

Komparasi Model Klasifikasi Data Online Retail dengan Seleksi Fitu

Rando¹, Muhammad Abdur Rasyid Asruddin²

^{1,2}Rekayasa Sistem Komputer, Universitas Muhammadiyah Buton

e-mail: rando@umbuton.ac.id

Abstrak

Transaksi yang sering dilakukan dalam aktivitas sehari-hari. aktivitas tersebut dapat dilakukan analisis terhadap barang yang sering dibeli dan pelanggan yang sering dibelanjakan. Setiap barang memiliki nilai penunjang dan nilai kepastian. Metode yang digunakan adalah Data Mining untuk mendapatkan hasil yang terbaik pada data yang tersedia. Penelitian ini melihat komparasi antara model klasifikasi yang akan diterapkan pada data transaksi belanja. Data yang diolah yaitu transaksi belanja online yang didapat pada UCI ML Repository. Penelitian ini akan melakukan komparasi dengan menggunakan fitur seleksi dan tidak menggunakan fitur seleksi. Model klasifikasi yang digunakan pada penelitian ini yaitu Naïve Bayes, J48 dan RandomForest. Hasil yang diharapkan akan mendapatkan model klasifikasi terbaik dalam menentukan model klasifikasi.

Kata kunci: *Data Mining, Transaksi, Klasifikasi*

Abstract

Transactions are frequently carried out in daily activities. These activities can be analyzed to identify frequently purchased items and customers with frequent spending patterns. Each item is associated with a supporting value and a certainty value. The method used in this research is Data Mining to achieve the best results from the available data. This study examines a comparison of classification models applied to shopping transaction data. The processed data consists of online shopping transactions obtained from the UCI ML Repository. The research will compare results with and without feature selection. The classification models used in this study are Naïve Bayes, J48, and Random Forest. The expected outcome is to identify the best classification model for determining the classification approach.

Keywords : *Data Mining, Transactions, Classification*

PENDAHULUAN

Informasi merupakan suatu hal yang sering ditemui dalam kehidupan manusia baik informasi tentang ekonomi, Pendidikan, kesehatan, pertanian, teknologi dan lain sebagainya. Informasi yang tersedia banyak tersedia, namun dalam proses pengolahan data masih kurang untuk meningkatkan aktivitas manusia. Manusia dalam memenuhi kebutuhan melakukan transaksi jual beli. Pengusaha untuk meningkatkan produksinya diperlukan selalu peningkatan proses bisnis untuk dapat bertahan dengan kondisi bisnis yang sangat dinamis.

Para pelaku usaha memiliki informasi berupa data transaksi yang dilakukan tiap ada transaksi jual beli. Data transaksi dapat diolah sebagai analisis kondisi keadaan usaha saat ini. Perkembangan teknologi yaitu big data dapat membantu dalam analisis data yang tersedia. Proses data mining dapat melihat kondisi nyata dari data yang tersedia. Menurut Liao 2020 (Liao & Yang, 2020) peningkatan teknologi akan membantu dalam proses bisnis yang lebih bervariasi dan inovatif. Menurut Abdurrezak Sener (Sener et al., 2019) data mining menjadi peran penting dalam penyelidikan informasi yang bervariasi agar mendapatkan efektivitas operasional.

Berdasarkan latar belakang diatas penelitian ini bertujuan untuk memanfaatkan informasi data transaksi untuk mendapatkan informasi yang berguna. Penelitian ini akan melakukan evaluasi dari data yang telah dilakukan normalisasi. Hasil penelitian ini akan melakukan komparasi dengan metode yang telah tersedia, menemukan metode terbaik untuk menemukan hasil terbaik. Data

penelitian ini diambil dari UCI Repository yaitu Online Retail (link: <http://archive.ics.uci.edu/ml/datasets/Online+Retail>).

Dataset

Data yang digunakan dalam penelitian ini adalah dataset transaksi yang terdapat pada UCI ML Repository. Data yang tersedia merupakan data transaksi antara 01/12/2010 dan 09/12/2011 untuk retail online non-store yang telah terdaftar di Inggris. Perusahaan ini menjual tentang barang dan pembeli mayoritas adalah pedagang grosir. Berikut ini penjelasan atribut yang terdapat pada dataset.

