DATA MINING TERAPAN

Konsep, Algoritma, dan Implementasi Berbasis Java



Tim Penulis:
Dody | Riri Fajriah | Haryanto | Nelis Febriani
Tb. Zaenal Muttaqin | Meta Amalya Dewi | Oleh Soleh
Alfredo Gormantara | Sucianna Ghadati Rabiha
Fandi Ali Mustika | Diva Syabina Putri | Dewi Immaniar Desrianti
Ali Herdian | Sekar Ayu Wijaya Kusumaningrum
Muhamad Yusup | Dedeh Supriyanti

Editor: Ajay Supriadi

DATA MINING TERAPAN

Konsep, Algoritma, dan Implementasi Berbasis Java

Dody
Riri Fajriah
Haryanto
Nelis Febriani
Tb. Zaenal Muttaqin
Meta Amalya Dewi
Oleh Soleh
Alfredo Gormantara
Sucianna Ghadati Rabiha
Fandi Ali Mustika
Diva Syabina Putri
Dewi Immaniar Desrianti
Ali Herdian
Sekar Ayu Wijaya Kusumaningrum
Muhamad Yusup
Dedeh Supriyanti

Editor: **Ajay Supriadi, M.Kom.**

sada
kurnia pustaka

DATA MINING TERAPAN

Konsep, Algoritma, dan Implementasi Berbasis Java

Tim Penulis:
Dody
Riri Fajriah
Haryanto
Nelis Febriani
Tb. Zaenal Muttaqin
Meta Amalya Dewi
Oleh Soleh
Alfredo Gormantara
Sucianna Ghadati Rabiha
Fandi Ali Mustika
Diva Syabina Putri
Dewi Immaniar Desrianti
Ali Herdian
Sekar Ayu Wijaya Kusumaningrum
Muhamad Yusup
Dedeh Supriyanti

Editor : Ajay Supriadi, M.Kom.
Tata Letak : Lilis Khalisatul Karimah, S.H.
Desain Cover : Asep Nugraha, S.Hum.
Ukuran : UNESCO 15,5 x 23 cm
Halaman : x, 272
ISBN : 978-634-7949-15-8
Terbit Pada : November 2026
Anggota IKAPI : No. 073/BANTEN/2023

Hak Cipta 2026 @ Sada Kurnia Pustaka dan Penulis
Hak cipta dilindungi undang-undang dilarang memperbanyak karya tulis ini dalam bentuk dan dengan cara apapun tanpa izin tertulis dari penerbit dan penulis.

PENERBIT PT SADA KURNIA PUSTAKA
Jl. Kramat, Panenjoan Kec. Carenang, Kab. Serang – Banten, 42195
Email : sadapenerbit@gmail.com
Website : sadapenerbit.com & repository.sadapenerbit.com
Telpon/WA : +62 838 1281 8431

KATA PENGANTAR

Puji syukur senantiasa kami panjatkan ke hadirat Allah SWT, Tuhan Yang Maha Esa, atas limpahan rahmat, karunia, dan petunjuk-Nya, sehingga penulisan buku yang berjudul "*DATA MINING* TERAPAN: Konsep, Algoritma, dan Implementasi Berbasis Java" ini dapat diselesaikan dengan baik.

Di era transformasi digital saat ini, data telah menjadi aset strategis yang sangat berharga bagi berbagai sektor, mulai dari bisnis, pendidikan, kesehatan, hingga pemerintahan. Namun, tumpukan data yang masif (*big data*) tidak akan memberikan dampak yang signifikan tanpa adanya teknik dan metode yang tepat untuk mengekstrak informasi serta pengetahuan tersembunyi di dalamnya. Di sinilah peran krusial *data mining* bekerja.

Buku ini hadir sebagai jawaban atas kebutuhan literatur yang komprehensif di bidang teknologi informasi. Berbeda dengan referensi yang hanya berfokus pada pendekatan teoretis, buku ini dirancang secara khusus untuk menjembatani kesenjangan antara teori dan praktik. Pembahasan dimulai dari konsep fundamental *data mining*, eksplorasi algoritma-algoritma utama seperti klasifikasi (misalnya *Decision Tree*, *Naïve Bayes*), klastering (*K-Means*, *Hierarchical*), hingga aturan asosiasi (*Apriori*).

Nilai tambah utama dari buku ini terletak pada pendekatan implementatifnya yang menggunakan bahasa pemrograman Java. Pemilihan Java didasarkan pada portabilitasnya yang tinggi, keandalannya dalam menangani sistem berskala besar (*enterprise*), serta ketersediaan pustaka (*library*) yang luas untuk pemrosesan data. Setiap algoritma yang dibahas disertai dengan contoh kasus, logika penyelesaian, dan potongan kode (*source code*) Java yang dapat langsung dipraktikkan oleh pembaca.

Buku ini sangat direkomendasikan bagi mahasiswa program studi Ilmu Komputer, Sistem Informasi, dan Teknik Informatika, serta para praktisi data, pengembang perangkat lunak, dan peneliti yang ingin memperdalam teknik penggalian data secara terstruktur dan aplikatif.

Penyusunan karya ini tentu tidak lepas dari dukungan, bimbingan, dan kontribusi dari berbagai pihak. Oleh karena itu, kami menyampaikan apresiasi yang setinggi-tingginya kepada seluruh rekan sejawat, akademisi, serta Penerbit Sada Kurnia Pustaka yang telah memfasilitasi lahirnya karya literasi ini agar dapat menjangkau khalayak luas.

Akhir kata, semoga buku ini mampu memberikan manfaat nyata, memperkaya wawasan, dan menjadi rujukan yang aplikatif bagi kemajuan ilmu pengetahuan dan teknologi di Indonesia.

Penulis

DAFTAR ISI

KATA PENGANTAR ..... **iii**
DAFTAR ISI ..... **v**
BAB 1 PENGANTAR *DATA MINING* TERAPAN DAN EKOSISTEM JAVA **1**
(Dody)
Data Mining di Era Transformasi Digital 2
Konsep Dasar *Data Mining* 8
Ekosistem Java untuk *Data Mining* 15
Daftar Pustaka 20
Profil Penulis 21
BAB 2 KARAKTERISTIK, JENIS, DAN REPOSITORI DATA **22**
(Riri Fajriah)
Pendahuluan 23
Konsep *Dataset* Dalam *Data Mining* 24
Karakteristik *Dataset* 25
Distribusi Data 27
Jenis Data Berdasarkan Struktur 28
Kualitas Data 29
Missing Value 30
Duplicate Data 30
Outlier 31
Class Imbalance 31
Repositori Dataset 32
Daftar Pustaka 35
Profil Penulis 36
BAB 3 DASAR PEMROGRAMAN JAVA UNTUK PENGOLAHAN DATA **37**
(Haryanto)
Java Virtual Machine (JVM) 38
Membuat Aplikasi Sederhana yang Berdiri Sendiri 39
javac HelloWorld.java 40
Membuat Applet Sederhana 41
Tipe Data, Variabel, dan Operator 42

Deklarasi dan Inisialisasi Variabel 43
Casting 44
Operator 46
Presedens 47
Ekspresi 50
Daftar Pustaka 55
Profil Penulis 56
BAB 4 PEMBERSIHAN DATA *(DATA CLEANING)* 58
(Nelis Febriani)
Pendahuluan Pembersihan Data 59
Mengapa *Data Cleaning* Penting? 60
Data Cleaning sebagai Bagian dari *Data Preprocessing* 61
Validasi Setelah *Data Cleansing* 67
Daftar Pustaka 71
Profil Penulis 72
BAB 5 TRANSFORMASI DAN NORMALISASI DATA 73
(Tb. Zaenal Muttaqin)
Pendahuluan 74
Pre-Processing Data 75
Transformasi dan Normalisasi Data 77
Daftar Pustaka 84
Profil Penulis 85
BAB 6 REDUKSI DIMENSI DAN SELEKSI FITUR 86
(Meta Amalya Dewi)
Pendahuluan 87
Dimensi Data dan Permasalahan *High-Dimensional* 87
Feature Selection 88
Dimensionality Reduction 90
Principal Component Analysis (PCA) 90
Feature Selection dan PCA: Kapan Memilih? 92
Implementasi Berbasis Java Menggunakan Weka 92
Studi Kasus Terapan: Data Aktivitas Pembelajaran 94
Evaluasi dan Risiko Penerapan 95
Pipeline Data Mining 96
Kesimpulan 96
Panduan Praktis Pemilihan Metode 97

Daftar Pustaka 98
Profil Penulis 100
BAB 7 EKSPLORASI DATA (EDA) DAN VISUALISASI DI JAVA ... 101
(Oleh Soleh)
Pendahuluan 102
Konsep dan Kedudukan EDA dalam Proses *Data Mining* 102
Ekosistem *Library* Java untuk EDA dan Visualisasi 103
Statistik Deskriptif dengan Java 105
Analisis Univariat: Distribusi Data 109
Analisis Bivariat dan Multivariat 113
Visualisasi Data Kategorikal dan Deret Waktu 118
Daftar Pustaka 122
Profil Penulis 123
BAB 8 KONSEP DAN ALGORITMA KLASIFIKASI BERBASIS ATURAN 124
(Alfredo Gormantara)
Konsep Dasar Klasifikasi Berbasis Aturan 125
Pembentukan Aturan Klasifikasi 126
Algoritma Aturan Klasifikasi 129
Studi Kasus dan Implementasi Klasifikasi Berbasis Aturan dengan Java 132
Daftar Pustaka 137
Profil Penulis 138
BAB 9 ALGORITMA KLASIFIKASI LANJUTAN 139
(Sucianna Ghadati Rabiha)
Pengantar Klasifikasi Lanjutan 140
Decision Tree 142
K-Nearest Neighbors (KNN) 146
Naive Bayes 148
Support Vector Machine (SVM) 150
Random Forest 151
Daftar Pustaka 153
Profil Penulis 155
BAB 10 IMPLEMENTASI ALGORITMA KLASIFIKASI MENGGUNAKAN JAVA 156
(Fandi Ali Mustika)

Pendahuluan 157
Konsep Dasar Klasifikasi 157
Persiapan Lingkungan Java 158
Persiapan dan Representasi Data 159
Pembagian Data Latih dan Data Uji 160
Implementasi Algoritma k-NN dengan Java 161
Menjalankan Program dan Menguji Prediksi 166
Mengukur Kinerja Model 167
Daftar Pustaka 169
Profil Penulis 170
BAB 11 KONSEP DAN ALGORITMA KLASTERISASI BERBASIS JARAK 171
(Diva Syabina Putri)
Pengantar Klasterisasi (*Unsupervised Learning*) 172
Metrik Pengukuran Jarak 174
Algoritma K-Means 179
Algoritma K-Medoids 183
Cara Evaluasi Model Klaster 186
Rangkuman 189
Daftar Pustaka 191
Profil Penulis 193
BAB 12 *ASSOCIATION RULE MINING* 194
(Dewi Immaniar Desrianti)
Pengertian *Associating Rule Mining* 195
Penerapan *Association Rule* Mining Di Berbagai Bidang 196
Terminologi Dan Struktur Data Pada *Association Rule* 196
Terminologi Preprocessing Data Transaksi 198
Pengukuran Support Dan Confidence 201
Pengukuran Lift Dan Evaluasi *Association Rule* 204
Daftar Pustaka 207
Profil Penulis 208
BAB 13 REGRESI LINIER DAN PREDIKSI BERBASIS DATA 209
(Ali Herdian)
Pendahuluan 210
Konsep Dasar Regresi Linier 211
Persamaan Regresi Linier 214

Perhitungan Regresi 216
Korelasi dan Hubungan Variabel 218
Daftar Pustaka 221
Profil Penulis 222
BAB 14 EVALUASI, VALIDASI, DAN OPTIMASI MODEL 223
(Sekar Ayu Wijaya Kusumaningrum)
Pengantar Evaluasi dan Validasi Model 224
Metrik Evaluasi untuk Model Klasifikasi 226
Metrik Evaluasi untuk Model Regresi 229
Teknik Validasi Model 231
Optimasi Model dan Penyetelan *Hyperparameter* 232
Implementasi Berbasis Java 233
Daftar Pustaka 237
Profil Penulis 239
BAB 15 *DEPLOYMENT MODEL DATA MINING* KE APLIKASI *ENTERPRISE* 240
(Muhamad Yusup)
Pengantar *Deployment* Model di *Enterprise* 241
Standarisasi dan Format Export Model 247
Daftar Pustaka 252
Profil Penulis 256
BAB 16 STUDI KASUS TERAPAN *DATA MINING* INDUSTRI 257
(Dedeh Supriyanti)
Pendahuluan 258
Konsep Dasar *Data Mining* Untuk Industri 261
Data Mining dan *Knowledge Discovery In Database* (KDD) 262
Karakteristik *Data Mining* Dalam Industri 263
Fungsi Utama *Data Mining* 265
Data Mining Dalam Industri Manufaktur 266
Contoh Kasus Sederhana Data Mining di Industri 267
Studi Kasus: Prediksi Kerusakan Mesin Produksi 268
Daftar Pustaka 269
Profil Penulis 271

BAB 1
PENGANTAR *DATA MINING* TERAPAN DAN EKOSISTEM JAVA

Dody, S.Kom., M.Kom.
Institut Teknologi Perusahaan Listrik Negara

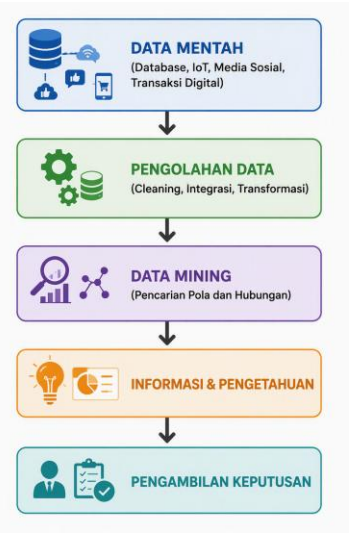
Data Mining di Era Transformasi Digital

1. Latar Belakang *Data Mining*

Perkembangan teknologi informasi dalam dua dekade terakhir telah membawa perubahan besar dalam cara individu, organisasi, maupun pemerintah mengelola dan memanfaatkan data. Aktivitas sehari-hari seperti penggunaan media sosial, transaksi digital, pembelajaran daring, sistem kesehatan elektronik, hingga perangkat *Internet of Things* (IoT) menghasilkan data dalam jumlah yang sangat besar. Fenomena ini semakin meningkat pada periode 2020-2025 ketika transformasi digital menjadi prioritas utama berbagai sektor akibat perubahan pola kerja, perkembangan layanan berbasis *cloud*, serta meningkatnya ketergantungan pada layanan digital (Sarker, 2021).

Data yang terkumpul sebenarnya memiliki potensi besar untuk menghasilkan informasi yang bernilai. Namun, keberadaan data dalam jumlah besar tidak secara otomatis menghasilkan pengetahuan yang berguna. Banyak organisasi menghadapi tantangan berupa data yang tersebar di berbagai sistem, memiliki format yang berbeda, serta terus bertambah setiap saat. Dalam kondisi seperti ini, analisis secara manual tidak lagi memadai untuk menemukan pola-pola penting yang tersembunyi di dalam data.

Data mining hadir sebagai solusi untuk mengekstraksi pengetahuan, pola, hubungan, dan informasi yang sebelumnya tidak terlihat secara langsung dari kumpulan data yang besar. Melalui berbagai teknik dan algoritma, *data mining* membantu organisasi memahami perilaku pelanggan, memprediksi tren masa depan, mendeteksi anomali, serta mendukung proses pengambilan keputusan yang lebih objektif dan berbasis fakta.



Gambar 1. 1: Transformasi Data Menjadi Pengetahuan Melalui Proses Data Mining
Sumber : diolah oleh Penulis

Pada era digital saat ini, data sering disebut sebagai "aset baru" yang memiliki nilai strategis setara bahkan melebihi aset fisik. Perusahaan digital seperti *marketplace*, layanan transportasi daring, perbankan digital, dan platform media sosial menjadikan data sebagai fondasi utama dalam menjalankan bisnis mereka. Oleh karena itu, kemampuan mengolah dan menganalisis data telah menjadi salah satu kompetensi penting dalam berbagai bidang pekerjaan modern.

2. Pengertian dan Tujuan *Data Mining*

Secara sederhana, *data mining* dapat diartikan sebagai proses menemukan pola, hubungan, atau informasi yang bermanfaat dari sejumlah besar data menggunakan teknik statistik, matematika, *machine learning*, dan basis data. Istilah "*mining*" atau penambangan menggambarkan proses pencarian sesuatu yang bernilai dari kumpulan data yang sangat besar, seperti halnya menambang emas dari dalam bumi.

