Anis Fitri Nur Masruriyah
Cici Emilia Sukmawati
Budi Arif Dermawan

Memahami DATA MINING dengan Python

IMPLEMENTASI PRAKTIS

Memahami DATA MINING dengan Python

IMPLEMENTASI PRAKTIS

Buku ini dimulai dengan penjelasan tahapan dalam beberapa pendekatan data mining populer seperti KDD, CRISP-DM, dan SEMMA, diikuti dengan diskusi tentang konsep dasar statistik, eksplorasi data, dan perbedaan antara machine learning dan data mining. Topik-topik lain yang dibahas meliputi pengumpulan data, pembersihan data, transformasi data, evaluasi model, dan visualisasi data.

Selanjutnya, membahas jenis algoritma clustering seperti K-Means, K-Medoids, Hierarchical Clustering, dan DBSCAN, serta penggunaan dan evaluasi algoritma tersebut dengan contoh data iris. Kemudian, menjelaskan algoritma pohon keputusan seperti CART, ID3, C4.5, CHAID, dan algoritma ensemble seperti Random Forest, XGBoost, dan LightGBM, beserta kode dan fungsi di Python.

Selain itu, membahas evaluasi dalam konteks klasifikasi, regresi, dan pengelompokan, serta algoritma regresi dan klasifikasi seperti Regresi Linear, Regresi Logistik, Support Vector Machine (SVM), dan K-Nearest Neighbors (KNN). Topik lain termasuk algoritma asosiasi seperti Apriori dan FP-Growth, evaluasi cross-validation, serta metrik evaluasi kinerja model.

Buku ini juga mencakup berbagai fungsi dan pustaka Python yang digunakan dalam data mining dan machine learning, serta penjelasan tentang evaluasi algoritma pohon keputusan, SVM, dan KNN. Terakhir, membahas algoritma regresi, metrik evaluasi, dan konsep MSE baseline.

eureka media aksara
Anggota IKAPI
No. 225/JTE/2021

0858 5343 1992
eurekamediaaksara@gmail.com
Jl. Banjaran RT.20 RW.10
Bojongsari - Purbalingga 53362

ISBN 978-623-120-424-0
9 786231 204240

MEMAHAMI DATA MINING DENGAN PYTHON: IMPLEMENTASI PRAKTIS

Anis Fitri Nur Masruriyah
Cici Emilia Sukmawati
Budi Arif Dermawan

PENERBIT CV. EUREKA MEDIA AKSARA

MEMAHAMI DATA MINING DENGAN PYTHON: IMPLEMENTASI PRAKTIS

Penulis : Anis Fitri Nur Masruriyah
Cici Emilia Sukmawati
Budi Arif Dermawan

Desain Sampul : Eri Setiawan

Tata Letak : Uli Mas'uliyah Indarwati

ISBN : 978-623-120-424-0

Diterbitkan oleh : **EUREKA MEDIA AKSARA, MARET 2024**
ANGGOTA IKAPI JAWA TENGAH
NO. 225/JTE/2021

Redaksi:
Jalan Banjaran, Desa Banjaran RT 20 RW 10 Kecamatan Bojongsari
Kabupaten Purbalingga Telp. 0858-5343-1992
Surel : eurekamediaaksara@gmail.com
Cetakan Pertama : 2024

All right reserved

Hak Cipta dilindungi undang-undang
Dilarang memperbanyak atau memindahkan sebagian atau seluruh isi buku ini dalam bentuk apapun dan dengan cara apapun, termasuk memfotokopi, merekam, atau dengan teknik perekaman lainnya tanpa seizin tertulis dari penerbit.

KATA PENGANTAR

Selamat datang dalam buku "Memahami Data Mining dengan Python: Implementasi Praktis". Buku ini merupakan sebuah panduan komprehensif yang dirancang untuk membantu pembaca memahami konsep dasar data mining dan menerapkannya secara praktis menggunakan bahasa pemrograman Python.

