

DATA MINING

DALAM ANALISIS DATA DAN
PENGAMBILAN KEPUTUSAN

Akbar Idaman, S.Kom., M.Kom., Amrullah, S.Kom.,
M.Kom., Dr. Firahmi Rizky, S.Kom., M.Kom., Hamjah
Arahma, S.Kom., M.Kom., M. Fakhru Hirzi, S.Kom.,
M.Kom., Muhammad Syahril, S.E., M.Kom., dan Mui
Sunjaya, S.Kom., M.Kom.

***Data Mining* dalam Analisis Data dan Pengambilan Keputusan**

Akbar Idaman, S.Kom., M.Kom.

Amrullah, S.Kom., M.Kom.

Hamjah Arahma, S.Kom., M.Kom.

M. Fakhrol Hirzi, S.Kom., M.Kom.

Dr. Firahmi Rizky, S.Kom., M.Kom.

Muhammad Syahril, S.E., M.Kom.

Muit Sunjaya, S.Kom., M.Kom.



PT YAPINDO JAYA ABADI

Anggota IKAPI: No. 627/DKI/2023

Data Mining dalam Analisis Data dan Pengambilan Keputusan

Penulis : Akbar Idaman, S.Kom., M.Kom., Amrullah, S.Kom., M.Kom., Hamjah Arahma, S.Kom., M.Kom., M. Fakhrol Hirzi, S.Kom., M.Kom., Dr. Firahmi Rizky, S.Kom., M.Kom., Muhammad Syahril, S.E., M.Kom., dan Muit Sunjaya, S.Kom., M.Kom.

ISBN : 978-634-7789-63-1 (PDF)

Penyunting Naskah : Witri Nur Annisa, S.Pd.

Tata Letak : Witri Nur Annisa, S.Pd.

Desain Sampul : Novikean Keysah Sanisri

Penerbit

PT Yapindo Jaya Abadi

Jl. Tanjung Duren Raya No. 89 C RT 06/RW 05, Kelurahan Tanjung Duren Selatan, Kecamatan Grogol Petamburan, Kota Jakarta Barat, Daerah Khusus Ibukota Jakarta 11470

E-Mail : yapjadi@gmail.com

Website : yapindo.co.id

© Hak cipta dilindungi oleh undang-undang

Berlaku selama 50 (lima puluh) tahun sejak Ciptaan tersebut pertama kali dilakukan pengumuman.

Dilarang mengutip atau memperbanyak sebagian atau seluruh isi buku ini tanpa izin tertulis dari penerbit. Ketentuan Pidana Sanksi Pelanggaran Pasal 2 UU Nomor 19 Tahun 2002 Tentang Hak Cipta.

Barang siapa dengan sengaja dan tanpa hak melakukan perbuatan sebagaimana dimaksud dalam Pasal 2 ayat (1) atau Pasal 49 ayat (1) dan ayat (2) dipidana dengan pidana penjara masing-masing paling singkat 1 (satu) bulan dan/atau denda paling sedikit Rp1.000.000,00 (satu juta rupiah), atau pidana penjara paling lama 7 (Tujuh) tahun dan/atau denda paling banyak Rp5.000.000.000,00 (lima miliar rupiah).

Barang siapa dengan sengaja menyerahkan, menyiarkan, memamerkan, mengedarkan atau menjual kepada umum suatu Ciptaan atau barang hasil pelanggaran Hak Cipta atau Hak Terkait sebagaimana dimaksud pada ayat (1) dipidana dengan pidana penjara paling lama 5 (lima) tahun dan/atau denda paling banyak Rp500.000.000,00 (lima ratus juta rupiah).

KATA PENGANTAR

Puji syukur ke hadirat Tuhan Yang Maha Esa atas rahmat dan karunia-Nya, sehingga buku referensi berjudul *Data Mining dalam Analisis Data dan Pengambilan Keputusan* dapat disusun sebagai bacaan yang memperkenalkan peran penting data dalam kehidupan modern. Perkembangan teknologi membuat jumlah data terus bertambah, sehingga kemampuan memahami dan memanfaatkannya menjadi kebutuhan yang semakin penting.

Buku ini membahas *Data Mining* sebagai cara untuk menggali informasi dari kumpulan data agar dapat digunakan dalam menentukan arah keputusan. Materi disajikan secara sederhana, mulai dari pengenalan konsep, proses pengolahan data, hingga penerapannya dalam berbagai bidang yang dekat dengan kehidupan masyarakat.

Buku ini tersedia untuk masyarakat umum, baik pembaca yang baru mengenal dunia data maupun pihak yang ingin memperluas wawasan mengenai analisis data. Diharapkan buku ini dapat memberikan manfaat dan membantu pembaca memahami pentingnya data sebagai dasar dalam mengambil keputusan yang lebih cermat dan terarah.