Daftar Pustaka

Cios, K. J., & Kurgan, L. A. (2005). Trends in *Data Mining* and Knowledge Discovery. In *Advanced Techniques in Knowledge Discovery and Data Mining* (pp. 1–26). Springer London. https://doi.org/10.1007/1-84628-183-0_1

Data Mining. (2012). Elsevier. https://doi.org/10.1016/C2009-0-61819-5

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A Survey on Bias and Fairness in *Machine Learning*. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Sarker, I. H. (2021). Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. *SN Computer Science*, *2*(5), 377. https://doi.org/10.1007/s42979-021-00765-8

Syahirah, D., Priati, P., & Martadireja, O. P. (2025). Association *Rule* Mining across Multiple Domains: Systematic Literature Review. *Sinkron*, *9*(4), 1953–1964. https://doi.org/10.33395/sinkron.v9i4.15227

Wan, C. (2019). *Data Mining Tasks and Paradigms* (pp. 7–15). https://doi.org/10.1007/978-3-319-97919-9_2

Zaki, M. J., & Meira, J. W. (2020). *Data Mining and Machine Learning*. Cambridge University Press. https://doi.org/10.1017/9781108564175

PROFIL PENULIS



Dody, S.Kom.,M.Kom

Penulis merupakan seorang akademisi dan praktisi di bidang Teknologi Informasi dengan fokus pada pengembangan Sistem Informasi, *Knowledge Management*, dan *E-Government*. Saat ini berafiliasi dengan Institut Teknologi PLN dan aktif dalam berbagai kegiatan akademik serta pengembangan keilmuan berbasis teknologi digital. Pendidikan formal ditempuh melalui jenjang Sarjana (S1) Sistem Informasi di Universitas Budi Luhur, yang kemudian dilanjutkan pada jenjang Magister (S2) Teknologi Sistem Informasi di universitas yang sama. Latar belakang pendidikan tersebut membentuk kompetensi dalam mengintegrasikan aspek teknis dan manajerial dalam pengelolaan sistem informasi.

Memiliki minat yang kuat pada bidang *Knowledge Management*, *Information System, Business Process Modeling, E-Business* serta Pengembangan Sistem Informasi pada Multidisplin Ilmu lainnya. Selain aktivitas akademik dan pengembangan keilmuan, penulis juga aktif sebagai peneliti pada C-Green dan BEAM di Institut Teknologi PLN, yang berfokus pada pengembangan kajian dan inovasi berbasis teknologi, keberlanjutan, serta pemanfaatan data dan sistem informasi untuk mendukung kebutuhan organisasi dan masyarakat.

Penulis secara konsisten berkontribusi dalam pengembangan konsep dan implementasi teknologi informasi yang memberikan nilai tambah bagi organisasi dan masyarakat. Melalui kegiatan penelitian, pengajaran, dan pengembangan keilmuan, penulis berupaya menghubungkan perspektif akademik dengan kebutuhan praktis di lapangan. Melalui buku ini, diharapkan pembaca dapat memperoleh wawasan, pemahaman, serta kontribusi nyata dalam pengembangan ilmu pengetahuan di bidang sistem informasi dan analitik data, khususnya dalam perspektif yang aplikatif dan relevan dengan kebutuhan masa kini.Email Penulis: dody@itpln.ac.id

BAB 2
KARAKTERISTIK, JENIS, DAN REPOSITORI DATA

Riri Fajriah, S.Kom., MM., M.Kom.
Universitas Universitas Mercu Buana

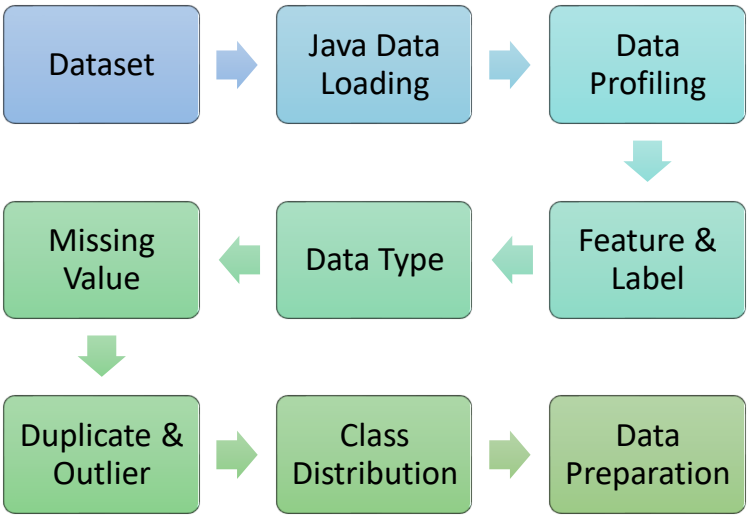
Pendahuluan

Data merupakan komponen utama dalam proses *data mining*. Algoritma seperti *Decision Tree*, K-Nearest Neighbor (KNN), Naive Bayes, K-Means, dan Apriori bekerja berdasarkan pola yang terdapat pada data. Oleh karena itu, keberhasilan proses *data mining* tidak hanya ditentukan oleh algoritma, tetapi juga oleh kualitas dan karakteristik *dataset* yang digunakan.

Sebelum membangun model, data perlu dipahami melalui proses *data understanding*. Tahap ini meliputi pemeriksaan jumlah data, atribut, tipe data, distribusi, *missing value*, duplikasi, *outlier*, dan distribusi kelas. Data preparation selanjutnya digunakan untuk membersihkan dan mengubah data agar sesuai dengan kebutuhan analisis (Fernandes et al., 2023).

Dalam buku ini, proses pemahaman data tidak hanya dilakukan secara konseptual, tetapi juga diimplementasikan menggunakan Java. Java dapat digunakan untuk membaca *dataset*, menghitung jumlah *record*, memeriksa atribut, mendeteksi nilai kosong, menghitung statistik dasar, dan membuat *data profiling* sederhana.

Alur berikut ini menggambarkan tahapan awal pengolahan *dataset* menggunakan Java, dimulai dari proses pemuatan data hingga pemeriksaan karakteristik dan kualitas data. Setiap tahap bertujuan memastikan data dipahami, dibersihkan, dan dipersiapkan dengan baik sebelum digunakan dalam proses analisis atau pembangunan model *Data Mining*, yaitu :



Gambar 2.1: Tahapan *Data Preparation*
Sumber : diolah oleh penulis

Pada Gambar 2.1, tahapan *Data Preparation* ini menunjukkan alur pengolahan data dalam proses *Data Mining*, yang dimulai dari *Dataset* sebagai sumber data, kemudian dilakukan *Java Data Loading* untuk memasukkan data ke dalam program dan *Data Profiling* untuk memahami karakteristik awal *dataset*. Selanjutnya, data diperiksa melalui identifikasi *Feature & Label*, *Data Type*, *Missing Value*, serta *Duplicate & Outlier* untuk memastikan kualitas dan struktur data, kemudian dianalisis *Class Distribution* untuk melihat keseimbangan kelas pada data. Setelah seluruh karakteristik dan permasalahan data diketahui, tahap terakhir adalah *Data Preparation,* yaitu mempersiapkan data agar siap digunakan dalam proses pemodelan *Data Mining* atau *Machine Learning*.

Konsep *Dataset* Dalam *Data Mining*

Dataset adalah kumpulan data yang terdiri atas sejumlah observasi atau *record* dan atribut.

Daftar Pustaka

Fernandes, A. A. A., Koehler, M., Konstantinou, N., Pankin, P., Paton, N. W., & Sakellariou, R. (2023). Data preparation: A technological perspective and *review*. *SN Computer Science, 4*, 425. https://doi.org/10.1007/s42979-023-01828-8

Geerkens, S., Siebericchs, C., Braun, A., & Waschulzik, T. (2024). QI^2: An interactive tool for data quality assurance. *AI and Ethics, 4*, 141–149. https://doi.org/10.1007/s43681-023-00390-6

Gerhart, N., Torres, R., & Giddens, L. (2023). Towards an agile CRISP-DM process. *MWAIS 2023 Proceedings*, 17.

Koukaras, P., & Tjortjis, C. (2025). Data preprocessing and *feature* engineering for *data mining*: Techniques, tools, and best practices. *AI, 6*(10), 257. https://doi.org/10.3390/ai6100257

Ma, Z., Jørgensen, B. N., & Ma, Z. G. (2024). A systematic data characteristic understanding framework towards physical-sensor *big data* challenges. *Journal of Big Data, 11*, 84. https://doi.org/10.1186/s40537-024-00942-5

Neutatz, F., Chen, B., Alkhatib, Y., Ye, J., & Abedjan, Z. (2022). Data cleaning and AutoML: Would an optimizer choose to clean? *Datenbank-Spektrum, 22*, 121–130. https://doi.org/10.1007/s13222-022-00413-2

Sancricca, C., Siracusa, G., & Cappiello, C. (2024). Enhancing data preparation: Insights from a time series case study. *Journal of Intelligent Information Systems, 62*, 1503–1530. https://doi.org/10.1007/s10844-024-00867-8

Schmetz, A., & Kampker, A. (2024). Inside production *data science*: Exploring the main tasks of data scientists in production environments. *AI, 5*(2), 873–886. https://doi.org/10.3390/ai5020043

PROFIL PENULIS



Riri Fajriah, S.Kom, MM, M.Kom.
Adalah seorang profesional dan akademisi yang memiliki keahlian di bidang Tata Kelola Teknologi Informasi dan Data Science. Beliau meraih gelar Sarjana Komputer pada tahun 2007 dari Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Bina Nusantara. Setelah menyelesaikan studi sarjananya, Riri memulai karier sebagai profesional IT di PT BNI Life Insurance, di mana posisi terakhir yang diembannya pada tahun 2016 adalah *Senior Assistant Manager Procurement*. Selama di perusahaan tersebut, beliau banyak terlibat dalam proyek pengadaan dan pengelolaan infrastruktur teknologi informasi. Pada tahun 2010, beliau melanjutkan pendidikan dan meraih gelar Magister Manajemen dari Universitas Mercu Buana. Untuk memperdalam keahliannya di bidang teknologi, Riri menyelesaikan program Magister Ilmu Komputer di Universitas Budi Luhur pada tahun 2024. Saat ini, beliau menjabat sebagai Dosen Tetap pada Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Mercu Buana. Dalam perannya sebagai akademisi, Riri aktif melakukan penelitian yang berfokus pada tata kelola teknologi informasi dan *data science*. Penelitiannya telah memberikan kontribusi signifikan terhadap pengembangan ilmu di bidang tersebut, serta mendukung inovasi dalam pengelolaan TI dan analisis data di lingkungan akademik maupun industri.

Email Penulis : riri.fajriah@mercubuana.ac.id

BAB 3
DASAR PEMROGRAMAN JAVA UNTUK PENGOLAHAN DATA

Haryanto, S.Kom., M.M., M.TI.
Universitas Raharja

Salah satu tantangan utama dalam pemrograman selama bertahun-tahun adalah masalah pembuatan program lintas platform. Bagaimana cara membuat perangkat lunak yang, setelah dikompilasi, dapat berjalan dengan cara yang persis sama di berbagai jenis komputer? Inilah masalah yang menjadi alasan diciptakannya Java. Java adalah bahasa pemrograman yang memungkinkan pemrogram untuk membuat dan mengompilasi program pada satu mesin, yang kemudian dapat dijalankan di komputer dengan sistem operasi yang berbeda-beda. Sebagai contoh, seseorang dapat membuat program di komputer Apple yang kemudian dapat dijalankan di Windows, Linux, atau Solaris. Ini merupakan keunggulan nyata dibandingkan bahasa pemrograman standar yang mengharuskan program dikompilasi secara terpisah untuk setiap sistem operasi. Dalam bab ini, kita akan membahas langkah-langkah instalasi dasar *Java Development Kit* (JDK) serta proses pembuatan, kompilasi, dan eksekusi sebuah program Java sederhana (Thomson:2000).

Java Virtual Machine (JVM)

Java *Virtual Machine* (JVM) membedakan Java dari sebagian besar bahasa pemrograman lainnya. Salah satu tujuan utama perancangan bahasa Java adalah agar kode yang sama dapat dijalankan di platform mana pun. JVM adalah program perangkat lunak yang berfungsi layaknya sebuah komputer utuh. Dengan menggunakan komputer virtual ini pada berbagai platform (UNIX, Win32, Macintosh, dan sebagainya), Anda dapat menggunakan kembali program tanpa perlu membuat versi khusus untuk setiap platform. Program Java akan selalu dijalankan di dalam JVM.

Java memungkinkan Anda untuk menulis aplikasi mandiri, program Java yang tidak perlu dijalankan di dalam aplikasi Java yang berjalan di server atau program bernama Appletviewer. Saat kode dikompilasi, kompiler menghasilkan yang disimpan dalam berkas . Mesin virtual bertugas menerjemahkan ini menjadi format yang dapat dipahami oleh mesin. Berkat kombinasi antara kompilasi dan interpretasi, Java dianggap sebagai bahasa hibrida. Perlu dicatat bahwa terdapat banyak kompiler yang dapat mengkompilasi kode

Java menjadi format. Jika hal ini dilakukan, program hasil kompilasi tersebut tidak akan bersifat portabel; namun, program itu dapat dijalankan tanpa memerlukan mesin virtual (Thomson:2000).

Platform pengembangan yang digunakan dalam kursus ini adalah Java Software Development Kit, Standard Edition atau yang lebih baru. Java sebelumnya dikenal sebagai Java Development dan memiliki kompatibilitas mundur dengan JDK 1.1. Perangkat lunak ini mencakup beberapa komponen, antara lain: Java Runtime Environment (JRE) yang menyediakan JVM. Perangkat pengembangan dan pustaka kompilasi yang diperlukan untuk membuat aplikasi Java. Kelompok ini mencakup kompiler Java, program, serta perangkat untuk menjalankan.

Membuat Aplikasi Sederhana yang Berdiri Sendiri

Anda akan mulai menulis program sederhana dengan mendefinisikan sebuah kelas. Berikan nama pada kelas tersebut serta sepasang tanda kurung kurawal untuk memuat isi kelas. Sesuai konvensi, nama kelas sebaiknya diawali dengan huruf kapital, seperti pada contoh berikut:

```
class HelloWorld
{
}
class HelloWorld
```

Saat Anda menulis aplikasi mandiri, *Java Virtual Machine* harus mengetahui di mana eksekusi kode dimulai. Untuk memulai eksekusi kode, JVM mencari dan memanggil sebuah metode khusus bernama `*public static void*. JVM menggunakan metode ini untuk memulai eksekusi program. Program ini sudah lengkap. Perlu diingat bahwa program Java bersifat membedakan huruf besar dan kecil; kata *Class* tidak sama dengan *class*. Anda harus menyimpan program tersebut menggunakan penyunting teks dengan nama `HelloWorld.java` dan mengompilasinya menggunakan kompiler Java. Anda dapat menggunakan program `javac` jika Anda memiliki Java 2 *Software Development Kit* (SDK) yang disediakan (Thomson:2000).

Daftar Pustaka

A Imran, A Zoha, Challenges in 5G: how to empower SON with *big data* for enabling 5G. IEEE Netw 28(6), 27–33 (2014)

El Amrani C., et al., Development of a real-time urban remote sensing initiative in the mediterranean region for early warning and mitigation of disasters,|| IEEE, pp. 2782–5, 2012.

H Hu, Y Wen, T Chua, X Li, Toward scalable systems for *big data* analytics: a technology tutorial. IEEE Access 2, 652–687 (2014) J

Manyika, M Chui, B Brown, J Bughin, R Dobbs, C Roxburgh, AH Byers, Big data: the next frontier for innovation, competition, and productivity (McKinsey Global Institute, USA, 2011)

Q Wu, G Ding, Y Xu, S Feng, Z Du, J Wang, K Long, Cognitive internet of things: a new paradigm beyond connection. IEEE Internet Things J 1(2), 129–143 (2014) CW Tsai, CF Lai, MC Chiang, LT Yang, Data *mining* for internet of things: a survey. IEEE Commun Surv Tut 16(1), 77–97 (2014)

Smith K. R., et al., Millions Dead: How Do We Know and What Does It Mean? Methods Used in the Comparative Risk Assessment of Household Air Pollution,|| Annu Rev Public Health, vol. 35, pp. 185–206, 2014.

Thomson , Java Programming Fundamentals JVPG-SSMN-0101A ©1988-2000 Wave Technologies International, Inc., a Thomson Learning company. Thomson Learning™ is a trademark used herein under license. All rights reserved.

X Wu, X Zhu, G Wu, W Ding, Data *mining* with *big data*. IEEE Trans Knowl Data Eng 26(1), 97–107 (2014)

PROFIL PENULIS



Haryanto, S.Kom., M.M., M.TI.

Penulis memiliki kepakaran di bidang keuangan perusahaan, manajemen, teknologi informasi, dan strategi pemasaran. Ketertarikan tersebut mendorong penulis untuk melanjutkan pendidikan di Universitas Budi Luhur pada Program Studi Komputerisasi Akuntansi dan berhasil menyelesaikan pendidikan Strata Satu (S1) pada tahun 2006.,Pada tahun 2008, penulis menyelesaikan pendidikan Magister Manajemen (S2) di Universitas Budi Luhur.

