

**Editor :**  
**Tundo, S.Kom., M.Kom.**



**FUTURE SCIENCE**

# **PENERAPAN APLIKASI SEDERHANA DATA MINING** (dengan Bahasa Pemrograman Python)

**Penulis :**

Norbertus Tri Suswanto Saptadi | Ade Oktarino | Deddy Kurniawan  
Tio Dharmawan | Erna Hudianti Pujiarini | Aditya Febrianto  
Nelly Oktavia Adiwijaya | Sofiansyah Fadli | Dwi Remawati  
Deborah Kurniawati | Saikin | Dadang Iskandar Mulyana



Bunga Rampai

**PENERAPAN APLIKASI SEDERHANA  
DATA MINING  
(dengan Bahasa Pemrograman Python)**

## **UU No 28 tahun 2014 tentang Hak Cipta**

### **Fungsi dan sifat hak cipta Pasal 4**

Hak Cipta sebagaimana dimaksud dalam Pasal 3 huruf a merupakan hak eksklusif yang terdiri atas hak moral dan hak ekonomi.

### **Pembatasan Pelindungan Pasal 26**

Ketentuan sebagaimana dimaksud dalam Pasal 23, Pasal 24, dan Pasal 25 tidak berlaku terhadap:

- i. Penggunaan kutipan singkat Ciptaan dan/atau produk Hak Terkait untuk pelaporan peristiwa aktual yang ditujukan hanya untuk keperluan penyediaan informasi aktual;
- ii. Penggandaan Ciptaan dan/atau produk Hak Terkait hanya untuk kepentingan penelitian ilmu pengetahuan;
- iii. Penggandaan Ciptaan dan/atau produk Hak Terkait hanya untuk keperluan pengajaran, kecuali pertunjukan dan Fonogram yang telah dilakukan Pengumuman sebagai bahan ajar; dan
- iv. Penggunaan untuk kepentingan pendidikan dan pengembangan ilmu pengetahuan yang memungkinkan suatu Ciptaan dan/atau produk Hak Terkait dapat digunakan tanpa izin Pelaku Pertunjukan, Produser Fonogram, atau Lembaga Penyiaran.

### **Sanksi Pelanggaran Pasal 113**

1. Setiap Orang yang dengan tanpa hak melakukan pelanggaran hak ekonomi sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf i untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 1 (satu) tahun dan/atau pidana denda paling banyak Rp100.000.000 (seratus juta rupiah).
2. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf c, huruf d, huruf f, dan/atau huruf h untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 3 (tiga) tahun dan/atau pidana denda paling banyak Rp500.000.000,00 (lima ratus juta rupiah).

# PENERAPAN APLIKASI SEDERHANA DATA MINING (dengan Bahasa Pemrograman Python)

Penulis:

Norbertus Tri Suswanto Saptadi  
Ade Oktarino  
Deddy Kurniawan  
Tio Dharmawan  
Erna Hudianti Pujiarini  
Aditya Febrianto  
Nelly Oktavia Adiwijaya  
Sofiansyah Fadli  
Dwi Remawati  
Deborah Kurniawati  
Saikin  
Dadang Iskandar Mulyana

Editor:

Tundo, S.Kom, M.Kom.



# **PENERAPAN APLIKASI SEDERHANA DATA MINING**

**(dengan Bahasa Pemrograman Python)**

**Penulis:**

**Norbertus Tri Suswanto Saptadi  
Ade Oktarino  
Deddy Kurniawan  
Tio Dharmawan  
Erna Hudianti Pujiarini  
Aditya Febrianto  
Nelly Oktavia Adiwijaya  
Sofiansyah Fadli  
Dwi Remawati  
Deborah Kurniawati  
Saikin  
Dadang Iskandar Mulyana**

**Editor: Tundo, S.Kom., M.Kom.**

**Desain Cover: Nada Kurnia, S.I.Kom.**

**Tata Letak: Samuel, S.Kom.**

**Halaman: A5 Unesco (15,5 x 23 cm)**

**Ukuran: xii, 214**

**e-ISBN: 978-634-7037-89-3**

**Terbit Pada: Februari 2025**

Hak Cipta 2025, Pada Penulis

Isi diluar tanggung jawab percetakan

---

**Copyright © 2025 by Future Science Publisher**  
All Right Reserved

Hak cipta dilindungi undang-undang  
Dilarang keras menerjemahkan, memfotokopi, atau  
memperbanyak sebagian atau seluruh isi buku ini  
tanpa izin tertulis dari Penerbit.

**PENERBIT FUTURE SCIENCE  
(CV. FUTURE SCIENCE)**  
Anggota IKAPI (348/JTI/2022)

Jl. Terusan Surabaya Gang 1 A No. 71 RT 002 RW 005, Kel. Summersari, Kec. Lowokwaru, Kota  
Malang, Provinsi Jawa Timur.  
[www.futuresciencepress.com](http://www.futuresciencepress.com)

## **KATA PENGANTAR**

Puji syukur kami panjatkan kepada Tuhan Yang Maha Esa atas rahmat dan karunia-Nya sehingga buku **PENERAPAN APLIKASI SEDERHANA DATA MINING** (dengan Bahasa Pemrograman Python) dapat terselesaikan dengan baik. Buku ini diharapkan dapat memberikan kontribusi yang berarti bagi pengembangan ilmu dan praktik dalam bidang ilmu komputer di Indonesia khususnya pada lingkup data mining.

Buku ini merupakan penerapan aplikasi sederhana teknik data mining dengan bahasa pemrograman python yang dirancang untuk pemula yang ingin belajar teknik data mining dalam aplikasinya. Terdiri dari 12 penulis dengan judul bab Pengantar Data Mining, Instalasi dan Library Pemrograman Python untuk Konsep Data Mining, Aplikasi Klasifikasi K-NN dengan Python, Aplikasi Klasifikasi Naïve Bayes dengan Python, Aplikasi Klasifikasi Decision Tree C 4.5 dengan Python, Aplikasi Forecasting Moving Average dengan Python, Aplikasi Forecasting Exponential Smoothing dengan Python, Aplikasi Forecasting Regresi Linear dengan Python, Aplikasi Forecasting ARIMA dengan Python, Aplikasi Association Apriori dengan Python, Aplikasi Clustering K-Means dengan Python, Aplikasi Clustering C-Means dengan Python.

Kami ingin mengucapkan terima kasih yang sebesar-besarnya kepada Penerbit Future Science yang telah memberikan dukungan dan kesempatan untuk menerbitkan buku ini.

Dukungan dari penerbit sangat berarti bagi kami dalam proses penyusunan dan penerbitan buku ini. Ucapan terima kasih juga kami sampaikan kepada para penulis yang telah berkontribusi dengan pengetahuan, pengalaman, dan dedikasi mereka.

Semoga buku ini dapat memberikan manfaat yang besar bagi para pembaca dan menjadi referensi yang berharga dalam penguasaan Aplikatif Data Mining dengan Bahasa Pemrograman Python. Terima kasih atas perhatian dan dukungannya.

Jakarta, Desember 2024

Editor,

Tundo

## DAFTAR ISI

KATA PENGANTAR.....	v
DAFTAR ISI .....	vii
BAB 1 PENGANTAR DATA MINING .....	1
Norbertus Tri Suswanto Saptadi .....	1
DEFINISI DAN KONSEP DASAR DATA MINING .....	1
SEJARAH DAN EVOLUSI DATA MINING .....	3
PROSES DATA MINING .....	4
TEKNIK-TEKNIK DATA MINING.....	6
APLIKASI DATA MINING .....	8
TANTANGAN DAN MASA DEPAN DATA MINING ...	10
KESIMPULAN.....	12
BAB 2 INSTALASI DAN LIBRARY PEMROGRAMAN PYTHON UNTUK KONSEP DATA MINING .....	17
Ade Oktarino.....	17
PENDAHULUAN .....	17
INSTALASI PYTHON.....	21
PUSTAKA PYTHON UNTUK DATA MINING .....	21
CONTOH WORKFLOW DATA MINING DENGAN PYTHON .....	25
TIPS DAN SUMBER DAYA BELAJAR DATA MINING.....	26
KESIMPULAN.....	26
BAB 3 APLIKASI KLASIFIKASI K-NN DENGAN PYTHON...	29
Deddy Kurniawan .....	29



	PENDAHULUAN .....	29
	KONSEP DASAR.....	30
	CARA KERJA KNN .....	31
	KELEBIHAN KNN .....	32
	KEKURANGAN KNN .....	33
	WAKTU PENGGUNAAN KNN .....	33
	PENERAPAN ALGORITMA KNN DENGAN PYTHON	34
	KESIMPULAN.....	41
BAB 4	APLIKASI KLASIFIKASI NAÏVE BAYES DENGAN PYTHON .....	45
	Tio Dharmawan.....	45
	PENDAHULUAN .....	45
	TEORI NAÏVE BAYES .....	46
	NAÏVE BAYES: DATA KATEGORIKAL .....	47
	NAÏVE BAYES: DATA NUMERIK .....	50
	NAÏVE BAYES: KODE PROGRAM MENGGUNAKAN SKLEARN .....	52
	KESIMPULAN.....	58
BAB 5	KLASIFIKASI DECISION TREE C 4.5 DENGAN PYTHON .....	61
	Erna Hudianti Pujiarini .....	61
	PENDAHULUAN .....	61
	ALGORITMA DECISION TREE C4.5 .....	64
	EVALUASI DAN VALIDASI MODEL .....	69
	IMPLEMENTASI DECISION TREE C4.5 DENGAN PYTHON .....	72
	KESIMPULAN.....	77

<b>BAB 6</b>	<b>APLIKASI FORECASTING MOVING AVERAGE DENGAN PYTHON .....</b>	<b>79</b>
	Aditya Febrianto .....	79
	PENDAHULUAN .....	79
	PENGANTAR FORECASTING MOVING AVERAGE ..	80
	PEMBUATAN ALGORITMA FORECASTING MOVING AVERAGE DENGAN PYTHON .....	84
	PEMBUATAN APLIKASI FORECASTING MOVING AVERAGE MENGGUNAKAN PYTHON .....	90
	KESIMPULAN.....	100
<b>BAB 7</b>	<b>APLIKASI FORECASTING EXPONENTIAL SMOOTHING DENGAN PYTHON .....</b>	<b>103</b>
	Nelly Oktavia Adiwijaya .....	103
	PENDAHULUAN .....	103
	DEFINISI DAN KONSEP DASAR FORECASTING.....	104
	Pengenalan Exponential Smoothing .....	106
	IMPLEMENTASI EXPONENTIAL SMOOTHING DENGAN PYTHON .....	110
	STUDI KASUS.....	112
	Proses Pembersihan dan Persiapan Data.....	112
	EVALUASI MODEL DAN HASIL .....	117
	Menghitung Error untuk Model SES, DES, dan TES.....	118
	Interpretasi Hasil Evaluasi .....	119
	Visualisasi Perbandingan Hasil.....	120
	Kesimpulan dari Evaluasi Model.....	121
	KESIMPULAN.....	121

BAB 8	APLIKASI FORECASTING REGRESI LINEAR DENGAN PYTHON.....	125
	Sofiansyah Fadli.....	125
	PENDAHULUAN .....	125
	PERSIAPAN DATA.....	127
	PENERAPAN FORECASTING REGRESI LINEAR .....	130
	KESIMPULAN .....	138
BAB 9	APLIKASI FORECASTING ARIMA DENGAN PHYTON .....	143
	Dwi Remawati.....	143
	PENDAHULUAN .....	143
	KONSEP DASAR.....	144
	Stasioneritas dan Nonstasioneritas .....	144
	Stasioneritas .....	145
	Nonstasioneritas .....	145
	Cara Mengatasi Nonstasioneritas .....	146
	Autocorrelation Function (ACF) dan Fungsi Autokorelasi Parsial (PACF) .....	146
	KELEBIHAN .....	148
	KEKURANGAN .....	148
	PENERAPAN FORECASTING ARIMA DENGAN PYTHON .....	149
	KESIMPULAN .....	157
BAB 10	APLIKASI ASSOCIATION APRIORI DENGAN PYTHON .....	161
	Deborah Kurniawati .....	161
	ATURAN ASOSIASI .....	161

ALGORITMA APRIORI.....	163
Definisi Algoritma Apriori .....	163
Cara Kerja Algoritma Apriori .....	163
Kelebihan Algoritam Apriori .....	166
Kekurangan Algoritma Apriori .....	166
DATASET UNTUK APRIORI .....	167
IMPLEMENTASI ASOSIASI APRIORI DENGAN PYTHON .....	168
Dataset berada pada program .....	168
Dataset dari file .....	176
KESIMPULAN.....	180
<b>BAB 11 APLIKASI CLUSTERING K-MEANS DENGAN PYTHON .....</b>	<b>183</b>
Saikin .....	183
PENDAHULUAN .....	183
K-Means.....	184
PENERAPAN K-MEANS MENGGUNAKAN PYTHON .....	185
KESIMPULAN.....	195
<b>BAB 12 APLIKASI CLUSTERING C-MEANS DENGAN PYTHON .....</b>	<b>199</b>
Dadang Iskandar Mulyana .....	199
PENDAHULUAN .....	199
METODE-METODE DALAM CLUSTERING .....	199
PENGERTIAN C-MEANS CLUSTERING (FUZZY C- MEANS) .....	203
PRINSIP DASAR FUZZY C-MEANS / FCM.....	203

LANGKAH-LANGKAH PENERAPAN ALGORITMA	
FUZZY C-MEANS.....	205

# **BAB 1**

## **PENGANTAR DATA MINING**

Norbertus Tri Suswanto Saptadi  
Universitas Atma Jaya Makassar, Makassar  
E-mail: ntsaptadi@gmail.com

### **DEFINISI DAN KONSEP DASAR DATA MINING**

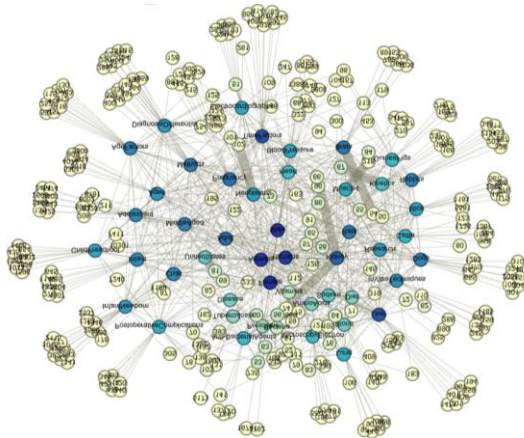
*Data mining* atau penambangan data merupakan tahapan eksplorasi dan analisis data dalam jumlah banyak dan besar yang bertujuan memperoleh suatu pola dan informasi yang berguna (Aras and Sarjono, 2016). Proses mengintegrasikan metode dari kecerdasan buatan (*artificial intelligence*), pembelajaran mesin (*machine learning*), statistik (*statistics*), dan sistem basis data (*database system*) untuk mengidentifikasi tren, hubungan, dan pola yang mungkin tidak tampak jelas dari data mentah yang besar, banyak, dan kompleks. Penggunaan teknik membuat *data mining* dapat mengungkap wawasan bernilai dan berharga yang dapat dimanfaatkan dalam mendukung pengambilan keputusan yang relevan, lebih baik dan strategis.

