



RAHMADYA TRIAS HANDAYANTO
HERLAWATI

DATA MINING DAN MACHINE LEARNING MENGUNAKAN MATLAB DAN PYTHON

Konsep Data Mining | Mempersiapkan Matlab | Mempersiapkan Python | Metode Data Mining Berbasis Statistika | Adaptive Neuro-fuzzy Inference System (ANFIS) | Klusterisasi (clustering) | Support Vector Machine (SVM) | Validasi dan Pengujian Model | Data Mining World Wide Web



Penerbit **INFORMATIKA**

Data Mining dan Machine Learning Menggunakan Matlab dan Python

Rahmadya Trias Handayanto
Herlawati



Penerbit **INFORMATIKA**

Data Mining dan
Machine Learning
Menggunakan
Matlab dan Python

**Data Mining dan Machine Learning
Menggunakan Matlab dan Python**

Penyusun : Rahmadya Trias Handayanto
Herlawati

Penerbit : Informatika Bandung

Pemasaran : **BI-Obses**
Pasar Buku Palasari No. 82
Bandung 40264
Telp.(022)7317812
Fax. (022)7317896

Cetakan : Oktober 2020

ISBN : 978-623-7131-32-8

Copyright © 2020 pada Penerbit **INFORMATIKA** Bandung

DAFTAR ISI

Kata Pengantar	v
Panduan Penggunaan CD	vii
Daftar Isi	ix
BAB 1. KONSEP DATA MINING	1
1.1. Pendahuluan.....	1
1.2. Manfaat <i>Data Mining</i>	3
1.3. Proses <i>Data Mining</i>	3
1.4. Jenis Permasalahan <i>Data Mining</i>	4
1.4.1 Klasifikasi.....	4
1.4.2 Analisis Klaster (<i>Cluster Analysis</i>)	6
1.4.3 Pencarian Hukum Asosiasi (<i>Association Rule Discovery</i>)	8
1.4.4 Pencarian Pola Berurutan (<i>Sequential Pattern Discovery</i>)	8
1.4.5 Regresi.....	9
1.4.6 Deteksi Penyimpangan (<i>Deviation/Anomaly Detection</i>)	9
1.5. Pemodelan <i>Data Mining</i>	10
1.6. Aplikasi-Aplikasi <i>Data Mining</i>	14
1.7. Bahasa Pemrograman untuk <i>Data Mining</i>	15
1.8. Terminologi <i>Data Mining</i>	16
BAB 2. MEMPERSIAPKAN MATLAB.....	19
2.1. Dasar-Dasar Penggunaan Matlab	19
2.1.1 <i>Command Window</i>	19
2.1.2 Kalkulus.....	20
2.1.3 Aljabar Linear	25
2.1.4 Statistik	30
2.1.5 Grafik.....	31
2.1.6 <i>Excel Link</i>	34
2.2. Kompilasi Program pada Matlab.....	37
2.3. Instalasi Matlab	44

