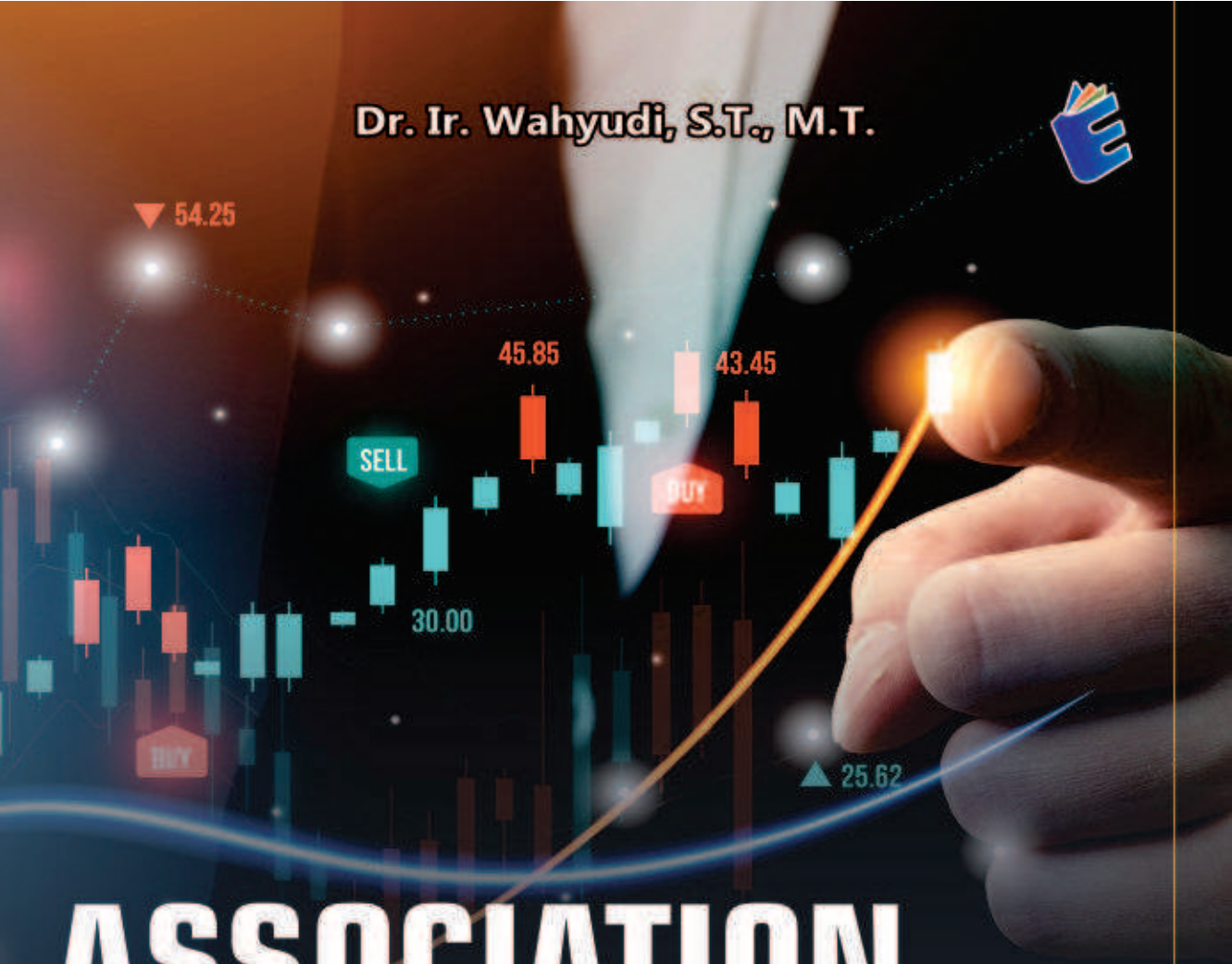
Dr. Ir. Wahyudi, S.T., M.T.