Teknik dalam *data mining* mencakup berbagai metode analisis, seperti klasifikasi, klusterisasi, asosiasi, dan prediksi (Ginantra *et al.*, 2021). Klasifikasi digunakan dalam mengelompokkan data berdasarkan kategori tertentu melalui fitur yang relevan. Klusterisasi membagi data menjadi kelompok yang mempunyai kesamaan. Analisis asosiasi mencari hubungan antara item dalam *dataset*. Prediksi dimanfaatkan dalam memproyeksikan nilai atau tren masa depan yang bersumber dari data historis. Semua teknik dalam penggunaan *data mining* akan bertujuan untuk mengeksplorasi dan mengekstrak informasi yang berguna dari data yang berjumlah besar, banyak,

dan sering kali dalam kondisi bentuk tidak terstruktur (Witten, Frank, & Hall, 2016).

Proses *data mining* dimulai dengan pemahaman tujuan bisnis yang relevan dan jelas untuk menentukan jenis analisis yang diperlukan. Data dikumpulkan dan dipersiapkan melalui proses pembersihan dan transformasi untuk memastikan kualitas dan konsistensi. Setelah data siap (*ready*), berbagai model *data mining* akan diterapkan untuk menemukan pola dan hubungan yang dapat terjadi. Hasil analisis kemudian dievaluasi untuk memastikan akurasi sebelum diterapkan agar dapat mendukung keputusan bisnis atau strategi.

Dalam era *digital* seperti saat ini keberadaan volume data yang terus berkembang dan meningkat membuat *data mining* menjadi sangat berperan dan menjadi alat yang sangat penting dalam penggunaan berbagai sektor industri. Teknik ini memungkinkan organisasi untuk memanfaatkan data secara efektif, menemukan pola tersembunyi, dan membuat keputusan yang lebih berbasis data dengan kemampuan mengidentifikasi wawasan mendalam untuk meningkatkan efisiensi operasional dan daya saing (Raschka, 2015).



Sumber: <https://runawayhorse001.github.io/DataminingTutorial/>

Gambar 1.1. Data Mining With Python and R

*data mining* membantu dalam optimalisasi dan manajemen persediaan (Chen, Chiang, & Storey, 2012; Jha & Deshmukh, 2016).

Kendati *data mining* menawarkan banyak manfaat, namun tantangan seperti kualitas data yang buruk, masalah privasi, dan interpretasi hasil yang kompleks harus dapat diatasi. Kualitas data yang tidak memadai dan memenuhi standar atau ketentuan dapat mengganggu hasil analisis, sementara masalah privasi memerlukan perhatian ekstra untuk memastikan bahwa data pribadi dilindungi sesuai dengan regulasi yang berlaku. Selain itu, interpretasi hasil yang kompleks dapat menjadi hambatan, terutama bagi pengguna non-teknis. Namun, perkembangan teknologi seperti kecerdasan buatan dan pembelajaran mesin menawarkan harapan baru untuk mengatasi tantangan ini dan memperluas kemampuan *data mining* (Goodfellow, Bengio, & Courville, 2016; Zhou & Pardo, 2020).

Buku ini akan memberikan pembelajaran bagaimana menerapkan teknik-teknik *data mining* dengan menggunakan bahasa pemrograman *Python* untuk membuat aplikasi yang sederhana namun efektif. *Python* merupakan alat yang sangat berguna dalam penggunaan *data mining* berkat pustaka-pustaka yang kuat seperti *Pandas*, *Scikit-Learn*, dan *TensorFlow*. Dengan memahami dan menerapkan teknik-teknik ini diharapkan dapat memanfaatkan potensi untuk menghasilkan wawasan yang berharga dan membuat keputusan yang lebih baik, tepat, dan berguna dalam berbagai konteks aplikasi (Kitchin, 2014).

## **DAFTAR PUSTAKA**

Aras, Z. and Sarjono (2016) ‘Analisis Data Mining Untuk Menentukan Kelompok Prioritas Penerima Bantuan Bedah Rumah Menggunakan Metode Clustering K-Means( Studi Kasus : Kantor Kecamatan Bahar Utara)’, *Jurnal*



- Manajemen Sistem Informasi*, 1(2), pp. 159–170.
- Arta, I.K.J., Indrawan, G. and Rasben Dantes, G. (2019) ‘Data Mining Rekomendasi Calon Mahasiswa Berprestasi di STMIK Denpasar Menggunakan Metode Technique For Other Reference By Similarity to Ideal Solution’, *Jurnal Ilmu Komputer Indonesia (JIKI)*, 4(1), pp. 11–21.
- Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Computing Surveys (CSUR)*, 41(3), 1-58.
- Chen, M., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, 36(4), 1165-1188.
- Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, 51(1), 107-113.
- Gilbert, N. (2014). *Data Mining for Business Intelligence: A Hands-On Guide for Managers*. Wiley.
- Ginantra, N.L.W.S.R. *et al.* (2021) *Data Mining dan Penerapan Algoritma*.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- Gunasekaran, A., Lai, K. H., & Cheng, T. C. E. (2008). Performance of Supply Chain Management Systems: The Case of a Large Multinational Corporation. *International Journal of Production Economics*, 113(1), 61-80.
- Hand, D. J. (2009). Classifier Technology and the Illusion of Progress. *Statistical Science*, 24(1), 23-28.
- Jha, K. N., & Deshmukh, S. G. (2016). Data Mining Techniques for Quality Improvement in Manufacturing Industry. *Journal of Quality in Maintenance Engineering*, 22(4), 400-416.
- Kitchin, R. (2014). Big Data and Human Geography: Opportunities, Challenges and Risks. *Dialogues in Human*

- Geography, 4(3), 262-267.
- Rani, M. U., Kumar, P., & Gupta, S. (2018). Data Mining Techniques for Healthcare Decision Making. *Journal of Health Informatics and Management*, 12(2), 85-93.
- Raschka, S. (2015). *Python Machine Learning*. Packt Publishing.
- Salsabila, F. and Maulida Intani, S. (2021) 'Implementasi Algoritma K-Means dan C4.5 Dalam Menentukan Tingkat Penyebaran Covid-19 di Indonesia', *Jurnal Siliwangi*, 7(1), pp. 25–30.
- Saptadi, N.T.S., Chyan, P. and Leda, J.M. (2023) 'Analysis of Supermarket Product Purchase Transactions With the Association Data Mining Method', *Rekayasa Sistem dan Teknologi Informasi (JURNAL RESTI)*, 5(158), pp. 2–6.
- Witten, I. H., Frank, E., & Hall, M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.
- Zhou, J., & Pardo, T. A. (2020). Privacy-Preserving Data Mining: Challenges and Solutions. *Data Mining and Knowledge Discovery*, 34(2), 198-224.

## PROFIL PENULIS



### **Norbertus Tri Suswanto Saptadi**

Lahir di Cirebon Jawa Barat, tanggal 7 Juni 1975. Memiliki Jabatan Fungsional Lektor Kepala, Pembina Tingkat I (IV/b). Berpendidikan Sarjana Komputer (S.Kom.) di Universitas Teknologi Digital Indonesia (UTDI) tahun 1998, Magister Manajemen (M.M.) di Universitas Hasanuddin (UNHAS) tahun 2004, Magister Teknologi Informasi (M.T.) di Universitas Gadjah Mada (UGM) tahun 2007, Insinyur (Ir.) di Pendidikan Profesi Insinyur UNHAS tahun 2020, Insinyur Profesional Madya (IPM.) di Persatuan Insinyur Indonesia (PII) tahun 2021,

Doktor (Dr.) di Fakultas Teknik UNHAS tahun 2023, dan Program Pendidikan Reguler Angkatan (PPRA) LX Lemhannas RI tahun 2020. Menjadi tenaga pengajar (Dosen) pada Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Atma Jaya Makassar (UAJM). Peraih Poster terbaik DPRM Dikti tahun 2016. Dosen berprestasi IKDKI tahun 2020 dan 2021. Pernah menjabat Kepala UPT Komputer, Kepala BAPSI, Wakil Dekan FT, Dekan FT dan FTI, Wakil Rektor III, Ketua Penjaminan Mutu. Tim PAK Dosen dan Asesor BKD UAJM. Reviewer International Conference dan Jurnal SINTA. Pemenang Hibah Kemdikbud Penelitian Dosen Pemula, Bersaing, Fundamental, dan Strategi Nasional. Penulis artikel media massa Tribun Timur, Koinonia, Bisnis Sulawesi, Sesawi.net, Mirifica.net, HidupKatolikCom, OMKNet, KatolikanaTV, Jalan Hidup Katolik, dll. Penulis Buku di Penerbit Kanisius, Sada Kurnia Pustaka, Aksara Sastra Media, Future Science, HEI Publishing, Mifandi Mandiri Digital, Rey Media Grafika, Widina Salemba, dan Cendikia Mulia Mandiri. Aktifis organisasi IKA Lemhannas RI LX, IARMI, DPP ISKA, BAPOMI Sulsel, LP3KD Sulsel, IKDKI SulselTraBar, Komkep KAMS, Komsos KAMS, PUKAT KAMS, TPP KAMS, FMKI KAMS, UPS KAMS, Pengurus Kebun Sawit Laimbo, FDI, PII Makassar, INAPR, Dewan Keuangan Paroki dan Program Ayo Sekolah Mariso, Animator Laudato Si', dan lain-lain.

## **BAB 2**

# **INSTALASI DAN LIBRARY PEMROGRAMAN PYTHON UNTUK KONSEP DATA MINING**

Ade Oktarino  
Universitas Adiwangsa Jambi, Kota Jambi  
E-mail: adeoktarino@unaja.ac.id

## **PENDAHULUAN**

Dalam era data besar (big data), data mining menjadi salah satu teknik paling penting untuk mendapatkan informasi berharga dari sejumlah besar data yang ada. Pemrograman Python adalah salah satu bahasa yang paling populer untuk melakukan data mining karena memiliki sintaks yang sederhana, pustaka yang kaya, dan komunitas yang besar. Python menyediakan berbagai alat yang memungkinkan kita untuk melakukan analisis data dengan lebih mudah dan cepat. Beberapa alasan mengapa Python populer untuk data mining adalah:

1. Sintaks yang Mudah Dipelajari: Python memiliki sintaks yang sederhana dan mirip dengan bahasa Inggris, sehingga mudah dipahami baik oleh pemula maupun oleh pengguna yang lebih berpengalaman. Python memungkinkan penulisan kode yang bersih dan ringkas, sehingga proses pengembangan dan eksperimen dapat dilakukan lebih cepat dibandingkan dengan bahasa pemrograman lain yang lebih kompleks.
2. Pustaka yang Kaya dan Beragam: Python menawarkan berbagai pustaka yang khusus dirancang untuk menangani data mining, mulai dari manipulasi data hingga penerapan algoritma machine learning. Pustaka seperti Pandas, Numpy, Scikit-Learn, dan TensorFlow sangat membantu

dalam mengimplementasikan konsep data mining. Dengan pustaka ini, pengguna dapat dengan mudah membersihkan data, memvisualisasikan pola, membangun model prediktif, dan bahkan menerapkan deep learning untuk menangani data yang lebih kompleks.

3. Komunitas yang Besar dan Dukungan yang Luas: Python memiliki komunitas yang sangat besar dan aktif. Hal ini berarti pengguna dapat dengan mudah menemukan tutorial, dokumentasi, serta mendapatkan bantuan dari pengguna lain ketika menemui masalah. Selain itu, komunitas yang aktif juga berarti adanya perkembangan pustaka dan alat yang selalu diperbarui, sehingga pengguna dapat terus mengakses teknologi terbaru dan praktik terbaik dalam data mining.
4. Alat yang Kuat untuk Analisis Data: Python didukung oleh berbagai alat yang memungkinkan untuk memproses, memvisualisasikan, dan menganalisis data dengan lebih efektif. Pustaka seperti Pandas dan Numpy membantu dalam manipulasi data yang efisien, sedangkan Matplotlib dan Seaborn membantu dalam memvisualisasikan data agar lebih mudah dipahami. Hal ini menjadikannya sangat cocok untuk data mining, mulai dari analisis sederhana hingga implementasi algoritma yang lebih kompleks. Python juga memungkinkan integrasi dengan alat lain seperti SQL, yang mempermudah dalam mengakses dan menganalisis data dari berbagai sumber.
5. Ekosistem yang Terintegrasi untuk Data Science: Python menyediakan ekosistem yang terintegrasi untuk data science, termasuk data mining. Misalnya, pengguna dapat menggunakan Jupyter Notebook yang merupakan platform interaktif untuk menjalankan kode Python, visualisasi, dan dokumentasi secara bersamaan. Hal ini memungkinkan data scientist untuk mencatat eksperimen mereka dengan cara

## 5. Evaluasi Model

```
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f'Akurasi Model: {accuracy * 100:.2f}%')
```

### TIPS DAN SUMBER DAYA BELAJAR DATA MINING

1. Kaggle: Platform untuk belajar data mining dan machine learning melalui notebook dan kompetisi.
2. Google Colab: Platform berbasis cloud yang dapat digunakan untuk eksperimen Python tanpa perlu instalasi di komputer lokal.
3. Dokumentasi Pustaka: Jangan lupa untuk sering membaca dokumentasi resmi dari pustaka seperti Pandas, Scikit-learn, dan lain-lain agar lebih memahami fungsi yang ada.

