di Lampung Barat tanggal 03 Maret 1989. Penulis adalah dosen tetap pada Program Studi Sistem Informasi Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang. Menyelesaikan pendidikan S1 pada Jurusan Teknik Informatika, Universitas Singaperbangsa Karawang dan melanjutkan S2 pada Jurusan Ilmu Komputer, Universitas Budi Luhur. Penulis merupakan dosen tetap di Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang sejak tahun 2016 sampai dengan saat ini.

BIODATA PENULIS

Tutuk Indriyani
Dosen Teknik Informatika bidang Kecerdasan Buatan
Institut Teknologi Adhi Tama Surabaya

Penulis lahir di Bojonegoro pada tanggal 10 Agustus 1977. Penulis merupakan dosen tetap di Institut Teknologi Adhi Tama Surabaya Program Studi Teknik Informatika bidang Kecerdasan Buatan. Penulis telah menyelesaikan pendidikan S1 Teknik Informatika di Institut Teknologi Adhitama Surabaya (ITATS), S2 Teknik Informatika di Institut Teknologi Surabaya (ITS) dan S3 Fakultas Sains dan Teknologi di Universitas Airlangga Surabaya (UNAIR)

BIODATA PENULIS

Leo Willyanto Santoso
Dosen Program Studi Informatika
Fakultas Teknologi Industri Universitas Kristen Petra

Penulis merupakan dosen tetap di Program Studi Informatika Universitas Kristen Petra Surabaya. Penulis telah menyelesaikan pendidikan S2 di University of Melbourne, Australia.