

DR. WAHYUDI, S.T., M.T.






ANALISIS
BIG DATA
MENGUNAKAN 

ANALISIS **BIG DATA** MENGUNAKAN



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

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

ISBN 978-623-151-729-6



9 786231 517296

ANALISIS BIG DATA MENGGUNAKAN R

Dr. Wahyudi, S.T., M.T.



eureka
media aksara

PENERBIT CV.EUREKA MEDIA AKSARA

ANALISIS BIG DATA MENGGUNAKAN R

Penulis : Dr. Wahyudi, S.T., M.T.

Desain Sampul : Ardyan Arya Hayuwaskita

Tata Letak : Rizki Rose Mardiana

ISBN : 978-623-151-729-6

Diterbitkan oleh : **EUREKA MEDIA AKSARA, OKTOBER 2023**
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 : eurekaediaaksara@gmail.com

Cetakan Pertama : 2023

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

Alhamdulillahirobbil'alamin. Penulis bersyukur kehadiran Allah SWT berkat rahmat, karunia dan pertolonganNya, penulis dapat menyelesaikan buku berjudul "**Analisis Big Data Menggunakan R**". Shalawat serta salam semoga senantiasa tercurah atas Nabi Muhammad SAW, para kerabat, serta pengikutnya hingga hari kiamat nanti.

Buku ini hadir untuk menambah literasi tentang teknologi informasi. Buku ini merupakan seri pertama dari beberapa buku analisis big data. Buku ini menjelaskan tentang munculnya bidang sains data, hubungan sains data dan big data, proses yang dilakukan di sains data dan pembelajaran mesin yang merupakan bagian paling penting di sains data.

Penulis menyadari bahwa dalam penulisan buku ini masih banyak terdapat kekurangan, untuk itu penulis mengharapkan kritik dan sarannya guna penyempurnaan buku ini di masa mendatang

Padang, September 2023

Penulis

DAFTAR ISI

KATA PENGANTAR.....	iii
DAFTAR ISI.....	iv
DAFTAR TABEL	v
DAFTAR GAMBAR.....	vi
BAB 1 REVIEW ULANG METODE ANALISIS DATA	
DASAR MENGGUNAKAN R	1
A. Pengantar ke R.....	2
B. Analisis Data Eksplorasi	29
C. Metode Statistik untuk Evaluasi.....	60
DAFTAR PUSTAKA.....	82
BAB 2 TEORI ANALISIS LANJUTAN DAN METODE:	
PENGELOMPOKAN.....	84
A. Gambaran Umum tentang Clustering	84
B. K-means.....	85
C. Algoritma Tambahan	110
DAFTAR PUSTAKA.....	112

DAFTAR TABEL

Tabel 1. 1	Default Fungsi Impor	10
Tabel 1. 2	Tipe Atribut Noir	13
Tabel 1. 3	Sifat Statistik Kuartet Anscombe	32
Tabel 1. 4	Contoh Fungsi untuk Memvisualisasikan Variabel Tunggal.....	41
Tabel 1. 5	Contoh Hipotesis Nol dan Hipotesis Alternatif	63
Tabel 1. 6	Kesalahan Tipe I dan Tipe II	74

DAFTAR GAMBAR

Gambar 1. 1	Memeriksa Data Secara Grafis	5
Gambar 1. 2	Bukti dari Residual yang Besar	7
Gambar 1. 3	GUI RStudio.....	8
Gambar 1. 4	Mengakses Bantuan di Rstudio.....	9
Gambar 1. 5	Scatterplot dapat dengan Mudah Menunjukkan Apakah x dan y Memiliki Hubungan.....	30
Gambar 1. 6	Kuartet Anscombe.....	31
Gambar 1. 7	Kuartet Anscombe Divisualisasikan Sebagai Diagram Pencar	33
Gambar 1. 8	Distribusi Usia Pemegang Rekening Bank.....	36
Gambar 1. 9	Distribusi KPR dalam Beberapa Tahun Sejak Awal dari Portofolio Kredit Perumahan Sebuah Bank	39
Gambar 1. 10	(a) Diagram Titik pada Jarak Tempuh Per Galon Mobil dan (b) Diagram Batang pada Distribusi Silinder Mobil	42
Gambar 1. 11	(a) Histogram dan (b) Plot Densitas Pendapatan Rumah Tangga	43
Gambar 1. 12	Plot Densitas dari (a) Harga Berlian dan (b) Logaritma Harga Berlian	46
Gambar 1. 13	Memeriksa Dua Variabel dengan Regresi.....	47
Gambar 1. 14	Dotplot untuk Memvisualisasikan Beberapa Variable	49
Gambar 1. 15	Barplot untuk Memvisualisasikan Beberapa Variable	50
Gambar 1. 16	Plot Kotak dan Kumis Rata-Rata Pendapatan Rumah Tangga dan Wilayah Geografis	51
Gambar 1. 17	(a) Scatterplot dan (b) Hexbinplot Pendapatan Rumah Tangga Terhadap Lama Pendidikan	53
Gambar 1. 18	Matriks Scatterplot dari Dataset Iris Fisher	55
Gambar 1. 19	Jumlah Penumpang Maskapai Penerbangan dari Tahun 1949 Hingga 1960.....	57