ASSOCIATION RULES dan SUPERVISED LEARNING MENGGUNAKAN R

ASSOCIATION RULES
dan SUPERVISED LEARNING
MENGGUNAKAN R

Buku Association rules dan Supervised Learning Menggunakan R terdiri dari tiga bab. Bab pertama yaitu Association Rules. Association Rules atau Aturan asosiasi adalah metode pembelajaran mesin berbasis aturan untuk menemukan hubungan yang menarik antara variabel dalam database yang besar. Metode ini dimaksudkan untuk mengidentifikasi aturan kuat yang ditemukan dalam basis data dengan menggunakan beberapa ukuran ketertarikan. Dalam transaksi tertentu dengan berbagai item, aturan asosiasi dimaksudkan untuk menemukan aturan yang menentukan bagaimana atau mengapa item tertentu terhubung. Salah satu algoritma Association Rules yang banyak digunakan adalah algoritma apriori yang dibahas dalam buku ini.

Bab ke dua dan ke tiga membahas tentang dua metode supervised learning yaitu regresi dan klasifikasi. Bab dua menceritan tentang regresi ada dua metode yang dibahas yaitu regresi linear dan regresi logistik. Bab tiga menjelaskan tentang metode klasifikasi yaitu decision tree dan Naïve bayes.

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 978-623-120-447-9
9 786231 204479

ASSOCIATION RULES DAN SUPERVISED LEARNING MENGGUNAKAN R

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

PENERBIT CV.EUREKA MEDIA AKSARA

ASSOCIATION RULES DAN SUPERVISED LEARNING MENGGUNAKAN R

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

Desain Sampul : Eri setiawan

Tata Letak : Amini Nur Ihwati

ISBN : 978-623-120-447-9

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

KATA PENGANTAR

Alhamdulillahirobbil'alamin. Penulis bersyukur kehadirat Allah SWT berkat rahmat, karunia dan pertolonganNya, penulis dapat menyelesaikan buku berjudul "**Association Rules dan Supervised Learning 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 ketiga dari beberapa buku sains data menggunakan R. Buku ini menjelaskan tentang Association rules atau aturan asosiasi dan beberapa model klasifikasi.

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, Februari 2024

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

ASSOCIATION RULES DAN SUPERVISED LEARNING MENGGUNAKAN R

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

BAB 1

TEORI DAN METODE ANALISIS TINGKAT LANJUT: ASSOCIATION RULES

Bab ini membahas metode pembelajaran tanpa pengawasan yang disebut *association rules*/aturan asosiasi. Ini adalah metode deskriptif, bukan prediktif, metode yang sering digunakan untuk menemukan hubungan yang menarik dan tersembunyi dalam kumpulan data besar. Hubungan yang diungkapkan dapat direpresentasikan sebagai *rules* atau *frequent itemsets*. *Association rules* biasanya digunakan untuk menambang transaksi dalam *database*. Berikut adalah beberapa pertanyaan yang dapat dijawab oleh association rules:
Produk mana yang cenderung dibeli bersamaan?
Dari para pelanggan yang mirip dengan orang ini, produk apa yang cenderung mereka beli?
Dari para pelanggan yang telah membeli produk ini, produk serupa apa yang cenderung mereka lihat atau beli?

A. Gambaran Umum

Gambar 1 menunjukkan logika umum dibalik association rules. Diberikan sebuah koleksi transaksi yang besar (digambarkan sebagai tiga tumpukan struk pada gambar), di mana setiap transaksi terdiri dari satu atau lebih item, association rules apa saja yang digunakan untuk melihat item yang sering dibeli bersamaan dan menemukan daftar rules yang menggambarkan perilaku pembelian. Tujuan dari association rules untuk menemukan hubungan yang menarik di antara item-item tersebut. (Hubungan yang terjadi sering menjadi acak dan bermakna dari perspektif bisnis, yang mungkin jelas atau tidak jelas). Hubungan yang menarik tergantung pada konteks bisnis dan sifat algoritma yang digunakan untuk discovery.

BAB 2

TEORI DAN METODE ANALISIS TINGKAT LANJUT: REGRESI

Secara umum, analisis regresi mencoba menjelaskan pengaruh dari satu set variabel terhadap hasil variabel lain yang diminati. Sering kali, variabel hasil disebut dependent variable karena hasilnya bergantung pada variabel lainnya. Variabel-variabel tambahan ini kadang-kadang disebut input variables atau independent variables. Analisis regresi berguna untuk menjawab pertanyaan-pertanyaan berikut:

1. Berapa pendapatan yang diharapkan dari seseorang? dan
2. Berapa besar kemungkinan pemohon akan gagal membayar pinjaman?