Tabel 1 Informasi Atribut

No	Nama Atribut	Keterangan
1.	Invoice	Nomor transaksi
2.	StockCode	Kode dari tiap produk
3.	Description	Nama dari Produk
4.	Quantity	Jumlah dari setiap produk per transaksi
5.	InvoiceDate	Tanggal dan waktu tiap melakukan transaksi
6.	UnitPrice	Harga dari 1 produk
7.	CustomerID	Nomor dari pembeli
8.	Country	Nama dari negara

Data Mining

Data mining digunakan untuk menemukan pola dari kumpulan data yang besar dari berbagai metode seperti *machine learning*, *database manipulation* dan *statistic* (Salo et al., 2018; Thakkar et al., 2020). Data banyak diberbagai aplikasi baik dari segi perekonomian, kesehatan, pertanian dan lain sebagainya. *Data mining* juga disebut sebagai *Knowledge Discovery in Database* (KDD), untuk membantu mengolah data menjadi informasi yang berguna dapat digunakan untuk pengambilan keputusan. Proses data mining memiliki beberapa tahapan yaitu sebagai berikut :

1. *Data Cleaning*
Pada tahapan ini adalah menelusuri dataset guna membersihkan data yang tidak konsisten dan tidak relevan. Tahapan data cleaning adalah menghapus data yang tidak berguna atau kosong yang menjadikan perhitungan pada dataset ini menjadi tidak konsisten.
2. *Data Integration*
Proses *data integration* dilakukan pada atribut yang terdapat pada dataset yang unik. Hasil dari integrasi data yaitu menambahkan data sehingga akan menghasilkan data.
3. *Data Transformation*
Data yang akan diubah dengan sesuai *tools* yang digunakan
4. *Pattern evaluation*
Pada proses *pattern evaluation* akan menemukan data yang bernilai. Hasil dari proses ini yaitu pola pola yang khas berdasarkan data yang dimiliki
5. *Knowledge Presentation*
Proses ini akan menemukan pengetahuana baru. Hasil dari analisis data yang dikembangkan menjadi pengetahuan. Selanjutnya dilakukan proses data mining.

Naïve Bayes

Naïve Bayes Classifier (NBC) merupakan salah satu teknik klasifikasi yang sangat berguna dalam berbagai bidang, termasuk untuk diagnosis medis, analisis sentimen, dan pengelolaan risiko. Metode ini berakar pada teorema Bayes yang dikembangkan oleh ilmuwan Inggris, Thomas Bayes, yang berfokus pada perhitungan probabilitas berdasarkan data yang ada. Asumsi utama dari Naïve Bayes adalah independensi antar fitur, yang menjadikan metode ini disebut "naïf" karena mengasumsikan bahwa tidak ada interaksi antara variabel input..

Penggunaan Naïve Bayes dalam mendeteksi penyakit, seperti *Mucopolysaccharidosis* tipe II (MPS II), menunjukkan bahwa klasifikasi ini dapat mencapai akurasi yang sangat tinggi, yaitu

99% dalam pengujian data, dengan tingkat sensitivitas dan spesifisitas yang juga sangat baik (Ehsani-Moghaddam et al., 2018). Ini menunjukkan efektivitas metode ini dalam memproses dan menganalisis data medis, memudahkan dokter untuk membuat diagnosis yang akurat dengan memanfaatkan catatan medis elektronik. Dalam konteks analisis data yang lebih luas, Naïve Bayes sering digunakan dalam klasifikasi teks dan analisis sentimen. Contohnya, metode Naïve Bayes telah terbukti mampu menghasilkan klasifikasi yang baik saat diterapkan pada analisis komentar dalam pemilihan umum (Fernando et al., 2019) dan pada prediksi pencapaian akademik siswa (Gata et al., 2019).