Jakarta, Juli 2026

Tim Penyusun

DAFTAR ISI

KATA PENGANTAR	iii
DAFTAR ISI	iv
Bab 1: Sistem Data dan Data Mining	1
1.1 Konsep Data dan Informasi	1
1.2 Perkembangan Data Mining	4
1.3 Peran Data Mining dalam Ilmu Data	8
1.4 Hubungan dengan Big Data.....	12
1.5 Aplikasi Data Mining	16
Bab 2: Konsep Dasar Data Mining.....	20
2.1 Definisi Data Mining	20
2.2 Proses Knowledge Discovery	23
2.3 Tahapan Data Mining	27
2.4 Jenis Data.....	30
2.5 Tantangan Data Mining	34
Bab 3: Persiapan dan Preprocessing Data	38
3.1 Pembersihan Data	38
3.2 Integrasi Data.....	41
3.3 Transformasi data	45
3.4 Reduksi Data	49
3.5 Seleksi Fitur.....	53
Bab 4: Statistik dan Analisis Data	57
4.1 Statistik Deskriptif.....	57
4.2 Distribusi Data.....	60
4.3 Korelasi dan Regresi.....	64
4.4 Analisis Variansi	68

4.5 Interpretasi Data	72
Bab 5: Teknik Klasifikasi	76
5.1 Konsep Klasifikasi.....	76
5.2 Decision Tree.....	79
5.3 Naive Bayes.....	83
5.4 K-Nearest Neighbor.....	87
5.5 Penilaian Klasifikasi	90
Bab 6: Teknik Clustering.....	95
6.1 Konsep Clustering	95
6.2 K-Means	98
6.3 Hierarchical Clustering.....	102
6.4 Density-Based Clustering	105
Bab 7: Association Rule Mining.....	110
7.1 Konsep Asosiasi	110
7.2 Apriori Algorithm.....	113
7.3 Support dan Confidence	117
7.4 Lift Ratio	121
7.5 Implementasi Asosiasi.....	124
Bab 8: Data Mining Berbasis Prediksi	129
8.1 Konsep Prediksi.....	129
8.2 Regresi Linear	132
8.3 Regresi non-Linear	136
8.4 Model Prediktif.....	140
8.5 Penilaian Model.....	143
Bab 9: Penilaian dan Validasi Model.....	148
9.1 Konsep Penilaian	148
9.2 Confusion Matrix.....	151
9.3 Cross validation	154

9.4 Overfitting dan Underfitting.....	158
9.5 Interpretasi Hasil	162
Bab 10: Data Mining dalam Big Data.....	166
10.1 Konsep Big Data.....	166
10.2 Karakteristik Big Data	169
10.3 Tools Data Mining.....	173
10.4 Pemrosesan Data Skala Besar.....	176
10.5 Tantangan Big Data	180
Daftar Pustaka	184
Profil Penulis	192

Daftar Pustaka

- Aggarwal, C. C. (2020). *Data Mining: The textbook*. Springer.
- Han, J., Kamber, M., & Pei, J. (2020). *Data Mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
- Agrawal, R., & Srikant, R. (2020). Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Very Large Data Bases*.
- Agrawal, R., & Srikant, R. (2020). Fast algorithms for mining association rules. *Proceedings of the VLDB Conference*, 487–499.
- Tan, P. N., Steinbach, M., & Kumar, V. (2020). *Introduction to Data Mining* (2nd ed.). Pearson.
- Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2), 207–216.
- Han, J., Kamber, M., & Pei, J. (2022). *Data Mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.
- Chandrashekar, G., & Sahin, F. (2020). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16–28.
- Chen, M., Mao, S., & Liu, Y. (2020). *Big data: A survey*. *Mobile Networks and Applications*, 25(1), 1–15.

- Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, W. K. M., Alam, M., Shiraz, M., & Gani, A. (2021). *Big data: Survey, technologies, opportunities, and challenges. The Scientific World Journal*, 2021, 1–18.
- Chen, M., Mao, S., & Liu, Y. (2020). *Big data: A survey. Mobile Networks and Applications*, 19(2), 171–209.
- Cover, T., & Hart, P. (2020). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21–27.
- Peterson, L. E. (2021). *K-Nearest Neighbor. Scholarpedia*, 4(2), 1883.
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (2020). A *Density-Based* algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the ACM SIGKDD Conference*, 226–231.
- Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2021). DBSCAN revisited, revisited: Why and how you should still use DBSCAN. *ACM Transactions on Database Systems*, 42(3), 1–21.
- Field, A. (2020). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage Publications.
- Montgomery, D. C. (2020). *Design and analysis of experiments* (10th ed.). Wiley.
- Gandomi, A., & Haider, M. (2020). Beyond the hype: *Big data* concepts, methods, and analytics. *International*

Journal of Information Management, 35(2), 137–144.

Zikopoulos, P., Eaton, C., deRoos, D., Deutsch, T., & Lapis, G. (2020). *Understanding big data: Analytics for enterprise class Hadoop and streaming data*. McGraw-Hill.

Gandomi, A., & Haider, M. (2020). Beyond the hype: *Big data* concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2020). *Big data Preprocessing: Methods and prospects*. *Big data Analytics*, 3(1), 1–22.

Rahm, E., & Do, H. H. (2020). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4), 3–13.

Goodfellow, I., Bengio, Y., & Courville, A. (2020). *Deep learning*. MIT Press.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer.