Data mining telah menjadi salah satu bidang yang semakin penting dalam dunia teknologi informasi, di mana kita berada di era di mana jumlah data yang dihasilkan terus meningkat secara eksponensial. Dalam menghadapi tantangan ini, pengetahuan tentang bagaimana menggali informasi berharga dari data menjadi semakin krusial.

Buku ini ditujukan untuk semua kalangan yang tertarik untuk memahami konsep data mining dan ingin menerapkannya dalam konteks praktis menggunakan Python. Mulai dari mahasiswa yang sedang belajar tentang data mining, profesional di bidang teknologi informasi, hingga para peneliti yang ingin memperluas pemahaman mereka tentang analisis data, buku ini akan memberikan wawasan yang berharga dan panduan langkah demi langkah untuk memulai.

Dalam buku ini, pembaca akan dibawa melalui berbagai konsep data mining yang penting, mulai dari pemrosesan data awal, pemilihan fitur, model pembelajaran mesin, hingga evaluasi kinerja model. Setiap konsep akan disertai dengan contoh implementasi praktis menggunakan bahasa pemrograman Python, sehingga pembaca dapat dengan mudah mengikuti dan memahami setiap langkahnya.

Kami berharap buku ini dapat menjadi sumber pengetahuan yang bermanfaat dan membantu pembaca memperluas pemahaman mereka tentang data mining serta menerapkannya dalam situasi dunia nyata. Terima kasih telah memilih buku ini, dan kami berharap Anda menemukan pengalaman belajar yang memuaskan.

Selamat membaca!

DAFTAR ISI

KATA PENGANTAR ..... **iii**
DAFTAR ISI ..... **iv**
DAFTAR TABEL ..... **vi**
DAFTAR GAMBAR ..... **vii**
BAB 1 PENDAHULUAN **1**
A. Pengantar Data Mining 1
B. Tujuan dan Manfaat Data Mining 1
C. Proses Data Mining 2
D. Tools yang Digunakan 10
BAB 2 KONSEP DASAR DAN PEMAHAMAN DATA **26**
A. Definisi dan Karakteristik Data Mining 26
B. Jenis-jenis Data dalam Data Mining 28
C. Pengumpulan dan Pemahaman Data 29
D. Konsep Dasar Statistik untuk Data Mining 31
BAB 3 EXPLARATORY DATA ANALYSIS **34**
A. Eksplorasi Data 34
B. Visualisasi Data 36
C. Integrasi Data 40
D. Pembersihan Data 42
E. Transformasi dan Reduksi Dimensi Data 46
BAB 4 UNSUPERVISED LEARNING **51**
A. Algoritma K-Means 51
B. Algoritma K-Medoids 54
C. Algoritma Density-Based Spatial Clustering of Applications with Noise (DBSCAN) 57
D. Algoritma Hierarchical Clustering 60
BAB 5 SUPERVISED LEARNING **63**
A. Algoritma Pohon Keputusan 64
B. Algoritma Support Vector Machine 68
C. Algoritma K-Nearest Neighbour 71
D. Algoritma Regresi Linear 73
E. Algoritma Regresi Logistik 75
F. Algoritma Asosiasi 77

BAB 6 EVALUASI MODEL **82**
A. Evaluasi Pada Algoritma Pengelompokan 82
B. Evaluasi Pada Algoritma Klasifikasi 83
C. Evaluasi pada Algoritma Regresi 85
DAFTAR PUSTAKA **87**
TENTANG PENULIS **88**

DAFTAR TABEL

Tabel 5. 1 Jenis Kernel pada SVM 68
Tabel 5. 2 Fungsi Jarak pada KNN.....72
Tabel 5. 3 Matrik Konfusi.....84