Hal ini menunjukkan fleksibilitas Naïve Bayes dalam menangani berbagai bentuk data, termasuk data yang bersifat kategorikal dan frekuensi. Kelebihan lain dari Naïve Bayes adalah efisiensinya dalam komputasi, yang memungkinkan penggunaannya dengan dataset yang lebih kecil tanpa memerlukan perhitungan kovarian yang kompleks antar fitur (Karandikar et al., 2015). Ini mengizinkan NBC untuk tetap relevan bahkan dalam kondisi di mana data yang tersedia terbatas. Namun, penting untuk dicatat bahwa meskipun Naïve Bayes efektif dalam banyak aplikasi, asumsi independensi atribut dapat mempengaruhi akurasi klasifikasi terutama ketika fitur saling terkait atau terdapat interaksi (Chen et al., 2021).

Tree J-48

J48, yang merupakan implementasi dari algoritma C4.5, adalah metode klasifikasi yang populer dalam data mining dan digunakan untuk membangun pohon keputusan. Dalam pendekatan ini, data dengan label digunakan untuk mengembangkan model prediksi yang berbasis pada struktur berhierarki dari keputusan yang terbentuk. J48 bekerja dengan cara pemilihan atribut yang paling prediktif pada setiap langkah, menggunakan pendekatan "*divide and conquer*" untuk membagi dataset berdasarkan atribut yang memaksimalkan rasio gain (Chabalenge et al., 2020; Rolón-González et al., 2021). Keunggulan dari J48 terletak pada mekanismenya yang transparan, serta kemudahan dalam menginterpretasikan hasil dari model pohon keputusan yang dihasilkan (Chabalenge et al., 2020; Zeng, 2013). Berikut ini tahapan algoritme J48 :

- 1) Menyiapkan data training.
- 2) Menentukan akar dari pohon.
- 3) Menghitung nilai Gain melalui Persamaan (1)

$$\text{Entropy} = - \sum_{i=1}^n p_i * \log_2 p_i$$

- 4) Ulangi langkah ke-2 hingga semua tupel terpartisi dengan menggunakan Persamaan (2).

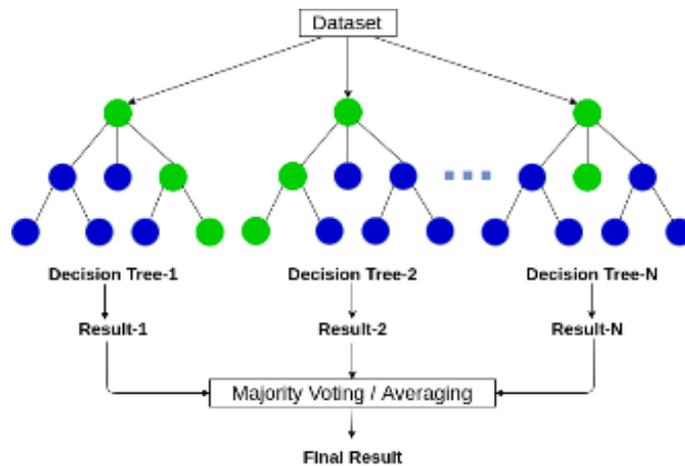
$$\text{Gain (S,A)} = S - \sum_{i=1}^n \frac{|S_i|}{|S|} * S_i$$

- 5) Proses partisi pohon keputusan akan berhenti saat semua tupel dalam node N mendapat kelas yang sama dan atau tidak ada atribut di dalam tupel yang dipartisi lagi dan atau tidak ada tupel di dalam cabang yang kosong

Random Forest

Random Forest (RF) adalah metode klasifikasi yang diperoleh dari penggabungan beberapa pohon keputusan (*decision trees*) dalam satu sistem yang disebut "hutan". Metode ini merupakan bagian dari teknik *ensemble learning*, yang bertujuan untuk meningkatkan akurasi prediksi melalui penggabungan hasil dari berbagai keputusan pohon. Penambahan jumlah pohon dalam hutan biasanya berbanding lurus dengan peningkatan akurasi model, serta kemampuannya untuk menghindari *overfitting* dibandingkan dengan penggunaan satu pohon tunggal (Ramdani et al., 2022; Singh et al., 2017). Random Forest tidak hanya efisien dalam menangani data berangkat dari campuran variabel kontinu dan kategorik, tetapi juga menunjukkan daya tahan yang baik terhadap noise dalam data.

