

# DATA MINING

## TEORI DAN IMPLEMENTASI



- Sigrid Tamriesfatno
- La Ode Muh. Armadi. AM
- Alman Arapa
- Sulkifli
- Andi Riski Ramadani
- Hafсах. HS
- Chandra Wisnu Nugroho

# **DATA MINING : TEORI DAN IMPLEMENTASI**

**Penulis:**

**Sigrid Tamriesfatno**

**La Ode Muh. Armadi. AM**

**Alman Arapa**

**Sulkifli**

**Andi Riski Ramadani**

**Hafsah. HS**

**Chandra Wisnu Nugroho**



**CV DIGITAL PUBLISHING LIBRARY**



## **DATA MINING : TEORI DAN IMPLEMENTASI**

### **Penulis :**

Sigrid Tamriesfatno

La Ode Muh. Armadi. AM

Alman Arapa

Sulkifli

Andi Riski Ramadani

Hafsah. HS

Chandra Wisnu Nugroho

**ISBN :** 978-634-04-6997-4

**Editor :** Alzukri Maiwisra, S.Tr.Kes

**Penyunting :** Afridon, ST. M.Si

**Desain Sampul dan Tata Letak :** Putri, S.E

**Penerbit :** DIGITAL PUBLISHING LIBRARY

### **Redaksi :**

Pasar Taratak, Kecamatan Sutera, Kabupaten Pesisir Selatan, Provinsi Sumatera Barat.

Website : <https://dplibrary.id/>

Email : [dplibrary.id@gmail.com](mailto:dplibrary.id@gmail.com)

Cetakan pertama, Januari 2025

Hak cipta dilindungi undang-undang Dilarang memperbanyak karya tulis ini dalam bentuk dan dengan cara apapun tanpa izin tertulis dari penerbit.

## KATA PENGANTAR

Dengan mengucapkan puji syukur kehadiran Allah SWT, atas limpahan rahmat dan hidayahNya, maka penulisan buku dengan judul **“Data Mining: Teori dan Implementasi”** dapat diselesaikan. Buku ini disusun sebagai bahan ajar dan referensi bagi mahasiswa, dosen, serta praktisi yang ingin memahami konsep dasar hingga penerapan teknik data mining secara sistematis dan aplikatif.

Buku ini masih banyak kekurangan dalam penyusunannya. Oleh karena itu, kami sangat mengharapkan kritik dan saran demi perbaikan dan kesempurnaan buku ini selanjutnya. Kami mengucapkan terima kasih kepada berbagai pihak yang telah membantu dalam proses penyelesaian Buku ini. Semoga Buku ini dapat menjadi sumber referensi dan literatur yang mudah dipahami.

Padang, Januari 2025

Penulis

# DAFTAR ISI

<b>KATA PENGANTAR .....</b>	<b>i</b>
<b>DAFTAR ISI .....</b>	<b>ii</b>
<b>DAFTAR GAMBAR .....</b>	<b>v</b>
<b>DAFTAR TABEL .....</b>	<b>vi</b>
<b>BAB 1 PENDAHULUAN DATA MINING .....</b>	<b>1</b>
1.1 Pendahuluan .....	1
1.2 Definisi Data Mining .....	3
1.3 Tujuan dan Manfaat Data Mining .....	6
1.4 Proses Dalam Data Mining (KDD Process) .....	9
1.5 Ruang Lingkup Data Mining .....	11
<b>BAB 2 KONSEP DASAR DATA MINING .....</b>	<b>15</b>
2.1 Pengantar Era Data .....	15
2.2 Apa itu Data Mining .....	18
2.3 Cara Kerja Data Mining .....	23
<b>BAB 3 PREPROCESSING DATA NOMINAL .....</b>	<b>33</b>
3.1 Pendahuluan .....	34
3.2 Karakteristik Data Nominal: Lebih Dari Sekadar Label .....	35

3.3 Tahapan Umum Preprocessing: Sebuah Orkestrasi Sistematis .....	36
3.4 Penanganan Missing pada Data Nominal: Seni Mengelola Ketidakterlengkapan .....	39
3.5 Teknik Encoding untuk Data Nominal: Dari Sederhana hingga Canggih .....	41
3.6 Menghadapi High-Cardinality dan Kategori Langka: Seni Berkompromi dengan Kompleksitas .....	46
3.7 Feature Selection & Evaluasi untuk Variabel Kategori: Beyond Correlation .....	48
3.8 Praktik Terbaik & Checklist Preprocessing Nominal: Dari Teori ke Praktek .....	49
3.9 Ringkasan dan Rekomendasi: The Strategic Landscape .....	51
<b>BAB 4 KONSEP DAN METODE <i>CLUSTERING</i> .....</b>	<b>65</b>
4.1 Definisi <i>Clustering</i> .....	65
4.2 Metode <i>Clustering</i> .....	68
4.3 Penerapan <i>Clustering</i> dalam kehidupan sehari-hari .....	81
<b>BAB 5 KONSEP DAN METODE KLASIFIKASI .....</b>	<b>85</b>
5.1 Pendahuluan .....	85