Gambar 1. 20	Plot Kepadatan Lebih Baik untuk Ditunjukkan Kepada Ilmuwan Data.....	58
Gambar 1. 21	Histogram Lebih Baik Ditunjukkan Kepada Pemangku Kepentingan.....	60
Gambar 1. 22	Distribusi Dua Sampel Data	61
Gambar 1. 23	Tumpang Tindih dari Dua Distribusi adalah Besar Jika $X1 \approx X2$	65
Gambar 1. 24	Area di Bawah Ekor (Diarsir) dari (Student's T-Distribution) Distribusi-T Siswa	68
Gambar 1. 25	Interval Kepercayaan 95% yang Melangkahi Rata-Rata Populasi yang Tidak Diketahui μ	71
Gambar 1. 26	Ukuran Sampel yang Lebih Besar Lebih Baik dalam Mengidentifikasi Ukuran Efek yang Tetap.....	75
Gambar 2. 1	Kemungkinan Cluster K-Means Untuk $K=3$	86
Gambar 2. 2	Titik Awal Awal untuk Centroids.....	88
Gambar 2. 3	Titik-Titik Ditetapkan ke Centroid Terdekat.....	89
Gambar 2. 4	Menghitung Rata-Rata dari Setiap Klaster	91
Gambar 2. 5	WSS dari Data Nilai Siswa	95
Gambar 2. 6	Plot Kelompok Siswa yang Teridentifikasi	99
Gambar 2. 7	Contoh dari Kelompok-Kelompok yang Berbeda.....	101
Gambar 2. 8	Contoh Kelompok yang Kurang Jelas.....	102
Gambar 2. 9	Enam Kelompok yang Diterapkan pada Titik-Titik pada Gambar 2.8.....	102
Gambar 2. 10	Matriks Scatterplot untuk Tujuh Atribut.....	105
Gambar 2. 11	Gugus Dengan Ketinggian yang Dinyatakan dalam Sentimeter	106
Gambar 2. 12	Klaster dengan Tinggi Badan yang Dinyatakan dalam Meter	106

Gambar 2. 13 Klaster dengan Atribut yang Diskalakan.....108

BAB

1

REVIEW ULANG METODE ANALISIS DATA DASAR MENGUNAKAN R

Buku sebelumnya telah menyajikan enam fase Siklus Hidup Analisis Data.

- Fase 1: Penemuan
- Fase 2: Persiapan Data
- Fase 3: Perencanaan Model
- Fase 4: Membangun Model
- Fase 5: Mengkomunikasikan Hasil
- Fase 6: Evaluasi: Mengoperasionalkan

Tiga fase pertama melibatkan berbagai aspek eksplorasi data. Secara umum, keberhasilan suatu data proyek analisis data membutuhkan pemahaman yang mendalam tentang data. Hal ini membutuhkan perangkat untuk menambang dan menyajikan data. Kegiatan ini mencakup studi data dalam ukuran statistik dasar dan pembuatan grafik dan plot untuk memvisualisasikan dan mengidentifikasi hubungan dan pola. Beberapa alat gratis atau komersial gratis tersedia untuk mengeksplorasi, mengkondisikan, membuat model, dan menyajikan data. Karena popularitas dan keserbagunaannya, bahasa pemrograman sumber terbuka R digunakan untuk mengilustrasikan banyak tugas dan model yang disajikan dalam buku ini.