Keuntungan utama dari Random Forest adalah kemampuannya untuk melakukan klasifikasi yang lebih akurat pada dataset besar dan beragam. Penelitian menunjukkan bahwa RF sering kali unggul dibandingkan dengan Naïve Bayes dalam tugas klasifikasi yang lebih kompleks, terutama ketika banyak fitur mempengaruhi keputusan klasifikasi hasil akhir (Imran et al., 2022; Yang, 2022). Namun, masih ada konteks di mana Naïve Bayes menunjukkan hasil yang baik, terutama ketika model dibutuhkan untuk komputasi yang lebih cepat dan efisien, lebih-lebih dalam situasi dengan dataset besar di mana waktu pemrosesan menjadi penting (Nalatissifa et al., 2021)

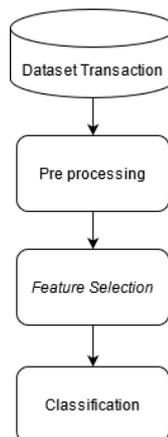


Gambar 1 Algoritma Random Forest

METODE

Perancangan Sistem

Perancangan Diagram alir merupakan gambaran proses satu dengan yang lain dalam sistem komputer. Pengguna akan memasukkan data aktivitas fisik ditambah dengan durasi dan kalori yang masuk ke dalam tubuh. Setelah pengguna memasukkan data yang diperlukan, sistem akan melakukan proses *fuzzy logic* dengan metode Mamdani dengan memasukkan fasilitas yang tersedia. Hasil akhir dari pengguna akan mendapatkan harga yang sesuai dengan harga pasaran.



Gambar 2 Diagram Alir

Penjelasan Gambar 2 adalah sebagai berikut:

1. Dataset yang digunakan yaitu UCI ML Repository
2. Sebelum dilakukan klasifikasi dilakukan pre processing yaitu dengan melakukan pembersihan data.
3. *Feature Selection* digunakan untuk meningkatkan hasil klasifikasi.

Perancangan Sistem

Dalam proses *data cleaning* ini, terdapat sejumlah data yang perlu dihapus agar analisis menjadi lebih akurat. Salah satu langkah yang dilakukan adalah menghilangkan data pengembalian barang, yang ditandai dengan harga bertanda minus (-), serta data yang memiliki nilai kosong. Selain itu, pada data transaksi ini, terdapat beberapa kolom yang dianggap tidak relevan atau tidak digunakan untuk analisis lebih lanjut. Kolom-kolom tersebut diidentifikasi dan dihapus untuk menyederhanakan struktur data dan mempermudah proses pengolahan di tahap berikutnya. Langkah-langkah ini dilakukan untuk memastikan bahwa data yang digunakan benar-benar bersih dan siap untuk dianalisis. Pada data transaksi ini beberapa kolom data yang tidak digunakan yaitu :

1. *Invoice*: tergabung pada data transaksi
2. *Description*: sudah terdapat pada stockcode

Integrasi dan Data Transformasi

Perancangan Diagram alir merupakan gambaran proses satu dengan yang lain dalam sistem komputer. Pengguna akan memasukkan data aktivitas fisik ditambah dengan durasi dan kalori yang masuk ke dalam tubuh. Setelah pengguna memasukkan data yang diperlukan, sistem akan melakukan proses *fuzzy logic* dengan metode Mamdani dengan memasukkan fasilitas yang tersedia. Hasil akhir dari pengguna akan mendapatkan harga yang sesuai dengan harga pasaran.