Selanjutnya, penulis melanjutkan pendidikan pada jenjang magister Teknologi Informasi di Universitas Raharja sebagai upaya untuk memperluas wawasan dan kompetensi di bidang manajemen serta teknologi informasi. Selain memiliki latar belakang akademik, penulis juga memiliki pengalaman profesional di bidang marketing dan pengembangan bisnis, khususnya dalam pemasaran produk-produk yang berkaitan dengan mesin industri. Bersama rekan-rekannya, penulis turut mengembangkan dan memimpin PT. Multi Sinar Teknik, berfokus pada pengembangan usaha, strategi pemasaran, serta perencanaan dan pengelolaan organisasi pemasaran. Dalam rangka mengembangkan kompetensi sebagai dosen dan profesional, penulis aktif melakukan penelitian dan publikasi ilmiah di berbagai bidang yang relevan dengan keilmuan dan pengalaman profesionalnya.

Sejumlah artikel dan hasil penelitian telah dipublikasikan sebagai bagian dari kontribusi penulis dalam pengembangan ilmu pengetahuan. Selain aktif sebagai dosen dan peneliti, penulis juga memiliki ketertarikan dalam menulis buku. Kegiatan menulis menjadi salah satu bentuk kontribusi penulis dalam menyebarluaskan pengetahuan dan pengalaman kepada mahasiswa, akademisi, praktisi, serta masyarakat luas. Penulis berharap karya-karya yang dihasilkan dapat memberikan manfaat dan kontribusi positif bagi perkembangan

pendidikan, ilmu pengetahuan, dunia usaha, serta kemajuan bangsa dan negara. Pengalaman penulis dalam bidang strategi marketing, perencanaan pemasaran, pengembangan bisnis, dan organisasi pemasaran menjadi landasan dalam mengembangkan berbagai gagasan dan karya yang dihasilkan.

Email: haryanto@raharja.info Email Penulis: haryanto@raharja.info

BAB 4
PEMBERSIHAN DATA
(DATA CLEANING)

N. Nelis Febriani SM, S.Kom., M.Kom
Universitas Cipasung Tasikmalaya

Pendahuluan Pembersihan Data

Data *mining* sering dipahami sebagai proses mencari pola atau pengetahuan tersembunyi dari sekumpulan data. Namun, sebelum algoritma seperti K-Means, K-Nearest Neighbor (KNN), *Naïve Bayes*, C4.5, Apriori, atau metode lainnya diterapkan, terdapat satu pekerjaan yang sering kali justru membutuhkan perhatian besar, yaitu pembersihan data (*data cleaning*). Data yang diperoleh dari dunia nyata hampir tidak pernah berada dalam kondisi sempurna. Data dapat mengandung nilai yang kosong, data ganda, kesalahan penulisan, format yang berbeda, nilai yang tidak masuk akal, data ekstrem, maupun ketidakkonsistenan antaratribut. Han, Kamber, dan Pei menjelaskan bahwa data dunia nyata cenderung tidak lengkap, mengandung *noise*, dan tidak konsisten sehingga diperlukan proses pembersihan sebelum digunakan dalam proses *data mining*.

Dengan demikian, *data cleaning* bukan sekadar menghapus data yang dianggap salah**.** Pembersihan data merupakan proses memahami kondisi data, menemukan masalah kualitas data, menentukan tindakan yang tepat, dan menghasilkan *dataset* yang lebih layak digunakan untuk proses analisis. Berikut adalah tahapan proses data *cleansing*.



Gambar 4.1: Tahapan Proses Data *Cleansing*
Sumber: diolah oleh penulis

Dalam proyek *data mining*, kualitas data sangat berpengaruh terhadap kualitas hasil analisis. IBM dalam dokumentasi CRISP-DM menempatkan pembersihan data sebagai bagian penting dari tahap *data preparation*, dengan tujuan meningkatkan kualitas data sesuai kebutuhan teknik analisis yang akan digunakan. Hal penting yang perlu dipahami adalah tidak semua data yang berbeda merupakan data yang salah**.** Keputusan untuk menghapus atau mengubah data harus mempertimbangkan konteks dan aturan dari bidang yang dianalisis.

Sebelum membersihkan data, kita perlu menentukan terlebih dahulu apa yang dimaksud dengan data berkualitas. Data yang berkualitas bukan berarti data yang harus selalu lengkap 100%. Data disebut berkualitas apabila data tersebut sesuai dengan tujuan penggunaannya. Pembersihan data atau *data cleaning* adalah serangkaian kegiatan untuk menemukan dan menangani permasalahan pada data sehingga *dataset* menjadi lebih sesuai untuk tujuan analisis atau pemodelan.

Pembersihan dapat meliputi menangani *missing value,* menghapus atau menggabungkan data duplikat, memperbaiki kesalahan penulisan, memeriksa nilai yang tidak masuk akal, menyeragamkan format, menangani *outlier,* mengurangi *noise,* memperbaiki inkonsistensi. Han, Kamber, dan Pei mengelompokkan aktivitas *data cleaning* antara lain pengisian nilai yang hilang, penghalusan data yang mengandung *noise*, identifikasi *outlier*, serta penyelesaian ketidakkonsistenan data (Jiawei Han, Micheline Kamber, 2011).

Mengapa *Data Cleaning* Penting?

Bayangkan sebuah perguruan tinggi mempunyai data 5.000 mahasiswa. Data tersebut digunakan untuk membangun model prediksi mahasiswa yang berpotensi terlambat lulus. Setelah diperiksa, ditemukan 125 data IPK kosong, 73 mahasiswa tercatat dua kali, 45 nilai SKS bernilai negatif, 60 tanggal lahir menggunakan format berbeda, 18 mahasiswa memiliki usia 150 tahun, nama program studi ditulis sebagai Sistem Informasi, SI, Sist. Informasi. Apabila data tersebut langsung digunakan untuk *machine learning*, algoritma akan memprosesnya sebagai data biasa. Masalah utamanya adalah algoritma tidak mengetahui bahwa data tersebut salah kecuali peneliti memberikan aturan yang tepat. Kualitas data menjadi penting karena kesalahan, duplikasi, nilai hilang, dan ketidakkonsistenan dapat memengaruhi analisis dan keputusan.

Daftar Pustaka

Cunningham, S. A., & Muir, J. A. (2023). *Data Cleaning*. Dalam *The Cambridge Handbook of Research Methods and Statistics for the Social and Behavioral Sciences*, pp. 443–467. Cambridge University Press.

Jiawei Han, Micheline Kamber, J. P. (2011). *Data Mining: Concepts and Techniques, 3rd Edition*. Morgan Kaufman.

Nguyen, T., Nguyen, H.-T., & Nguyen-Hoang, T.-A. (2025). *Data quality management in big data: Strategies, tools, and educational implications*. Journal of Parallel and Distributed Computing, 200, 105067.

Zhu, J., Zhao, X., Sun, Y., Song, S., et al. (2024/2025). *Relational Data Cleaning Meets Artificial Intelligence: A Survey*. Data Science and Engineering, 10, 147–174

PROFIL PENULIS



N. Nelis Febriani S.M., S.Kom., M.Kom.

Penulis lahir di Tasikmalaya pada tanggal 16 Februari 1995. Penulis menyelesaikan pendidikan S1 di STMIK Tasikmalaya jurusan Teknik Informatika dan pendidikan S2 di Universitas Dian Nuswantoro (UDINUS) Semarang jurusan Magister Teknik Informatika. Saat ini penulis merupakan dosen tetap Program Studi S1 Sistem Informasi di Universitas Cipasung Tasikmalaya serta aktif sebagai tutor Universitas Terbuka sejak tahun 2024. Penulis aktif melakukan penelitian yang berkaitan dengan Data Mining, Sistem Penunjang Keputusan dan Sistem Pakar. Selain mempublikasikan artikel ilmiah, juga terlibat sebagai penulis dalam beberapa buku yang berkaitan dengan Sistem Pendukung Keputusan, Data Science dan Data Warehouse. Penulis juga memiliki minat dan komitmen dalam pengembangan pembelajaran STEAM, khususnya media ajar digital dan proyek berbasis AR. Dengan pendekatan berbasis teknologi dan riset aplikatif, penulis telah menghasilkan banyak karya ilmiah yang relevan untuk kebutuhan praktis dan pengembangan ilmu komputer, terutama dalam konteks lokal seperti kesehatan masyarakat dan edukasi berbasis teknologi.

Email Penulis: nelis.sm@uncip.ac.id.

BAB 5
TRANSFORMASI DAN NORMALISASI DATA

Tb. Zaenal Muttaqin, S.Kom., S.E., M.M.
Universitas Serang Raya

Pendahuluan

Data Mining lahir dari pertumbuhan data yang semakin besar yang berasal dari berbagai sumber. Data yang dihasilkan dapat berasal dari kegiatan transaksi di suatu pusat perbelanjaan, di mana data barang dan data pembayaran terkoleksi. Data dapat juga diperoleh dari aktivitas olahraga seperti pertandingan sepak bola di mana seorang pemain melakukan berapa kali operan dan tendangan ke gawang. Contoh lain dapat kita lihat dari bidang kesehatan, di mana terdapat data rekam medis seorang pasien untuk melihat riwayat penyakit apa saja yang pernah diderita dan berapa kali dilakukan rawat jalan atau rawat inap.

Dengan *Data Mining* dapat diperoleh manfaat untuk memudahkan pengambilan keputusan, membuat prediksi akurat untuk perencanaan, otomatisasi pengambilan keputusan yang berimbas pada efisiensi biaya dan pengetahuan (*knowledge*) bagi pimpinan perusahaan untuk melihat karakteristik pelanggan (Rinci K.H., dkk., 2024).

Data mining merupakan teknik penggalian data yang tersimpan dan berharga dari kumpulan data besar dan kompleks. Dalam proses *data mining* terdapat berbagai metode yang dapat digunakan untuk menemukan pola dan hubungan dalam data yang sebelumnya belum diketahui (Muttaqin, et al., 2023).

Jenis-jenis *data mining* mencakup metode sebagai berikut:

1. Asosiasi – Teknik *data mining* yang digunakan untuk menemukan korelasi antara item dalam sebuah kumpulan data.
2. Klasifikasi – Teknik *data mining* yang biasa digunakan untuk memprediksi kelas atau label dari suatu objek berdasarkan karakteristiknya.
3. Klastering – Teknik *data mining* yang digunakan untuk mengelompokan objek-objek berdasarkan pada atribut atau fitur yang serupa, mirip ataupun yang terdekat.
4. Regresi – Teknik *data mining* yang digunakan untuk memprediksi nilai numerik dari sebuah variabel berdasarkan pada variabel lainnya (variabel X dan Y).

Pre-Processing Data

Data merupakan objek utama dalam proses *data mining*, *di* mana data-data tersebut diolah dengan menggunakan metode tertentu dan dapat menghasilkan output sesuai dengan yang diharapkan oleh seorang analis. Tujuan utama dari *Data Mining* adalah untuk menemukan pola atau informasi yang berguna yang tidak secara langsung terlihat atau dapat diprediksi oleh manusia (Kharis dan Zili, 2022).

Data Collection merupakan langkah awal dalam perjalan proses pengolahan data. Kualitas data yang dikumpulkan akan sangat memengaruhi kualitas model dan output yang dihasilkan. Himpunan Data (*dataset*) merupakan kumpulan dari objek dan atributnya. Atribut merupakan sifat atau karakteristik dari suatu objek (Fajar Astuti H., 2009). Secara umum, karakteristik yang dimiliki oleh suatu data antara lain *dimensionality* (dimensi), *sparsity* dan *resolution* (resolusi).

Data Collection dapat berasal dari berbagai macam sumber seperti dari hasil database *extraction, Web Scrapping, Collecting by application Programming Interface (API), Internet of Thing (IoT) Sensory Data* dan pengumpulan data eksternal oleh pengguna secara manual. Terdapat beberapa hal yang harus dipertimbangkan dalam memilih sumber data antara lain:

1. **Relevansi**
   Memastikan bahwa data yang akan kita kumpulkan merupakan data yang sesuai (relevan) dengan kebutuhan dan tujuan dari aplikasi model *data mining*. Jika tidak sesuai, maka hasil akhir yang diperoleh akan mengakibatkan buruknya tingkat akurasi, bahkan akan menyesatkan bagi penggunanya.
2. **Kualitas**
   Kualitas data sangat mendukung tingkat reputasi hasil akhir pengolahan data. Pastikan data yang diperoleh memiliki kualitas dan konsistensi tinggi untuk mendukung tingkat akurasi.
3. **Aksesibilitas**
   Akses yang mudah dan legal akan sangat membantu dalam pengolahan data. Hindari pengambilan data dari sumber yang memiliki *restricted area* yang dapat mempersulit dalam pengumpulan data dan menghambat dalam pemrosesan data.

Daftar Pustaka

Afrilia, Y. (2025). *Data Processing Dalam Data Mining*, Universitas Malikussaleh Aceh Press.

Aristo Felix, (2023). *Exploratory Data Analysis with Python*. One GML Digital Learning Solutions.

Astuti, F.H, (2013). *Data Mining,* Penerbit Andi Jakarta

Han, J., Pei J., & Tong, H. (2022). *Data Mining: Concepts and Techniques.* Morgan Kaufman

Hapsari, R.K, dkk (2022). *Data Mining,* Indie Press Bandung

Ridho, Ali B. (2019). *Normalisasi Data*, Electronic Engineering Polytechnic Institute of Surabaya.

Tim Dosen Sarjana Terapan, (2022). *Modul Data Mining*. Fakultas Vokasi Universitas Indonesia Maju.

Yulia, R.H, (2022). *Data Mining dan Kecerdasan Bisnis*, Nusa Mandiri Publisher Jakarta.

PROFIL PENULIS



Tb. Zaenal Muttaqin, S.Kom, S.E., M.M.
Lahir di Kabupaten Pandeglang pada 04 Juni 1980. Menyelesaikan Pendidikan Strata 1 (S1) di Universitas Bina Nusantara Jakarta pada tahun 2003 dengan mengambil program studi Sistem Informasi. Melanjutkan pendidikan Strata 2 (S2) di Universitas Budi Luhur Jakarta dan lulus pada tahun 2006 dengan konsentrasi Manajemen Sistem Informasi. Saat ini aktif sebagai Dosen Profesional yang telah memiliki Sertifikat Pendidik Nasional dan mengajar di salah satu perguruan tinggi terkemuka dengan mengajar untuk mata kuliah Sistem Informasi Manajemen, Analisa dan Perancangan Sistem Informasi, Perancangan Database, *Data Mining* dan Data Analitik.

Saat ini sedang mendalami keilmuan di bidang *Supply Chain Management* berbasis *Blockchain Technology*. Menjabat sebagai Ketua Program Studi Sistem Informasi di Universitas Serang Raya periode 2018 – 2023. Selain aktif sebagai dosen, beliau juga aktif sebagai asesor Sertifikasi Kompetensi di bawah naungan Badan Nasional Sertifikasi Profesi (BNSP) dan tergabung dalam organisasi Ikatan Asesor Profesional (IASPRO). Sebagai asesor beliau menguji dibidang Informatika khususnya *Data Science, Junior Networking*, *Graphic Design, Multimedia* dan *Digital Marketing*.
Moto: *Scientia et Moribus Praeditus Vincit*

Email Penulis: d.zaey.vu@gmail.com

BAB 6
REDUKSI DIMENSI DAN SELEKSI FITUR

Meta Amalya Dewi, S.Kom., M.Kom.
Universitas Bina Nusantara

Pendahuluan

Dataset pada aplikasi *data mining* semakin sering memiliki jumlah fitur yang besar. Kondisi ini muncul pada data transaksi, sensor, log aplikasi, teks, citra, maupun data aktivitas pengguna. Banyaknya fitur tidak selalu berarti informasi yang tersedia semakin berkualitas. Sebagian fitur dapat tidak relevan terhadap tujuan analisis, sedangkan sebagian lainnya dapat membawa informasi yang berulang. Ruang fitur yang besar juga dapat meningkatkan beban komputasi dan menyulitkan visualisasi serta interpretasi model. Permasalahan tersebut berkaitan dengan fenomena *curse of dimensionality*, yaitu meningkatnya kesulitan pembelajaran ketika dimensi data bertambah (Han et al., 2023 ; Moslemi, 2023).

Dua strategi yang umum digunakan untuk mengatasi persoalan tersebut adalah *feature selection* dan *dimensionality reduction*. *Feature selection* memilih sebagian fitur asli yang dianggap paling relevan, sedangkan *dimensionality reduction* membentuk representasi baru dengan dimensi yang lebih rendah. Keduanya dapat mengurangi kompleksitas data, tetapi menghasilkan konsekuensi yang berbeda terhadap interpretabilitas dan informasi yang dipertahankan (Moslemi, 2023 ; Villa-Blanco et al., 2023).

Bab ini membahas prinsip, jenis metode, *Principal Component Analysis* (PCA), kriteria evaluasi, serta implementasi berbasis Java menggunakan Weka. Pembahasan difokuskan pada penerapan sehingga pembaca tidak hanya memahami konsep, tetapi juga mampu menentukan metode dan mengevaluasi dampaknya terhadap proses *data mining*.