Regresi linier adalah alat yang berguna untuk menjawab pertanyaan pertama, dan regresi logistik adalah metode populer untuk menjawab pertanyaan kedua. Bab ini membahas kedua teknik regresi tersebut dan menjelaskan kapan satu teknik tepat digunakan dibandingkan teknik lainnya.

Analisis regresi adalah alat penjelas yang berguna untuk mengidentifikasi **variable input** yang memiliki pengaruh statistik terbesar terhadap hasil. Dengan pengetahuan dan wawasan seperti itu, perubahan lingkungan dapat diupayakan untuk menghasilkan nilai yang lebih baik dari variable input. Contoh, jika ditemukan tingkat membaca siswa yang sangat baik pada usia 10 tahun, itu menentukan keberhasilan siswa di sekolah menengah dan faktor keberhasilan mereka masuk perguruan tinggi, dan dievaluasi untuk meningkatkan kemampuan membaca siswa, diimplementasikan, dan dievaluasi untuk meningkatkan tingkat membaca siswa pada age yang lebih muda.

A. Regresi Linier

Regresi linier adalah teknik analisis yang digunakan untuk memodelkan hubungan antara beberapa variabel input dan variabel hasil yang berkelanjutan. Asumsi utamanya adalah hubungan antara variabel input dan variabel hasil adalah linier. Meskipun asumsi ini terlihat membatasi, namun sering kali dimungkinkan untuk mengubah variabel input atau variabel hasil dengan benar untuk mencapai hubungan linier antara variabel input dan variabel hasil yang

BAB 3 TEORI ANALISIS LANJUTAN DAN METODE: KLASIFIKASI

Bab ini terutama berfokus pada dua metode klasifikasi dasar: *decision trees* (pohon keputusan) dan *naïve Bayes*.

A. Decision Trees

Decision trees (disebut juga *prediction tree*) menggunakan struktur *tree* untuk menentukan urutan *decision* dan konsekuensi-konsekuensinya. Diberikan input $X = \{x_{1,}x_{2,\ldots}\, x_n\}$ tujuannya untuk memprediksi sebuah respon atau variabel *output* Y. Setiap anggota himpunan $\{x_{1,}x_{2,\ldots}\, x_n\}$, disebut variable input. Prediksi dapat dicapai dengan membangun *decision trees* dengan titik uji dan cabang. Pada akhirnya, titik akhir tercapai, dan prediksi dapat dibuat. Setiap pengujian *decision trees* melibatkan pengujian variabel input (atau atribut) tertentu, dan setiap cabang mewakili *decision* yang dibuat. Karena fleksibilitas dan visualisasinya yang mudah, *decision trees* biasanya digunakan dalam aplikasi data mining untuk tujuan klasifikasi.

Nilai input dari *decision trees* dapat berupa kategorikal atau kontinu. *Decision trees* menggunakan struktur titik uji (disebut *nodes*) dan cabang, yang mewakili *decision* yang dibuat. Sebuah *node* tanpa cabang lebih lanjut disebut *leaf node* (simpul daun). *Leaf nodes* mengembalikan label kelas dan, dalam beberapa implementasi, mereka mengembalikan nilai probabilitas. Sebuah *decision tree* dapat dikonversi menjadi satu set *decision rules* (aturan keputusan). Dalam contoh aturan berikut ini, pendapatan dan *mortgage_amount* (jumlah_kredit) adalah variabel input, dan responnya adalah output variable *default* (standar) dengan nilai probabilitas.

IF income < $50,000 **AND** mortgage_amount > $100K
THEN default = True **WITH PROBABILITY** 75%

Decision trees memiliki dua jenis: klasifikasi trees dan regresi trees. Klasifikasi trees biasanya berlaku untuk variabel output yang bersifat kategorikal-sering kali biner, seperti yes (ya) atau no (tidak), beli atau tidak beli,

DAFTAR PUSTAKA

B. Mobasher, H. Dai, T. Luo, and M. Nakagawa, "Effective Personalization Based on Association Rule Discovery from Web Usage Data," in ACM, 2011.

C. C. Aggarwal and P. S. Yu, "A New Framework for Itemset Generation," in Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'98), Seattle, Washington, USA, 1998.