Perancangan Sistem

Perancangan Pada proses ini mendapatkan data baru yaitu jumlah pembelian (*Amount*), Kebaruan (*Recency*) dan Jumlah Transaksi tiap customer (*Frequency*). Hasil berasal dari jumlah barang dapat dilihat pada rumus dibawah

$$\text{Amount} = \text{Jumlah item yang dibeli} * \text{Harga per item}$$

Quantity	InvoiceDate	Price	CustomerID	Country	Amount
6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30
6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00

Gambar 3 Hasil dari biaya yang dikeluarkan (*Amount*)

Berikut ini adalah data kebaruan. Data kebaruan ditentukan dengan menghitung selisih antara tanggal transaksi maksimal dan tanggal transaksi tertentu. *Recency* merepresentasikan jarak waktu dari data kebaruan tersebut. Data ini menunjukkan selang waktu hingga transaksi terbaru.

$$\text{Recency} = \text{Tanggal Paling Baru} - \text{Tanggal Invoice}$$

CustomerID	Recency
0	326.0
1	2.0
2	75.0

Gambar 4 Hasil dari Recency

Data *Frequency* menunjukkan jumlah transaksi yang dilakukan oleh setiap pelanggan dalam periode tertentu. Dengan menganalisis data ini, kita dapat memahami seberapa sering seorang pelanggan melakukan transaksi. Frekuensi transaksi ini menjadi indikator penting untuk menilai tingkat keterlibatan pelanggan dan membantu dalam segmentasi pelanggan berdasarkan perilaku pembelian.

	CustomerID	Recency	Amount	Frequency
0	12346.0	326.0	77183.601	1
1	12347.0	2.0	4310.001	182
2	12348.0	75.0	1797.241	31
3	12349.0	19.0	1757.551	73
4	12350.0	310.0	334.401	17

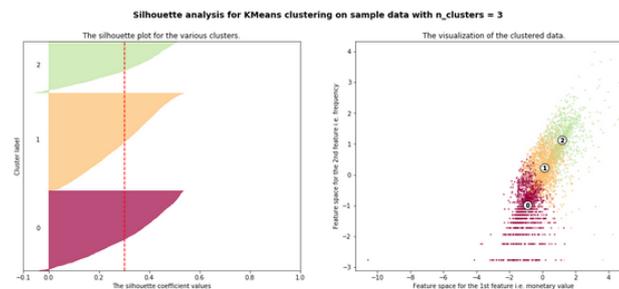
Gambar 5 Hasil dari Frequency

Perancangan Diagram alir merupakan gambaran proses satu dengan yang lain dalam sistem komputer. Pengguna akan memasukkan data aktivitas fisik ditambah dengan durasi dan kalori yang masuk ke dalam tubuh. Setelah pengguna memasukkan data yang diperlukan, sistem akan melakukan proses *fuzzy logic* dengan metode Mamdani dengan memasukkan fasilitas yang tersedia. Hasil akhir dari pengguna akan mendapatkan harga yang sesuai dengan harga pasaran

Representasi Pengetahuan

Dari data yang telah tersedia, proses analisis menghasilkan wawasan baru yang bernilai. Pengetahuan baru ini diperoleh melalui tahapan *data cleaning*, transformasi data, dan evaluasi pola (*pattern evaluation*), yang memastikan data dalam kondisi optimal untuk analisis. Pada tahap ini, dilakukan *clustering* untuk mengelompokkan data dan memperoleh label berdasarkan karakteristik yang telah ada.

Dalam proses *clustering* pada dataset ini, dilakukan pembagian data ke dalam tiga kelompok, menghasilkan tiga label berdasarkan pola-pola yang ditemukan selama proses *data mining*. Penentuan label baru ini didasarkan pada kedekatan atau kemiripan antar data, sehingga setiap kelompok menggambarkan karakteristik yang serupa. Metode *clustering* menggunakan algoritma KMeans, yang telah terbukti efektif untuk analisis seperti ini. Visualisasi hasil *clustering* dapat dilihat pada gambar di bawah, yang memberikan representasi grafis dari pengelompokan data berdasarkan label yang dihasilkan.



Gambar 6 Hasil Representasi pengetahuan

HASIL DAN PEMBAHASAN

Hasil yang didapatkan pada proses pembersihan data yaitu dengan menggabungkan beberapa dari tabel yang diperlukan. pada penelitian ini menghasilkan data baru dengan 3 atribut dan 1 kelas yaitu 2.