BAB 3. MEMPERSIAPKAN PYTHON

3.1 Mengunduh dan Menginstal Paket Anaconda

3.2 Menjalankan Anaconda Navigator

3.3 Menggunakan Jupyter Notebook

 3.3.1 Menjalankan Jupyter Notebook

 3.3.2 Menjalankan Jupyter Notebook *Online* (Google Colab)

3.4 Memasang TensorFlow

3.5 Mengelola Data dengan Python

 3.5.1 Mengimpor Data Berformat Excel

 3.5.2 Mengimpor Data berformat CSV

 3.5.3 Mengimpor Data CSV dengan Google Colab

 3.5.4 Mengunggah Data CSV pada Google Colab

3.6 Menggunakan Prosesor-Prosesor Google

3.7 *Graphical User Interface* (GUI) Berbasis Desktop Python

3.8 *Graphical User Interface* (GUI) Berbasis Web Pada Python

 3.8.1 Mengunduh Pustaka Flask dan Jinja2

 3.8.2 Menguji Framework Flask

BAB 4. METODE DATA MINING BERBASIS STATISTIKA

4.1 Regresi dan Peramalan

 4.1.1 Menyiapkan Dataset

 4.1.2 Proses Pemodelan

 4.1.3 Pembuatan Grafik/Plot

 4.1.4 Pembuatan *Graphical User Interface* (GUI)

4.2 Pohon Keputusan

 4.2.1 Studi Kasus Pohon Keputusan

 4.2.2 Pembentukan Pohon Keputusan

 4.2.3 Membuat Aplikasi Berbasis Pohon Keputusan

4.3 Penyelesaian Klasifikasi dengan Naïve Bayes menggunakan Python

 4.3.1 Menyiapkan Dataset

 4.3.2 Memanggil Naïve Bayes Classifier

4.4. Penyelesaian Klasifikasi dengan KNN

 4.4.1 Klasifikasi dengan KNN Matlab

 4.4.2 Klasifikasi dengan KNN Python

 4.4.3 Menggunakan KNN dengan Google Colab

BAB 5. ADAPTIVE NEURO-FUZZY INFERENCE SYSTEM (ANFIS)	151
5.1 Teori Dasar	151
5.2 Studi Kasus Data Mining dengan ANFIS	153
5.3 Membuat FIS melalui Pelatihan (<i>Learning</i>).....	154
5.4 Membuat Aplikasi Berbasis ANFIS	161
BAB 6. JARINGAN SYARAF TIRUAN (JST).....	167
6.1 Teori Dasar	170
6.1.1 Fungsi Aktivasi.....	170
6.1.2 Metode Pembelajaran	171
6.1.3 Tipe Jaringan Syaraf Tiruan	172
6.1.4 Perambatan Balik (<i>Backpropagation</i>)	173
6.1.5 <i>Fuzzy Neural Network</i> (FNN)	175
6.1.6 <i>Radial Basis Function</i> (RBF).....	176
6.2 Studi Kasus Data Mining dengan JST	178
6.2.1 Membuat Aplikasi untuk <i>Backpropagation Learning</i>	179
6.2.2 Membuat Aplikasi Pengambilan Keputusan	186
6.2.3 Memadukan Sistem Pembelajaran dengan <i>Decision Support System</i> (DSS)	189
6.3 Jaringan Syaraf Tiruan (JST) dengan Python.....	191
6.3.1 Menyiapkan Pustaka (<i>Library</i>)	191
6.3.2 Menyiapkan Dataset	194
6.3.3 Menyiapkan Model JST	197
6.3.4 Validasi Model JST	199
6.3.5 Menyimpan Model JST	201
6.3.6 Menjalankan Model JST Hasil Pelatihan (<i>Training</i>)	201
6.4 JST dengan <i>Google Interactive Notebook</i> (<i>Google Colab</i>) .	203
6.4.1 Pelatihan JST	204
6.4.2 Menggunakan JST Hasil Pelatihan dengan GUI Desktop	208
6.5 Menggunakan JST Hasil Pelatihan dengan Web-based Python	212
6.6 Pemodelan JST Untuk Kasus Kelas Jamak (<i>Multi-class</i>)...	219
BAB 7. KLASTERISASI (CLUSTERING)	223
7.1 Teori Dasar	223
7.1.1 Klasterisasi Halus dengan <i>Fuzzy C-Means</i> (FCM).....	226

7.1.2	FCM dengan Matlab.....
7.1.3	Mencari Pusat Klaster
7.2	Klasterisasi K-Means dengan Python.....
7.2.1	Menyiapkan Dataset
7.2.2	Memanggil K-Means
7.3	Klasterisasi Sebagai Alat Bantu Aplikasi Klasifikasi
7.3.1	Membuat Aplikasi Berbasis Matlab
7.3.2	Aplikasi Berbasis Text
7.3.3	Aplikasi Berbasis GUI

BAB 8. SUPPORT VECTOR MACHINE (SVM)

8.1	Teori Dasar
8.1.1	Kasus Data yang Terpisah secara Linear
8.1.2	Kasus Data yang Tidak Terpisah Secara Linear.....
8.2	Studi Kasus <i>Data Mining</i> dengan SVM
8.2.1	Kasus Kelas Biner (Dua Kelas).....
8.2.2	Kasus Kelas Jamak (<i>Multiclass</i>)
	a. Membuat persamaan svmStruct1 antara kelas IPA dan IPS.
	b. Membuat Persamaan svmStruct2 antara IPA dan Bahasa.
	c. Membuat Persamaan svmStruct3 antara IPS dan Bahasa
8.2.3	Membuat Kode Program untuk Klasifikasi dengan <i>Basis Text (Text-based)</i>
8.2.4	Membuat Kode Program untuk Klasifikasi dengan <i>Graphical User Interface (GUI)</i>
8.3.	Support Vector Machine (SVM) dengan Python.....
8.3.1	Menyiapkan Dataset
8.3.2	Memanggil SVM.....
8.3.3	SVM Kelas Jamak (<i>Multi-Class</i>).....
8.3.4	Klasifikasi dengan SVM dengan GUI Desktop
8.3.5	Klasifikasi dengan SVM dengan <i>Web-based Python</i>
8.3.6	<i>Support Vector Regression (SVR)</i>