DAFTAR GAMBAR

Gambar 1. 1 Tahapan KDD 3
Gambar 1. 2 Tahapan CRISP-DM 5
Gambar 1. 3 Tahapan SEMMA 8
Gambar 1. 4 Halaman download anaconda navigator 10
Gambar 1. 5 Tahap pertama installasi 11
Gambar 1. 6 Tahap ke dua installasi 11
Gambar 1. 7 Tahap ke tiga proses installasi 12
Gambar 1. 8 Tahap ke empat proses installasi 13
Gambar 1. 9 Tahap ke lima proses installasi 13
Gambar 1. 10 Tahap ke enam proses installasi 14
Gambar 1. 11 Tahap ke tujuh proses installasi 14
Gambar 1. 12 Tahap ke delapan proses installasi 15
Gambar 1. 13 Tahap pertama setup anaconda 16
Gambar 1. 14 Tahap ke dua setup anaconda 16
Gambar 1. 15 Tahap ke tiga setup anaconda 17
Gambar 1. 16 Tahap ke empat setup anaconda 17
Gambar 1. 17 Tahap ke lima setup anaconda 18
Gambar 1. 18 Tahap ke enam setup anaconda 19
Gambar 1. 19 Tahap pertama setup jupyter notebook 19
Gambar 1. 20 Tahap ke dua setup jupyter notebook 20
Gambar 1. 21 Tahap ke tiga setup jupyter notebook 20
Gambar 1. 22 Tahap ke empat setup jupyter notebook 21
Gambar 1. 23 Tahap ke lima setup jupyter notebook 22
Gambar 3. 1 Ilustrasi Visualisasi Data 37
Gambar 6. 1 Ilustrasi Cross-Validation 85

MEMAHAMI DATA MINING DENGAN PYTHON: IMPLEMENTASI PRAKTIS

Anis Fitri Nur Masruriyah
Cici Emilia Sukmawati
Budi Arif Dermawan

BAB 1 PENDAHULUAN

A. Pengantar Data Mining

Data mining adalah suatu proses ekstraksi pengetahuan yang bermanfaat atau pola yang menarik dari suatu kumpulan data yang besar, kompleks, dan umumnya tidak terstruktur. Tujuan utama dari data mining adalah menemukan informasi yang tersembunyi, berpotensi memberikan wawasan, dan mendukung pengambilan keputusan. Dalam data mining, digunakan berbagai teknik dan algoritma untuk mengidentifikasi pola atau hubungan yang tidak terlihat secara langsung oleh manusia. Metode ini melibatkan analisis statistik, pembelajaran mesin, kecerdasan buatan, dan pemodelan matematis. Data mining dapat diterapkan dalam berbagai bidang seperti bisnis, ilmu pengetahuan, kesehatan, keuangan, dan lainnya.

B. Tujuan dan Manfaat Data Mining

Tujuan utama Data Mining adalah mengeksplorasi data dengan tujuan menemukan wawasan yang bermanfaat dan mendukung pengambilan keputusan yang lebih baik. Melalui teknik seperti klastering, klasifikasi, dan regresi, Data Mining dapat membantu mengidentifikasi pola tersembunyi, memprediksi perilaku masa depan, dan mengoptimalkan strategi bisnis. Manfaatnya meluas dari peningkatan efisiensi operasional hingga pengembangan strategi pemasaran yang lebih cerdas.

BAB 2

KONSEP DASAR DAN PEMAHAMAN DATA

A. Definisi dan Karakteristik Data Mining

Data mining bukan hanya tentang aplikasi algoritma dan teknik analisis, tetapi juga melibatkan pemahaman konteks bisnis dan interpretasi hasil untuk pengambilan keputusan yang efektif. Data Mining adalah proses ekstraksi pengetahuan yang bermanfaat atau pola tersembunyi dari suatu kumpulan data besar menggunakan berbagai teknik statistik, matematika, dan kecerdasan buatan. Tujuan utama dari data mining adalah mengidentifikasi hubungan atau pola yang dapat membantu dalam pengambilan keputusan dan memahami fenomena yang terkandung dalam data.