Dimensi Data dan Permasalahan *High-Dimensional*

Sebuah *dataset* dapat direpresentasikan sebagai matriks X berukuran n × p, dengan n sebagai jumlah observasi dan p sebagai jumlah fitur. Dengan demikian, dimensi data pada konteks ini merujuk pada jumlah fitur yang digunakan untuk menggambarkan setiap observasi. *Dataset* dengan 10 fitur memiliki ruang fitur berdimensi 10, sedangkan *dataset* dengan 1.000 fitur berada dalam ruang berdimensi 1.000.

Masalah utama bukan semata-mata jumlah kolom, tetapi kualitas informasi yang dibawa oleh kolom tersebut. Fitur dapat dibedakan secara praktis menjadi fitur relevan, tidak relevan, dan redundan. Misalnya, pada data penjualan, Harga, Jumlah, dan Total_Penjualan dapat membawa informasi yang sangat tumpang tindih apabila Total_Penjualan dihitung dari Harga × Jumlah. Sebaliknya, Customer_ID umumnya merupakan *identifier* dan bukan karakteristik substantif pelanggan.

High-dimensional data juga dapat meningkatkan waktu pemrosesan, kebutuhan penyimpanan, dan kompleksitas model. Literatur mutakhir menempatkan *feature selection* sebagai salah satu strategi penting untuk mengurangi fitur tidak relevan dan redundan sekaligus menjaga efisiensi dan interpretabilitas model (Zaman et al., 2022 ; Theng & Bhoyar, 2024). Oleh karena itu, tujuan reduksi dimensi bukan memperoleh jumlah fitur sekecil mungkin, melainkan memperoleh representasi yang cukup informatif untuk tujuan *data mining*.

Feature Selection

1. Konsep dan Tujuan

Feature selection adalah proses memilih subset fitur dari keseluruhan fitur tanpa mengubah fitur yang dipertahankan menjadi variabel baru. Jika himpunan awal fitur adalah F = {f1, f2, ..., fp}, maka proses seleksi menghasilkan F' ⊂ F. Dengan demikian, makna substantif fitur tetap dapat dipertahankan. Pendekatan ini dapat mengurangi redundansi, mempercepat pembelajaran, menyederhanakan model, dan membantu interpretasi hasil (Villa-Blanco et al., 2023 ; Theng & Bhoyar, 2024).

Feature selection juga perlu dinilai dari sisi stabilitas. Subset yang dipilih sebaiknya tidak berubah secara ekstrem hanya karena terjadi perubahan kecil pada data. Penelitian terbaru menunjukkan bahwa stabilitas merupakan dimensi evaluasi yang berbeda dari sekadar performa prediksi dan penting untuk membangun model yang lebih dapat dipercaya (BÜYÜKKEÇECİ & OKUR, 2023; Lazebnik & Rosenfeld, 2024).

Daftar Pustaka

Barbieri, M. C., Grisci, B. I., & Dorn, M. (2024). Analysis and comparison of *feature selection method*s towards performance and *stability*. *Expert Systems with Applications*, *249*, 123667. https://doi.org/10.1016/j.eswa.2024.123667

BÜYÜKKEÇECİ, M., & OKUR, M. C. (2023). A Comprehensive Review of *Feature Selection* and *Feature Selection* Stability in Machine Learning. *Gazi University Journal of Science*, *36*(4), 1506–1520. https://doi.org/10.35378/gujs.993763

Greenacre, M., Groenen, P. J. F., Hastie, T., D'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100. https://doi.org/10.1038/s43586-022-00184-w

Han, J. H., Pei, J., & Tong, H. (2023). *Data Mining*. Elsevier. https://doi.org/10.1016/C2013-0-18660-6

Lazebnik, T., & Rosenfeld, A. (2024). A new definition for *feature selection stability* analysis. *Annals of Mathematics and Artificial Intelligence*, *92*(3), 753–770. https://doi.org/10.1007/s10472-024-09936-8

Marukatat, S. (2023). Tutorial on PCA and approximate PCA and approximate kernel PCA. *Artificial Intelligence Review*, *56*(6), 5445–5477. https://doi.org/10.1007/s10462-022-10297-z

Moslemi, A. (2023). A tutorial-based survey on *feature selection*: Recent advancements on *feature selection*. *Engineering Applications of Artificial Intelligence*, *126*, 107136. https://doi.org/10.1016/j.engappai.2023.107136

Theng, D., & Bhoyar, K. K. (2024). *Feature selection* techniques for *machine learning*: a survey of more than two decades of research. *Knowledge and Information Systems*, *66*(3), 1575–1637. https://doi.org/10.1007/s10115-023-02010-5

Villa-Blanco, C., Bielza, C., & Larrañaga, P. (2023). Feature subset selection for data and feature streams: a review. *Artificial Intelligence Review*, *56*(S1), 1011–1062.

https://doi.org/10.1007/s10462-023-10546-9

Zaman, E. A. K., Mohamed, A., & Ahmad, A. (2022). *Feature selection* for online streaming high-dimensional data: A state-of-the-art review. *Applied Soft Computing*, *127*, 109355. https://doi.org/10.1016/j.asoc.2022.109355

PROFIL PENULIS



Meta Amalya Dewi, S.Kom., M.Kom.

Meta Amalya Dewi lahir di Purwokerto. Penulis menyelesaikan pendidikan Sarjana (S1) pada bidang Sistem Informasi dan melanjutkan pendidikan Magister (S2) pada bidang Sistem Informasi di Universitas Budi Luhur. Selanjutnya, penulis menempuh pendidikan Doctor of Computer Science (DCS) di Universitas Bina Nusantara. Saat ini, penulis merupakan dosen tetap di Universitas Bina Nusantara dan aktif dalam kegiatan pendidikan, penelitian, serta pengembangan keilmuan di bidang sistem informasi dan teknologi informasi.

Bidang keahlian dan minat penelitian penulis meliputi Artificial Intelligence in Online Learning, Learning Analytics, Information System Strategy, Enterprise Information System, Business Application Development, *Data Mining*, serta User Interface/User Experience (UI/UX). Ketertarikan tersebut diarahkan pada pengembangan dan penerapan teknologi informasi untuk mendukung pengambilan keputusan, peningkatan kualitas sistem informasi, serta pemanfaatan data dalam lingkungan pendidikan dan organisasi. Penulis juga aktif dalam penulisan dan publikasi artikel ilmiah pada forum dan publikasi bereputasi internasional.

Melalui pengalaman akademik dan penelitian tersebut, penulis berkontribusi dalam pengembangan materi pembelajaran dan penulisan ilmiah yang mengintegrasikan konsep, metode, dan penerapan teknologi informasi secara praktis. Pembaca dapat berkomunikasi dengan penulis melalui email meta.dewi@binus.edu.

BAB 7
EKSPLORASI DATA (EDA) DAN VISUALISASI DI JAVA

Oleh Soleh S.Kom., M.MSI
Universitas Raharja

Pendahuluan

Setelah penjelasan data melalui tahap pembersihan (Bab 4), transformasi dan normalisasi (Bab 5), serta reduksi dimensi dan seleksi fitur (Bab 6), langkah berikutnya sebelum data digunakan untuk membangun model *data mining* adalah memahami karakteristik data tersebut secara mendalam. Tahap ini dikenal sebagai eksplorasi data, atau lebih dikenal dengan istilah *Exploratory Data Analysis* (EDA). EDA merupakan proses menyelidiki data secara sistematis untuk menemukan pola, mendeteksi anomali, menguji asumsi, dan meringkas karakteristik utama data, umumnya dengan bantuan metode statistik deskriptif dan visualisasi.

Dalam praktik *data mining*, kesalahan yang sering terjadi adalah melompat langsung ke tahap pemodelan—baik klasifikasi, klasterisasi, maupun regresi—tanpa terlebih dahulu memahami sebaran, hubungan antarvariabel, dan potensi anomali yang terkandung dalam data. Padahal, model sebaik apa pun tidak akan memberikan hasil yang bermakna apabila dibangun di atas data yang belum dipahami dengan baik. Dengan demikian, EDA berperan sebagai jembatan antara tahap persiapan data pada Bab 4–6 dan tahap pemodelan pada Bab 8 dan seterusnya.

Bahasa pemrograman Java, meskipun tidak sepopuler Python dalam ekosistem *data science,* memiliki sejumlah *library* matang yang dapat dimanfaatkan untuk melakukan EDA dan visualisasi data secara efektif, mulai dari perhitungan statistik deskriptif, pembuatan histogram dan *boxplot*, hingga *heatmap* korelasi. Bab ini membahas cara memanfaatkan *library-library* tersebut secara terpadu.

Konsep dan Kedudukan EDA dalam Proses *Data Mining*

Istilah *Exploratory Data Analysis* pertama kali dipopulerkan oleh statistikawan John W. Tukey melalui bukunya *Exploratory Data Analysis* (Tukey, 1977). Tukey menekankan filosofi "biarkan data berbicara" (*let the data speak for itself*)—yaitu pendekatan yang mengutamakan eksplorasi terbuka terhadap data menggunakan ringkasan numerik dan representasi visual, dibandingkan langsung menguji hipotesis yang telah ditetapkan sebelumnya.

Pendekatan ini berbeda dengan analisis konfirmatori (*confirmatory data analysis*), yang bertujuan menguji hipotesis tertentu menggunakan metode statistik inferensial seperti uji-t atau ANOVA. EDA bersifat lebih fleksibel dan eksploratif: analis tidak dibatasi oleh hipotesis awal, melainkan membiarkan pola-pola dalam data memunculkan pertanyaan dan hipotesis baru yang dapat diuji lebih lanjut.

Dalam kerangka kerja *data mining* yang lebih luas, seperti KDD (*Knowledge Discovery in Databases*) maupun CRISP-DM (*Cross Industry Standard Process for Data Mining*), EDA umumnya ditempatkan setelah tahap persiapan data (*data preparation*) dan sebelum tahap pemodelan (*modeling*). Pada posisi ini, EDA berfungsi memastikan bahwa data yang telah dibersihkan dan ditransformasikan benar-benar siap dan sesuai untuk algoritma yang akan digunakan.

Berdasarkan jumlah variabel yang dianalisis, EDA dapat dibedakan menjadi tiga jenis: (1) analisis univariat, yang menyelidiki distribusi satu variabel pada satu waktu, misalnya distribusi nilai akhir mahasiswa; (2) analisis bivariat, yang menyelidiki hubungan antara dua variabel, misalnya hubungan waktu belajar dengan nilai akhir; dan (3) analisis multivariat, yang menyelidiki hubungan antara tiga variabel atau lebih secara simultan, misalnya korelasi antarseluruh variabel akademik dalam bentuk matriks.

Secara praktis, EDA bertujuan untuk mendeteksi pola dan tren tersembunyi dalam data, mengidentifikasi anomali atau *outlier* yang berpotensi mengganggu proses pemodelan, memeriksa keberadaan nilai hilang (*missing value*) yang tersisa setelah tahap pembersihan data, serta menguji asumsi statistik dasar (misalnya normalitas atau linearitas) yang mendasari sejumlah algoritma *data mining*. Bab ini menunjukkan bagaimana keempat tujuan tersebut dapat dicapai menggunakan *library* Java, dengan data akademik mahasiswa sebagai konteks penerapan.

Ekosistem *Library* Java untuk EDA dan Visualisasi

Salah satu tantangan utama ketika melakukan EDA dengan Java adalah tidak adanya satu *library* tunggal yang menyediakan seluruh fungsi

Daftar Pustaka

Cleveland, W. S. (1993). Visualizing data. Hobart Press.

Cortez, P. (2008). Student performance [*Dataset*]. UCI *Machine Learning Repository*. https://doi.org/10.24432/C5TG7T

Cortez, P., & Silva, A. M. G. (2008). Using *data mining* to predict secondary school student performance. In A. Brito & J. Teixeira (Eds.), Proceedings of the 5th Annual Future Business Technology Conference (pp. 5–12). EUROSIS.

Han, J., Kamber, M., & Pei, J. (2012). Data *mining*: Concepts and techniques (3rd ed.). Morgan Kaufmann.

JFree.org. (2024). JFreeChart: A free Java chart *library* for all kinds of charts. https://www.jfree.org/jfreechart/

Knowm Inc. (2024). XChart: A light-weight Java *library* for plotting data. GitHub. https://github.com/knowm/XChart

Tablesaw Development Team. (2024). Tablesaw: Java dataframe and visualization *library*. https://jtablesaw.github.io/tablesaw/

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). Introduction to *data mining* (2nd ed.). Pearson.

The Apache Software Foundation. (2024). Apache Commons Math: The Apache Commons Mathematics Library. https://commons.apache.org/proper/commons-math/

Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.

University of Waikato, Machine Learning Group. (2024). Weka 3: Machine *learning* software in Java. https://www.cs.waikato.ac.nz/ml/weka/

PROFIL PENULIS



Oleh Soleh S.Kom., M.MSI.

Ketertarikan penulis terhadap ilmu komputer dimulai pada tahun 1996. Hal tersebut membuat penulis memilih untuk masuk ke Universitas Gunadarma dengan memilih Jurusan Manajemen Informatika (MI) pada Fakultas Ilmu Komputer dan berhasil menyelesaikan program studi S1 dengan kelulusan pada tahun 2000. Penulis kemudian melanjutkan pendidikan ke jenjang master yang masih di kampus yang sama pada tahun 2004 berhasil menyelesaikan studi S2 di prodi Ilmu Komputer dengan Jurusan Manajemen Informasi Bisnis. Sejak tahun 2022 sampai saat ini, sedang menempuh pendidikan doktoral di Universitas Kristen Satya Wacana Salatiga melalui prodi Ilmu Komputer dengan konsentrasi Deep Learning & Computer Vision.

Penulis memiliki kepakaran di bidang Business Intelligence, *Data Mining*, Manajemen Informasi serta Deep Learning. Dan untuk mewujudkan karier sebagai dosen profesional, penulis pun aktif sebagai peneliti di bidang kepakarannya tersebut. Beberapa penelitian yang telah dilakukan didanai oleh internal perguruan tinggi.

Email Penulis: oleh.soleh@raharja.info

BAB 8
KONSEP DAN ALGORITMA KLASIFIKASI BERBASIS ATURAN

Alfredo Gormantara, S. Kom., M.Kom.
STMIK Kharisma Makassar

Konsep Dasar Klasifikasi Berbasis Aturan

Klasifikasi merupakan salah satu permasalahan penting dalam *data mining* dan *machine learning* yang bertujuan untuk menentukan kategori atau kelas dari suatu data berdasarkan karakteristik yang dimilikinya. Pada proses klasifikasi, data yang telah memiliki label digunakan sebagai dasar untuk membangun suatu model. Model tersebut kemudian digunakan untuk menentukan kelas dari data baru yang belum diketahui labelnya.

Dalam literatur *rule learning*, model klasifikasi dapat direpresentasikan dalam bentuk sejumlah aturan sederhana yang menghubungkan kondisi tertentu dengan kelas hasil prediksi. Bentuk ini membuat hubungan antara karakteristik data dan keputusan klasifikasi dapat ditelusuri secara langsung. Rapp, Fürnkranz, dan Hüllermeier menjelaskan bahwa *rule learning* merupakan salah satu pendekatan dalam *symbolic learning* yang menghasilkan model dalam bentuk aturan *if–then*. Setiap aturan memiliki sejumlah kondisi yang diperiksa terhadap atribut suatu data dan menghasilkan prediksi kelas apabila kondisi tersebut terpenuhi (Nössig et al., 2024).

Pendekatan tersebut dapat dituliskan secara sederhana sebagai:

IF kondisi *THEN* kelas

Aturan tersebut menyatakan bahwa suatu data akan dikategorikan sebagai layak apabila memenuhi dua kondisi, yaitu memiliki pendapatan tinggi dan riwayat kredit yang baik.

Klasifikasi berbasis aturan memiliki karakteristik yang berbeda dari beberapa model yang menghasilkan representasi matematis atau struktur yang sulit ditelusuri. Aturan dapat dibaca secara langsung oleh manusia sehingga proses pengambilan keputusan relatif mudah dipahami. Karakteristik interpretabilitas tersebut menjadi salah satu alasan *rule learning* tetap menjadi bidang penelitian dalam klasifikasi, termasuk pada penelitian-penelitian terbaru (Rapp et al., 2023).

Klasifikasi berbasis aturan (*rule-based classification*) menggunakan sekumpulan aturan sebagai representasi model. Setiap aturan berisi kondisi tertentu dan kelas yang akan diberikan apabila kondisi tersebut terpenuhi.

Aturan tersebut terdiri atas dua bagian utama:

1. *Antecedent*, yaitu bagian kondisi atau bagian *IF*.
2. *Consequent*, yaitu bagian kesimpulan atau bagian *THEN*.

Dalam penerapannya, sebuah sistem tidak hanya terdiri atas satu aturan. Model biasanya memiliki sejumlah aturan yang bersama-sama membentuk *rule set*. Setiap aturan dapat menangani kelompok data yang berbeda. Oleh karena itu, proses pembentukan *rule set* menjadi bagian penting dalam *rule learning*.

Pembentukan Aturan Klasifikasi

Pembentukan aturan klasifikasi dimulai dari *dataset* yang memiliki atribut sebagai karakteristik data dan satu atribut target sebagai kelas yang akan diprediksi. Setiap baris *dataset* merepresentasikan satu objek atau *instance*, sedangkan setiap kolom merepresentasikan atribut.