5.2 Konsep Dasar Klasifikasi.....	86
5.3 Metode Klasifikasi .....	89
5.4 Prapemrosesan dan Rekayasa Fitur .....	91
5.5 Evaluasi Model Klasifikasi.....	92
5.6 Strategi Penyetelan Hiperparameter (Tuning) .....	92
5.7 Interpretabilitas, Etika, dan Fairness .....	93
5.8 Studi Kasus Implementasi: Pipeline Klasifikasi (Python) .....	93
5.9 Rangkuman .....	94
<b>BAB 6 EVALUASI DAN VALIDASI MODEL <i>DATAMINING</i>.....</b>	<b>99</b>
6.1 Konsep Dasar Evaluasi dan Validasi Model Datamining....	100
6.2 Metrik Evaluasi Untuk Klasifikasi .....	103
6.3 Metrik Evaluasi Untuk Regresi .....	108
6.4 Metrik Evaluasi Untuk <i>Clustering</i> .....	112
6.5 Contoh Studi Kasus .....	115
<b>BAB 7 KEAMANAN DATA DAN PRIVASI .....</b>	<b>127</b>
7.1 Konsep Dasar Keamanan Data.....	127
7.2 Privasi dalam Data Mining .....	137
<b>BIODATA PENULIS</b>	

## DAFTAR GAMBAR

Gambar 3.1. Proses Preprocessing.....	58
Gambar 4.1. simulasi single linkage.....	70
Gambar 4.2. ilustrasi perhitungan complete .....	71
Gambar 4.3. ilustrasi simulasi average linkage.....	72
Gambar 4.4. ilustrasi simulasi Distance Between Centroid .....	73
Gambar 4.5. simulasi K-Means.....	74
Gambar 4.6. Core Point, Border Point, dan Outlier dengan MintPts = 4 dan $\epsilon$ = .....	79
Gambar 5.1. Skema pohon keputusan (Decision Tree) .....	95
Gambar 5.2. Garis pemisah SVM dengan margin dan support vectors.....	95

## DAFTAR TABEL

Tabel 5.1. Perbandingan Metode Klasifikasi .....	91
Tabel 5.2. Evaluasi Model Klasifikasi .....	92
Tabel 6.1. Struktur <i>confussion matrix</i> .....	105

# BAB 1

## PENDAHULULAN DATA MINING

Oleh Sigrid Tamriesfatno

### 1.1 Pendahuluan

Pentingnya data dalam era digital tidak bisa diragukan lagi, karena data telah menjadi aset berharga yang setara dengan sumber daya ekonomi lainnya. Di era transformasi digital, hampir semua aktivitas manusia mulai dari berbelanja, belajar, bekerja, hingga berinteraksi sosial menghasilkan jejak data yang terus meningkat jumlahnya setiap detik. Perusahaan, institusi pendidikan, lembaga pemerintahan, hingga sektor kesehatan menggunakan data sebagai dasar pengambilan keputusan yang lebih cepat, tepat, dan akurat. Data yang diolah dengan tepat akan mampu memberikan pengetahuan baru mengenai perilaku pengguna, tren pasar, efisiensi operasional, serta peluang inovasi. Dengan kata lain, data adalah "minyak baru" yang menggerakkan perkembangan teknologi modern, seperti kecerdasan buatan (AI) dan sistem berbasis big data. Jika data tidak dianalisis dengan tepat, organisasi akan kesulitan bersaing di era global.

Namun, tantangan dalam pengelolaan dan pemanfaatan data yang besar (big data) juga semakin kompleks seiring dengan pertumbuhannya yang eksponensial. Volume data yang sangat besar, variasi format data yang beragam (teks, gambar, video, suara, sensor), serta kecepatan aliran data yang terus-menerus menjadi hambatan utama dalam proses pengolahan. Selain itu, isu keamanan dan privasi data juga menjadi perhatian penting karena kebocoran informasi dapat merugikan individu maupun organisasi. Infrastruktur teknologi yang memadai, seperti kapasitas penyimpanan yang besar, sistem komputasi yang cepat, serta algoritma yang efisien, mutlak diperlukan agar data benar-benar dapat dimanfaatkan dengan maksimal. Di sisi lain, kualitas data yang buruk, seperti data yang tidak lengkap, redundan, atau bias, dapat menghasilkan analisis yang menyesatkan sehingga mengurangi nilai strategis dari data itu sendiri. Oleh karena itu, pengelolaan data membutuhkan pendekatan yang terintegrasi, mulai dari pengumpulan, pembersihan, penyimpanan, hingga analisis.

Dalam konteks inilah muncul kebutuhan terhadap teknik penggalian informasi dari data atau data mining sebagai solusi untuk memanfaatkan data secara efektif. Data mining memungkinkan ditemukannya pola, tren, dan hubungan tersembunyi dalam kumpulan data yang besar dan kompleks, yang tidak mungkin diperoleh hanya dengan analisis manual. Teknik ini