BAB 9. VALIDASI DAN PENGUJIAN MODEL	289
9.1 Pendahuluan.....	289
9.2 Data Training dan <i>Data Testing</i>	290
9.3 Validasi Silang (<i>Cross Validation</i>)	291
9.4 Kurva <i>Receiver Operating Characteristic</i> (ROC).....	292
9.5 Studi Kasus Validasi Silang dengan Matlab	294
9.6 Studi Kasus Validasi Silang dengan Python.....	298
9.7 Kalkulasi Akurasi pada Python	300
BAB 10. DATA MINING WORLD WIDE WEB.....	301
10.1 Pendahuluan.....	301
10.2 Menggunakan Google Colab untuk Menggali Data Twitter	302
10.3 Mengambil Data Twitter	302
10.4 Sentimen Analysis	304
10.5 Mengambil Data Web.....	307
DAFTAR PUSTAKA	313

1.1 Pendahuluan

Salah satu tantangan utama pada pembelajaran mesin adalah bagaimana memastikan bahwa model yang dibangun dapat bekerja dengan baik pada data baru yang belum pernah dilihat sebelumnya. Untuk mengatasi masalah ini, teknik validasi silang dan pengujian silang digunakan untuk mengevaluasi kinerja model secara lebih akurat dan konsisten. Dalam bab ini, kita akan membahas konsep-konsep dasar dari validasi silang dan pengujian silang, serta bagaimana menerapkannya dalam praktik menggunakan Matlab dan Python. Kita juga akan melihat bagaimana menghitung akurasi model pada Python.

Salah satu tantangan utama dalam pembelajaran mesin adalah bagaimana memastikan bahwa model yang dibangun dapat bekerja dengan baik pada data baru yang belum pernah dilihat sebelumnya. Untuk mengatasi masalah ini, teknik validasi silang dan pengujian silang digunakan untuk mengevaluasi kinerja model secara lebih akurat dan konsisten. Dalam bab ini, kita akan membahas konsep-konsep dasar dari validasi silang dan pengujian silang, serta bagaimana menerapkannya dalam praktik menggunakan Matlab dan Python. Kita juga akan melihat bagaimana menghitung akurasi model pada Python.

BAB

1

KONSEP DATA MINING

1.1 Pendahuluan

Beberapa *literature* yang membahas *data mining* banyak kita jumpai saat ini dan layak untuk dibaca (lihat daftar referensi). Tidak ada buku yang sanggup memberikan informasi lengkap karena Data Mining merupakan bidang yang sangat luas penerapannya. Untuk pembahasan ini beberapa materi diambil dari buku referensi tersebut (kompilasi) sesuai dengan perkembangan terkini yang cenderung mengarah ke sistem *Soft Computing* yang lebih baik dalam mengatasi permasalahan yang ada saat ini.

Data mining menurut David Hand, Heikki Mannila, dan Padhraic Smyth dari MIT adalah analisis terhadap data (biasanya data yang berukuran besar) untuk menemukan hubungan yang jelas serta menyimpulkannya yang belum diketahui sebelumnya dengan cara terkini dipahami dan berguna bagi pemilik data tersebut (Larose, 2006). Disebutkan bahwa cara yang digunakan adalah cara terkini (Novel) karena perkembangan teknologi yang terus berkembang dan database administrator harus mampu mengikuti perkembangan yang

BAB

3

MEMPERSIAPKAN PYTHON

Semua bahasa pemrograman dapat digunakan untuk *data science*, salah satunya adalah bahasa pemrograman terkenal python. Bahasa ini diluncurkan pertama kali pada tahun 1991 di *Scitiching Mathematisch Centrum*, Belanda. Walaupun menggunakan nama spesies ular, sesungguhnya perancang Python, Guido van Rossum, mengambil nama bahasa pemrograman tersebut dari nama komedi Inggris terkenal Monty Python. Kemudian dengan sejalannya waktu, Python beralih dari satu laboratorium ke laboratorium lain dan satu perusahaan ke perusahaan lain.