Tabel 2 Atribut Baru

No	Nama Atribut	Keterangan
1.	Recency	Keterkinian pembeli untuk melakukan pembelian
2.	Amount	Jumlah Harga yang dibeli
3.	Frequency	Jumlah CustomerID dalam transaksi

Sistem ini Dari hasil analisis ini, model *Random Tree* dianggap sebagai algoritma yang paling sesuai untuk dataset yang digunakan. Nilai evaluasi yang konsisten dan unggul pada sebagian besar metrik menunjukkan bahwa *Random Tree* mampu mengelola data dengan baik dan menghasilkan prediksi yang akurat. Namun, meskipun *Random Tree* menunjukkan hasil terbaik, algoritma *Random Forest* juga menjadi kandidat yang kuat dengan keunggulan pada nilai ROC Area dan PRC Area yang sedikit lebih tinggi. Hal ini menunjukkan bahwa *Random Forest* memiliki kemampuan yang sedikit lebih baik dalam menangani prediksi probabilitas. Hasil dari seleksi fitur dapat dilihat pada tabel 3.

Tabel 3 Hasil Klasifikasi

Metode	MAE	Kappa Statistic	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Naïve Bayes	0.0942	0.8712	0.918	0.917	0.917	0.869	0.985	0.976
Random Forest	0.0386	0.9546	0.971	0.971	0.971	0.954	0.998	0.997
J-48	0.0372	0.9217	0.95	0.95	0.949	0.92	0.964	0.93
Random Tree	0.032	0.9257	0.952	0.952	0.952	0.924	0.962	0.924

Rekomendasi selanjutnya adalah untuk mempertimbangkan kebutuhan spesifik dari penelitian ini. Jika akurasi prediksi menjadi prioritas utama, *Random Tree* adalah pilihan ideal. Namun, jika stabilitas dalam berbagai kondisi dataset lebih diutamakan, *Random Forest* bisa menjadi alternatif yang baik.

SIMPULAN

Dalam Proses data cleaning yang dilakukan berhasil menghasilkan dataset yang lebih terstruktur dengan tiga atribut dan satu kelas. Langkah-langkah ini memastikan data yang digunakan sudah optimal untuk analisis lebih lanjut. Dalam penelitian ini, model machine learning diuji untuk mengevaluasi performanya, dan hasilnya menunjukkan bahwa model *Random Tree* memberikan hasil terbaik di sebagian besar metrik evaluasi. Kinerja yang unggul pada model *Random Tree* menunjukkan bahwa algoritma ini mampu menangani dataset dengan baik, menghasilkan prediksi yang akurat, dan menunjukkan konsistensi yang tinggi. Meskipun demikian, model *Random Forest* juga mencatat keunggulan tertentu pada beberapa metrik, menjadikannya alternatif yang patut dipertimbangkan, terutama jika stabilitas dan prediksi probabilitas yang lebih presisi menjadi prioritas.

Dengan demikian, hasil penelitian ini memberikan kontribusi penting bagi optimalisasi analisis data, khususnya dalam proses klasifikasi dan pengelompokan data menggunakan algoritma machine learning. Model *Random Tree* direkomendasikan sebagai pilihan utama untuk kasus ini, dengan mempertimbangkan hasil evaluasi metrik yang telah diperoleh.

DAFTAR PUSTAKA

- Chabalenge, B., Korde, S., Kelly, A. L., Neagu, D., & Paradkar, A. (2020). Understanding Matrix-Assisted Continuous Co-crystallization Using a Data Mining Approach in Quality by Design (QbD). *Crystal Growth & Design*, 20(7), 4540–4549. <https://doi.org/10.1021/acs.cgd.0c00338>
- Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved naïve Bayes classification algorithm for traffic risk management. *EURASIP Journal on Advances in Signal Processing*, 2021(1), 30. <https://doi.org/10.1186/s13634-021-00742-6>
- Ehsani-Moghaddam, B., Queenan, J. A., MacKenzie, J., & Birtwhistle, R. V. (2018). Mucopolysaccharidosis type II detection by Naïve Bayes Classifier: An example of patient classification for a rare disease using electronic medical records from the Canadian Primary Care Sentinel Surveillance Network. *PLOS ONE*, 13(12). <https://doi.org/10.1371/journal.pone.0209018>
- Fernando, J. R., Budiraharjo, R., & Haganusa, E. (2019). Spam Classification on 2019 Indonesian President Election Youtube Comments Using Multinomial Naïve-Bayes. *Indonesian Journal*