Guyon, I., & Elisseeff, A. (2020). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.

Han, J., Kamber, M., & Pei, J. (2020). *Data Mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for *Data Mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.*
- Han, J., Kamber, M., & Pei, J. (2020). *Data Mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
- Han, J., Kamber, M., & Pei, J. (2022). *Data Mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.
- Han, J., Kamber, M., & Pei, J. (2022). *Data Mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.
- Provost, F., & Fawcett, T. (2013). *Data science for business.* O'Reilly Media.
- Han, J., Kamber, M., & Pei, J. (2022). *Data Mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2020). *Data Mining: Practical machine learning Tools and techniques* (4th ed.). Morgan Kaufmann.
- Han, J., Kamber, M., & Pei, J. (2022). *Data Mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer.
- Han, J., Kamber, M., & Pei, J. (2022). *Data Mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.

- Tan, P. N., Steinbach, M., & Kumar, V. (2020). *Introduction to Data Mining* (2nd ed.). Pearson.
- Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2021). The rise of “big data” on cloud computing: Review and open research issues. *Information Systems*, 47, 98–115.
- Marr, B. (2021). *Big data in practice: How 45 successful companies used big data analytics to deliver extraordinary results*. Wiley.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.
- Kohavi, R. (1995). A study of *Cross-validation* and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*.
- Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2020). *Big data* and its technical challenges. *Communications of the ACM*, 57(7), 86–94.
- Kitchin, R. (2021). The real-time city? *Big data* and smart urbanism. *GeoJournal*, 79(1), 1–14.
- Jain, A. K. (2021). Data *Clustering: 50 years beyond K-Means*. *Pattern Recognition Letters*, 31(8), 651–666.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer.

- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear regression analysis* (6th ed.). Wiley.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer.
- Kuhn, M., & Johnson, K. (2020). *Applied predictive modeling*. Springer.
- Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2020). Fundamentals of machine learning for predictive data analytics. MIT Press.
- Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Ali, W. K. M., Alam, M., & Gani, A. (2021). *Big data: Survey, technologies, opportunities, and challenges. The Scientific World Journal*.
- Kitchin, R. (2020). The data revolution: *Big data*, open data, data infrastructures and their consequences. Sage Publications.
- Knafllic, C. N. (2020). Storytelling with data: A data visualization guide for business professionals. Wiley.
- Kuhn, M., & Johnson, K. (2020). *Applied predictive modeling*. Springer.
- Shmueli, G., & Koppius, O. (2020). Predictive analytics in information systems research. *MIS Quarterly*, 35(3), 553–572.

- Lenzerini, M. (2020). Data integRation: A theoretical perspective. *Proceedings of the ACM Symposium on Principles of Database Systems*.
- MacQueen, J. (2020). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear regression analysis* (5th ed.). Wiley.
- Motulsky, H., & Christopoulos, A. (2004). *Fitting models to biological data using Linear and nonLinear regression*. Oxford University Press.
- Seber, G. A. F., & Wild, C. J. (2003). *NonLinear regression*. Wiley.
- Murtagh, F., & Contreras, P. (2012). Algorithms for *Hierarchical Clustering*: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1), 86–97.
- Özsu, M. T., & Valduriez, P. (2020). *Principles of distributed database systems* (4th ed.). Springer.
- Powers, D. M. W. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*.

- Provost, F., & Fawcett, T. (2020). Data science for business: What you need to know about *Data Mining* and data-analytic thinking. O'Reilly Media.
- Quinlan, J. R. (2020). *C4.5: Programs for machine learning*. Morgan Kaufmann.
- Rahm, E., & Do, H. H. (2020). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23(4), 3–13.
- Sokolova, M., & Lapalme, G. (2020). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.
- Stonebraker, M., & Hellerstein, J. M. (2020). What goes around comes around. *Readings in Database Systems*.
- Tan, P. N., Steinbach, M., & Kumar, V. (2021). *Introduction to Data Mining* (2nd ed.). Pearson.
- Triola, M. F. (2020). *Elementary statistics* (13th ed.). Pearson.
- Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2021). *Probability and statistics for engineers and scientists* (10th ed.). Pearson.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2021). *Data Mining: Practical machine learning Tools and techniques* (4th ed.). Morgan Kaufmann.
- Zhang, H. (2004). The optimality of *Naive Bayes*. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*.

Buku referensi berjudul **Data Mining dalam Analisis Data dan Pengambilan Keputusan** membahas cara memahami data agar dapat diolah menjadi informasi yang berguna. Melalui bahasa yang jelas dan mudah diikuti, buku ini mengenalkan proses penggalian pola, hubungan, dan kecenderungan dari data untuk mendukung keputusan yang lebih tepat.

Buku ini cocok untuk masyarakat umum yang ingin memahami bagaimana data digunakan dalam kehidupan modern, mulai dari dunia bisnis, layanan publik, kesehatan, pendidikan, hingga aktivitas sehari-hari. Dengan penyajian yang praktis dan mudah dipahami, buku ini dapat membantu pembaca melihat data sebagai aset penting dalam menentukan langkah dan strategi.