Sebagai contoh, *dataset* mahasiswa dapat ditunjukkan sebagai berikut.

Tabel 7.1: Contoh *Dataset*

IPK	Kehadiran	Jumlah SKS	Kelas
3,70	90%	120	Tinggi
3,20	85%	110	Sedang
2,70	75%	100	Rendah

Sumber: diolah oleh penulis

Pada tabel tersebut, IPK, Kehadiran, dan Jumlah SKS merupakan atribut prediktor, sedangkan Kelas merupakan atribut target. Aturan klasifikasi dibentuk dengan mencari hubungan antara atribut prediktor dan kelas target.

Contohnya:

IF IPK >= 3.50 AND Kehadiran >= 85%
THEN Kelas = "Tinggi"

Bagian setelah *IF* berisi kondisi yang harus dipenuhi oleh data, sedangkan bagian setelah *THEN* menunjukkan kelas yang diberikan apabila kondisi tersebut terpenuhi.

Daftar Pustaka

Huynh, V. Q. P., Fürnkranz, J., & Beck, F. (2023). Efficient *learning* of large sets of locally optimal *class*ification *rule*s. *Machine Learning 2023 112:2*, *112*(2), 571–610. https://doi.org/10.1007/S10994-022-06290-W

Kliegr, T., & Izquierdo, E. (2023). QCBA: improving *rule class*ifiers learned from quantitative data by recovering information lost by discretisation. *Applied Intelligence 2023 53:18*, *53*(18), 20797–20827. https://doi.org/10.1007/S10489-022-04370-X

Maszczyk, C., Sikora, M., & Wróbel, Ł. (2024). Classification, Regression, and Survival *Rule* Induction with Complex and M-of-N Elementary Conditions. *Machine Learning and Knowledge Extraction 2024, Vol. 6, Pages 554-579*, *6*(1), 554–579. https://doi.org/10.3390/MAKE6010026

Nössig, A., Hell, T., & Moser, G. (2024). *Rule learning* by modularity. *Machine Learning 2024 113:10*, *113*(10), 7479–7508. https://doi.org/10.1007/S10994-024-06556-5

Rapp, M., Fürnkranz, J., & Hüllermeier, E. (2023). On the efficient implementation of *class*ification *rule learning*. *Advances in Data Analysis and Classification 2023 18:4*, *18*(4), 851–892. https://doi.org/10.1007/S11634-023-00553-7

Zhang, Y., Gong, M., Li, J., Feng, K., & Zhang, M. (2023). Autonomous perception and adaptive standardization for few-shot *learning*. *Knowledge-Based Systems*, *277*, 110746. https://doi.org/10.1016/J.KNOSYS.2023.110746

PROFIL PENULIS



Alfredo Gormantara S.Kom., M.Kom.
Adalah dosen dengan minat riset di bidang Information Systems, Machine Learning, dan Parallel Computing. Ia menyelesaikan pendidikan sarjana di bidang Teknik Informatika di Universitas Atma Jaya Makassar dan melanjutkan studi magister di bidang yang sama di Universitas Atma Jaya Yogyakarta. Kepakarannya mencakup pengembangan perangkat keras dan komputasi cerdas, yang sejalan dengan pokok bahasan dalam buku ini.

Sebagai dosen profesional dan peneliti aktif, ia telah menghasilkan berbagai penelitian di bidang kepakarannya, baik yang didanai secara internal oleh perguruan tinggi maupun eksternal oleh Kemenristek DIKTI. Penulis juga aktif dalam kegiatan pengabdian kepada masyarakat melalui penerapan teknologi berbasis riset. Buku ini disusun berdasarkan pengalaman akademik dan profesional, serta diharapkan dapat menjadi kontribusi nyata dalam pengembangan ilmu pengetahuan di bidang informatika. Penulis terbuka untuk berkolaborasi dalam penelitian dan pengabdian, dan dapat dihubungi melalui

email: alfredogormantara0105@gmail.com.

BAB 9
ALGORITMA KLASIFIKASI LANJUTAN

Dr. Sucianna Ghadati Rabiha, S.Kom., M.Kom.
Universitas Bina Nusantara

Pengantar Klasifikasi Lanjutan

Klasifikasi merupakan salah satu teknik penting dalam *data mining* dan *machine learning* yang digunakan untuk membangun model berdasarkan data yang telah memiliki label kelas. Model tersebut kemudian digunakan untuk memprediksi kelas dari data baru. Dalam konteks *data mining*, klasifikasi termasuk pendekatan *supervised learning* karena proses pembentukan model menggunakan data pelatihan yang telah diketahui kategorinya. Klasifikasi banyak digunakan dalam berbagai permasalahan, seperti prediksi *customer churn*, deteksi transaksi berisiko, klasifikasi dokumen, identifikasi tingkat risiko, dan sistem pendukung keputusan. Han, Pei, dan Tong (2022) menempatkan klasifikasi sebagai salah satu tugas utama *data mining* yang bertujuan membangun model untuk menemukan pola dan menghasilkan prediksi dari data (Han et al., 2022).

Seiring meningkatnya kompleksitas dan volume data, pendekatan klasifikasi perlu berkembang dari sekadar penggunaan aturan keputusan menuju algoritma yang mampu mempelajari pola secara lebih adaptif dari data. Klasifikasi menjadi salah satu pendekatan penting dalam *data mining* karena memungkinkan pola yang diperoleh dari data historis digunakan untuk menentukan kelas atau kategori pada data baru. Perkembangan algoritma klasifikasi menghasilkan berbagai pendekatan dengan karakteristik dan mekanisme yang berbeda, seperti pembentukan struktur keputusan, pengukuran kedekatan antarobjek, pendekatan probabilistik, pembentukan batas pemisah, hingga penggabungan beberapa model. Perbedaan tersebut memberikan fleksibilitas dalam menangani karakteristik data dan kebutuhan permasalahan yang beragam (Tan et al., 2022).

Berbagai algoritma dapat digunakan untuk melakukan klasifikasi, dan masing-masing memiliki mekanisme serta karakteristik yang berbeda. *Decision Tree*, misalnya, membangun struktur keputusan berdasarkan atribut tertentu sehingga hasil klasifikasi dapat direpresentasikan dalam bentuk struktur pohon. K-Nearest Neighbors (KNN) menentukan kelas berdasarkan kedekatan suatu data dengan sejumlah data lain yang telah memiliki label. Naïve Bayes

menggunakan pendekatan probabilistik untuk menentukan kelas yang paling mungkin, sedangkan *Support Vector Machine* (SVM) membangun batas pemisah antarkelas berdasarkan karakteristik data. Selain itu, pendekatan *ensemble learning* seperti Random Forest menggabungkan sejumlah *Decision Tree* untuk menghasilkan prediksi yang lebih *robust*. Berbagai algoritma tersebut masih menjadi bagian penting dalam pembelajaran dan penerapan klasifikasi pada *dataset* terstruktur(Lu et al., 2022).

Studi komparatif yang dilakukan dalam beberapa tahun terakhir juga menunjukkan bahwa tidak terdapat satu algoritma yang selalu memberikan hasil terbaik untuk semua *dataset*. Kinerja algoritma dapat berubah bergantung pada karakteristik data, jumlah fitur, distribusi kelas, ukuran *dataset*, serta tujuan analisis. Sebuah studi komparatif terhadap *Decision Tree*, SVM, Naïve Bayes, dan KNN pada beberapa *dataset* menunjukkan adanya perbedaan performa antaralgoritma dan menegaskan pentingnya mempertimbangkan karakteristik *dataset* dalam pemilihan metode (Sheth et al., 2022). Studi lain yang membandingkan beberapa algoritma klasifikasi pada *dataset* dari UCI *Machine Learning Repository* juga menunjukkan bahwa pemilihan algoritma merupakan bagian penting dalam membangun model klasifikasi karena algoritma yang berbeda dapat menghasilkan performa yang berbeda pada *dataset* yang berbeda (V et al., 2023).

Dengan demikian, pembelajaran klasifikasi lanjutan tidak hanya bertujuan untuk mengetahui nama dan definisi berbagai algoritma. Hal yang lebih penting adalah memahami bagaimana suatu algoritma bekerja, karakteristik data yang sesuai, kelebihan dan keterbatasannya, serta alasan pemilihannya untuk suatu permasalahan. Dalam konteks Sistem Informasi, kemampuan tersebut menjadi penting karena model klasifikasi pada akhirnya digunakan untuk membantu organisasi menghasilkan informasi dan mendukung pengambilan keputusan.

Sebagai contoh, perusahaan dapat memiliki *dataset* pelanggan yang berisi informasi mengenai usia, lama berlangganan, frekuensi penggunaan layanan, jumlah keluhan, dan riwayat pembayaran. Berdasarkan data tersebut, perusahaan dapat membangun model

Daftar Pustaka

Blockeel, H., Devos, L., Frénay, B., Nanfack, G., & Nijssen, S. (2023). *Decision Trees*: from efficient prediction to responsible AI. In *Frontiers in Artificial Intelligence* (Vol. 6). Frontiers Media SA. https://doi.org/10.3389/frai.2023.1124553

Han, J., Pei, J., & Tong, H. (2022). *Data Mining: Concepts and Techniques* (4th ed.). Morgan Kaufmann.

Liu, Y., & Yang, S. (2022). Application of *Decision Tree*-Based Classification Algorithm on Content Marketing. *Journal of Mathematics*, *2022*. https://doi.org/10.1155/2022/6469054

Lu, W., Zhang, Y., Wen, W., Yan, H., & Li, C. (Eds.). (2022). *Cyber Security* (Vol. 1506). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-9229-1

Montesinos López, O. A., López, A. M., & Crossa, J. (2022). Multivariate Statistical Machine Learning Methods for Genomic Prediction. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer International Publishing. https://doi.org/10.1007/978-3-030-89010-0

Piccialli, V., & Sciandrone, M. (2022). Nonlinear optimization and support vector *machine*s. *Annals of Operations Research*, *314*(1), 15–47. https://doi.org/10.1007/s10479-022-04655-x

Rivera-Lopez, R., Canul-Reich, J., Mezura-Montes, E., & Cruz-Chávez, M. A. (2022). Induction of *Decision Tree*s as *class*ification models through metaheuristics. *Swarm and Evolutionary Computation*, *69*, 101006. https://doi.org/https://doi.org/10.1016/j.swevo.2021.101006

Sheth, V., Tripathi, U., & Sharma, A. (2022). A Comparative Analysis of Machine Learning Algorithms for Classification Purpose. *Procedia Computer Science*, *215*, 422–431. https://doi.org/10.1016/j.procs.2022.12.044

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2022). *Introduction to Data Mining* (2nd ed.). Pearson.

V, K., G, D., Sarveshwaran, V., Jayakanth, J. J., M, K., & Renu, Y. (2023).

Comparative Analysis of Diverse Classification Algorithms of Machine Learning by Using Various Quality Metrics. *2023 IEEE 5th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*, 551–556. https://doi.org/10.1109/ICCCMLA58983.2023.10346683

Zhou, Z.-H. (2021). Ensemble Learning. In *Machine Learning* (pp. 181–210). Springer Singapore. https://doi.org/10.1007/978-981-15-1967-3_8

PROFIL PENULIS



Dr. Sucianna Ghadati Rabiha, S.Kom., M.Kom.

Sucianna Ghadati Rabiha merupakan akademisi dan peneliti di bidang Sistem Informasi yang telah bergabung sebagai faculty member di BINUS Online, BINUS University sejak tahun 2015. Ketertarikannya terhadap ilmu komputer dan teknologi informasi berkembang melalui perjalanan akademik dan profesional yang berfokus pada pemanfaatan teknologi untuk mendukung kebutuhan bisnis, pendidikan, dan organisasi.

Penulis menyelesaikan pendidikan sarjana pada bidang Sistem Informasi dan magister pada bidang Teknik Informatika di Universitas Dian Nuswantoro. Penulis kemudian melanjutkan pendidikan doktor pada Program Studi Computer Science di Bina Nusantara University dan telah menyelesaikan studinya. Bidang keahlian dan minat penulis meliputi Analytical Information Systems, Data Mining, Computer Vision, Digital Business, serta Business Process Management. Selain aktif dalam kegiatan penelitian dan publikasi ilmiah, penulis juga berperan dalam pengembangan bahan ajar dan modul pembelajaran pada bidang Sistem Informasi dan Transformasi Digital. Melalui kegiatan akademik, penelitian, dan penulisan buku, penulis berkomitmen untuk berbagi pengetahuan serta memberikan kontribusi dalam pengembangan kompetensi digital dan pemanfaatan teknologi yang inovatif, khususnya dalam menghadapi perkembangan transformasi digital.

Email Penulis: sucianna.rabiha@binus.ac.id

BAB 10
IMPLEMENTASI ALGORITMA KLASIFIKASI MENGGUNAKAN JAVA

Fandi Ali Mustika, S.Kom., M.T.
Universitas Mercu Buana

Pendahuluan

Klasifikasi adalah salah satu tugas dalam pembelajaran mesin yang bertujuan menempatkan suatu data ke dalam kelas atau kategori yang telah ditentukan. Contohnya antara lain mengelompokkan pesan sebagai spam atau bukan spam, mengidentifikasi jenis bunga berdasarkan ukuran bagian-bagiannya, atau menentukan apakah transaksi berpotensi mencurigakan. Sistem klasifikasi mempelajari pola dari contoh yang labelnya sudah diketahui, lalu menggunakan pola tersebut untuk memperkirakan label data baru.

Bab ini membahas implementasi klasifikasi menggunakan bahasa Java. Algoritma yang digunakan sebagai contoh adalah k-Nearest Neighbors (k-NN), atau k tetangga terdekat. k-NN dipilih karena prinsip kerjanya mudah diikuti dan dapat diimplementasikan dengan struktur data sederhana. Pembahasan mencakup representasi data, pembagian data latih dan data uji, implementasi program, pengukuran kinerja, serta beberapa cara menyempurnakan hasil.

Contoh program di sini sengaja dibuat dari pustaka standar Java. Dengan demikian, pembaca dapat melihat bagaimana jarak dihitung dan bagaimana kelas dipilih. Dalam proyek nyata, pustaka pembelajaran mesin dapat membantu menangani *dataset* besar, prapemrosesan yang lebih kompleks, dan evaluasi yang lebih lengkap.

Konsep Dasar Klasifikasi

Satu baris data umumnya terdiri atas fitur dan label. Fitur adalah informasi yang digunakan untuk membuat prediksi, sedangkan label adalah kelas yang ingin diprediksi. Misalnya, untuk mengklasifikasikan bunga, fitur dapat berupa panjang dan lebar sepal serta panjang dan lebar petal. Labelnya dapat berupa nama spesies.

Misalkan suatu data memiliki tiga fitur, data tersebut dapat ditulis sebagai vektor:

$$x = [x_1, x_2, x_3]$$

Label kelasnya ditulis sebagai y Misalnya “A” atau “B”. Kumpulan pasangan fitur-label yang digunakan untuk membangun atau menjalankan model tersebut adalah data berlabel. Dalam pembelajaran terawasi, data dibagi menjadi:

1. Data Latih, yaitu contoh berlabel yang digunakan untuk mempelajari pola.
2. Data Uji, yaitu contoh terpisah yang digunakan untuk mengukur kemampuan model terhadap data yang belum dipakai saat pelatihan.

Pemisahan tersebut penting. Jika kinerja diukur menggunakan data yang sama dengan data yang digunakan untuk membangun model, hasil evaluasi dapat tampak terlalu baik dan tidak mencerminkan kinerja pada kasus baru.

Cara Kerja k-NN

k-NN termasuk algoritma berbasis kedekatan. Untuk, memprediksi kelas sebuah data baru, algoritma menghitung jarak data tersebut terhadap seluruh data latih, memilih sebanyak k tetangga dengan jarak terkecil, lalu menentukan kelas berdasarkan label para tetangga itu. Pada klasifikasi sederhana, kelas denga suara terbanyak dipilih.

Untuk dua vector numerik x dan z dengan m fitur, jarak Euclidean dihitung sebagai berikut:

$$d(x, z) = \sqrt{\sum_{j=1}^{m} (x_j - z_j)^2}$$

Sebagai ilustrasi, jarak antara [2, 3] dan [5, 7] adalah:

$$\sqrt{(2-5)^2 + (3-7)^2} = 5$$

Nilai k menentukan banyaknya tetangga yang memberikan suara. Nilai k yang kecil membuat predikasi lebih dipengaruhi contohnya terdekat, sedangkan nilai yang lebih besar mempertimbangkan lebih banyak contoh. Tidak ada satu nilai k yang selalu terbaik; pemilihannya perlu diuji pada data validasi.

Persiapan Lingkungan Java

Untuk menjalankan contoh, siapkan Java Development Kit (JDK) dan editor atau IDE yang mendukung Java, misalnya IntelliJ IDEA, Eclipse, atau Visual Studio Code dengan ekstensi Java. Pastikan perintah java -version dan javac -version dapat dijalankan dari terminal.