C. Phua, V. C. S. Lee, S. Kate, and R. W. Gayler, "A Comprehensive Survey of Data Mining-Based Fraud Detection," CoRR, vol. abs/1009.6119, 2010.

D. Michie, D. J. Spiegelhalter, and C. C. Taylor, Machine Learning, Neural and Statistical Classification, New York: Ellis Horwood, 1994.

G. Piatetsky-Shapiro, "Discovery, Analysis and Presentation of Strong Rules," Knowledge Discovery inDatabases, pp. 229–248, 1991.

I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropoulos, "An Evaluation of Naïve Bayesian Anti-Spam Filtering," in Proceedings of the Workshop on Machine Learning in the New Information Age, Barcelona, Spain, 2000.

I. H. Witten, E. Frank, and M. A. Hall, "The Bootstrap," in Data Mining, Burlington, Massachusetts, Morgan Kaufmann, 2011, pp. 155–156.

J. R. Quinlan, "Bagging, Boosting, and C4. 5," AAAI/IAAI, vol. 1, 1996.

J. R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.

J. R. Quinlan, C4. 5: Programs for Machine Learning, Morgan Kaufmann, 1993.

L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.

L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees, Belmont, CA: Wadsworth International Group, 1984.

M. Hahsler, "A Comparison of Commonly Used Interest Measures for Association Rules," 9 March 2011. [Online]. Available: http://michael.hahsler.net/research/Association rules/measures.html. [Accessed 4 March 2014].

M. Shouman, T. Turner, and R. Stocker, "Using Decision Tree for Diagnosing Heart Disease Patients," in Australian Computer Society, Inc., Ballarat, Australia, in Proceedings of the Ninth Australasian Data Mining Conference (AusDM '11).

M. Thomas, B. Pang, and L. Lee, "Get Out the Vote: Determining Support or Opposition from Congressional Floor-Debate Transcripts," in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 2006.

M.-S. Chen, J. S. Park, and P. Yu, "Efficient Data Mining for Path Traversal Patterns," IEEE Transactionson Knowledge and Data Engineering, vol. 10, no. 2, pp. 209–221, 1998.

N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge, United Kingdom: Cambridge university press, 2000.

P. Hájek, I. Havel, and M. Chytil, "The GUHA Method of Automatic Hypotheses Determination," Computing, vol. 1, no. 4, pp. 293–308, 1966.

R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," in Proceedings of the 20th International Conference on Very Large Data Bases, San Francisco, CA, USA, 1994.

R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules Between Sets of Items in LargeDatabases," SIGMOD '93 Proceedings of the 1993 ACM SIGMOD International Conference onManagement of Data, pp. 207–216, 1993.

R. Bhowmik, "Detecting Auto Insurance Fraud by Data Mining Techniques," Journal of Emerging Trends in Computing and Information Sciences, vol. 2, no. 4, pp. 156–162, 2011.

R. Cooley, B. Mobasher, and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web," Proceedings of the 9th IEEE International Conference on Tools with ArtificialIntelligence, pp. 558–567, 1997.

S. Brin, R. Motwani, and C. Silverstein, "Beyond Market Baskets: Generalizing Association Rules to Correlations," Proceedings of the ACM SIGMOD/PODS '97 Joint Conference, vol. 26, no. 2, pp. 265–276, 1997.

S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data," SIGMOD, vol. 26, no. 2, pp. 255–264, 1997.

S. Moro, P. Cortez, and R. Laureano, "Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology," in Proceedings of the European Simulation and Modelling Conference - ESM'2011, Guimaraes, Portugal, 2011.

T. M. Mitchell, "Decision Tree Learning," in Machine Learning, New York, NY, USA, McGraw-Hill, Inc., 1997, p. 68.

W. Lin, S. A. Alvarez, and C. Ruiz, "Collaborative Recommendation via Adaptive Association Rule Mining," in Proceedings of the International Workshop on Web Mining for E-Commerce (WEBKDD), Boston, MA, 2000.

W. Lin, S. A. Alvarez, and C. Ruiz, "Efficient Adaptive-Support Association Rule Mining for Recommender Systems," Data Mining and Knowledge Discovery, vol. 6, no. 1, pp. 83–105, 2002.

Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997.