- of Artificial Intelligence and Data Mining*, 2(1). <https://doi.org/10.24014/ijaidm.v2i1.6445>
- Gata, W., Basri, H., Hidayat, R., Patras, Y. E., Baharuddin, B., Fatmasari, R., Tohari, S., & Wardhani, N. K. (2019). Algorithm Implementations Naïve Bayes, Random Forest. C4.5 on Online Gaming for Learning Achievement Predictions. *Proceedings of the 2nd International Conference on Research of Educational Administration and Management (ICREAM 2018)*. <https://doi.org/10.2991/icream-18.2019.1>
- Imran, B., Hambali, H., Subki, A., Zaeniah, Z., Yani, A., & Alfian, M. R. (2022). DATA MINING USING RANDOM FOREST, NAÏVE BAYES, AND ADABOOST MODELS FOR PREDICTION AND CLASSIFICATION OF BENIGN AND MALIGNANT BREAST CANCER. *Jurnal Pilar Nusa Mandiri*, 18(1), 37–46. <https://doi.org/10.33480/pilar.v18i1.2912>
- Karandikar, J., McLeay, T., Turner, S., & Schmitz, T. (2015). Tool wear monitoring using naïve Bayes classifiers. *The International Journal of Advanced Manufacturing Technology*, 77(9–12), 1613–1626. <https://doi.org/10.1007/s00170-014-6560-6>
- Liao, S. H., & Yang, L. L. (2020). Mobile payment and online to offline retail business models. *Journal of Retailing and Consumer Services*, 57(151), 102230. <https://doi.org/10.1016/j.jretconser.2020.102230>
- Nalatissifa, H., Gata, W., Diantika, S., & Nisa, K. (2021). Perbandingan Kinerja Algoritma Klasifikasi Naive Bayes, Support Vector Machine (SVM), dan Random Forest untuk Prediksi Ketidakhadiran di Tempat Kerja. *Jurnal Informatika Universitas Pamulang*, 5(4), 578. <https://doi.org/10.32493/informatika.v5i4.7575>
- Ramdani, N., Prasetyowati, S. S., & Sibaroni, Y. (2022). Performance Analysis of Bandung City Traffic Flow Classification with Machine Learning and Kriging Interpolation. *Building of Informatics, Technology and Science (BITS)*, 4(2), 694–704. <https://doi.org/10.47065/bits.v4i2.1972>
- Rolón-González, C. A., Castañón-Méndez, R., Alarcón-Paredes, A., López-Yáñez, I., & Yáñez-Márquez, C. (2021). Improving the Performance of an Associative Classifier in the Context of Class-Imbalanced Classification. *Electronics*, 10(9), 1095. <https://doi.org/10.3390/electronics10091095>
- Salo, F., Injadat, M., Nassif, A. B., Shami, A., & Essex, A. (2018). Data mining techniques in intrusion detection systems: A systematic literature review. *IEEE Access*, 6, 56046–56058. <https://doi.org/10.1109/ACCESS.2018.2872784>
- Sener, A., Barut, M., Oztekin, A., Avcilar, M. Y., & Yildirim, M. B. (2019). The role of information usage in a retail supply chain: A causal data mining and analytical modeling approach. *Journal of Business Research*, 99(November 2017), 87–104. <https://doi.org/10.1016/j.jbusres.2019.01.070>
- Singh, A., N., M., & Lakshmiganthan, R. (2017). Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms. *International Journal of Advanced Computer Science and Applications*, 8(12). <https://doi.org/10.14569/IJACSA.2017.081201>
- Thakkar, H., Shah, V., Yagnik, H., & Shah, M. (2020). Comparative Anatomization of Data Mining and Fuzzy Logic Techniques Used In Diabetes Prognosis. *Clinical EHealth*, 4, 12–23. <https://doi.org/10.1016/j.ceh.2020.11.001>
- Yang, R. (2022). Detection of Fake News Based on Typical Machine Learning Models. *BCP Business & Management*, 23, 348–352. <https://doi.org/10.54691/bcpbm.v23i.1371>