Daftar Pustaka

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). Pattern *class*ification (2nd ed.). Wiley.

Han, J., Kamber, M., & Pei, J. (2011). Data *mining*: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Japkowicz, N., & Shah, M. (2011). Evaluating *learning* algorithms: A *class*ification perspective. Cambridge University Press.

Mitchell, T. M. (1997). Machine *learning*. McGraw-Hill.

Witten, I. H., Frank, E., Trigg, L., Hall, M. A., Holmes, G., & Cunningham, S. J. (1999). Weka: Practical *machine learning* tools and techniques with Java implementations (Computer Science Working Paper 99/11). Department of Computer Science, University of Waikato. https://hdl.handle.net/10289/1040

PROFIL PENULIS



Fandi Ali Mustika, S.Kom., MT.

Ketertarikan penulis terhadap ilmu komputer dimulai pada tahun 2008 silam. Hal tersebut membuat penulis memilih untuk masuk ke Pondok Pesantren Daar El-Qolam dengan memilih Jurusan Ilmu Pengetahuan Alam (IPA) dan berhasil lulus pada tahun 2009. Penulis kemudian melanjutkan pendidikan ke Perguruan Tinggi dan berhasil menyelesaikan studi S1 di prodi Teknik Informatika Universitas Mercu Buana pada tahun 2013. Di tahun 2014, penulis baru masuk S2 di Universitas Mercu Buana Program Pasca Sarjana. Dua tahun kemudian, penulis menyelesaikan studi S2 di prodi Teknik Elektro Program Pasca Sarjana Universitas Mercu Buana.

Penulis memiliki kepakaran dibidang Wireless Sensor, Cloud Computing dan Cyber Security. Dan untuk mewujudkan karir sebagai dosen profesional, penulis pun aktif sebagai peneliti dibidang kepakarannya tersebut. Selain peneliti, penulis juga aktif sebagai praktisi di salah satu Perusahaan swasta sebagai Operation Manager.

Email Penulis: fandi.ali@mercubuana.ac.id

BAB 11
KONSEP DAN ALGORITMA KLASTERISASI BERBASIS JARAK

Diva Syabina Putri
Universitas Raharja

Pengantar Klasterisasi (*Unsupervised Learning*)

Dalam dunia data, tidak semua persoalan datang dengan jawaban yang sudah tersedia. Pada klasifikasi, misalnya, setiap baris data memiliki label yang jelas, seperti "spam" atau "bukan spam", sehingga mesin dapat belajar dari contoh-contoh yang kebenarannya sudah diketahui. Namun, di lapangan banyak sekali data yang tidak memiliki label sama sekali. Perusahaan menyimpan jutaan catatan transaksi, tetapi tidak ada kolom yang menyatakan pelanggan tertentu termasuk tipe apa. Pada situasi seperti inilah klasterisasi (*clustering*) berperan (Han et al., 2012b).

Klasterisasi adalah teknik pembelajaran tanpa pengawasan (*unsupervised learning*) yang bertujuan membagi sekumpulan data menjadi beberapa kelompok atau klaster, sehingga data dalam satu klaster saling mirip, sedangkan data dari klaster yang berbeda saling berbeda (Jain, 2010). Mesin tidak diberi tahu ada berapa kelompok yang benar dan apa nama masing-masing kelompok. Mesin hanya diminta menemukan struktur tersembunyi di dalam data berdasarkan ukuran kemiripan.

1. Perbedaan dengan Pembelajaran Terawasi

Pada pembelajaran terawasi (*supervised learning*), seperti pohon keputusan atau regresi, model dilatih menggunakan pasangan fitur dan target, lalu dievaluasi dengan membandingkan prediksi terhadap target yang sebenarnya. Pada klasterisasi, hanya fitur yang tersedia. Akibatnya, tidak ada jawaban "benar" yang bisa dijadikan pembanding. Kualitas hasil dinilai dari sifat geometris klaster itu sendiri, yaitu seberapa rapat anggota dalam satu klaster dan seberapa terpisah klaster satu dengan lainnya. Perbedaan ini ringkas disajikan pada Tabel 11.1.

Tabel 11.1: Perbedaan Klasifikasi dan Klasterisasi

Aspek	Klasifikasi	Klasterisasi
Jenis pembelajaran	Terawasi (*supervised*)	Tanpa pengawasan (*unsupervised*)
Ketersediaan label	Ada, sebagai target	Tidak ada
Tujuan	Memprediksi kelas data baru	Menemukan kelompok alami pada data

Aspek	Klasifikasi	Klasterisasi
Evaluasi	Akurasi, presisi, recall	Elbow, Silhouette, dan sejenisnya

Sumber: Diolah penulis dari Han et al. (2012) dan Jain (2010).

2. Analogi Dunia Nyata

Bayangkan seorang pemilik toko swalayan yang memiliki catatan belanja ribuan pelanggan. Ia ingin menyusun promosi yang tepat sasaran, tetapi belum pernah mengelompokkan pelanggannya. Dengan klasterisasi, data seperti seberapa sering pelanggan berbelanja dan berapa rata-rata nominal belanjanya dapat dianalisis. Hasilnya mungkin memperlihatkan kelompok pelanggan yang sering datang dengan belanja kecil, kelompok yang jarang datang tetapi belanjanya besar, dan kelompok yang jarang datang dengan belanja kecil. Nama kelompok tersebut baru diberikan oleh manusia setelah klaster terbentuk, bukan sebelumnya.

3. Penerapan Klasterisasi

Klasterisasi digunakan di banyak bidang. Beberapa contoh yang umum ditemui adalah sebagai berikut:

a. Segmentasi pelanggan pada pemasaran untuk menentukan strategi promosi yang berbeda bagi tiap kelompok.
b. Pengelompokan dokumen atau berita berdasarkan kemiripan topik.
c. Kompresi dan segmentasi citra, yaitu mengelompokkan piksel dengan warna yang serupa.
d. Analisis biologi, misalnya mengelompokkan gen atau spesies dengan pola yang mirip.
e. Deteksi anomali, yaitu data yang tidak cocok dengan klaster mana pun dicurigai sebagai penyimpangan, misalnya transaksi yang tidak wajar.

Kunci dari seluruh penerapan tersebut sama: mesin harus punya cara untuk mengukur seberapa dekat atau seberapa mirip dua data. Itulah sebabnya pembahasan jarak menjadi fondasi bab

Meskipun bab ini berfokus secara eksklusif pada fondasi matematis dan konseptual, seluruh logika pengukuran jarak dan iterasi centroid ini akan menjadi basis pengetahuan untuk penulisan kode sumber (source code).

Daftar Pustaka

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). *On the Surprising Behavior of Distance Metrics in High Dimensional Space* (pp. 420–434). https://doi.org/10.1007/3-540-44503-X_27

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, *07-09-January-2007*, 1027–1035.

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Han, J., Kamber, M., & Pei, J. (2012a). 8 - Classification: Basic Concepts. In *Data Mining: Concepts and Techniques (Third Edition)* (Third Edition, pp. 327–391). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-381479-1.00008-3

Han, J., Kamber, M., & Pei, J. (2012b). 9 - Classification: Advanced Methods. In *Data Mining: Concepts and Techniques (Third Edition)* (Third Edition, pp. 393–442). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-381479-1.00009-5

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley. https://books.google.co.id/books?id=YeFQHiikNo0C

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/10.1109/TIT.1982.1056489

MacQueen, J. (1967). Multivariate observations. *Proceedings Ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability*, *1*(1967), 281–297.

Ng, R. T., & Jiawei Han. (2002). CLARANS: a method for clustering objects for spatial *data mining*. *IEEE Transactions on Knowledge and Data Engineering*, *14*(5), 1003–1016. https://doi.org/10.1109/TKDE.2002.1033770

Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, *36*(2), 3336–3341. https://doi.org/10.1016/j.eswa.2008.01.039

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Tan, P. N., Steinbach, M., Kumar, V., & Karpatne, A. (2019). *Introduction to Data Mining eBook: Global Edition*. Pearson Education. https://books.google.co.id/books?id=274oEAAAQBAJ

Thorndike, R. L. (1953). Who Belongs in the Family? *Psychometrika*, *18*(4), 267–276. https://doi.org/10.1007/BF02289263

PROFIL PENULIS



Diva Syabina Putri

Ketertarikan penulis terhadap sains data dan kecerdasan buatan bermula dari antusiasmenya mengeksplorasi logika sistem sejak menempuh studi sarjana. Saat ini, penulis merupakan mahasiswi tingkat akhir (Semester 7) program studi Sistem Informasi di Universitas Raharja (angkatan 2023) dengan rekam jejak akademik yang gemilang. Fokus kepakarannya mencakup pengembangan *Machine Learning*, analisis sistem, serta visualisasi data.

Pada tahun 2026, penulis terpilih sebagai *AI Engineering Fellow* dalam program beasiswa intensif Coding Camp 2026 powered by DBS Foundation x Dicoding, di mana ia memotori pengembangan *capstone project* sistem cerdas bernama "Road2Work.id". Kapasitas riset inovatifnya juga dibuktikan lewat raihan Juara 2 Lomba Karya Tulis Ilmiah Nasional (LKTI) BAPPEDA Kota Tangerang 2025 kategori Sarana Prasarana melalui inovasi berbasis IoT, serta partisipasi aktifnya di program ASEAN *Data Science* sejak 2024.

Tidak hanya berfokus pada baris kode dan algoritma, penulis memiliki *passion* besar dalam menjembatani teknologi dengan masyarakat. Hal inilah yang membawanya mengemban amanah sebagai *Google Student Ambassador* 2025 dan tampil sebagai *content talent* di akun resmi Google Indonesia. Kepeduliannya pada ekosistem digital lokal turut diwujudkan lewat perannya bersama Diskominfo dalam *Training of Trainer* (ToT) Aplikasi Tangerang LIVE. Melalui karya di buku ini, penulis berharap dapat menghadirkan referensi *Data Mining* yang aplikatif dan berdampak nyata.

Email Penulis: divacansu2609@gmail.com

BAB 12
ASSOCIATION RULE MINING

Dewi Immaniar Desrianti, S.Kom., M.TI.
Universitas Raharja

Pertemuan ini membahas konsep dasar, ruang lingkup, dan penerapan Data Mining serta Association *Rule* Mining dalam pengolahan data. Materi difokuskan pada pemahaman proses Knowledge Discovery in Databases (KDD), berbagai teknik *data mining*, serta pengenalan Association *Rule* Mining sebagai teknik untuk menemukan keterkaitan antaritem dalam data transaksi. Data Mining adalah proses menemukan pola, hubungan, atau pengetahuan baru yang tersembunyi dari sekumpulan data besar, sedangkan Association *Rule* Mining adalah teknik *data mining* yang digunakan untuk menemukan hubungan atau keterkaitan antaritem dalam suatu kumpulan data transaksi.

Pengertian *Associating Rule Mining*

Association Rule Mining (ARM) adalah teknik *data mining* yang digunakan untuk menemukan hubungan atau keterkaitan antaritem dalam suatu kumpulan data transaksi (Kusrini & Luthfi, 2009). Teknik ini sangat populer dengan istilah *Market Basket Analysis* (Djamaludin & Nursikuwagus, 2017), karena awalnya dikembangkan untuk menganalisis "keranjang belanja" pelanggan. Contoh kasus klasik (yang sering dikutip dalam literatur *data mining*):

Konsumen yang membeli popok bayi pada malam hari, cenderung juga membeli bir.

Meskipun terkesan tidak berhubungan secara logis, pola semacam ini benar-benar dapat ditemukan dari data transaksi nyata, dan itulah kekuatan ARM, menemukan hubungan yang tidak selalu terlihat jelas secara intuisi, namun terbukti secara statistik dari data.

Dalam notasi association *rule*, kasus toko roti kita gunakan sebagai contoh utama sepanjang modul ini:

{Roti, Selai} → {Susu}

Artinya: jika seorang pelanggan membeli Roti dan Selai, maka pelanggan tersebut cenderung juga membeli Susu.

Penerapan *Association Rule* Mining Di Berbagai Bidang

Tabel 12.1 Contoh Penerapan ARM di Berbagai Bidang

Bidang	Contoh Penerapan
Retail	Rekomendasi produk, penataan rak/tata letak barang di toko
E-Commerce	Fitur "Pelanggan yang membeli ini juga membeli..."
Perpustakaan	Rekomendasi buku yang sering dipinjam bersamaan
Kesehatan	Menemukan pola gejala atau obat yang sering muncul bersamaan pada pasien
Pendidikan	Analisis pola mata kuliah yang sering diambil bersamaan oleh mahasiswa
Perbankan	Analisis pola penggunaan produk perbankan (tabungan, kartu kredit, dsb.)

Sumber: diolah oleh penulis

Penerapan ARM di Indonesia sendiri banyak ditemukan pada studi kasus retail, kesehatan, dan perpustakaan (Kumalasari et al., 2020; Ritha et al., 2021; Mugnia & Mutoffar, 2024).

Terminologi Dan Struktur Data Pada *Association Rule*

Terminologi dan Struktur Data pada Association *Rule* adalah kumpulan istilah dan format data yang digunakan untuk merepresentasikan item, itemset, serta hubungan antaritem dalam bentuk yang dapat diolah oleh algoritma association *rule mining*.

1. Istilah Istilah Dasar

Sebelum masuk ke perhitungan, mahasiswa perlu memahami istilah dasar berikut karena akan terus digunakan di seluruh pertemuan:

a. Item: satu produk atau barang tunggal, misalnya "Roti".
b. Itemset: kumpulan satu atau lebih item yang dianggap sebagai satu kesatuan, misalnya {Roti, Susu}.
c. K-itemset: itemset yang terdiri dari tepat k item. Contoh: {Roti} adalah 1-itemset, {Roti, Susu} adalah 2-itemset, {Roti, Susu, Telur} adalah 3-itemset.

Daftar Pustaka

Dewi, H. F., Handayani, H. H., & Indra, J. (2022). Implementasi algoritma apriori terhadap market basket analysis pada data penjualan retail. Jurnal Informatika Teknologi dan Sains (Jinteks), 4(4), 432–436.

Hermawati, F. A. (2013). Data *mining*. Andi Offset.

Kumalasari, U. N., Putra, I. K. G. D., & Dharmaadi, I. P. A. (2020). Implementasi algoritma apriori untuk menemukan pola pembelian konsumen pada perusahaan retail. Jurnal Ilmiah Teknologi dan Komputer, 1(2). https://doi.org/10.24843/jitter.v1i2.69743

Kusrini, & Luthfi, E. T. (2009). Algoritma *data mining*. Andi Offset.

Mugnia, A., & Mutoffar, M. M. (2024). Implementasi algoritma apriori untuk sistem rekomendasi buku pada perpustakaan digital. Jurnal Teknik Informatika dan Sistem Informasi (JATISI), 11(1), 1–12.

Riszky, A. R., & Sadikin, M. (2019). Data *mining* menggunakan algoritma apriori untuk rekomendasi produk bagi pelanggan. Jurnal Teknologi dan Sistem Komputer, 7(3), 103–108. https://doi.org/10.14710/jtsiskom.7.3.2019.103-108

Ritha, N., Suswaini, E., & Pebriadi, W. (2021). Penerapan association *rule* menggunakan algoritma apriori pada poliklinik penyakit dalam (Studi kasus: Rumah Sakit Umum Daerah Bintan). Jurnal Sistem Informasi (JSI), 7(2), 224–231.

PROFIL PENULIS



Dewi Immaniar Desrianti, S.Kom., M.TI.

Merupakan dosen tetap pada Program Studi Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Raharja. Beliau menyelesaikan pendidikan Sarjana (S.Kom.) bidang Teknik Informatika di STMIK Raharja pada tahun 2010 dan Magister (M.T.I.) bidang Teknik Informatika pada tahun 2015.

Saat ini, beliau masih aktif sebagai dosen di Universitas Raharja. Selain menjalankan tugas akademik, beliau aktif dalam kegiatan pendidikan, penelitian, publikasi ilmiah, pengabdian kepada masyarakat, serta sebagai sekretaris redaksi jurnal ilmiah pada CCIT Journal, selain itu juga aktif menjadi reviewer jurnal dan tim penguji/asesor kompetensi di bidang Teknologi Informatika dan Komunikasi.

Bidang keahlian dan pengajaran yang ditekuni meliputi Pemrograman, Rekayasa Perangkat Lunak, Machine Learning, Data Warehouse, Mobile Computing, Komunikasi Data, Sistem Digital, Organisasi Komputer, Interaksi Manusia dan Komputer, serta Computer Generated Imagery (CGI).

Email Penulis: dewi.immaniar@raharja.info

BAB 13
REGRESI LINIER DAN PREDIKSI BERBASIS DATA

Ali Herdian, S.Kom., M.M.
Universitas Mercu Buana

Pendahuluan

1. Data sebagai Dasar Pengambilan Keputusan

Perkembangan teknologi informasi telah menghasilkan data dalam jumlah yang sangat besar. Data dapat diperoleh dari berbagai sumber, seperti aktivitas pengguna internet, transaksi bisnis, sistem pendidikan, perangkat sensor, media sosial, hingga berbagai aplikasi digital. Keberadaan data tersebut memberikan peluang untuk memperoleh informasi yang dapat digunakan dalam proses pengambilan keputusan.