### KESIMPULAN

Pemrograman Python sangat kaya akan pustaka yang membantu untuk proses data mining, mulai dari pengolahan data, analisis, hingga visualisasi. Instalasi pustaka seperti **Numpy**, **Pandas**, **Scikit-learn**, dan lainnya sangatlah mudah, baik melalui Anaconda atau menggunakan pip. Pahami penggunaan setiap pustaka sesuai dengan kebutuhan spesifik dalam data mining agar dapat melakukan analisis secara efisien dan efektif.

### DAFTAR PUSTAKA

- Ayu ramadhanty, D., Renita Syafitri, Errissya Rasywir, & Despita Meisak (2022). Implementasi data mining untuk menentukan persediaan stok obat di apotek k-24 menggunakan metode k-means clustering. Jurnal Informatika Dan Rekayasa
- Azhari, P., Taufik, F., & Murniyanti, S. (2023). Penerapan data

- mining untuk analisa transaksi penjualan menggunakan metode apriori. Jurnal Sistem Informasi Triguna Dharma (JURSI TGD), 2(3), 373. <https://doi.org/10.53513/jursi.v2i3.6401>
- Erawati, N. L., Adi Jaya, I. K. N., & Asmarajaya, I. K. A. (2023). Visualisasi data pemetaan nasabah kredit pada lpd desa adat kesiman menggunakan python. RESI : Jurnal Riset Sistem Informasi, 2(1), 129-138. <https://doi.org/10.32795/resi.v2i1.4266>
- Komputer(JAKAKOM), 2(1), 155-160. <https://doi.org/10.33998/jakakom.2022.2.1.31>
- Muhammad, L. J., Islam, M. M., Usman, S. S., & Ayon, S. I. (2020). Predictive data mining models for novel coronavirus (covid-19) infected patients' recovery. SN Computer Science, 1(4). <https://doi.org/10.1007/s42979-020-00216-w>
- Luthfiansyah, R. and Wasito, B. (2023). Penerapan teknik deep learning (long short term memory) dan pendekatan klasik (regresi linier) dalam prediksi pergerakan saham bri. Jurnal Informatika Dan Bisnis, 12(2), 42-54. <https://doi.org/10.46806/jib.v12i2.1059>
- Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in python: main developments and technology trends in data science, machine learning, and artificial intelligence. Information, 11(4), 193. <https://doi.org/10.3390/info11040193>
- Sholeh, M., Andayati, D., & Rachmawati, R. Y. (2022). Data mining model klasifikasi menggunakan algoritma k-nearest neighbor dengan normalisasi untuk prediksi penyakit diabetes. TeIKa, 12(02), 77-87. <https://doi.org/10.36342/teika.v12i02.2911>
- Takdirillah, R. (2020). Penerapan data mining menggunakan algoritma apriori terhadap data transaksi sebagai pendukung

informasi strategi penjualan. Edumatic : Jurnal Pendidikan Informatika, 4(1), 37-46.

<https://doi.org/10.29408/edumatic.v4i1.2081>

Wulandari, P. R., Suarjaya, I. M. A. D., & Rusjyanthi, N. K. D. (2022). Komparasi algoritma pincer search dan algoritma fp-growth. Techno.Com, 21(2), 280-291. <https://doi.org/10.33633/tc.v21i2.5803>

## PROFIL PENULIS



### **Ade Oktarino, S.Kom., M.S.I**

Penulis merupakan alumni S1 Teknik Informatika Universitas Surakarta Tahun 2010, Alumni S2 Magister Sistem Informasi Pada Sekolah Tinggi Ilmu Komputer Dinamika Bangsa Jambi Tahun 2013 dan saat ini sedang menempuh Program Doctoral Information Technology di Universitas Putra Indonesia YPTK Padang pada Tahun 2022, sebelumnya pernah sebagai dosen tetap di Politeknik Jambi hingga Tahun 2014 dan sekarang aktif mengajar di Universitas Adiwangsa Jambi pada Program Studi Teknologi Informasi Fakultas Teknik dan Ilmu Komputer. Sebelumnya Dosen Tugas Tambahan sebagai Wakil Rektor II Universitas Adiwangsa periode 2017 – 2023 dan saat ini sebagai Dosen Tugas Tambahan pada posisi Dekan Fakultas Keguruan Ilmu Pendidikan Universitas Adiwangsa Jambi. Aktif dalam organisasi Relawan TIK Provinsi Jambi sebagai salah satu pengurus dan juga aktif sebagai konsultasi IT di instansi Polda Jambi, Dinas Lingkungan Hidup Kota Jambi dan Kantor Imigrasi Provinsi Jambi. Kosentrasi penelitian dengan topik artificial intelligence, machine learning, deep learning dengan bidang data science, computer vision dan natural language processing.



## **BAB 3**

### **APLIKASI KLASIFIKASI K-NN DENGAN PYTHON**

Deddy Kurniawan  
Universitas Mulia, Samarinda  
E-mail: deddy.kurniawan@universitasmulia.ac.id

#### **PENDAHULUAN**

Kita seringkali tidak menyadari bahwa teknologi yang kita gunakan dalam kehidupan sehari-hari seringkali didukung oleh algoritma canggih yang dapat mengklasifikasikan dan memprediksi banyak hal. Salah satu contoh yang paling umum adalah sistem peringatan cuaca yang digunakan oleh pihak meteorologi. Ketika kita ingin tahu cuaca hari ini, kita seringkali melihat informasi tentang kemungkinan hujan, suhu, dan lain-lain. Namun, apa yang terjadi di balik layar? Bagaimana sistem tersebut dapat memprediksi cuaca? Jawabannya adalah algoritma *k-Nearest Neighbor* (k-NN).

Algoritma KNN adalah salah satu algoritma dalam *supervised learning* yang dapat mengklasifikasikan objek baru berdasarkan atribut dan sampel-sampel dari *training* data. Algoritma ini bekerja dengan cara mencari k-nearest neighbor (jumlah tetang terdekat) dari objek baru tersebut dan kemudian mengklasifikasikan objek tersebut berdasarkan mayoritas kategori dari *k-nearest neighbor* tersebut. Dalam kasus sistem peringatan cuaca, KNN dapat digunakan untuk memprediksi cuaca hari ini berdasarkan data cuaca sebelumnya. Dengan demikian, KNN dapat membantu kita dalam membuat keputusan yang lebih tepat tentang apa yang harus kita lakukan hari ini.

Dalam bab ini, kita akan membahas lebih lanjut tentang algoritma KNN dan bagaimana cara mengaplikasikan KNN

dengan Python. Kita akan mencoba melihat contoh sederhana dari penggunaan KNN dan bagaimana cara mengintegrasikan KNN dengan data yang ada.

## KONSEP DASAR

Dalam ilmu statistik, algoritma *K-Nearest Neighbor* (KNN) merupakan metode pembelajaran yang di awasi non-parametrik (*non-parametric supervised learning*) termasuk kedalam teknik klasifikasi dan juga regresi yang dikembangkan pertama kali oleh Evelyn Fix dan Joseph Hodges pada tahun 1951 (Rehan Guha, 2020). Pada kedua kasus, penggunaan KNN menggunakan nilai  $k$  yang berasal dari jumlah data pelatihan terdekat/tetangga terdekat (*nearest neighbor*) dalam suatu dataset.

KNN dapat dikatakan sebagai algoritma yang bekerja secara lokal dan semua proses perhitungan ditangguhkan hingga evaluasi fungsi. Dalam algoritma k-NN proses perhitungan tidaklah dilakukan secara intensif selama proses model klasifikasi dibangun atau dilatih. Sebaliknya, semua perhitungan yang dilakukan untuk klasifikasi data akan dilakukan hanya ketika ada permintaan untuk mengevaluasi fungsi atau pada saat model digunakan untuk mengklasifikasikan data baru. Dengan sifat ini KNN dikenal sebagai algoritma dengan sifat “*lazy*” karena tidak melakukan perhitungan intensif saat pelatihan, tetapi menunggu sampai ada data baru masuk untuk dilakukan proses perhitungan jarak dan menentukan kelas berdasarkan tetangga terdekatnya (Han et al., 2022).

Mari kita asumsikan bahwa kita memiliki dua kelas, bintang biru dan segitiga hijau. Gambar 3.1 berikut menunjukkan bagaimana KNN menggunakan data pelatihan yang tersimpan (yang merupakan sekumpulan objek bintang dan segitiga) dan mencari pola yang cocok dengan titik data yang baru.

dimana KNN dapat di implementasikan dengan mudah baik dalam tujuan klasifikasi ataupun regresi. KNN bekerja dengan memperhitungkan kondisi jumlah tetangga (*k neighbors*) terdekat sebagai pertimbangan dalam menentukan label (*class*) data untuk data baru dengan melihat label data yang menjadi label data mayoritas dari kumpulan tetangga terdekat. Nilai *k* dalam KNN menjadi sangat penting karena penentuan nilai *k* yang baik akan menentukan hasil klasifikasi data yang optimal dari algoritma KNN.

## DAFTAR PUSTAKA

- Fisher, R. A. (1936). *Iris [Dataset]*. UCI Machine Learning Repository. <https://doi.org/https://doi.org/10.24432/C56C76>
- GeeksforGeeks. (2024, July 15). *K-Nearest Neighbor(KNN) Algorithm*. GeeksforGeeks.
- Han, J., Peo, J., & Tong, H. (2022). Classification: basic concepts and methods. In *Data Mining Concepts and Techniques* (4th ed., pp. 266–268). Morgan Kaufmann.
- Rehan Guha. (2020). Iris Species. In *Machine Learning Cookbook with Python* (Vol. 11). bponline.
- Say, D., Zidi, S., Qaisar, S. M., & Krichen, M. (2023). Automated Categorization of Multiclass Welding Defects Using the X-ray Image Augmentation and Convolutional Neural Network. *Sensors*, 23(14), 6422. <https://doi.org/10.3390/s23146422>
- Tavish. (2024, August 18). *Guide to K-Nearest Neighbors Algorithm in Machine Learning*. Analytich Vidhya.

## **PROFIL PENULIS**

### **Deddy Kurniawa, S.Kom., M.Kom.**

Lahir di Kota Samarinda 08 Oktober 1996 dan saat ini telah berusia 28 tahun. Ia merupakan anak ke dua dari dua bersaudara. Ia merupakan lulusan program Sarjana S1 Teknik Informatik, Sekolah Tinggi Manajemen Informatika dan Komputer Widya Cipta Dharma (STMIK WICIDA) pada tahun 2018, ia juga merupakan alumni program Magister (S2) Teknik Informatika pada Universitas Bina Nusantara pada tahun 2021. Saat ini berprofesi sebagai Dosen tetap di Universitas Mulia, Program Studi Diluar Kampus Utama (PSDKU) Universitas Mulia Samarinda. Memiliki ketertarikan penelitian pada bidang keilmuan data mining, analisis big data, dan model pendukung keputusan. Memiliki 5 artikel jurnal yang di publikasi pada jurnal nasional internasional bereputasi, 1 buah artikel pengabdian kepada masyarakat (PkM) dan 1 book chapter berjudul Sistem Basis Data pada chapter Database Manajemen Sistem (DBMS).

## **BAB 4**

# **APLIKASI KLASIFIKASI NAÏVE BAYES DENGAN PYTHON**

Tio Dharmawan  
Universitas Jember, Jember  
E-mail: tio.pssi@unej.ac.id

### **PENDAHULUAN**

Naive Bayes (Webb, 2011) adalah algoritma yang disusun berdasarkan Teorema Bayes, yang merupakan model matematika dengan dasar statistik dan probabilitas. Teorema Bayes membantu dalam menghitung probabilitas suatu kejadian dengan menghitung probabilitas lain yang berkaitan. Algoritma ini menggunakan rumus umum Teorema Bayes untuk menghasilkan asumsi dasar bahwa setiap fitur mandiri, setara, dan memiliki kontribusi pada hasil.

Naive Bayes adalah metode klasifikasi yang efisien dan efektif, memiliki beberapa keunggulan dan karakteristik khusus:

1. Efisiensi Komputasi: Waktu pelatihan Naive Bayes linear terhadap jumlah contoh pelatihan dan jumlah atribut. Waktu klasifikasi linear terhadap jumlah atribut dan tidak dipengaruhi oleh jumlah contoh pelatihan.
2. Varians Rendah: Naive Bayes memiliki varians yang rendah karena tidak menggunakan pencarian, meskipun ini datang dengan biaya bias yang tinggi.
3. Pembelajaran Inkremental: Naive Bayes beroperasi dari estimasi probabilitas orde rendah yang berasal dari data pelatihan. Estimasi ini dapat dengan mudah diperbarui saat data pelatihan baru diperoleh.

4. **Prediksi Langsung Probabilitas Posterior:** Metode ini memberikan prediksi langsung tentang probabilitas posterior, yang berguna dalam banyak aplikasi praktis.
5. **Ketahanan terhadap Noise:** Naive Bayes menggunakan semua atribut untuk semua prediksi, sehingga relatif tidak sensitif terhadap noise dalam contoh yang diklasifikasikan. Karena menggunakan probabilitas, metode ini juga relatif tidak sensitif terhadap noise dalam data pelatihan.

