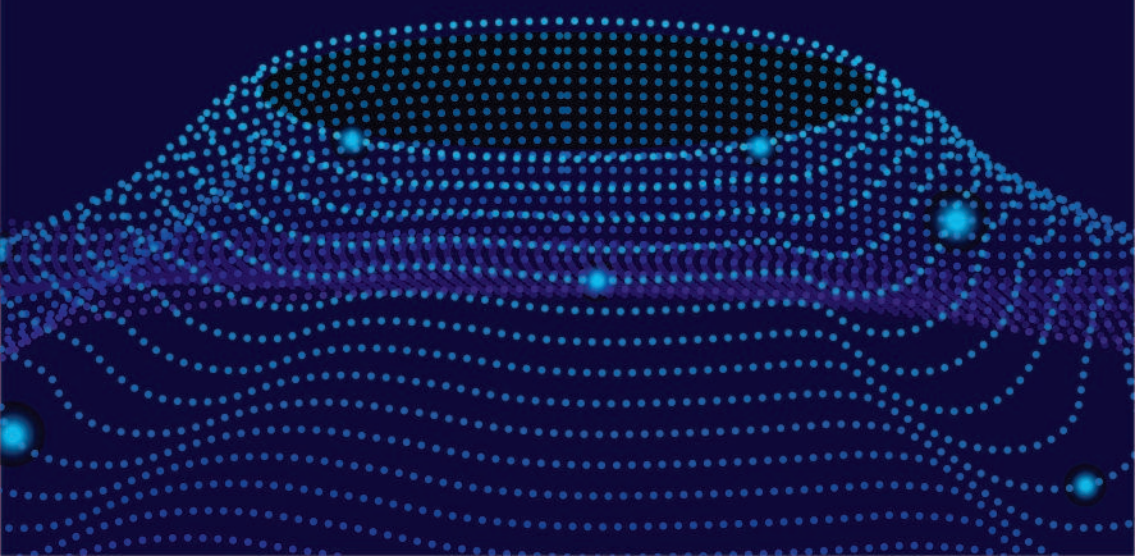
DR. WAHYUDI, S.T., M.T.

ANALISIS BIG DATA MENGGUNAKAN R



ANALISIS BIG DATA MENGGUNAKAN R

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 978-623-151-729-6

ANALISIS BIG DATA MENGGUNAKAN R

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

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 : eurekamediaaksara@gmail.com

Cetakan Pertama : 2023

KATA PENGANTAR

Alhamdulillahirobbil'alamin. Penulis bersyukur kehadirat 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

DAFTAR TABEL

DAFTAR GAMBAR

BAB 1

REVIEW ULANG METODE ANALISIS DATA DASAR MENGGUNAKAN 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-roject.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-roject.org/web/packages/rattle/index.html.[Accessed 12 December 2013].

H. Wickham, "ggplot2," 2013. [Online]. Available: http://ggplot2.org/. [Accessed 8January 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 January2014].

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 sebelumya, "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 Groupsin 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: ttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.83&rep=rep1&type=pdf. [Accessed 13 March 2014].