Karakteristik Data Mining:

1. Non-Trivial:
   Analisis data mining mencari pola atau pengetahuan yang tidak dapat dengan mudah atau langsung ditemukan secara manual. Solusi yang ditemukan oleh data mining seringkali bersifat non-trivial dan dapat menyediakan wawasan baru.
2. Implicit, Previously Unknown, and Potentially Useful:
   Data mining dapat mengidentifikasi informasi implisit atau pengetahuan yang sebelumnya tidak diketahui. Hasilnya memiliki potensi untuk memberikan pemahaman baru yang dapat digunakan untuk pengambilan keputusan atau perbaikan proses.

BAB 3

EXPLARATORY DATA ANALYSIS

A. Eksplorasi Data

Eksplorasi data (data exploration) adalah tahap awal dalam proses data mining di mana analis data mencari pemahaman yang lebih dalam terhadap karakteristik data yang ada. Tujuan eksplorasi data adalah mengidentifikasi pola, tren, anomali, dan wawasan awal sebelum proses analisis lebih lanjut. Deskripsi umum dari eksplorasi data adalah memahami sifat dan karakteristik umum dari data, seperti tipe variabel, ukuran data, dan format. Secara statistik eksplorasi data menggunakan metrik seperti rata-rata, median, dan deviasi standar untuk merangkum data secara deskriptif.

Python menyediakan berbagai pustaka dan fungsi yang sangat berguna untuk melakukan Exploratory Data Analysis (EDA), yang membantu untuk memahami dataset Anda dengan lebih baik. Beberapa pustaka utama dan fungsi yang digunakan dalam proses EDA di Python antara lain:

1. Pustaka Umum
   a. Pandas
      Pandas adalah salah satu pustaka yang sangat penting dalam ekosistem Python untuk melakukan manipulasi dan analisis data. Dalam konteks data mining, Pandas berfungsi sebagai struktur data yang kuat dan alat analisis yang fleksibel.
   b. NumPy
      NumPy (Numerical Python) adalah pustaka yang sangat penting dalam ekosistem Python untuk melakukan operasi numerik dan manipulasi array

BAB 4

UNSUPERVISED LEARNING

Unsupervised learning adalah salah satu paradigma machine learning di dalam data mining yang tidak melibatkan variabel target atau label pada data yang digunakan. Dalam unsupervised learning, tujuan utamanya adalah mengeksplorasi struktur atau pola yang ada dalam data tanpa memiliki panduan yang jelas tentang hasil yang diinginkan. Unsupervised learning sangat berguna ketika tujuan analisis adalah mengeksplorasi struktur internal data, menemukan pola tersembunyi, atau membuat pengelompokan berdasarkan kesamaan antar data. Dengan menggunakan teknik-teknik seperti clustering dan reduksi dimensi, unsupervised learning dapat memberikan wawasan yang berharga dalam analisis data mining.

A. Algoritma K-Means

K-Means adalah salah satu algoritma pengelompokan (clustering) yang digunakan untuk mengelompokkan data ke dalam k kelompok berdasarkan kesamaan antar data. Tujuannya adalah untuk meminimalkan varians intra-klaster, yaitu jarak antara data dalam suatu klaster dengan pusat klasternya. Algoritma ini bekerja dengan cara mengiterasi antara dua langkah:

1. Inisialisasi Pusat Klaster (Centroid): Pusat klaster awalnya dipilih secara acak dari dataset.
2. Assigning Data ke Klaster: Setiap data diatributkan ke klaster yang memiliki pusat terdekat.