Naive Bayes telah terbukti efektif dalam berbagai aplikasi praktis seperti klasifikasi teks, diagnosis medis, dan manajemen kinerja sistem. Meskipun memiliki beberapa keterbatasan, seperti suboptimal untuk konsep yang tidak dapat dipisahkan secara linear, metode ini tetap menjadi pilihan yang kuat karena sifatnya yang sederhana dan efisien.

## **TEORI NAÏVE BAYES**

Naïve Bayes dibentuk berdasarkan dari Teorema Bayes. Terdapat tiga persamaan yang digunakan pada Naïve Bayes (Webb, 2011) , yaitu:

1. Probabilitas posterior

Probabilitas Posterior merupakan probabilitas suatu kelas  $y$  terjadi yang berkaitan dengan vektor fitur  $x$ . Persamaan probabilitas posterior seperti berikut.

$$P(y|x) = \frac{P(y)P(x|y)}{P(x)}$$

dimana:

- $P(y)$  merupakan probabilitas prior dari kelas  $y$ , yang menggambarkan seberapa sering kelas  $y$  muncul dalam dataset secara keseluruhan.

Dalam implementasinya menggunakan Python, Naive Bayes dapat diterapkan dengan mudah melalui berbagai pustaka seperti `scikit-learn`. Proses pengembangan model dapat meliputi persiapan data, pembagian data menjadi data latih dan data uji, pelatihan model, serta evaluasi performa model menggunakan metrik-metrik yang relevan seperti akurasi, precision, recall, dan F1-score.

## DAFTAR PUSTAKA

- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. In *Nature* (Vol. 585, Issue 7825). <https://doi.org/10.1038/s41586-020-2649-2>
- McKinney, W., & Team, P. D. (2015). Pandas-Powerful python data analysis toolkit. *Pandas—Powerful Python Data Analysis Toolkit*, 1625.
- 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.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3). <https://doi.org/10.1038/s41592-019-0686-2>

- Webb, G. I. (2011). Naïve Bayes. In *Encyclopedia of Machine Learning* (pp. 713–714). Springer US.  
[https://doi.org/10.1007/978-0-387-30164-8\\_576](https://doi.org/10.1007/978-0-387-30164-8_576)
- Zhang, Y.-C., & Sakhanenko, L. (2019). The naive Bayes classifier for functional data. *Statistics & Probability Letters*, 152, 137–146.  
<https://doi.org/10.1016/j.spl.2019.04.017>

## PROFIL PENULIS

### Tio Dharmawan

Penulis adalah seorang dosen di Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Jember. Ia memiliki latar belakang akademik yang kuat, dengan gelar Sarjana pada Program Studi Sistem Informasi di Fakultas Ilmu Komputer, Universitas Jember (2009-2013) dan Magister Teknik Informatika dari Institut Teknologi Sepuluh Nopember (2014-2016). Fokus penelitian Tio terletak pada bidang visi cerdas (computer vision), terutama dalam pengembangan teknologi interaksi manusia dan komputer (human-computer interaction). Ia telah berkontribusi dalam berbagai penelitian dan publikasi yang bertujuan untuk meningkatkan cara manusia berinteraksi dengan teknologi melalui aplikasi visi cerdas. Sebagai seorang akademisi, Tio Dharmawan memiliki dedikasi untuk membagikan ilmunya kepada para mahasiswa dan berperan aktif dalam pengembangan dunia pendidikan di bidang informatika. Ia juga sering menjadi pembicara pada berbagai seminar dan konferensi, baik di tingkat nasional maupun internasional, membahas topik-topik terkait visi cerdas dan interaksi manusia dengan teknologi.



## **BAB 5**

### **KLASIFIKASI DECISION TREE C 4.5**

### **DENGAN PYTHON**

Erna Hudianti Pujiarini  
Universitas Teknologi Digital Indonesia, Yogyakarta  
E-mail: ernahudi@utdi.ac.id

#### **PENDAHULUAN**

Decision tree merupakan salah satu metode klasifikasi yang paling populer dan banyak digunakan dalam data mining dan machine learning. Decision tree merupakan model prediktif yang menggunakan struktur pohon keputusan untuk memetakan observasi tentang suatu item ke dalam kesimpulan mengenai nilai target dari item tersebut (Jiawei Han, 2011). Decision tree dapat digunakan untuk masalah klasifikasi maupun regresi, dengan target variabel berupa kategori atau numerik.

Algoritma decision tree dimulai dari root node, kemudian secara rekursif membagi data berdasarkan atribut terbaik hingga mencapai leaf nodes. Pemilihan atribut terbaik untuk pembagian didasarkan pada kriteria seperti information gain, gain ratio, atau Gini index, yang mengukur kemampuan atribut dalam membedakan kelas target. Struktur decision tree ini memungkinkan proses pengambilan keputusan yang bertahap dan sistematis, sehingga menghasilkan model prediktif yang dapat diinterpretasikan dengan mudah. Berikut adalah penjelasan lebih rinci mengenai struktur decision tree :

#### **1. Root Node**

Simpul akar yang mewakili atribut/fitur terbaik untuk membagi dataset. Tidak memiliki cabang masuk, hanya cabang keluar. Simpul akar memiliki kriteria pemilihan

atribut yang optimal, seperti information gain, gain ratio, atau gini index.

2. Internal Nodes

Simpul-simpul yang berada di dalam pohon keputusan. Mewakili atribut/fitur yang digunakan untuk membuat keputusan. Memiliki minimal satu cabang masuk dan minimal satu cabang keluar. Setiap cabang keluar mewakili nilai atau rentang nilai dari atribut yang digunakan untuk pembelahan.

3. Edges

Cabang atau garis penghubung antara node. Merepresentasikan nilai-nilai atribut yang menghubungkan antar node. Setiap cabang menghubungkan satu node ke node lainnya berdasarkan nilai atribut.

4. Leaf Nodes (Terminal Nodes)

Simpul akhir dari decision tree. Mewakili kelas target atau nilai prediksi akhir. Tidak memiliki cabang keluar, hanya cabang masuk. Leaf node merepresentasikan hasil klasifikasi atau prediksi akhir.

5. Paths

Jalur dari root node ke leaf node. Merepresentasikan aturan klasifikasi atau prediksi. Setiap path mencerminkan serangkaian keputusan yang mengarah ke kelas target atau nilai prediksi tertentu.

Decision tree merupakan teknik machine learning yang populer karena kesederhanaan dan kemudahan interpretasinya, namun terdapat kekurangan terutama terkait overfitting. Berikut adalah kelebihan dari decision tree:

- a. Mudah diinterpretasikan dan dipahami. Decision tree menyediakan model yang mudah diinterpretasikan dalam bentuk aturan if-then-else. Setiap path dari root ke leaf node merepresentasikan aturan klasifikasi yang jelas.

```

|--- Age <= 44.50
|   |--- AnnualSalary <= 90750.00
|   |   |--- AnnualSalary <= 69750.00
|   |   |   |--- class: 0
|   |   |   |--- AnnualSalary > 69750.00
|   |   |       |--- Age <= 36.50
|   |   |       |   |--- class: 0
|   |   |       |   |--- Age > 36.50
|   |   |           |--- AnnualSalary <= 84750.00
|   |   |           |   |--- Age <= 42.50
|   |   |           |   |   |--- Gender_Enc <= 0.50
|   |   |           |   |   |   |--- AnnualSalary <= 77000.00
|   |   |           |   |   |   |   |--- class: 0
|   |   |           |   |   |   |   |--- AnnualSalary > 77000.00
|   |   |           |   |   |       |--- AnnualSalary <= 80250.00
|   |   |           |   |   |       |   |--- AnnualSalary <= 78250.00
|   |   |           |   |   |       |   |   |--- class: 1
|   |   |           |   |   |       |   |   |--- AnnualSalary > 78250.00
|   |   |           |   |   |       |   |   |   |--- AnnualSalary <= 79250.00
|   |   |           |   |   |       |   |   |   |   |--- class: 0
|   |   |           |   |   |       |   |   |   |   |--- AnnualSalary > 79250.00
|   |   |           |   |   |           |--- truncated branch of depth 3
|   |   |           |   |   |       |--- AnnualSalary > 80250.00
|   |   |           |   |   |       |   |--- class: 0
|   |   |           |   |   |       |   |--- Gender_Enc > 0.50
|   |   |           |   |   |           |--- AnnualSalary <= 73250.00
|   |   |           |   |   |           |   |--- AnnualSalary <= 72250.00
|   |   |           |   |   |           |   |   |--- AnnualSalary <= 70500.00
|   |   |           |   |   |           |   |   |   |--- class: 1
|   |   |           |   |   |           |   |   |   |--- AnnualSalary > 70500.00
|   |   |           |   |   |           |   |   |       |--- Age <= 39.50
|   |   |           |   |   |           |   |   |       |   |--- class: 0
|   |   |           |   |   |           |   |   |       |   |--- Age > 39.50

```

Dari output terlihat dari tree yang dibangun beberapa path yang nantinya menjadi aturan pengetahuan dari klasifikasi, diantaranya :

1. Jika Age <= 45.50, AnnualSalary <= 90750,00, AnnualSalary <= 69750,00 maka Class = 0
2. Age <= 44,50, AnnualSalary > 69750,00, Age <= 36,50 maka Class = 0 dan seterusnya

## KESIMPULAN

Decision tree merupakan model prediktif yang menggunakan struktur pohon keputusan untuk memetakan observasi tentang suatu item ke dalam kesimpulan mengenai nilai target dari item tersebut. Algoritma decision tree dimulai dari root node, kemudian secara rekursif membagi data berdasarkan atribut terbaik hingga mencapai leaf nodes. Pemilihan atribut terbaik untuk pembagian didasarkan pada

kriteria seperti information gain, gain ratio, atau Gini index, yang mengukur kemampuan atribut dalam membedakan kelas target. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 (Iterative Dichotomiser 3). ID3 memiliki beberapa kekurangan, di antaranya tidak dapat menangani atribut kontinu, tidak dapat menangani atribut bernilai hilang (missing value), cenderung memilih atribut dengan banyak nilai (high branching factor). Di Python, library c45-decision-tree atau scikit-learn digunakan untuk implementasi dari algoritma C4.5.

## DAFTAR PUSTAKA

- Alpaydin, E. (2020). *Introduction to Machine Learning*. Cambridge, Massachusetts: MIT Press.
- Jiawei Han, M. K. (2011). *Data Mining: Concepts and Techniques*. San Francisco, California: Morgan Kaufmann.
- Santello, G. (2022, September 8). *gabrielsantello*. Retrieved from Kaggle: <https://www.kaggle.com/>

## PROFIL PENULIS



### **Erna Hudianti Pujiarini, S.Si, M.Si.**

Lahir di Magetan, 28 September 1971. Lulus Sarjana di Program Studi Statistik Universitas Gadjah Mada tahun 1990, lulus program studi Matematika Program Pasca Sarjana Universitas Gadjah Mada tahun 2000. Mulai mengajar tahun 1996 di STMIK AKAKOM yang sekarang berubah menjadi Universitas Teknologi Digital Indonesia. Saat ini menjadi dosen pada program studi Informatika, fakultas Teknologi Informasi, Universitas Teknologi Digital Indonesia. Matakuliah yang diampu Statistika, Statistika Terapan, Matematika, Data Mining, Jaringan Syarat Tiruan. Bidang konsentrasi penelitian Statistika Terapan dan Data Science. Buku yang ditulis “SPSS: Analisis Data Statistik”, “Data Mining : Pola dibalik Angka”.

## **BAB 6**

# **APLIKASI FORECASTING MOVING AVERAGE DENGAN PYTHON**

Aditya Febrianto  
ADW Consulting, Jakarta  
E-mail: a.feb28@gmail.com

### **PENDAHULUAN**

*Moving average* adalah salah satu metode *forecasting*/peramalan data dalam bidang statistika. *Moving average* memiliki berbagai manfaat yang penting, di antaranya adalah memudahkan identifikasi tren jangka panjang, mendukung perencanaan strategis, dan pengambilan keputusan yang lebih baik dalam suatu bisnis (Rozikin et al., 2021). *Moving average* memiliki beberapa jenis, pada *chapter* ini akan membahas jenis *Simple Moving Average* dan *Weighted Moving Average*. *Chapter* ini dibuat dengan tujuan sebagai berikut:

1. Memahami perhitungan *forecasting* menggunakan *Simple Moving Average* dan *Weighted Moving Average* dan penerapannya menggunakan python,
2. Dapat membandingkan metode *Simple Moving Average* dan *Weighted Moving Average* menggunakan RMSE (*Root Mean Square Error*) dan penerapannya menggunakan python,
3. Dapat membuat aplikasi *forecasting Simple Moving Average* dan *Weighted Moving Average* dengan python menggunakan *library* tkinter, dan
4. Dapat menerapkan data sendiri dalam aplikasi *forecasting moving average* yang telah dibuat.

## PENGANTAR FORECASTING MOVING AVERAGE

Pada subbab pengantar ini akan dibahas tentang *Simple Moving Average* dan *Weighted Moving Average* dalam perhitungan statistika. Dan dilakukan perbandingan antara metode *Simple Moving Average* dan *Weighted Moving Average* menggunakan RMSE (*Root Mean Squared Error*) untuk menentukan metode yang terbaik.