Tulisan ini dibuat, Python sudah masuk ke versi 3 dengan alat terkenal, TensorFlow, yang tangguh untuk mengelola data *big data*. Dengan karakteristiknya yang *open source* dengan lisensi *GPL - compatible*, bahasa ini sangat diminati oleh pengembang perangkat lunak di seluruh dunia. Beberapa literatur mengungkapkan beberapa kelebihan Python, di antaranya: *Readability*, efisien, multifungsi, interoperabilitas yang baik, dan dukungan komunitas yang kuat. Oleh karena itu, diharapkan pembaca menginstal terlebih dahulu *software* Python.

```
y_pred=regressor.predict(X_uji)
print(y_pred)
```

```
[655]: In [55]: #Fitting pada Data Training
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)
#Memprediksi hasil dengan Data Tes
y_pred=regressor.predict(X_uji)
print(y_pred)

[3.162 3.376 3.59 3.804 4.018 4.232]
```

Gambar 4.3 Tampilan hasil fitting pada data training

amat, Anda berhasil memproyeksikan jumlah penduduk dengan regresi linear untuk enam tahun ke depan. Berikutnya adalah proses menyimpan hasilnya ke dalam format CSV yang dapat dibuka dengan Excel atau Text Editor. Buat sel baru dan ikuti instruksi berikut untuk menyimpan hasil proyeksi ke file CSV.

```
#Menyimpan Hasil ke File CSV
tahunuji=np.array(X_uji)
tahun=pd.DataFrame(tahunuji)
prediksi=pd.DataFrame(y_pred)
hasil=pd.concat([tahun, prediksi],axis=1)
np.savetxt('hasil.csv', hasil, delimiter=',')
hasil
```

Anda perlu melakukan operasi "concatenation" dalam rangka menggabungkan antara tahun uji dengan hasil proyeksi sebelumnya.

BAB

5

ADAPTIVE NEURO-FUZZY INFERENCE SYSTEM (ANFIS)

1 Teori Dasar

beda dengan *Fuzzy Inference System* (FIS), atau yang lebih dikenal dengan fuzzy saja, ANFIS membuat rule berdasarkan data yang ditraining lewat mekanisme mirip Jaringan Syaraf Tiruan (JST). Rule yang bisa dilayani hanyalah yang bertipe **Takagi-Sugeno** (TSK), atau dikenal dengan istilah Sugeno saja. Jenis Mamdani tidak dapat diterapkan pada ANFIS. Untuk mengenal lebih dekat tentang fuzzy TSK berikut ini ringkasannya.

Metode FSK diprakarsai oleh Takagi, Sugeno, dan Kang pada tahun 1975. Tujuannya adalah untuk memperoleh rule yang berasal dari pengalaman masukan dan keluaran suatu sistem. Prinsipnya antara

$$\text{If } x \text{ is } A \text{ and } y \text{ is } B \text{ then } z = f(x, y) \quad (1)$$

di mana A dan B merupakan set Fuzzy sedangkan z merupakan nilai dalam bentuk *crisp* (bukan *fuzzy*). Jadi perbedaan yang

BAB

6

JARINGAN SYARAF TIRUAN (JST)

1 Teori Dasar

Jaringan Syaraf Tiruan (JST) atau dalam istilah internasionalnya Artificial Neural Networks, bermaksud meniru cara kerja otak makhluk hidup. Komponen utama dari JST antara lain neuron dan sinaps. Neuron menerima informasi apakah satu informasi diteruskan atau tidak ke neuron lainnya. Sinaps berisi hubungan antara satu neuron dengan neuron lainnya. Diteruskan atau tidak pada suatu neuron didasarkan pada pengaturan bobot (pengali) dan bias (penambah). Di awal-awal perkembangannya JST, bias dan bobot diatur manual, tetapi sejak berkembangnya prinsip pembelajaran, misalnya propagasi balik (backpropagation) maka bobot dan bias diset berdasarkan pelatihan pada data yang ada. Data yang ada tersebut memiliki label atau target yang harus dicapai oleh model JST. Biasanya ada error antara hasil output model JST dengan label/target. Tetapi jika error cukup kecil, model yang diperoleh dari proses pelatihan dapat digunakan untuk memprediksi data lain di luar data pelatihan. Karena adanya prinsip pembelajaran maka JST merupakan salah satu metode dalam ilmu pembelajaran (*machine learning*). Juga karena berusaha

DATA MINING DAN MACHINE LEARNING MENGGUNAKAN MATLAB DAN PYTHON

Buku ini merupakan kelanjutan dari buku sebelumnya tentang soft computing. Banyaknya pembaca yang berminat membuat sistem berbasis mesin pembelajaran (Machine Learning) untuk menggali data (Data Mining) membuat kami melanjutkan membuat buku khusus tentang Data Mining. Tema-tema yang ditulis sebagian besar diambil dari kasus-kasus yang menjadi topik riset mahasiswa-mahasiswa tingkat sarjana dan pascasarjana.