BAB 5

SUPERVISED LEARNING

Supervised Learning adalah paradigma dalam machine learning di mana model atau algoritma dilatih menggunakan dataset yang telah diberi label. Dalam supervised learning, dataset pelatihan mencakup pasangan input dan output yang sudah diketahui. Tujuan utama adalah untuk membuat model yang dapat mempelajari hubungan atau pola antara input dan output sehingga dapat membuat prediksi atau klasifikasi yang akurat untuk data baru yang belum pernah dilihat sebelumnya. Secara umum tahapan dalam supervised learning melibatkan:

1. Pemilihan Model: Memilih jenis model atau algoritma yang cocok untuk tugas yang diinginkan, seperti regresi linear, regresi logistik, atau berbagai jenis model dalam machine learning yang lebih kompleks.
2. Pengumpulan Dataset: Mengumpulkan dataset yang memuat pasangan input dan output yang sudah diketahui. Dataset ini biasanya dibagi menjadi dua bagian: satu untuk pelatihan model (training set) dan satu untuk menguji performa model (test set).
3. Pemisahan Data: Memisahkan dataset menjadi dua bagian, yaitu data pelatihan untuk melatih model dan data pengujian untuk mengukur seberapa baik model dapat melakukan prediksi pada data baru.
4. Pelatihan Model: Memasukkan data pelatihan ke dalam model untuk melatihnya. Model memperbarui parameter internalnya selama proses pelatihan untuk mengoptimalkan performa.

BAB 6

EVALUASI MODEL

A. Evaluasi Pada Algoritma Pengelompokan

Evaluasi pada pengelompokan (clustering) merupakan langkah penting dalam menilai kualitas hasil pengelompokan yang dihasilkan oleh algoritma clustering. Berbagai metrik evaluasi digunakan untuk mengukur sejauh mana pengelompokan mencerminkan struktur sebenarnya dalam data. Dalam mengevaluasi pengelompokan, pemilihan metrik evaluasi yang sesuai dengan karakteristik data dan tujuan analisis clustering menjadi kunci untuk mendapatkan pemahaman yang akurat terhadap kualitas dan relevansi hasil clustering. Beberapa jenis evaluasi clustering yang umum digunakan melibatkan metrik-metrik berikut:

1. Silhouette Score:
   a. Mengukur seberapa baik setiap titik dalam sebuah klaster dari klaster tetangganya dan seberapa baik klaster tersebut terpisah.
   b. Rentang nilai antara -1 hingga 1, dengan nilai yang lebih tinggi menunjukkan hasil clustering yang lebih baik.
2. Davies-Bouldin Index:
   a. Mengukur seberapa baik setiap klaster dipisahkan satu sama lain dan seberapa kompak klaster tersebut.
   b. Nilai yang lebih rendah menunjukkan hasil clustering yang lebih baik.

DAFTAR PUSTAKA

Porcu, V. (2018). Python for Data Mining Quick Syntax Reference. In *Python for Data Mining Quick Syntax Reference*. https://doi.org/10.1007/978-1-4842-4113-4

Raja, R., Nagwanshi, K. K., Kumar, S., & Laxmi, K. R. (2022). *Data Mining and Machine Learning Applications*.

Swamynathan, M. (2019). Mastering Machine Learning with Python in Six Steps. In *Mastering Machine Learning with Python in Six Steps*. Apress. https://doi.org/10.1007/978-1-4842-4947-5

Wang, L. (2017). Data Mining, Machine Learning and Big Data Analytics. *International Transaction of Electrical and Computer Engineers System*, *4*(2), 55–61. https://doi.org/10.12691/iteces-4-2-2

Zaki, M. J., & Meira, W. J. (2020). Data Mining and Machine Learning Fundamental Concepts and Algorithms. In *Journal of Chemical Information and Modeling* (Vol. 53, Issue 9).

Zaki, M. J., & Wagner Jr, M. (2020). *Data Mining and Machine Learning Fundamental Concepts and Algorithms*.