Berikut adalah penjelasan tentang kedua jenis *moving average* (Goul et al., 2020):

### 1. *Simple Moving Average* (SMA)

*Simple Moving Average* adalah jenis *moving average* yang memiliki perhitungan paling sederhana dibandingkan dengan jenis-jenis *moving average* lainnya. SMA memiliki perhitungan sebagai berikut:

$$SMA_t = \frac{D_{(t-n)} + D_{(t-(n-1))} + D_{(t-(n-2))} + \dots + D_{(t-1)}}{n}$$

Di mana:

$SMA_t \rightarrow$  Data Prediksi pada Waktu  $t$  menggunakan SMA

$D_t \rightarrow$  Data yang Sebenarnya pada Waktu  $t$

$n \rightarrow$  Jumlah Periode yang Digunakan

Pada *chapter* ini akan mencoba jumlah periode yang digunakan ( $n$ ) adalah 3, sehingga formula SMA yang digunakan adalah sebagai berikut:

$$SMA_t = \frac{D_{(t-3)} + D_{(t-2)} + D_{(t-1)}}{3}$$

3. Pembuatan aplikasi *forecasting moving average* menggunakan python dapat dilakukan menggunakan *library* tkinter (*library* standar milik python), dan
4. Pengguna dapat menerapkan data sendiri dalam aplikasi *forecasting moving average* yang telah dibuat lalu mendapatkan perbandingan metode SMA dan WMA menggunakan RMSE serta dapat menampilkan grafik perbandingan data asli, SMA, dan WMA berdasarkan *input* data sendiri oleh pengguna.

## DAFTAR PUSTAKA

- Alan D. Moore. (2018). *Python GUI Programming with Tkinter - Second Edition* (Vol. 1). Packt Publishing.
- Goul, M., Raghu, T. S., & Li, Z. (2020). Procurement Analytics. In *The Oxford Handbook of Supply Chain Management*. <https://doi.org/10.1093/oxfordhb/9780190066727.013.5>
- Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, 15(14), 5481–5487. <https://doi.org/10.5194/gmd-15-5481-2022>
- Klein, B. (2021). *Data Analysis Numpy, Matplotlib and Pandas*. 1–514. [www.python-course.eu](http://www.python-course.eu)
- Ofosu, K. (2019). *Python Programming for Engineers-Part 4: Graphical User Interfaces II*. [www.SunCam.com](http://www.SunCam.com)[www.SunCam.com](http://www.SunCam.com)
- Rozikin, K., Rudjiono, D., & Setiawan, N. (2021). Pemanfaatan Metode Moving Average Dalam Sistem Informasi Pendukung Keputusan Pembelian Barang Berdasarkan Peramalan Penjualan Dengan Berbasis Web. *Elkom : Jurnal Elektronika Dan Komputer*, 14(2), 198–207. <https://doi.org/10.51903/elkom.v14i2.540>
- Prayogo, N.A. (2023). Dasar Pemrograman Python. Retrieved from

<https://raw.githubusercontent.com/novalagung/dasarpemrogramanpython/ebooks/dasarpemrogramanpython.pdf>

## PROFIL PENULIS



### **Aditya Febrianto**

Lahir di Kota Bandung, 28 Februari 1997. Ia merupakan alumnus S1 Fisika FMIPA UNPAD dan S2 Instrumentasi dan Kontrol FTI ITB. Kini bekerja di perusahaan ADW Consulting di Jakarta Selatan sebagai business dan data analyst. Perusahaan tersebut bergerak di bisnis pengadaan barang/jasa untuk client-client dari BUMN, BUMD, dan swasta di Indonesia. Di perusahaan tersebut, ia mengerjakan projek-projek dengan menggunakan excel, power point, word, dan python. Selain di perusahaan tersebut, ia juga seorang guru les mata pelajaran matematika dan bahasa inggris untuk anak SD hingga SMA. Sedari dulu ia menyukai dan mempelajari berbagai macam bahasa pemrograman, mulai dari java, PHP hingga python. Dengan ia mempelajari suatu bahasa pemrograman, ia akan mendapatkan kepuasan tersendiri karena mampu memahami ilmu tersebut. Dari semua bahasa pemrograman yang ada, ia paling tertarik kepada bahasa pemrograman python karena saat ini python adalah bahasa pemrograman yang paling top. Python sedang top karena python mampu merancang *Artificial Intelligence* dan *Big Data Analytics* yang sedang tren pada zaman digital ini. Awal mula tertarik dalam dunia menulis yaitu sejak ia mengikuti lomba menulis artikel tentang Pengadaan Barang/Jasa yang diadakan oleh IAPI. Dari ajakan sang istri untuk mengikuti lomba tersebut, Alhamdulillah artikel yang dibuat telah meraih juara 15 Besar Artikel Terbaik. Saat ini, ia sedang berkontribusi menulis *chapter* buku di penerbit *future science* untuk dapat mengembangkan dan membagikan ilmu-ilmu kepada para pembaca.



## **BAB 7**

### **APLIKASI FORECASTING EXPONENTIAL SMOOTHING DENGAN PYTHON**

Nelly Oktavia Adiwijaya  
Universitas Jember, Jember  
E-mail: nelly.oa@unej.ac.id

#### **PENDAHULUAN**

*Forecasting* atau peramalan, adalah proses penting dalam berbagai bidang seperti bisnis, ekonomi, dan teknologi informasi (Taylor, 2003). Dalam konteks bisnis, kemampuan untuk memprediksi permintaan produk atau jasa di masa depan memungkinkan perusahaan untuk mengoptimalkan persediaan, mengatur strategi pemasaran, dan merencanakan produksi (Ferbar Tratar et al., 2016). Di sektor ekonomi, *forecasting* membantu dalam pengambilan keputusan kebijakan dengan menyediakan proyeksi pertumbuhan ekonomi, inflasi, dan indikator ekonomi lainnya (Shvaiba, 2019). Sementara dalam teknologi informasi, *forecasting* digunakan untuk merencanakan kapasitas server dan jaringan berdasarkan tren penggunaan data. Dengan demikian, *forecasting* menjadi alat yang vital untuk meningkatkan efisiensi operasional dan meminimalkan risiko.

Salah satu metode *forecasting* yang populer dan efektif adalah *Exponential Smoothing*. Metode ini dikembangkan untuk mengatasi kelemahan metode *moving average* yang tidak responsif terhadap perubahan data terbaru. *Exponential Smoothing* memberikan bobot yang lebih besar pada data terbaru, sehingga lebih adaptif terhadap perubahan pola data (R. G. Brown & Meyer, 1961). Ada tiga tipe utama dari metode ini: *Single Exponential Smoothing* yang cocok untuk data tanpa tren atau musiman, *Double Exponential Smoothing* yang

memperhitungkan tren, dan *Triple Exponential Smoothing* (atau *Holt-Winters*) yang juga mempertimbangkan faktor musiman. Kelebihan utama dari metode ini adalah kesederhanaannya dalam implementasi dan kemampuannya memberikan hasil yang akurat dengan sedikit data historis.

Bab ini bertujuan untuk memberikan pemahaman yang komprehensif mengenai aplikasi *Exponential Smoothing* dalam *forecasting* menggunakan *Python*. Melalui penjelasan teoritis dan studi kasus praktis, pembaca akan belajar bagaimana mengimplementasikan berbagai tipe *Exponential Smoothing* dengan bahasa pemrograman *Python*.

## **DEFINISI DAN KONSEP DASAR *FORECASTING***

*Forecasting*, atau peramalan, adalah proses untuk memprediksi kejadian atau nilai di masa depan berdasarkan data historis dan analisis tren. Dalam konteks data, *forecasting* menggunakan berbagai teknik statistik dan model matematis untuk mengidentifikasi pola dalam data masa lalu dan memproyeksikannya ke masa depan. Proses ini mencakup pengumpulan data, analisis data, modelisasi, dan validasi hasil untuk menghasilkan prediksi yang akurat (Gardner, 2006).

Konsep dasar dari *forecasting* meliputi:

- a) Data Historis: Data masa lalu yang digunakan sebagai dasar untuk analisis dan prediksi.
- b) Tren: Pola umum dalam data yang menunjukkan arah pergerakan data dari waktu ke waktu.
- c) Musiman: Fluktuasi periodik dalam data yang terjadi secara reguler dalam interval waktu tertentu, seperti mingguan, bulanan, atau tahunan.
- d) Variasi Acak: Komponen yang tidak terduga atau tidak dapat dijelaskan oleh tren atau musiman, sering dianggap sebagai '*noise*'.

menggunakan *Python*. Kami memanfaatkan beberapa *library* populer seperti *Pandas*, *Numpy*, *Statsmodels*, dan *Matplotlib* untuk membantu dalam Pengolahan dan pembersihan data, Penerapan model, Visualisasi hasil. Dengan *Python*, proses *forecasting* menjadi lebih mudah dan dapat dilakukan secara efisien untuk dataset yang besar. Untuk menentukan model yang paling sesuai dengan data, evaluasi menggunakan berbagai metrik error seperti MAE, MSE, RMSE, dan MAPE. Metrik-metrik ini membantu menilai seberapa jauh hasil prediksi dari data sebenarnya. Visualisasi hasil prediksi juga memperlihatkan bagaimana masing-masing model menangani pola dalam data.

Dari evaluasi ini, dapat disimpulkan bahwa Triple Exponential Smoothing (TES) biasanya memberikan hasil terbaik untuk data yang memiliki tren dan musiman, sedangkan Single Exponential Smoothing (SES) lebih cocok untuk data yang stabil tanpa pola yang kompleks.

## DAFTAR PUSTAKA

- B, P., B.D., N., M., H., & D., E. (2019). The Exponential Smoothing Methods (Double-Triple) and Its Applications On Time Series Data. *International Journal of Engineering and Technology*, 11(6), 1123–1127. <https://doi.org/10.21817/ijet/2019/v11i6/191106011>
- Belas, A. O., & Bidyuk, P. I. (2021). CHOOSING A QUALITY CRITERION FOR EVALUATING THE FORECAST OF NONLINEAR NON-STATIONARY PROCESSES. *KPI Science News*, 2. <https://doi.org/10.20535/kpispn.2021.2.236936>
- Brown, B., Jensen, T., Gotway, J. H., Bullock, R., Gilleland, E., Fowler, T., Newman, K., Adriaansen, D., Blank, L., Burek, T., Harrold, M., Hertneky, T., Kalb, C., Kucera, P., Nance, L., Opatz, J., Vigh, J., & Wolff, J. (2021). The Model Evaluation Tools (MET): More than a Decade of

- Community-Supported Forecast Verification. *Bulletin of the American Meteorological Society*, 102(4), E782–E807. <https://doi.org/10.1175/BAMS-D-19-0093.1>
- Brown, R. G., & Meyer, R. F. (1961). The Fundamental Theorem of Exponential Smoothing. *Operations Research*, 9(5), 673–685. <https://doi.org/10.1287/opre.9.5.673>
- Cipra, T. (2006). Exponential smoothing for irregular data. *Applications of Mathematics*, 51(6), 597–604. <https://doi.org/10.1007/s10492-006-0023-9>
- Ferbar Tratar, L., Mojšker, B., & Toman, A. (2016). Demand forecasting with four-parameter exponential smoothing. *International Journal of Production Economics*, 181, 162–173. <https://doi.org/10.1016/j.ijpe.2016.08.004>
- Fu, T. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1), 164–181. <https://doi.org/10.1016/j.engappai.2010.09.007>
- Gardner, E. S. (2006). Exponential smoothing: The state of the art—Part II. *International Journal of Forecasting*, 22(4), 637–666. <https://doi.org/10.1016/j.ijforecast.2006.03.005>
- Kazempour, M. K. (1990). Stationary Forecasting; Using Holt-Winter and a Modification of Holt-Winter. *Biometrical Journal*, 32(3), 347–356. <https://doi.org/10.1002/bimj.4710320317>
- Maydeu-Olivares, A., Shi, D., & Fairchild, A. J. (2020). Estimating causal effects in linear regression models with observational data: The instrumental variables regression model. *Psychological Methods*, 25(2), 243–258. <https://doi.org/10.1037/met0000226>
- McPherson, S., Reese, C., & Wendler, M. C. (2018). Methodology Update. *Nursing Research*, 67(5), 404–410. <https://doi.org/10.1097/NNR.0000000000000297>
- Milniadi, A. D., & Adiwijaya, N. O. (2023). ANALISIS PERBANDINGAN MODEL ARIMA DAN LSTM

DALAM PERAMALAN HARGA PENUTUPAN SAHAM  
(STUDI KASUS : 6 KRITERIA KATEGORI SAHAM  
MENURUT PETER LYNCH). SIBATIK JOURNAL:  
Jurnal Ilmiah Bidang Sosial, Ekonomi, Budaya, Teknologi,  
Dan Pendidikan, 2(6), 1683–1692.  
<https://doi.org/10.54443/sibatik.v2i6.798>

Ostertagová, E., & Ostertag, O. (2012). Forecasting using simple exponential smoothing method. *Acta Electrotechnica et Informatica*, 12(3). <https://doi.org/10.2478/v10198-012-0034-2>

Shvaiba, D. (2019). Forecasting of socio-economic security indicators by means of exponential smoothing. *Bulletin of Science and Practice*, 5(3), 241–249.  
<https://doi.org/10.33619/2414-2948/40/30>

Taylor, J. W. (2003). Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting*, 19(4), 715–725. [https://doi.org/10.1016/S0169-2070\(03\)00003-7](https://doi.org/10.1016/S0169-2070(03)00003-7)

## PROFIL PENULIS



**Nelly Oktavia Adiwijaya, S.Si, MT, CEC**

Menyelesaikan Sarjana Matematika di Fakultas MIPA Universitas Jember pada tahun 2006; Magister Game Teknologi di STEI, ITB Bandung pada tahun 2009; serta saat ini sedang menempuh Program Doktorat di Informatika FTCIS ITS Surabaya. Pernah menjabat sebagai Kepala Laboratorium Pemrograman di Prodi Sistem Informasi UNEJ. Kemudian saat ini menjadi KoProdi Informatika Fakultas Ilmu Komputer UNEJ. Mendapatkan sertifikasi Internasional dari AAPM (American Association of Project Management) di bidang e-commerce. Beberapa kali mengikuti berbagai kegiatan seminar baik nasional maupun internasional.