Namun, data dalam jumlah besar tidak akan memberikan manfaat apabila tidak diolah dan dianalisis dengan baik. Oleh karena itu, diperlukan metode yang dapat membantu mengidentifikasi pola, hubungan, serta kecenderungan yang terdapat di dalam data. Salah satu metode yang banyak digunakan adalah regresi linier**.**

Regresi linier merupakan metode statistik yang digunakan untuk mengetahui hubungan antara satu atau beberapa variabel serta melakukan estimasi terhadap suatu nilai berdasarkan data yang telah tersedia. Dengan menggunakan regresi linier, data historis dapat dimanfaatkan untuk membuat suatu prediksi terhadap kondisi yang akan datang atau nilai yang belum diketahui.

2. Prediksi Berbasis Data

Prediksi berbasis data merupakan proses memperkirakan suatu nilai atau kondisi berdasarkan pola yang ditemukan dari data sebelumnya. Prediksi tidak berarti memberikan hasil yang pasti, tetapi memberikan estimasi terbaik berdasarkan informasi yang tersedia**.**

Sebagai contoh, sebuah sekolah dapat menggunakan data jam belajar dan nilai ujian siswa untuk memperkirakan nilai ujian berdasarkan jumlah jam belajar. Dalam bidang bisnis, data penjualan sebelumnya dapat digunakan untuk memperkirakan penjualan pada periode berikutnya. Sementara dalam bidang teknologi informasi, data penggunaan jaringan dapat digunakan untuk memperkirakan kebutuhan bandwidth.

Proses tersebut menunjukkan bahwa data dapat digunakan bukan hanya untuk memahami apa yang telah terjadi, tetapi juga untuk membantu memperkirakan apa yang mungkin terjadi.

3. Mengapa Regresi Linier?

Regresi linier menjadi salah satu metode yang penting untuk dipelajari karena konsepnya relatif sederhana, mudah diinterpretasikan, dan dapat diterapkan pada berbagai permasalahan. Model regresi linier menggambarkan hubungan antara variabel prediktor dengan variabel yang ingin diprediksi melalui suatu persamaan matematis.

Secara sederhana, hubungan tersebut dapat digambarkan sebagai:

Data → Analisis Hubungan → Model Regresi → Prediksi → Pengambilan Keputusan

Keunggulan lain dari regresi linier adalah hasil modelnya relatif mudah dijelaskan. Pengguna dapat mengetahui bagaimana perubahan suatu variabel berkaitan dengan perubahan variabel lainnya. Hal ini menjadikan regresi linier tidak hanya berguna untuk melakukan prediksi, tetapi juga untuk membantu memahami hubungan antarvariabel.

Konsep Dasar Regresi Linier

1. Pengertian Regresi Linier

Regresi linier merupakan salah satu metode statistik yang digunakan untuk menganalisis hubungan antara variabel dan memperkirakan nilai suatu variabel berdasarkan variabel lainnya. Disebut linier karena hubungan yang dimodelkan dapat direpresentasikan dalam bentuk garis lurus.

Dalam regresi linier, terdapat dua jenis variabel utama, yaitu variabel independen dan variabel dependen. Variabel independen merupakan variabel yang digunakan sebagai prediktor, sedangkan variabel dependen merupakan variabel yang nilainya ingin dijelaskan atau diprediksi.

Daftar Pustaka

Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed.). SAGE Publications.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed.). OTexts. https://otexts.com/fpp3/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). Applied linear statistical models (5th ed.). McGraw-Hill/Irwin.

McKinney, W. (2022). Python for data analysis: Data wrangling with pandas, NumPy, and Jupyter (3rd ed.). O'Reilly Media.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis (6th ed.). Wiley.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. O'Reilly Media.

Wooldridge, J. M. (2020). Introductory econometrics: A modern approach (7th ed.). Cengage.

PROFIL PENULIS



Ali Herdian, S.Kom., M.M.

Ketertarikan penulis terhadap ilmu komputer dimulai pada tahun 2001. Hal tersebut membuat penulis memilih untuk melanjutkan pendidikan ke Perguruan Tinggi dan berhasil menyelesaikan studi S1 di prodi Sistem Informasi Universitas Gunadarma pada tahun 2005. Di tahun 2011, penulis baru melanjutkan kembali S2 di Universitas Trisakti program Pasca Sarjana. Dua tahun kemudian, penulis menyelesaikan studi S2 di prodi Magister Management Program Pasca Sarjana Universitas Trisakti.

Penulis memiliki kepakaran dibidang Perancangan Basis Data, Pemograman Basis Data dan Pengembangan Sistem Informasi. Dan untuk mewujudkan karir sebagai dosen professional, penulis pun aktif sebagai peneliti dibidang kepakaran tersebut. Selain peneliti, penulis juga aktif sebagai praktisi disalah satu Perusahaan swasta sebagai Data Analyst.

Email Penulis: ali.herdian@mercubuana.ac.id

BAB 14
EVALUASI, VALIDASI, DAN OPTIMASI MODEL

Sekar Ayu Wijaya Kusumaningrum
Universitas Raharja

Pengantar Evaluasi dan Validasi Model

Dalam siklus hidup pengembangan sistem *Data Mining* dan *Machine Learning* berbasis *enterprise*, fase konstruksi model seperti algoritma klasifikasi, regresi, maupun klasterisasi hanyalah separuh dari keseluruhan estimasi pekerjaan (Aggarwal, 2015; Zhou, 2021). Tantangan paling krusial terletak pada bagaimana memastikan bahwa model yang dibangun baik menggunakan pustaka mandiri Java maupun ekosistem eksternal seperti Weka atau Deep*learning*4j yang memiliki performa konsisten dan dapat diandalkan saat dipublikasikan ke lingkungan produksi *(production environment)* (Witten et al., 2016). Di sinilah urgensi dari Evaluasi dan Validasi Model berperan sebagai fondasi ilmiah untuk mengukur kualitas, keandalan, serta objektivitas dari pengetahuan *(knowledge)* yang diekstrak dari data latih.

Tanpa proses validasi yang ketat, model *Machine Learning* yang tampak superior pada data latih sering kali mengalami penurunan performa secara drastis ketika dihadapkan pada data baru di dunia nyata akibat fenomena *overfitting*. Oleh karena itu, pemahaman mendalam mengenai berbagai metrik evaluasi seperti akurasi, presisi, *recall*, hingga teknik validasi silang (*cross-validation*) menjadi kompetensi esensial bagi praktisi *data mining* untuk menjamin bahwa pengetahuan dan keputusan berbasis data yang dihasilkan benar-benar objektif, akurat, serta dapat dipertanggungjawabkan secara ilmiah.

1. Urgensi Kinerja Model dalam Data Mining

Suatu algoritma *data mining* tidak akan memberikan nilai tambah bagi organisasi jika gagal menyajikan prediksi yang akurat pada instance data baru (Kelleher et al., 2020). Sering kali, praktisi pemula terjebak dalam ilusi "akurasi pelatihan tinggi" *(high training accuracy)*. Ketika model mencapai tingkat akurasi 99% pada data latih, muncul asumsi keliru bahwa model telah sempurna. Padahal, fenomena tersebut umumnya mengindikasikan bahwa model hanya menghafal pola derau *(noise)* pada data latih alih-alih memahami pola esensial di dalamnya (Goodfellow et al., 2016). Evaluasi kinerja model berfungsi sebagai

prosedur uji kendali mutu *(quality control)* secara kuantitatif yang memberikan tiga manfaat utama:

a. Objektivitas Pengukuran: Memberikan landasan angka matematis yang objektif (seperti akurasi, presisi, *recall*, MAE, dan RMSE) untuk membandingkan beberapa algoritma pembelajaran secara fair.
b. Mitigasi Risiko Bisnis: Mengidentifikasi potensi bias model dan meminimalkan kegagalan prediksi pada data operasional riil yang dapat berdampak finansial atau operasional.
c. Efisiensi Sumber Daya: Memastikan model yang di-deploy pada lingkungan enterprise Java memiliki kompleksitas komputasi dan memori yang seimbang dengan akurasi yang dihasilkan.

2. Konsep Generalisasi: Model yang Baik untuk Data Baru

Tujuan utama dari setiap model *Machine Learning* adalah mencapai daya generalisasi *(generalization capacity)* yang tinggi (Goodfellow et al., 2016). Generalisasi didefinisikan sebagai kemampuan model untuk memberikan prediksi yang akurat pada data baru yang belum pernah ditemui sebelumnya (*unseen data* atau *test set*) (Sammut & Webb, 2017). Guna mengukur daya generalisasi secara objektif, praktisi membagi *dataset* secara sistematis menjadi dua bagian independen:

a. Data Latih (Training Data): Digunakan oleh algoritma untuk mempelajari atribut data dan menyesuaikan parameter internal model.
b. Data Uji (Testing Data): Disembunyikan sepenuhnya selama proses pelatihan dan hanya digunakan di akhir untuk menguji daya generalisasi model.

Sebuah model dikatakan memiliki kapabilitas generalisasi yang unggul apabila tingkat kesalahan *(error rate)* pada data uji berbanding lurus dan sama rendahnya dengan tingkat kesalahan pada data latih (James et al., 2021).

3. Mengatasi *Overfitting* dan *Underfitting*

Dua kendala utama dalam mencapai generalisasi optimal adalah fenomena *underfitting* dan *overfitting* (James et al., 2021).

and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Lever, J., Krzywinski, M., & Altman, N. (2016). Classification evaluation. *Nature Methods*, *13*(8), 603–604. https://doi.org/10.1038/nmeth.3945

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2018). Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, *18*, 1–52. http://arxiv.org/abs/1603.06560

Raschka, S. (2020). Model Evaluation, Model Selection, and Algorithm Selection in *Machine Learning*. *ArXiv*. http://arxiv.org/abs/1811.12808

Sammut, C., & Webb, G. I. (2017). Encyclopedia of *Machine Learning* and Data Mining. In *Encyclopedia of Machine Learning and Data Mining* (2nd ed.). Springer Science+Business Media. https://doi.org/10.1007/978-1-4899-7687-1

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.

Zhou, Z.-H. (2021). *Machine Learning*. https://doi.org/10.1007/978-981-15-1967-3

PROFIL PENULIS



Sekar Ayu Wijaya Kusumaningrum

Perjalanan penulis dalam dunia teknologi berakar sejak masa sekolah menengah di SMA Negeri 18 Kabupaten Tangerang pada tahun 2019, ketika ketertarikannya pada bidang sains dan perangkat keras, khususnya eksplorasi Arduino melalui kegiatan Karya Ilmiah Remaja (KIR), mulai terbentuk. Saat ini, penulis tercatat sebagai mahasiswa aktif Program Studi Teknik Informatika di Universitas Raharja sejak tahun 2022.

Tidak hanya berfokus pada perkuliahan dan dunia pemrograman, penulis juga memiliki ketertarikan yang besar terhadap dunia tulis-menulis. Paduan antara minat riset teknologi dan literasi ini membawanya untuk aktif menuangkan gagasan serta hasil penelitian ke dalam berbagai karya tulis ilmiah. Fokus riset yang ia geluti utamanya mencakup bidang kecerdasan buatan seperti *machine learning*, *deep learning*, hingga sistem pakar (*expert system*). Berbagai karya tulis ilmiah tersebut dapat ditemui melalui jurnal terindeks seperti *G-Tech*, *JISA*, *ABDINE*, dan *CCIT*.

Email Penulis: asekar771@gmail.com

BAB 15
DEPLOYMENT MODEL DATA MINING KE APLIKASI *ENTERPRISE*

Muhamad Yusup, S.Kom., M.Kom.
Politeknik Negeri Jember

Pengantar *Deployment* Model di *Enterprise*

1. Kesenjangan Ekosistem: Memahami Perbedaan Lingkungan Eksperimen Data Science (Python/R) Dengan Lingkungan Produksi Enterprise (Java**).**

Akar tantangan dari *deployment machine learning* di perusahaan skala besar bermula dari fakta bahwa model diciptakan di satu dunia, tetapi harus hidup dan beroperasi di dunia yang sama sekali berbeda. Data Scientist dan Software Engineer menggunakan *tools*, bahasa, dan metrik kesuksesan yang bertolak belakang (Amershi dkk., 2019; Paleyes dkk., 2022).

Berikut adalah rincian kesenjangan mendasar antara lingkungan eksperimen (Data Science) dan lingkungan produksi (Enterprise IT).

a. Perbedaan Tujuan Utama (Metrik Kesuksesan)
   Lingkungan Eksperimen (Python/R): Fokus utamanya adalah Akurasi dan Eksplorasi. Data Scientist bereksperimen dengan berbagai algoritma untuk menemukan pola terbaik dari data historis statis. Waktu komputasi (apakah model butuh 2 detik atau 2 menit untuk inferensi) seringkali bukan prioritas utama selama proses training.

   Lingkungan Produksi (Java): Fokus utamanya adalah Stabilitas, Skalabilitas, dan Latensi. Model yang sangat akurat tidak akan berguna jika membutuhkan waktu 2 detik untuk merespons sistem Core Banking yang menuntut SLA (Service Level Agreement) di bawah 50 milidetik. Sistem harus mampu menangani ribuan request per detik (TPS) tanpa crash (Dean & Barroso, 2013; Paleyes dkk., 2022).

b. Paradigma Bahasa Pemrograman (Dinamis vs. Statis)
   Python/R (Dinamis & Fleksibel): Menggunakan duck-typing dan tipe data dinamis. Hal ini memungkinkan manipulasi matriks dan dataframe (seperti Pandas) dengan sangat cepat dan intuitif. Namun, fleksibilitas ini rawan memunculkan runtime error di tahap produksi jika bentuk data input sedikit berubah.

Java (Statis & Ketat): Bahasa yang strongly-typed dan berorientasi objek (OOP) secara terstruktur. Setiap variabel dan objek (seperti DTO) harus didefinisikan secara eksplisit. Kontrak data yang ketat ini membuat aplikasi Java sangat stabil, tetapi membutuhkan lapisan transformasi data (dari DTO Java ke vektor numerik) sebelum model bisa memprosesnya.

c. Arsitektur Konkurensi (Concurrency & Threading)
Batasan Eksperimen: Python memiliki Global Interpreter Lock (GIL), yang pada dasarnya membatasi eksekusi murni menjadi single-threaded (Python Software Foundation, n.d.). Saat Data Scientist melakukan scoring data, mereka biasanya menggunakan pendekatan batch processing (memasukkan ribuan baris data sekaligus ke dalam model).

Tuntutan Enterprise: Aplikasi Java Spring Boot beroperasi secara asinkron dan multi-threaded. Ratusan pengguna dapat menekan endpoint API secara bersamaan. Jika model Python di-API-kan secara asal-asalan tanpa load balancer dan worker yang tepat, sistem akan mengalami bottleneck. Oleh karena itu, menjalankan model secara native di dalam Java Virtual Machine (JVM) (menggunakan JPMML atau ONNX) (Ruusmann, n.d.; ONNX Runtime developers, n.d.) sering kali menjadi solusi agar prediksi bersifat thread-safe (Goetz dkk., 2006).

d. Siklus Kerja dan Ketergantungan (Dependency Hell)
Notebook vs. IDE: Eksperimen dilakukan di Jupyter Notebooks yang bersifat interaktif dan menyimpan state (variabel tetap ada di memori). Sementara itu, produksi berjalan di lingkungan stateless (layanan mikro) di mana setiap request independen satu sama lain.

Manajemen Dependensi: Ekosistem Python bergantung pada pip atau conda, yang seringkali diikat ke pustaka C/C++ spesifik (seperti TensorFlow atau PyTorch) yang bergantung pada arsitektur OS dan versi CUDA GPU tertentu (Sculley dkk., 2015). Memindahkan ini ke server produksi sering menimbulkan masalah "it works on my *machine*". Java menggunakan Maven atau Gradle yang mengemas semuanya dalam sebuah file .jar, berjalan di atas JVM, sehingga menjamin

Daftar Pustaka

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software engineering for *machine learning*: A case study. *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, 291–300. https://doi.org/10.1109/ICSE-SEIP.2019.00042

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

Burns, B., Grant, B., Oppenheimer, D., Brewer, E., & Wilkes, J. (2016). Borg, Omega, and Kubernetes. *ACM Queue*, *14*(1), 70–93. https://doi.org/10.1145/2898442.2898444

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable *tree* boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Data Mining Group. (n.d.). *PMML 4.4.1 specification: General structure*. https://dmg.org/pmml/v4-4-1/GeneralStructure.html

Dean, J., & Barroso, L. A. (2013). The tail at scale. *Communications of the ACM*, *56*(2), 74–80. https://doi.org/10.1145/2408776.2408794

Eclipse Deep*learning*4j Development Team. (n.d.). *Deeplearning4j* [Computer software]. GitHub. https://github.com/deep*learning*4j/deep*learning*4j

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, *46*(4), Article 44. https://doi.org/10.1145/2523813

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design patterns: Elements of reusable object-oriented software*. Addison-Wesley.