TENTANG PENULIS

Anis Fitri Nur Masruriyah, seorang akademisi dan peneliti yang berperan aktif dalam ranah pendidikan tinggi di Indonesia, telah lahir pada tahun 1992. Menekuni karir pengajarannya sejak tahun 2019 di salah satu perguruan tinggi di Indonesia, Anis memperoleh gelar Magister Komputer (M.Kom) dari Institut Pertanian Bogor (IPB) pada tahun 2018. Sebelumnya, pada tahun 2015, ia meraih gelar Sarjana Komputer (S.Kom) dari Universitas Brawijaya.

Selama perjalanan karirnya di bidang pendidikan, Anis telah berperan secara aktif dalam kegiatan penelitian dan pengabdian kepada masyarakat sejak tahun 2019. Ia turut memberikan kontribusi dalam upaya pengabdian masyarakat, dengan tujuan memanfaatkan pengetahuannya demi kemajuan masyarakat dan negara.

Sebagai seorang pendidik dan peneliti, buku "Data Mining" merupakan salah satu karya Anis yang bertujuan untuk berbagi ilmu pengetahuan. Anis berkomitmen untuk terus menghasilkan karya-karya berikutnya yang sejalan dengan misi ini.

Informasi kontak Anis yang dapat dihubungi melalui email adalah anis.masruriyah@gmail.com.

Cici Emilia Sukmawati, lahir di Kuningan tahun 1994 merupakan seorang yang aktif di bidang akademik pada salah satu universitas swasta di Indonesia.

Penulis adalah seorang pemula yang bersemangat dalam dunia data mining. Meskipun masih baru dalam bidang ini, Penulis memiliki minat yang besar dalam memahami dan mengaplikasikan konsep-konsep data mining dalam analisis data.

Dengan latar belakang pendidikan di bidang Ilmu Komputer, Penulis mulai tertarik pada data mining ketika belajar tentang kekuatan analisis data untuk mengungkap pola-pola yang berharga dalam set data.

Buku ini merupakan hasil dari eksplorasi dalam memahami konsep-konsep dasar data mining dan mengaplikasikannya dalam kasus-kasus sederhana. Penulis berharap bahwa buku ini dapat menjadi panduan yang bermanfaat bagi pembaca yang juga tertarik untuk memulai perjalanan mereka dalam dunia data mining.

Meskipun masih dalam tahap awal, Penulis berkomitmen untuk terus belajar dan berkembang dalam bidang data mining. Ia percaya bahwa dengan kesungguhan dan semangat belajar, ia dapat memberikan kontribusi yang berarti dalam pengembangan bidang ini di masa depan. Informasi kontak Cici dapat dihubungi melalui email: ciciemiliasukmawati@gmail.com

Budi Arif Dermawan, seorang akademisi dan peneliti yang berperan aktif dalam dunia pendidikan tinggi di Indonesia, lahir pada tahun 1990. Ia telah memulai karir pengajarannya sejak tahun 2018 di salah satu perguruan tinggi di Indonesia. Gelar Magister Komputer (M.Kom) diperolehnya dari Institut Pertanian Bogor (IPB) pada tahun 2017, setelah sebelumnya meraih gelar Sarjana Komputer (S.Kom) dari Universitas Singaperbangsa Karawang pada tahun 2013.

Selama perjalanan karirnya di dunia pendidikan, Budi telah aktif terlibat dalam kegiatan penelitian dan pengabdian kepada masyarakat sejak tahun 2018. Kontribusinya juga terlihat dalam upaya pengabdian kepada masyarakat, dimana tujuan utamanya adalah memanfaatkan pengetahuannya untuk kemajuan masyarakat dan negara.

Sebagai seorang pendidik dan peneliti, buku ini merupakan salah satu karya Budi yang bertujuan untuk berbagi ilmu pengetahuan. Dengan komitmennya yang kuat, Budi berencana untuk terus menghasilkan karya-karya berikutnya yang sejalan dengan misi ini.

Informasi kontak Budi dapat dihubungi melalui email budi.arif@staff.unsika.ac.id.