## **BAB 8**

# **APLIKASI FORECASTING REGRESI LINEAR DENGAN PYTHON**

Sofiansyah Fadli  
Program Studi Teknik Informatika, STMIK Lombok, Praya  
E-mail: sofiansyah182@gmail.com

### **PENDAHULUAN**

Dalam dunia modern yang serba cepat, kemampuan untuk memprediksi kejadian di masa depan berdasarkan data historis telah menjadi aset yang sangat berharga (Rukmana et al., 2023). *Forecasting* atau peramalan adalah proses yang memungkinkan individu dan organisasi untuk membuat keputusan yang lebih baik dengan memanfaatkan data masa lalu untuk memproyeksikan hasil yang akan datang (Fadli et al., 2022). Aplikasi *forecasting* dapat ditemukan di berbagai bidang seperti bisnis, ekonomi, meteorologi, dan kesehatan. Dalam ranah *forecasting*, metode *regresi linear* merupakan salah satu teknik yang paling sederhana dan efektif (Hasanah et al., 2020). *Regresi linear* adalah model statistik yang memprediksi nilai variabel *dependen* (Y) dengan menggunakan satu atau lebih variabel *independen* (X) (Id, 2021). Metode ini sangat populer karena kemudahannya dalam interpretasi dan implementasi, serta kemampuannya dalam memberikan gambaran yang jelas mengenai hubungan antara variabel (Rozak, 2024).

*Regresi linear* bekerja berdasarkan prinsip bahwa hubungan antara variabel dapat digambarkan dengan sebuah garis lurus. Dalam kasus *regresi linear* sederhana, model ini mencoba menemukan garis yang paling sesuai dengan data *historis* yang ada, dengan meminimalkan jarak antara data dan garis tersebut (Fengki, 2018). Metode ini menghasilkan dua parameter utama:

*intercept* dan *slope*, yang memberikan informasi tentang titik awal dan kemiringan garis (Rozak, 2024). Di era digital saat ini, *Python* telah menjadi salah satu bahasa pemrograman yang paling banyak digunakan untuk analisis data dan *machine learning*. Dengan berbagai *library* dan alat yang tersedia, *Python* mempermudah implementasi model *regresi linear* dan analisis data (Kurniawan, 2022). Library seperti '*scikit-learn*', '*pandas*', dan '*matplotlib*' menyediakan fungsi dan metode yang memungkinkan pengguna untuk membangun, melatih, dan mengevaluasi model *regresi linear* dengan mudah (Rifky et al., 2024). Penggunaan *Python* untuk *forecasting regresi linear* tidak hanya menghemat waktu dan usaha, tetapi juga memberikan fleksibilitas dan akurasi dalam membuat prediksi (Hermansyah et al., 2024). Dalam *forecasting, regresi linear* memungkinkan kita untuk memprediksi nilai variabel dependen berdasarkan variabel independen (Anjeliyani et al., 2024). Pada bagian ini, menjelaskan dasar-dasar *regresi linear*, termasuk model sederhana dan berganda. *Regresi linear* adalah metode yang digunakan untuk menemukan hubungan *linear* antara dua variabel: satu variabel *independen* (X) dan satu variabel *dependen* (Y). Tujuannya adalah untuk memodelkan variabel *dependen* sebagai fungsi dari variabel *independen* (Sunge & Zy, 2023).

*Regresi linear* berasumsi bahwa hubungan antara (X) dan (Y) adalah *linier*, artinya perubahan pada (X) akan menghasilkan perubahan proporsional pada (Y). Model *regresi linear* sederhana adalah bentuk paling dasar dari regresi linear yang hanya melibatkan satu variabel *independen* (Trianggana, 2020). Model ini digunakan ketika kita ingin menganalisis hubungan antara dua variabel dan membuat prediksi berdasarkan variabel *independen* tersebut. Secara grafis, model *regresi linear* sederhana dapat digambarkan sebagai garis lurus pada *plot scatter*, di mana sumbu *horizontal* mewakili variabel

## DAFTAR PUSTAKA

- Anjeliyani, A., Fatimah, F., & Casmat, M. (2024). Pemodelan Regresi Linear Untuk Memprediksi Nilai Penjualan Di Pt Goodiebag Custom Indonesia Tanpa Mempertimbangkan Uji Asumsi. *DESANTA (Indonesian of Interdisciplinary Journal)*, 5(1), 1-13.
- Aryani, Y. (2020). Sistem informasi penjualan barang dengan metode regresi linear berganda dalam prediksi pendapatan perusahaan. *Jurnal Riset Sistem Informasi Dan Teknologi Informasi (JURSISTEKNI)*, 2(2), 39-51.
- Elfaladonna, F., Isa, I. G. T., Sartika, D., & Putra, A. M. (2024). *Buku Ajar Dasar Exploratory Data Analysis (EDA)*. Penerbit NEM.
- Fadli, S., Saikin, S., & Ashari, M. (2022). Implementation of Data Mining on Tourist Visits Patterns on Lombok Island Tourism Objects. *JISA (Jurnal Informatika dan Sains)*, 5(1), 12-18.
- Fengki, F. (2018). *Implementasi Regresi Linear untuk memprediksi lama waktu pengiriman Catering kepada konsumen studi kasus home catering Malang Universitas Islam Negeri Maulana Malik Ibrahim*].
- Hairani, H., Saputro, K. E., & Fadli, S. (2020). K-means-SMOTE for handling class imbalance in the classification of diabetes with C4. 5, SVM, and naive Bayes. *J. Teknol. dan Sist. Komput*, 8(2), 89-93.
- Hartatik, H., Kwintiana, B., Nengsih, T. A., Baradja, A., Harto, B., Sudipa, I. G. I., Handika, I. P. S., Adhicandra, I., & Gugat, R. M. D. (2023). *DATA SCIENCE FOR BUSINESS: Pengantar & Penerapan Berbagai Sektor*. PT. Sonpedia Publishing Indonesia.
- Hasanah, M., Harani, N. H., & Riza, N. (2020). *Implementasi Barcode Dan Algoritma Regresi Linear Untuk Memprediksi Data Persediaan Barang*. Kreatif.



- Hermansyah, H., Abdullah, A., & Utami, P. Y. (2024). Penerapan Metode Regresi Linier Berganda Untuk Memprediksi Panen Kelapa Sawit. *Progresif: Jurnal Ilmiah Komputer*, 20(1), 540-554.
- Id, I. D. (2021). *Machine Learning: Teori, Studi Kasus dan Implementasi Menggunakan Python* (Vol. 1). Unri Press.
- Irrawati, M. D., & Mukaramah, M. (2024). Implementasi Metode Regresi Linear Berganda untuk Mengatasi Pelanggaran Asumsi Klasik. *Studi Akuntansi, Keuangan, Dan Manajemen*, 3(2), 83-94.
- Kurniawan, D. (2022). *Pengenalan machine learning dengan python*. Elex Media Komputindo.
- Kurniawan, R. (2016). *Analisis regresi*. Prenada Media.
- Ridwan, A. M., & Setyawan, G. D. (2023). Perbandingan Berbagai Model Machine Learning Untuk Mendeteksi Diabetes. *TEKNOKOM*, 6(2), 127-132.
- Rifky, S., Kharisma, L. P. I., Afendi, H. A. R., Napitupulu, S., Ulina, M., Lestari, W. S., Maysanjaya, I. M. D., Kelvin, K., Sinaga, F. M., & Muchtar, M. (2024). *Artificial Intelligence: Teori dan Penerapan AI di Berbagai Bidang*. PT. Sonpedia Publishing Indonesia.
- Rozak, F. (2024). Prediksi Jumlah Kunjungan Pasien Rawat Jalan Menggunakan Metode Resgresi Linier Sederhana. *PROSISKO: Jurnal Pengembangan Riset dan Observasi Sistem Komputer*, 11(2), 284-293.
- Rukmana, A. Y., Rahman, R., Afriyadi, H., Moeis, D., Setiawan, Z., Subchan, N., Magdalena, L., Singadji, M., El Rayeb, A., & Kusuma, A. T. A. P. (2023). *PENGANTAR SISTEM INFORMASI: Panduan Praktis Pengenalan Sistem Informasi & Penerapannya*. PT. Sonpedia Publishing Indonesia.
- Sunge, A. S., & Zy, A. T. (2023). Analisis Prediksi Penjualan Dengan Metode Regresi Linear Di Pt. Eagle Industry

- Indonesia. *Jurnal Informatika Teknologi dan Sains (Jinteks)*, 5(3), 398-403.
- Tabroni, T. (2022). Demand Forecasting Analysis With Descriptive Methods And Qualitative Approaches To The Sale Of Batik Clothing Products: Analisis Forecasting Demand Dengan Metode Deskriptif Dan Pendekatan Kualitatif Pada Penjualan Produk Baju Batik. *Indonesian Journal of Economy, Business, Entrepreneurship and Finance*, 2(3), 374-380.
- Trianggana, D. A. (2020). Peramalan Jumlah Siswa-Siswi Melalui Pendekatan Metode Regresi Linear. *Jurnal Media Infotama*, 16(2).

## PROFIL PENULIS



### **Sofiansyah Fadli, S.Kom., M.Kom.**

Memperoleh gelar Magister Teknik Informatika dari Program Pascasarjana Universitas Islam Indonesia (UII Yogyakarta) pada tahun 2017. Sejak tahun 2017 tercatat sebagai Dosen Tetap Program Studi Teknik Informatika STMIK Lombok dan memiliki tugas tambahan sebagai Ketua Program Studi Teknik Informatika Tahun 2018-2020 dan Tahun 2021-2025. Selain mengajar, menulis, meneliti, pemateri workshop dan seminar dalam bidang ilmu komputer, penulis juga pernah menjabat sebagai Kepala Sekolah SMK-IT Generasi Muslim Cendekia Tahun 2018-2021 salah satu jenjang pendidikan di Yayasan Generasi Muslim Cendekia yang didirikan oleh Prof. Ir. H. SUnarpi, PhD (Alm). Link Sinta: <https://sinta.kemdikbud.go.id/authors/profile/6073057>

## **BAB 9**

### **APLIKASI FORECASTING ARIMA DENGAN PHYTON**

Dwi Remawati  
STMIK Sinar Nusantara, Surakarta  
E-mail: dwirema@sinus.ac.id

#### **PENDAHULUAN**

Data merupakan hal yang penting di era teknologi informasi ini. Data lampau dan data saat ini dapat digunakan untuk peramalan atau *forecasting* untuk memprediksi atau memperkirakan nilai di masa depan. (Kusnawi et al., 2023). Tujuan dari forecasting adalah untuk memberikan wawasan atau informasi yang membantu dalam membuat keputusan yang lebih baik tentang hal-hal yang akan terjadi di masa (Zidan Rusminto et al., 2024). Misalnya, perusahaan dapat menggunakan forecasting untuk memprediksi tren penjualan, tingkat produksi, permintaan produk, atau harga komoditas di masa depan.

Dalam konteks deret waktu (*time series*), forecasting bertujuan memahami pola dalam data seperti tren, musiman, tidak musiman atau siklus, agar dapat memperkirakan nilai di periode mendatang (Tan & Astuti, 2020). ARIMA (*AutoRegressive Integrated Moving Average*) adalah model yang populer untuk forecasting deret waktu yang tidak musiman (Mahardika & Triloka, 2024). Misalnya untuk memprediksi jumlah penjualan produk berdasarkan data bulanan atau harian sebelumnya. Metode ARIMA efektif dalam memprediksi penjualan makanan dan menunjukkan bagaimana hal itu dapat digunakan untuk memperkirakan penjualan di masa depan dan mengambil tindakan yang tepat untuk mengoptimalkannya (Oktavianti & Ali, 2023). Untuk menyelesaikan masalah yang terjadi di taman bermain, metode

ARIMA digunakan sebagai solusi. Hasilnya menunjukkan bahwa model ARIMA optimal dari plotting ACF dan PACF dan dievaluasi menggunakan RMSE. Uji signifikan menghasilkan model ARIMA optimal (1,0,0) dengan nilai RMSE 28566 dan nilai p-value untuk masing-masing orde (Cherrly & Somya, 2023).

## **KONSEP DASAR**

Untuk peramalan data sequence time univariat, model yang dikenal sebagai Arima (Auto Regressive Integrated Moving Average) banyak digunakan. Model ARIMA terdiri dari tiga komponen: Auto Regressive (AR), Integrated (I), dan Moving Average (MA), yang disebut ARIMA (p, d, q). Komponen p menunjukkan jumlah orde AR pada model, dan model AR dapat dianggap sebagai model regresi dengan peubah penjelasnya adalah data dari periode sebelumnya. Sebagai contoh, jika nilai  $p=2$  menunjukkan bahwa model dapat memprediksi data pada waktu ke-t, maka model menggunakan peubah penjelas yaitu data pada waktu ke- $t-1$  dan  $t-2$ . Jumlah proses diferensiasi (d) diwakili oleh komponen d. Proses ini dibutuhkan untuk menangani kondisi non-stasioner (seperti trend) hingga mencapai stasioner. Kondisi data yang sudah stasioner ditunjukkan oleh nilai  $d=0$  tanpa perlu melakukan proses diferensiasi (Sinaga, 2023).

Untuk memperoleh model ARIMA yang ideal, sangat penting untuk memilih kombinasi yang tepat untuk p, d, dan q. Nilai p dan q biasanya dapat dihitung dengan plot Fungsi Autokorelasi (ACF) dan Fungsi Partial Autokorelasi (PACF), sementara untuk nilai d, uji Augmented Dickey–Fuller (ADF) dapat digunakan (Kurniawati et al., 2023).

## **Stasioneritas dan Nonstasioneritas**

Stasioneritas dan Nonstasioneritas adalah konsep dasar dalam analisis deret waktu yang menggambarkan stabilitas pola

## KESIMPULAN

Model peramalan penjualan dengan ARIMA yang telah dibuat menunjukkan bahwa model ini mampu memberikan prediksi yang layak, namun masih ada potensi perbaikan yang bisa dilakukan untuk meningkatkan akurasi. Model ARIMA berhasil menangkap pola data penjualan dengan hasil yang cukup akurat, ditunjukkan oleh MAPE sekitar 15%. Program ini dapat digunakan sebagai dasar untuk membuat prediksi penjualan di masa depan. Evaluasi menggunakan RMSE dan MAPE memberikan gambaran bahwa model masih bisa disesuaikan atau ditingkatkan jika diperlukan, terutama jika hasil akurasi ini belum memenuhi standar. Hasil prediksi ini berguna untuk memberikan prediksi penjualan di masa depan yang dapat membantu perencanaan bisnis, tetapi hasilnya perlu dikaji ulang untuk memastikan relevansi dengan pola bisnis yang dinamis.