Data mining mengharuskan tersedianya data yang akan digali dan dicari informasi-informasi tersembunyi yang bermanfaat bagi pengambil keputusan. Namun untuk mempermudah proses belajar, dalam buku ini kami hanya menyediakan data-data sederhana yang sengaja dirancang mirip dengan data-data yang jumlahnya banyak yang diperoleh lewat riset di institusi tertentu. Akan tetapi, walaupun sederhana prinsipnya dapat diterapkan langsung dengan data riil.

Buku ini disusun dalam bentuk teori dasar singkat yang dilanjutkan dengan terapannya untuk kasus tertentu dengan bahasa pemrograman Matlab dan Python. Diharapkan pembaca melihat referensi yang kami cantumkan di akhir tulisan jika ingin memperdalam teori dasarnya. Setelah Bab I membahas tinjauan singkat mengenai konsep Data Mining, kami lanjutkan dengan Bab II yang berisi pengenalan terhadap bahasa pemrograman Matlab dilanjutkan dengan bahasa Python di bab III untuk melakukan proses data mining. Kedua bab tersebut menjelaskan teknik-teknik kompilasi, menggunakan Integrated Development Environment (IDE) dan pembuatan Graphic User Interface (GUI) baik dengan bahasa Matlab maupun Python. Bab IV membahas data mining berbasis statistika klasik, antara lain regresi, pohon keputusan dan Naive Bayes. Bab V dan Bab VI membahas metode-metode yang lebih modern, yaitu ANFIS dan Jaringan Syaraf Tiruan. Berikutnya Bab VII berisi metode-metode dalam klusterisasi (clustering). Bab VIII membahas metode yang saat ini banyak diteliti, yaitu Support Vector Machint (SVM), baik untuk kasus bi-class maupun multi-class. Bab IX disertakan pula teknik-teknik validasi dan pengujian model. Bab X merupakan bab tambahan khusus untuk menggali data dari world wide web yang saat ini sedang ramai diteliti, baik berupa situs web maupun sosial media.



Penerbit **INFORMATIKA**

Pemasaran: **BI-OBSES**
Pasar buku Palasari 82 Bandung 40264
Tel.(022) 7317812 Fax.(022) 7317896
www.biobses.com

BIOGRAFI PENULIS



Rahmadya Trias Handayanto,
S.T., M.Kom., Ph.D.

Praktisi yang sempat bekerja sebagai staff IT di bank berskala nasional kini menjadi pengajar sejak Tahun 2002 di beberapa Universitas. Penulis sudah tersertifikasi dosen sejak tahun 2011, dan sebagai Asesor Sertifikasi Dosen. Bidang yang digeluti adalah data mining, sistem cerdas, Information Management, Computer Science, Data Spatial, Optimalisasi dan Sistem informasi Geografis. Saat ini aktif melakukan beberapa penelitian dengan sponsor dari Kementerian Riset dan Teknologi. Penulis menyelesaikan program doctoral dari Asian Institute of Technology (AIT) Thailand. **E-mail:** rahmadya.trias@gmail.com, website:rahmadya.com.



Herlawati, S.Si., M.M., M.Kom

Penulis sejak Tahun 1999 menjadi pengajar di beberapa Universitas. Penulis sudah tersertifikasi dosen sejak tahun 2009, dan sebagai Asesor Sertifikasi Dosen. Mata kuliah yang diampu Kalkulus I, Kalkulus II, Aljabar Linier dan Matriks, Statistika, Struktur Diskrit, Data Mining, dan Analisa Perancangan Sistem Informasi. Bidang yang digeluti adalah Data Mining, Computer Science, Data Spatial, Optimalisasi dan Sistem Informasi Geografis serta Metaheuristics. Penulis aktif melakukan beberapa penelitian dengan sponsor dari Kementerian Riset dan Teknologi. **E-mail:** mrs.herlawati@gmail.com, website:herlawati.com.

Komputer



Harga P. Jawa Rp. 85.000,-