Goetz, B., Peierls, T., Bloch, J., Bowbeer, J., Holmes, D., & Lea, D. (2006). Java *concurrency in practice*. Addison-Wesley.

gRPC Authors. (n.d.). *Introduction to gRPC*. https://grpc.io/docs/what-is-grpc/introduction/

Guazzelli, A., Zeller, M., Lin, W.-C., & Williams, G. (2009). PMML: An open standard for sharing models. *The R Journal*, *1*(1), 60–65. https://doi.org/10.32614/RJ-2009-010

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), Article 93. https://doi.org/10.1145/3236009

H2O.ai. (n.d.). *Productionizing H2O*. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/productionizing.html

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA *data mining* software: An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10–18. https://doi.org/10.1145/1656274.1656278

Humble, J., & Farley, D. (2010). *Continuous delivery: Reliable software releases through build, test, and deployment automation*. Addison-Wesley.

Kleppmann, M. (2017). *Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems*. O'Reilly Media.

Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine *learning* operations (MLOps): Overview, definition, and architecture. *IEEE Access*, *11*, 31866–31879. https://doi.org/10.1109/ACCESS.2023.3262138

Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, *31*(12), 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D. B., Amde, M., Owen, S., Xin, D., Xin, R., Franklin, M. J., Zadeh, R., Zaharia, M., & Talwalkar, A. (2016). MLlib: Machine *learning* in Apache Spark. *Journal of Machine Learning Research*, *17*(34), 1–7. https://jmlr.org/papers/v17/15-237.html

Newman, S. (2021). *Building microservices: Designing fine-grained*

systems (2nd ed.). O'Reilly Media.

Nygard, M. T. (2018). *Release it! Design and deploy production-ready software* (2nd ed.). Pragmatic Bookshelf.

ONNX Runtime developers. (n.d.). *Get started with ORT for* Java. https://onnxruntime.ai/docs/get-started/with-java.html

ONNX. (n.d.). *Open Neural Network Exchange*. https://onnx.ai/

Oracle. (n.d.). *HotSpot virtual machine garbage collection tuning guide*. https://docs.oracle.com/en/java/javase/21/gctuning/

Otoritas Jasa Keuangan. (2022). *Peraturan Otoritas Jasa Keuangan Nomor 11/POJK.03/2022 tentang penyelenggaraan teknologi informasi oleh bank umum*. https://ojk.go.id/id/regulasi/Pages/Penyelenggaraan-Teknologi-Informasi-oleh-Bank-Umum.aspx

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in deploying *machine learning*: A survey of case studies. *ACM Computing Surveys*, *55*(6), Article 114. https://doi.org/10.1145/3533378

Python Software Foundation. (n.d.). *Glossary: Global interpreter lock*. https://docs.python.org/3/glossary.html#term-global-interpreter-lock

Resilience4j. (n.d.). *CircuitBreaker*. https://resilience4j.readme.io/docs/circuitbreaker

Ruusmann, V. (n.d.). *JPMML-Evaluator* [Computer software]. GitHub. https://github.com/jpmml/jpmml-evaluator

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in *machine learning* systems. *Advances in Neural Information Processing Systems*, *28*, 2503–2511.

Spring. (n.d.-a). *Spring Boot reference documentation*. https://docs.spring.io/spring-boot/index.html

Spring. (n.d.-b). *Spring Framework reference documentation*. https://docs.spring.io/spring-framework/reference/index.html

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., & Stoica, I. (2016). Apache Spark: A unified engine for *big data* processing. *Communications of the ACM*, *59*(11), 56–65. https://doi.org/10.1145/2934664

PROFIL PENULIS



Muhamad Yusup, S.Kom.,M.Kom.

Ketertarikan penulis terhadap bidang ilmu komputer dimulai sejak tahun 2007. Minat tersebut mendorong penulis untuk menempuh pendidikan di SMAN 4 Tangerang dengan memilih jurusan Ilmu Pengetahuan Alam (IPA) dan berhasil menyelesaikan pendidikan menengah pada tahun yang sama. Selanjutnya, penulis melanjutkan studi ke jenjang pendidikan tinggi dan berhasil meraih gelar Sarjana Komputer (S.Kom.) dari Program Studi Sistem Informasi di STMIK Raharja. Tidak berhenti sampai di situ, penulis kemudian melanjutkan pendidikan magister di Program Studi Ilmu Komputer Universitas Budi Luhur dan berhasil meraih gelar Magister Komputer (M.Kom.) pada tahun 2014. Saat ini, penulis sedang menempuh pendidikan program doktor (S3) di Program Studi Ilmu Komputer Universitas Kristen Satya Wacana, Salatiga.

Penulis memiliki kepakaran di bidang Sains Data melalui workshop digitalent Departemen Kominfo RI selama 3 bulan. Saat ini penulis aktif sebagai dosen pada Program Studi Bisnis Digital Fakultas Ekonomi dan Bisnis Universitas Raharja. Dalam perannya sebagai akademisi, penulis terus mengembangkan keilmuan melalui kegiatan penelitian, pengajaran, dan penulisan ilmiah yang berfokus pada bidang kepakarannya. Penulis memiliki komitmen untuk berkontribusi dalam pengembangan teknologi dan pendidikan di Indonesia, khususnya dalam era transformasi digital.

Email Penulis: yusup@raharja.info

BAB 16
STUDI KASUS TERAPAN *DATA MINING* INDUSTRI

Dedeh Supriyanti, S.Kom., M.TI.
Universitas Raharja

Pendahuluan

Perkembangan teknologi *digital* telah mendorong industri menghasilkan data dalam jumlah yang semakin besar dari berbagai aktivitas operasional, mulai dari proses produksi, transaksi, persediaan, kualitas produk, pemeliharaan mesin, hingga interaksi dengan pelanggan. Data tersebut tidak lagi hanya berfungsi sebagai catatan operasional, tetapi dapat dimanfaatkan sebagai sumber informasi dan pengetahuan untuk mendukung pengambilan keputusan. Dalam kondisi persaingan industri yang semakin dinamis, kemampuan organisasi dalam mengolah data menjadi informasi yang bernilai menjadi salah satu faktor penting dalam meningkatkan efisiensi, produktivitas, kualitas, dan daya saing. Data *mining* merupakan salah satu pendekatan yang dapat digunakan untuk menemukan pola, hubungan, kecenderungan, dan pengetahuan yang tersembunyi di dalam kumpulan data. Witten, Frank, Hall, dan Pal (2025) menjelaskan bahwa *data mining* mencakup berbagai teknik untuk memperoleh pengetahuan dari data melalui proses seperti *classification*, *regression*, *clustering*, *association rules*, dan visualisasi data. Penerapannya tidak berhenti pada pengolahan data, tetapi diarahkan untuk menghasilkan pengetahuan yang dapat digunakan dalam penyelesaian masalah nyata. Dalam lingkungan industri, penerapan *data mining* menjadi semakin relevan seiring berkembangnya konsep Industry 4.0, *Internet of Things* (IoT), kecerdasan buatan, dan sistem produksi berbasis data. Mesin dan peralatan industri dapat menghasilkan data secara terus-menerus melalui sensor, sehingga perusahaan memiliki peluang untuk melakukan analisis terhadap kondisi peralatan dan proses produksi secara lebih cepat. Esteban dan kolega menunjukkan bahwa kemajuan *artificial intelligence* dan *data mining* memungkinkan pengolahan data dalam jumlah besar untuk menghasilkan model prediktif yang dapat digunakan dalam *predictive maintenance*. Pendekatan tersebut bertujuan mengoptimalkan waktu dan jenis pemeliharaan sehingga ketersediaan sistem dapat ditingkatkan dan sumber daya pemeliharaan dapat digunakan secara lebih efisien (Esteban, C., *et al.* 2022). Salah satu contoh penerapan *data mining* yang penting dalam

industri adalah prediksi dan diagnosis kerusakan mesin. Kerusakan peralatan dapat menyebabkan berhentinya proses produksi, meningkatnya biaya operasional, keterlambatan produksi, serta menurunnya produktivitas. Melalui pemanfaatan data historis dan data sensor, teknik *machine learning* dapat digunakan untuk mengidentifikasi pola yang berkaitan dengan kondisi normal maupun kondisi kerusakan.

Wuest dan kolega (2016) menunjukkan bahwa metode *machine learning* memiliki kemampuan untuk menangani data yang kompleks dan multivariat dalam lingkungan industri, sehingga dapat dimanfaatkan dalam berbagai aplikasi *predictive maintenance*. Selain pemeliharaan mesin, *data mining* juga dapat diterapkan untuk pengendalian dan peningkatan kualitas produk. Data yang berasal dari proses produksi dapat dianalisis untuk menemukan faktor-faktor yang berkaitan dengan produk cacat, mengelompokkan karakteristik produk, serta mengidentifikasi pola tertentu yang sulit ditemukan melalui pengamatan manual. Dengan demikian, *data mining* dapat membantu perusahaan mengembangkan proses pengawasan kualitas yang lebih berbasis bukti dan data. Kajian mengenai penerapan *machine learning* pada lingkungan manufaktur menunjukkan bahwa metode tersebut dapat digunakan untuk mengidentifikasi hubungan nonlinier yang kompleks dalam permasalahan industri, meskipun penerapan pada kondisi nyata tetap menghadapi tantangan seperti perubahan karakteristik data (*concept drift*). Penerapan lainnya dapat ditemukan pada optimasi proses produksi dan pengambilan keputusan. Data produksi yang dikumpulkan secara berkala dapat digunakan untuk mengetahui pola permintaan, tingkat penggunaan bahan baku, waktu produksi, performa mesin, tingkat cacat produk, hingga produktivitas tenaga kerja. Hasil analisis tersebut dapat menjadi dasar bagi manajemen dalam menentukan strategi operasional. Dalam konteks industri 4.0, arus data dari sensor dan mesin yang semakin besar menjadikan analisis berbasis data semakin penting untuk menjaga kesinambungan proses produksi, mengurangi gangguan, dan meningkatkan kinerja sistem. Studi kasus terapan menjadi penting dalam pembelajaran *data mining* karena memungkinkan konsep dan algoritma yang dipelajari dikaitkan secara

Daftar Pustaka

Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.

Choudhary, A. K., Harding, J. A., & Tiwari, M. K. (2009). Data Mining In Manufacturing: A Review Based on The Kind of Knowledge. *Journal of Intelligent Manufacturing,* 20(5), 501–521. https://doi.org/10.1007/s10845-008-0145-x.

Dogan, A., & Birant, D. (2021). *Machine Learning And Data Mining In Manufacturing. Expert Systems With Applications*, 166, 114060. https://doi.org/10.1016/j.eswa.2020.114060.

Duan, X., Vasudevan, A., Bekar, E. T., Gandhi, K., Skoogh, A., et al. (2022). *A Data Scientific Approach Towards Predictive Maintenance Application in Manufacturing Industry*. Advances in Transdisciplinary Engineering. https://doi.org/10.3233/ATDE22014.

Esteban, C., et al. (2022). *Data Mining In Predictive Maintenance Systems: A Taxonomy And Systematic Review*. Wiley Interdisciplinary Reviews: Data Mining And Knowledge Discovery. https://doi.org/10.1002/widm.1471. (Wiley Interdisciplinary Reviews).

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts And Techniques (3rd ed.).* Morgan Kaufmann.

Khan, C. N., & Kribs, J. D. (2026). Data-Driven Analytics In Smart Manufacturing: A Review of Methods, Applications, And Adoption Constraints. *The International Journal of Advanced Manufacturing Technology,* 144, 3061–3076.

Köksal, G., Batmaz, İ., & Testik, M. C. (2011). A Review of Data Mining Applications For Quality Improvement In Manufacturing Industry. *Expert Systems With Applications*, 38(10), 13448–13467. https://doi.org/10.1016/j.eswa.2011.04.063.

Mallioris, P., Aivazidou, E., & Bechtsis, D. (2024). Predictive Maintenance In Industry 4.0: A Systematic Multi-Sector Mapping. *CIRP Journal of Manufacturing Science And Technology, 50*, 80–103. https://doi.org/10.1016/j.cirpj.2024.02.003. (DOI).

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint For Data Mining. *Journal of Data Warehousing, 5*(4), 13–22.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools And Techniques (4th ed.).* Morgan Kaufmann. (Elsevier Educate).

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & Foulds, J. (2025). *Data Mining: Practical Machine Learning Tools And Techniques (5th ed.).* Morgan Kaufmann/Elsevier. (Elsevier Educate).

Wuest, T., Weimer, D., Irgens, C., & Thoben, K. D. (2016). Machine Learning In Manufacturing: Advantages, Challenges, And Applications. *Production & Manufacturing Research, 4*(1), 23–45.

Zonta, T., da Costa, C. A., da Rosa Righi, R., de Lima, M. J., da Trindade, E. S., & Li, G. (2020). Predictive Aaintenance In The Industry 4.0: A Systematic Literature Review. *Computers & Industrial Engineering, 150*, 106889. (ScienceDirect).

PROFIL PENULIS



Dedeh Supriyanti, S.Kom., M.TI.

Penulis merupakan seorang akademisi, Mulai Lulusan Magister Teknik Informatika Di Universitas Raharja, yang memiliki ketertarikan pada bidang pendidikan, teknologi informasi, pengembangan kurikulum, dan pengembangan program studi**.** Dalam menjalankan aktivitas akademiknya, penulis aktif dalam kegiatan pengajaran, penelitian, pengembangan bahan ajar, serta berbagai kegiatan pengembangan dan peningkatan mutu Pendidikan, Selain kegiatan pembelajaran, penulis juga memiliki pengalaman dalam melakukan kajian dan pengembangan kurikulum agar sesuai dengan perkembangan ilmu pengetahuan, teknologi, kebutuhan dunia kerja, serta tuntutan kompetensi lulusan.

Kajian kurikulum dilakukan melalui analisis kebutuhan, capaian pembelajaran, struktur mata kuliah, bahan kajian, metode pembelajaran, sistem asesmen, serta keterkaitan antara kompetensi lulusan dengan kebutuhan masyarakat dan dunia industri, penulis juga berkontribusi dalam pengembangan program studi**,** mulai dari perencanaan program akademik, pengembangan kurikulum, penyusunan dokumen akademik, pengembangan kegiatan pembelajaran, hingga peningkatan mutu dan relevansi program studi. Pengembangan tersebut diarahkan untuk membangun program studi yang adaptif terhadap perubahan teknologi, memiliki tata kelola akademik yang baik, serta mampu menghasilkan lulusan yang memiliki kompetensi sesuai dengan kebutuhan zaman. Dalam bidang teknologi, penulis memiliki minat pada teknologi informasi, *data mining*, kecerdasan buatan, sistem informasi, dan pemanfaatan teknologi dalam pendidikan**.** Penulis berupaya mengintegrasikan perkembangan teknologi dengan kebutuhan pembelajaran dan pengembangan pendidikan melalui pendekatan yang aplikatif dan berbasis data. Melalui berbagai kegiatan akademik dan karya yang

dihasilkan, penulis berkomitmen untuk mendukung peningkatan kualitas pendidikan melalui inovasi pembelajaran, pengembangan kurikulum, penguatan program studi, penelitian, dan pemanfaatan teknologi**.** Penulis berharap setiap karya yang dihasilkan dapat memberikan manfaat bagi mahasiswa, pendidik, akademisi, praktisi, serta masyarakat secara luas.

Email Penulis: dedeh@raharja.info.

DATA MINING TERAPAN

Konsep, Algoritma, dan Implementasi Berbasis Java

Di era transformasi digital saat ini, data telah menjadi aset strategis yang sangat berharga bagi berbagai sektor, mulai dari bisnis, pendidikan, kesehatan, hingga pemerintahan. Namun, tumpukan data yang masif (*big data*) tidak akan memberikan dampak yang signifikan tanpa adanya teknik dan metode yang tepat untuk mengekstrak informasi serta pengetahuan tersembunyi di dalamnya. Di sinilah peran krusial data mining bekerja. Buku ini hadir sebagai jawaban atas kebutuhan literatur yang komprehensif di bidang teknologi informasi. Berbeda dengan referensi yang hanya berfokus pada pendekatan teoretis, buku ini dirancang secara khusus untuk menjembatani kesenjangan antara teori dan praktik. Pembahasan dimulai dari fundamental konsep data mining, eksplorasi algoritma-algoritma utama seperti klasifikasi (misalnya *Decision Tree, Naïve Bayes*), klastering (*K-Means, Hierarchical*), hingga aturan asosiasi (*Apriori*). Nilai tambah utama dari buku ini terletak pada pendekatan implementatifnya yang menggunakan bahasa pemrograman Java. Pemilihan Java didasarkan pada portabilitasnya yang tinggi, keandalannya dalam menangani sistem berskala besar (*enterprise*), serta ketersediaan pustaka (*library*) yang luas untuk pemrosesan data. Setiap algoritma yang dibahas disertai dengan contoh kasus, logika penyelesaian, dan potongan kode (*source code*) Java yang dapat langsung dipraktikkan oleh pembaca.



sada kurnia pustaka

ISBN 978-634-7949-16-8

9 786347 949158