## DAFTAR PUSTAKA

- Cherrly, A., & Somya, R. (2023). Prediksi Penjualan Tiket Wisata Taman Bermain Menggunakan Metode ARIMA. *Techno.Com*, 22(2), 312–322. <https://doi.org/10.33633/tc.v22i2.7950>
- Kurniawan, A. F., Pane, S. F., & Awangga, R. M. (2021). Prediksi Jumlah Penjualan Rumah di Bojongsoang ditengah Pandemi Covid-19 dengan Metode ARIMA. *Jurnal Media Informatika Budidarma*, 5(4), 1479. <https://doi.org/10.30865/mib.v5i4.3121>
- Kurniawati, A., Ahmad, M. S., Fhadli, M., & Lutfi, S. (2023). Analisis Perbandingan Metode Time Series Forecasting Untuk Prediksi Penjualan Obat Di Apotek ( Studi Kasus : Kimia Farma Apotek Takoma ) Comparative Analysis of Methodstime Series Forecasting for Prediction of Drug Sales in Pharmacy ( Case Study : Chemicala. *Jurnal Jaringan*

- Dan ..., 3(1), 96–106. <https://doi.org/00.0000/jati>
- Kusnawi, K., Eka Putra, M. A. F., & Ipmawati, J. (2023). Price Prediction Of Basic Material Using ARIMA Forecasting Method Through Open Data Sumedang District. *Sistemasi*, 12(2), 293. <https://doi.org/10.32520/stmsi.v12i2.2282>
- Maharani, N. S., Angraini, Y., Rahmawan, M. A., Putri, O. A., Kurniawan, S., Safitri, T. A., Rizki, A., Ningsih, W. A. L., Hidayatulloh, N. G. T., & Ratnasari, A. P. (2023). Aplikasi Model Arima Garch Dalam Peramalan Data Nilai Tukar Rupiah Terhadap Dolar Tahun 2017-2022. *Jurnal Matematika Sains Dan Teknologi*, 24(1), 37–50. <https://doi.org/10.33830/jmst.v24i1.4875.2023>
- Mahardika, B. T., & Triloka, C. B. (2024). *Analisis Pola Penjualan dan Prediksi Permintaan Produk Parfum Toko Kayyasa Menggunakan Model FP-Growth dan Arima*. XIV(1), 77–86.
- Oktavianti, E., & Ali, H. (2023). Prediksi Pendapatan Penjualan Makanan Menggunakan Algoritma Autoregressive Integrated Moving Average (ARIMA). *Jurnal Multinetics*, 9(1), 2–3.
- Putri, E. S., & Sadikin, M. (2021). Prediksi Penjualan Produk Untuk Mengestimasi Kebutuhan Bahan Baku Menggunakan Perbandingan Algoritma LSTM dan ARIMA. *Format Jurnal Ilmiah Teknik Informatika*, 10(2), 162. <https://doi.org/10.22441/format.2021.v10.i2.007>
- Ryan, J., & Wijaya, H. (2024). Implementasi Data Mining untuk Sales Forecasting Berbasis Website dengan Metode ARIMA. *Bit-Tech*, 7(1), 19–27. <https://doi.org/10.32877/bt.v7i1.1332>
- Sihombing, E. I., Suhendra, C. D., & Marini, L. F. (2024). KLIK: Kajian Ilmiah Informatika dan Komputer Analisis Data Time Series Untuk Prediksi Harga Komoditas Pangan Menggunakan Autoregressive Integrated Moving Average.

- Media Online*), 4(6), 2711–2720.  
<https://doi.org/10.30865/klik.v4i6.1863>
- Sinaga, S. A. (2023). Implementasi Metode Arima (Autoregressive Moving Average) Untuk Prediksi Penjualan Mobil. *Journal Global Technology Computer*, 2(3), 102–109. <https://doi.org/10.47065/jogtc.v2i3.4013>
- Tan, E., & Astuti, I. (2020). Metode Autoregressive Integrated Moving Average untuk Meramalkan Penjualan. *EKOMABIS: Jurnal Ekonomi Manajemen Bisnis*, 1(02), 149–158. <https://doi.org/10.37366/ekomabis.v1i02.43>
- Zidan Rusminto, M., Adi Wibowo, S., & Santi Wahyuni, F. (2024). Peramalan Harga Saham Menggunakan Metode Arima (Autoregressive Integrated Moving Average) Time Series. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 8(2), 1263–1270. <https://doi.org/10.36040/jati.v8i2.9089>



## **PROFIL PENULIS**

### **Dwi Remawati**

Pendidikan S1 dari Teknik Informatika IST Palapa Malang Jawa Timur, S2 Teknik Informatika UDINUS Semarang-Jawa Tengah. Saat ini sebagai Dosen PNS Dpk di Program Studi Teknologi Informasi STMIK Sinar Nusantara Surakarta-Jawa Tengah – Indonesia.

## **BAB 10**

### **APLIKASI ASSOCIATION APRIORI DENGAN PYTHON**

Deborah Kurniawati  
Universitas Teknologi Digital Indonesia  
E-mail: debbie@utdi.ac.id

#### **ATURAN ASOSIASI**

Aturan asosiasi (*Association rule*) adalah salah satu teknik utama dalam data mining yang digunakan untuk menemukan hubungan antar item di dalam dataset, terutama dalam konteks transaksi. Misalnya, aturan asosiasi dapat digunakan untuk mengidentifikasi pola "jika item X dibeli, maka item Y cenderung juga dibeli" (Agrawal, R., et al., 1993). Aturan ini dinyatakan dalam bentuk implikasi, seperti "Jika X, maka Y". Misalnya, jika seseorang membeli roti (X), maka mereka juga membeli mentega (Y). Teknik ini sering digunakan dalam market basket analysis di mana aturan asosiasi membantu mengidentifikasi produk yang sering dibeli bersamaan.

Algoritma ini didasarkan pada prinsip bahwa setiap subset dari itemset yang sering muncul (*frequent itemset*) juga harus sering ditemukan. Dalam era big data, aturan asosiasi dapat diterapkan diberbagai bidang mengingat bahwa analisis data menjadi elemen yang penting dalam berbagai sektor dengan memberikan kontribusi yang signifikan pada proses pengambilan keputusan yang lebih baik dan efisien. Aturan asosiasi dapat digunakan dalam mendeteksi pola dan hubungan penting di dunia pendidikan yang kemudian dapat digunakan untuk pengembangan kebijakan pendidikan yang lebih baik (Yang L., & Zhang, H., 2024), bahkan dalam memahami perilaku siswa dalam penggunaan ponsel di ruang



kelas dan dampak penggunaan ponsel terhadap efektivitas pembelajaran di kelas (Gao, 2022).

Selain dimanfaatkan untuk segmentasi pelanggan (Silva, J., et al., 2019) algoritma Apriori juga dapat digunakan untuk menganalisis perilaku pengguna perpustakaan, yang kemudian dapat membantu pengelola perpustakaan dalam menyusun strategi layanan yang lebih personal dan efektif (Zhang, X., & Zhang, J., 2023). Di bidang kesehatan, aturan asosiasi dapat digunakan untuk menganalisis pola kesehatan mental berdasarkan data sosial dan demografis (Hassan, M.M., et al., 2023), serta menganalisa terkait penyebaran virus sehingga mampu membantu pengambilan keputusan yang lebih cepat dan akurat dalam situasi darurat kesehatan global (Kumar, S., et al., 2023).

#### Komponen Penting dalam Association Rule

1. **Support.** *Support* dihitung sebagai rasio jumlah transaksi yang mengandung kombinasi item tertentu terhadap total jumlah transaksi (Han, J., et al., 2011). Misalnya, jika 10% dari seluruh transaksi mengandung kombinasi roti dan mentega, maka support dari aturan ini adalah 0,10.
2. **Confidence.** *Confidence* mengukur seberapa sering item Y muncul dalam transaksi yang mengandung item X. *Confidence* dihitung dengan membagi jumlah transaksi yang mengandung X dan Y dengan jumlah transaksi yang hanya mengandung X (Tan, P.N., et al., 2002). Jika *confidence* suatu aturan adalah 0,8, ini berarti 80% dari transaksi yang mengandung roti juga mengandung mentega.
3. **Lift.** *Lift* mengukur sejauh mana kehadiran X meningkatkan peluang munculnya Y. *Lift* dihitung dengan membandingkan *confidence* suatu aturan dengan *support* dari item Y (Brin, S., et al., 1997). *Lift* digunakan untuk mengevaluasi apakah suatu aturan benar-benar signifikan atau hanya kebetulan.

## KESIMPULAN

1. Algoritma Apriori digunakan dalam data mining untuk menemukan pola hubungan antar item dalam dataset yang dapat diimplementasikan di berbagai bidang.
2. Analisis asosiasi mencakup beberapa komponen penting, seperti *support*, *confidence*, dan *lift*, yang digunakan untuk mengukur seberapa kuat asosiasi antara item-item dalam dataset. *Antecedents* dan *consequents* adalah bagian-bagian dari aturan asosiasi. *Lift*, *leverage*, dan *conviction* adalah metrik yang digunakan untuk mengevaluasi kekuatan hubungan antara *antecedents* dan *consequents*.
3. Langkah-langkah utama dalam algoritma Apriori meliputi penentuan *support* minimum, pembentukan *frequent itemsets*, dan pembuatan aturan asosiasi yang relevan. Dengan properti Apriori yang menghapus itemset yang tidak memenuhi *support*, algoritma ini menjadi lebih efisien dalam memproses data besar.

## DAFTAR PUSTAKA

- Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2), 207-216.
- Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules*. In Proceedings of the 20th International Conference on Very Large Data Bases (pp. 487-499).
- Borgelt, C. (2005). *An implementation of the FP-growth algorithm*. In Workshop Open Source Data Mining Software (pp. 1-5)
- Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). *Dynamic itemset counting and implication rules for market basket data*. In *ACM SIGMOD Record* (Vol. 26, No. 2, pp. 255-264).

- Gao, P. (2022). *Research on Analysis of Students Using Mobile Phones in Ideological and Political Classrooms by Apriori Algorithm of Association Rules*.
- Goethals, B. (2010). Survey on frequent pattern mining. In *Lecture Notes in Computer Science* (pp. 86-114). Springer.
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
- Hassan, M. M., Karim, A., & Al Haque, A. S. M. F. (2023). *An Apriori Algorithm-Based Association Rule Analysis to detect Human Suicidal Behaviour*
- Hipp, J., Güntzer, U., & Nakhaeizadeh, G. (2000). Algorithms for association rule mining—A general survey and comparison. *ACM SIGKDD Explorations Newsletter*, 2(1), 58-64.
- Kumar, S., Chuli, A., & Prasanth, N. (2023). *Data Analytics for Pandemic Management using MapReduce and Apriori Algorithm*.
- Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. Jr. (2018). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. John Wiley & Sons, Inc
- Silva, J., Varela, N., & Rojas Millán, R. H. (2019). *Association Rules Extraction for Customer Segmentation in the SMEs Sector Using the Apriori Algorithm*.
- Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley.
- Yang, L., & Zhang, H. (2024). Big data technology for teaching quality monitoring and improvement in higher education - joint K-means clustering algorithm and Apriori algorithm. *Systems and Soft Computing*.
- Zaki, M. J., Parthasarathy, S., Ogihara, M., & Li, W. (1997). *New algorithms for fast discovery of association rules*.

Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, 283-286.

Zhang, X., & Zhang, J. (2023). Analysis and research on library user behavior based on apriori algorithm. *Measurement: Sensors*. Elsevier.

<https://doi.org/10.1016/j.sensors.2023.06.01>

## PROFIL PENULIS



### **Deborah Kurniawati**

Penulis merupakan salah satu pengajar di Program Studi Sistem Informasi Fakultas Teknologi Informasi Universitas Teknologi Digital Indonesia (dahulu STMIK AKAKOM Yogyakarta). Penulis menyelesaikan S1 Teknik Informatika di STMIK AKAKOM Yogyakarta, dan S2 Ilmu Komputer Di Universitas Gadjah Mada Yogyakarta.

## **BAB 11**

### **APLIKASI CLUSTERING K-MEANS DENGAN PYTHON**

Saikin

Program Studi Sistem Informasi, STMIK Lombok, Praya

E-mail: eken.apache@gmail.com

#### **PENDAHULUAN**

*Clustering* atau klasterisasi adalah Teknik untuk mengelompokan data berdasarkan kesamaan atau perbedaan yang ada dari data tersebut (Sulistiyawati & Supriyanto, 2021). Disebut juga sebagai suatu proses mengelompokan data menjadi beberapa kelompok atau sering disebut *cluster* (Sari & Sutabri, 2023). Dalam *cluster* data yang dikelompokan dalam satu *cluster* memiliki Tingkat kemiripan karakteristik yang tinggi, sedangkan data antar cluster memiliki Tingkat kemiripan karakteristik ketatan yang rendah (Khormarudin, 2016). Kesamaan yang dimaksud merupakan pengukuran secara numeric terhadap dua buah objek data (Hidayatullah et al., 2018). *Cluster* tidak memerlukan data pelatihan (training data) untuk mengelompokan data karena *cluster* adalah termasuk tipe *unsupervised learning* (Herdiana et al., 2021).