Bab ini memperkenalkan fungsionalitas dasar bahasa pemrograman R dan lingkungannya. Bagian pertama memberikan gambaran umum tentang bagaimana menggunakan R untuk memperoleh, mengurai, dan menyaring data serta bagaimana memperoleh beberapa statistik deskriptif dasar pada sebuah

DAFTAR PUSTAKA

- B. L. Welch, "The Generalization of "Student's" Problem When Several Different Population Variances Are Involved," *Biometrika*, vol. 34, no. 1-2, pp. 28-35, 1947.
- B. Ripley, "RODBC: ODBC Database Access," CRAN. [Online]. Available: <http://cran.r-project.org/web/packages/RODBC/index.html>. [Accessed 13 December 2013].
- D. C. Hoaglin, F. Mosteller, and J. W. Tukey, *Understanding Robust and Exploratory Data Analysis*, New York: Wiley, 1983.
- F. J. Anscombe, "Graphs in Statistical Analysis," *The American Statistician*, vol. 27, no. 1, pp. 17-21, 1973.
- F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80-83, 1945.
- G. Williams, M. V. Culp, E. Cox, A. Nolan, D. White, D. Medri, and A. Waljee, "Rattle: Graphical User Interface for Data Mining in R," CRAN. [Online]. Available: <http://cran.r-project.org/web/packages/rattle/index.html>. [Accessed 12 December 2013].
- H. Wickham, "ggplot2," 2013. [Online]. Available: <http://ggplot2.org/>. [Accessed 8 January 2014].
- J. Fox and M. Bouchet-Valat, "The R Commander: A Basic-Statistics GUI for R," CRAN. [Online]. Available: <http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/>. [Accessed 11 December 2013].
- J. J. Faraway, "Practical Regression and Anova Using R," July 2002. [Online]. Available: <http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf>. [Accessed 22 January 2014].
- R Special Interest Group on Databases (R-SIG-DB), "DBI: R Database Interface." CRAN [Online]. Available: <http://cran.r->

project.org/web/packages/DBI/index.html.[Accessed 13 December 2013].

R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

RStudio, "RStudio IDE" [Online]. Available: <http://www.rstudio.com/ide/>. [Accessed 11 December 2013].

S. S. Stevens, "On the Theory of Scales of Measurement," *Science*, vol. 103, no. 2684, p. 677–680, 1946.

The R Project for Statistical Computing, "R Licenses." [Online]. Available: <http://www.rproject.org/Licenses/>. [Accessed 10 December 2013].

The R Project for Statistical Computing, "The Comprehensive R Archive Network." [Online]. Available: <http://cran.r-project.org/>. [Accessed 10 December 2013].

W. S. Cleveland, *Visualizing Data*, Lafayette, IN: Hobart Press, 1993.

BAB 2

TEORI ANALISIS LANJUTAN DAN METODE: PENGELOMPOKAN

Berdasarkan pengenalan terhadap R yang disajikan dalam buku sebelumnya, "Tinjauan Metode Analisis Data Dasar Menggunakan R, " dan buku sebelumnya, "Teori dan Metode Analitik Tingkat Lanjut: Pengelompokan" hingga buku berikutnya, "Teori dan Metode Analitik Tingkat Lanjut Teori dan Metode Analisis Lanjutan: Analisis Teks" menjelaskan beberapa metode analisis yang umum digunakan untuk dapat dipertimbangkan pada fase Perencanaan Model dan Eksekusi (Fase 3 dan 4) dari Siklus Hidup Analisis Data. Bab ini membahas teknik dan algoritme pengelompokan.

A. Gambaran Umum tentang Clustering

Secara umum, pengelompokan adalah penggunaan teknik (*unsupervised*) tanpa pengawasan untuk mengelompokkan objek-objek yang serupa. Dalam pembelajaran mesin, *unsupervised* mengacu pada masalah menemukan struktur tersembunyi di dalam data yang tidak berlabel. Pengelompokan tidak diawasi, artinya bahwa ilmuwan data tidak menentukan, terlebih dahulu label yang akan diterapkan pada cluster. Struktur data menggambarkan objek yang diminati dan menentukan cara terbaik untuk mengelompokkan objek-objek tersebut. Misalnya, berdasarkan pendapatan pribadi pelanggan, sangat mudah untuk dibagi pelanggan menjadi tiga kelompok tergantung pada nilai yang dipilih secara acak. Pelanggan dapat dibagi menjadi tiga kelompok sebagai berikut:

DAFTAR PUSTAKA

- J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, 1967.
- L. Kaufman and P. J. Rousseeuw, "Partitioning Around Medoids (Program PAM)," in Finding Groups in Data: An Introduction to Cluster Analysis, Hoboken, NJ, John Wiley & Sons, Inc, 2008, p. 68-125, Chapter 2.
- P.-N. Tan, V. Kumar, and M. Steinbach, Introduction to Data Mining, Upper Saddle River, NJ: Person, 2013.
- Z. Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining," 1997. [Online]. Available:
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.83&rep=rep1&type=pdf>. [Accessed 13 March 2014].