*Clustering* merupakan metode segmentasi yang memudahkan dalam menganalisis pola, tren dan hubungan antara data (Hairani et al., 2020). Pengelompokan ini berguna dalam prediksi dan analisis tertentu misalnya dalam analisis segemen pasar, marketing dan zona wilayah (Sulistyowati, 2018). Teknik *clustering* juga dapat digunakan untuk mendeteksi anomali yaitu dapat mengidentifikasi pola data yang tidak sama dengan data pada umumnya (Chafid & Wibowo, 2018). Selain itu *clustering* juga bisa dimanfaatkan sebagai

Teknik reduksi dimensi data dengan mengelompokan variable serupa (Hendrastuty, 2024).

### ***K-Means***

*K-Means* merupakan algoritma yang bersifat *unsupervised learning* yang memiliki fungsi untuk mengelompokan data kedalam *cluster* (Khormarudin, 2016). *K-means* dapat mengolah data tanpa label data, yang mengelompokan data menjadi beberapa cluster yang menjelaskan data yang ada dalam satu kelompok memiliki Tingkat kemiripan yang tinggi dan memiliki Tingkat kemiripan yang rendah pada cluster yang berbeda (Irianto et al., 2022). *Cluster* menggunakan algoritma *K-means* memiliki tujuan untuk meminimalisasikan fungsi objek yang telah diset kedalam *cluster*, tujuan tersebut dilakukan dengan meminimalkan variasi data yang ada dalam cluster dan memaksimalkan variasi data yang ada di *cluster* lain (Muningsih & Kiswati, 2018).

Algoritma *K-Means* memiliki dua tugas utama yakni menentukan nilai terbaik untuk titik pusat K atau *centeroid* dengan proses *iterative* (perulangan) dan menetapkan setiap titik data pada pusat K terdekat (Rahmadhani, 2022). Titik-titik data yang dekat dengan pusat K tertentu lalu dibuat sebuah cluster.

*K-Means clustering* merupakan algoritma dasar yang diterapkan untuk menentukan jumlah kelompok atau *cluster* data, menghitung rata-rata data yang ada di *cluster* (Putra et al., 2023). Tahapan *clustering* data mulai dari menentukan K (nilai bebas) sebagai *cluster* yang ingin dibentuk. Memuat nilai random sebagai titik *centeroid* data *cluster* menghitung setiap jarak data input dengan masing-masing *centeroid* data (Setiawan et al., 2023). Berikut formula yang digunakan menghitung jarak data dengan *centeroid* data.

## KESIMPULAN

K-means merupakan algoritma pembelajaran tanpa pengawasan yang umum digunakan untuk mengelompokkan data, seperti pengelompokan data pelanggan berdasarkan usia, pendidikan, tahun bekerja, pendapatan, dan rasio utang. Menggunakan metode elbow untuk menentukan jumlah cluster optimal, diperoleh tiga cluster. Cluster 0 terdiri dari pelanggan dengan usia rata-rata 40 tahun, pendidikan rata-rata level 3, tahun bekerja di atas 20 tahun, pendapatan di atas 50, dan rasio utang minimal 4.0. Cluster 1 mencakup pelanggan dengan usia minimal 20 tahun, pendidikan level 1 dan 2, tahun bekerja minimal 4 tahun, pendapatan di atas 10, dan rasio utang minimal 1.5. Cluster 2 berisi pelanggan dengan usia rata-rata 47 tahun, pendidikan level 1, tahun bekerja antara 20 hingga 31 tahun, pendapatan sebesar 253, dan rasio utang 7.2.

## DAFTAR PUSTAKA

- Adani, N. F., Boy, A. F., & Syahputra, R. (2019). Implementasi Data Mining Untuk Pengelompokan Data Penjualan Berdasarkan Pola Pembelian Menggunakan Algoritma K-Means Clustering Pada Toko Syihan. *Jurnal Cyber Tech*, 2(5).
- Chafid, N., & Wibowo, I. A. (2018). Implementasi Data Mining Untuk Clustering Daerah Penyebaran Penyakit Demam Berdarah Di Kota Tangerang Selatan Menggunakan Algoritma K-Means. *vol*, 3, 13.
- Erfina, A. (2021). *Buku Ajar Data Mining*. Nusa Putra Press.
- Hairani, H., Saputro, K. E., & Fadli, S. (2020). K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4. 5, SVM, dan naive Bayes. *Jurnal Teknologi dan Sistem Komputer*, 8(2), 89-93.
- Hendrastuty, N. (2024). Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil

- Pembelajaran Siswa. *Jurnal Ilmiah Informatika dan Ilmu Komputer (JIMA-ILKOM)*, 3(1), 46-56.
- Herdiana, O., Maulani, S., & Firdaus, E. A. (2021). Strategi Pemasaran Produk Industri Kreatif Menggunakan Algoritma K-Means Clustering Berbasis Particle Swarm Optimization. *Nuansa Informatika*, 15(2), 1-13.
- Hidayatullah, D. P., Rokhmawati, R. I., & Perdanakusuma, A. R. (2018). Analisis Pemetaan Pelanggan Potensial Menggunakan Algoritma K-Means dan LRFM Model Untuk Mendukung Strategi Pengelolaan Pelanggan (Studi Pada Maninjau Center Kota Malang). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(8), 2406-2415.
- Irianto, M. R., Maududie, A., & Arifin, F. N. (2022). Implementation of K-Means Clustering Method for Trend Analysis of Thesis Topics (Case Study: Faculty of Computer Science, University of Jember). *Berkala Sainstek*, 10(4), 210-226.
- Khormarudin, A. N. (2016). Teknik data mining: Algoritma K-Means clustering. *J. Ilmu Komput*, 1-12.
- Muningsih, E., & Kiswati, S. (2018). Sistem aplikasi berbasis optimasi metode elbow untuk penentuan clustering pelanggan. *Joutica*, 3(1), 117-124.
- Pribadi, W. W., Yunus, A., & Wiguna, A. S. (2022). Perbandingan Metode K-Means Euclidean Distance Dan Manhattan Distance Pada Penentuan Zonasi Covid-19 Di Kabupaten Malang. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 6(2), 493-500.
- Putra, R. F., Zebua, R. S. Y., Budiman, B., Rahayu, P. W., Bangsa, M. T. A., Zulfadhilah, M., Choirina, P., Wahyudi, F., & Andiyan, A. (2023). *Data Mining: Algoritma dan Penerapannya*. PT. Sonpedia Publishing Indonesia.



- Rahmadhani, A. (2022). Pendekatan Clustering untuk Menganalisis Efisiensi dan Kinerja Mahasiswa Berdasarkan Data Menerapkan Metode K-Means. *Jurnal Media Informatika Budidarma*, 6(4), 2461-2468.
- Saikin, S., Fadli, S., & Ashari, M. (2021). Optimization of Support Vector Machine Method Using Feature Selection to Improve Classification Results. *JISA (Jurnal Informatika dan Sains)*, 4(1), 22-27.
- Saikin, S., & Kusrini, K. (2019). Model Data Mining Untuk Karakteristik Data Traveller Pada Perusahaan Tour and Travel. *Jurnal Manajemen Informatika dan Sistem Informasi*, 2(2), 61-68.
- Sari, W. P., & Sutabri, T. (2023). Analisa Cluster Dengan K-Mean Clustering Untuk Pengelompokan Data Cybercrime. *Jurnal Informatika Teknologi dan Sains (Jinteks)*, 5(1), 49-53.
- Setiawan, Z., Fajar, M., Priyatno, A. M., Putri, A. Y. P., Aryuni, M., Yuliyanti, S., Widi Putra, H., Meilani, B. D., Ibrahim, R. N., & Azdy, R. A. (2023). *Buku Ajar Data Mining*. PT. Sonpedia Publishing Indonesia.
- Sulistiyawati, A., & Supriyanto, E. (2021). Implementasi Algoritma K-means Clustering dalam Penentuan Siswa Kelas Unggulan. *Jurnal Tekno Kompak*, 15(2), 25-36.
- Sulistyowati, S. (2018). Analisa Segmentasi Konsumen Menggunakan Algoritma K-Means Clustering. Prosiding Seminar Nasional Sains dan Teknologi Terapan, <https://www.kaggle.com/>

## PROFIL PENULIS



### **Saikin, S.Kom., M.Kom**

Penulis adalah seorang profesional di bidang data science dengan pengalaman lebih dari lima tahun dalam analisis data dan pengembangan model machine learning. Lulusan dari Universitas Amikom Yogyakarta, ia memiliki latar belakang akademis yang kuat di bidang ilmu komputer. Keahliannya mencakup pengolahan data,

eksplorasi data, serta penerapan algoritma pembelajaran mesin seperti K-means, regresi, dan jaringan saraf tiruan. Selain bekerja sebagai data scientist dan dosen pada salah satu perguruan tinggi swasta yang ada di Pulau Lombok Nusa Tenggara barat yaitu STMIK Lombok, penulis juga aktif menulis artikel tentang analisis data dan pembelajaran mesin, serta memberikan pelatihan dan seminar kepada profesional muda yang ingin mendalami bidang data science. Dengan komitmen tinggi terhadap inovasi teknologi dan perkembangan ilmu data, penulis terus berupaya membawa dampak positif bagi dunia bisnis dan teknologi.

## **BAB 12**

### **APLIKASI CLUSTERING C-MEANS DENGAN PYTHON**

Dadang Iskandar Mulyana

E-mail: mahvin2012@gmail.com

#### **PENDAHULUAN**

Klusterisasi adalah teknik analisis data yang bertujuan untuk mempartisi data menjadi sejumlah kelompok (kluster) sesuai dengan derajat keterkaitan antar data. Dari banyak nama yang digunakan untuk menggambarkan klusterisasi, salah satu istilah yang paling populer adalah C-Means Clustering yang digunakan di berbagai bidang, termasuk segmentasi gambar, analisis pasar dan bentuk, pemasaran target, dan dalam klusterisasi, diasumsikan bahwa semua unit dalam sebuah kluster serupa sementara unit dalam kluster lainnya semuanya berbeda. Klusterisasi adalah metode yang digunakan dalam pembelajaran tanpa pengawasan yang merupakan pembelajaran di mana tidak ada label atau kelas yang telah ditentukan untuk data yang diberikan. Ini adalah algoritma yang berusaha menemukan pola dasar dalam data dan menggunakan itu untuk mengelompokkan data bersama.

#### **METODE-METODE DALAM CLUSTERING**

Clustering memiliki beberapa metode yaitu :

##### ***1. Partition-Based Clustering***

Metode ini mengelompokkan data ke dalam sejumlah cluster tertentu. Algoritma partisi atau partition based biasanya membutuhkan sejumlah cluster yang harus diperoleh dalam bentuk input dan beroperasi pada pengoptimalan letak pusat

cluster sedemikian rupa sehingga data yang berada dalam satu cluster memiliki kesamaan tertentu.

Contoh Algoritma:

- ❖ **K-Means**: Mengelompokkan data berdasarkan jarak Euclidean ke pusat cluster yang diperbarui secara iteratif.
- ❖ **K-Medoids**: Mirip dengan K-Means, tetapi pusat cluster adalah titik data nyata (medoid), membuatnya lebih tahan terhadap outlier.

Contoh Penggunaan: Segmentasi pelanggan dalam pemasaran, analisis pola penggunaan pada aplikasi.

## 2. *Hierarchical Clustering*

Teknik ini membangun hierarki informasi dengan mengatur informasi dalam bentuk pohon (dendrogram). Klusterisasi hierarkis tidak memerlukan jumlah kluster sebagai masukan karena ia membuat kluster pada berbagai tingkatan.

Contoh Algoritma:

- ❖ **Agglomerative Hierarchical Clustering**: Setiap data sebagai unit penghasil kluster, kemudian secara berangsur-angsur terintegrasi dalam satu kluster yang besar. Jarak antar kluster dapat ditentukan dengan cara single linkage (menghubungkan titik-titik terdekat), complete linkage (titik-titik terjauh), dan average linkage (jarak rata-rata).
- ❖ **Divisive Hierarchical Clustering**: Dimulai dari satu cluster besar yang berisi semua data, kemudian dipecah bertahap menjadi cluster kecil.

Contoh Penggunaan: Analisis evolusi biologi, pengelompokan dokumen dalam analisis teks, pemetaan struktur organisasi.

# **PENERAPAN APLIKASI SEDERHANA DATA MINING**

## **(dengan Bahasa Pemrograman Python)**

Buku Penerapan Aplikasi Sederhana Data Mining dengan Bahasa Pemrograman Python merupakan penerapan aplikasi sederhana teknik data mining dengan menggunakan pemrograman python yang dirancang untuk pemula yang ingin belajar teknik data mining dalam aplikasinya. Buku ini mencakup berbagai konsep teknik data mining seperti teknik klasifikasi, forecasting, association, dan clustering hingga setting library python untuk melakukan penerapan aplikasi sederhana dalam teknik data mining. Berdasarkan pendekatan yang praktis dan disertai dengan contoh aplikasi sederhana teknik data mining yang mudah diikuti, buku ini membantu pembaca memahami cara memproses suatu data menggunakan teknik data mining yang dijalankan dengan Bahasa Pemrograman Python. Setiap bab dilengkapi dengan contoh aplikasi sederhana teknik data mining yang bertujuan untuk menguatkan penalaran dan logika pemahaman pembaca terhadap materi yang telah dipelajari. Selain itu, buku ini juga menyajikan dasar data science berupa visualisasi data dari teknik data mining, yaitu Teknik Klasifikasi (K-NN, Naïve Bayes, Decision tree C 4.5), Teknik Forecasting (Moving Average, Exponential Smoothing, Regresi Linear, ARIMA), Teknik Association dengan menggunakan Apriori, Teknik Clustering (K-Means, C-Means).



**FUTURE SCIENCE**

Jl. Terusan Surabaya, Gang 1 A No. 71 RT 002 RW 005,  
Kel. Sumbarsari, Kec. Lowokwaru, Kota Malang,  
Provinsi Jawa Timur.  
Website : [www.futuresciencepress.com](http://www.futuresciencepress.com)



**IKAPI**  
IKATAN PENULIS INDONESIA

No. 348/JTI/2022

ISBN 978-634-7037-89-3 (PDF)



9

786347

037893