JURNAL SIMPUL INOVASI
Journal of Innovation – Hub
Pusat Inkubator Bisnis – Universitas Jenderal Soedirman
ISSN: 0000-0000 E-ISSN: 0000-0000
Volume (2), Issue (1), Halaman 36-40, Juni 2025
Pusat Inkubator Bisnis Unsoed

Analisis Komparatif Tool Data Mining WEKA dan RapidMiner dalam Klasifikasi Penyakit Ginjal Kronis (PGK) Berbasis Dataset UCI

[1]Ennes Pratiwi, [2*]Agus Wantoro

[1]Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Teknokrat Indonesia
Bandar Lampung, Lampung, Indonesia
[2]Teknik Informatika, Fakultas Teknologi dan Informatika, Universitas Aisyah Pringsewu
Pringsewu, Lampung, Indonesia
*E-mail: aguswantoro@aisyahuniversity.ac.id (*corresponding author*)

Abstrak
Penyakit ginjal kronis (PGK) merupakan salah satu penyakit tidak menular dengan tingkat prevalensi dan mortalitas yang terus meningkat secara global. Deteksi dini PGK sangat penting untuk mencegah komplikasi dan memperpanjang harapan hidup pasien. Penelitian ini bertujuan untuk membandingkan performa algoritma klasifikasi yang diterapkan pada dua platform data mining populer, yaitu WEKA dan RapidMiner, dalam menganalisis *dataset* penyakit ginjal kronis dari UCI *Machine Learning* (ML) Repository. Tiga algoritma klasifikasi digunakan dalam eksperimen, yaitu Decision Tree, Naive Bayes, dan Support Vector Machine (SVM), dengan skema validasi silang 10-fold. Kinerja model dievaluasi berdasarkan Confusion Matrix berupa nilai *accuracy, precission*, dan *recall*. Hasil menunjukkan bahwa terdapat perbedaan performa antar algoritma pada masing-masing platform. Pada *tools* WEKA, algoritma Decision Tree menunjukkan performa terbaik dengan akurasi 99%, diikuti oleh SVM dan Naive Bayes. Untuk nilai precisson secara umum, RapidMiner menunjukkan kinerja lebih baik, namun untuk nilai racall, WEKA lebih unggul. Pada tools RapidMiner, Naive Bayes memberikan hasil paling akurat dengan nilai akurasi mencapai 99,5%, sedangkan SVM menyusul di bawahnya. Secara umum *tools* RapidMiner memiliki kinerja akurasi yang lebih baik. Namun, setiap platform memiliki keunggulan masing-masing. WEKA unggul dari segi fleksibilitas eksperimen dan dukungan terhadap tuning parameter yang lebih teknis, sedangkan RapidMiner lebih ramah pengguna berkat antarmuka grafis yang intuitif dan kemampuan visualisasi proses yang baik

Kata kunci: *Penyakit Ginjal kronis, WEKA, RapidMiner, Klasifikasi, Data mining, Decision Tree, SVM, Naive Bayes*

1 Pendahuluan

Penyakit Ginjal Kronis (PGK) merupakan salah satu masalah kesehatan global yang semakin meningkat prevalensinya [1]. Menurut laporan terbaru dari Global Burden of Disease Study (2023), PGK menempati peringkat ke-8 penyebab utama kematian di dunia, dengan angka kejadian yang terus bertambah akibat faktor seperti hipertensi, diabetes, dan gaya hidup tidak sehat [2]. Deteksi dini PGK sangat penting untuk mencegah komplikasi lanjut dan menurunkan angka kematian. Namun, diagnosis yang akurat dan cepat seringkali terhambat oleh keterbatasan sumber daya medis dan variabilitas dalam interpretasi data klinis[3].

Dalam konteks ini, data mining dan *Machine Learning* (ML) telah menjadi pendekatan yang menjanjikan dalam membantu diagnosis dan prediksi penyakit secara lebih objektif dan sistematis [4]. Dengan menggunakan dataset yang tersedia secara publik seperti Dataset Ginjal Kronis dari UCI Machine Learning Repository, berbagai algoritma klasifikasi dapat diterapkan untuk membangun model prediktif berbasis data laboratorium dan klinis pasien [5]

Dua platform atau *tools* yang paling populer digunakan untuk eksplorasi dan penerapan algoritma data mining adalah WEKA *(Waikato Environment for Knowledge Analysis)* dan RapidMiner. Keduanya menyediakan antarmuka pengguna grafis, fitur preprocessing, pemodelan, dan evaluasi kinerja model [6]. Namun demikian, perbedaan dalam arsitektur sistem, cara kerja workflow, serta dukungan terhadap jenis algoritma dan parameter *tuning* menjadikan keduanya memiliki karakteristik unik yang dapat memengaruhi hasil analisis dan klasifikasi [7]

Meski beberapa penelitian telah memanfaatkan WEKA maupun RapidMiner secara terpisah dalam klasifikasi berbagai penyakit, masih jarang dilakukan studi komparatif langsung antara keduanya dalam konteks klasifikasi PGK. Oleh karena itu, penelitian ini bertujuan untuk mengevaluasi perbedaan hasil klasifikasi kinerja algoritma ML populer seperti Decision Tree, Naive Bayes, dan SVM yang diterapkan pada dataset PGK di kedua platform. Evaluasi kinerja

dilakukan berdasarkan metrik akurasi, precision, dan recall untuk memberikan wawasan praktis mengenai keunggulan masing-masing tools dalam konteks klasifikasi dataset medis.

Dengan membandingkan performa WEKA dan RapidMiner secara langsung pada task yang sama, penelitian ini diharapkan dapat memberikan kontribusi nyata dalam pemilihan platform data mining yang lebih efisien dan akurat dalam pengembangan sistem pendukung keputusan medis, khususnya untuk diagnosis PGK

2 Metodologi

Tahapan penelitian adalah langkah-langkah yang disusun secara terstruktur dan sistematis untuk mencapai tujuan penelitian. Tahapan ini meliputi berbagai aktivitas, mulai dari identifikasi masalah, pengumpulan data, analisis data, hingga evaluasi penelitian. Penelitian ini menggunakan pendekatan eksperimental kuantitatif dengan beberapa tahapan terstruktur. Tahapan penelitian ditampilkan pada Gambar 1.

Gambar 1. Tahapan penelitian

2.1 Pengumpulan Data

Pengumpulan data adalah proses mengumpulkan, mengukur, dan menganalisis informasi dari berbagai sumber untuk menjawab pertanyaan penelitian, menguji hipotesis, atau mengevaluasi hasil. Proses ini penting dalam berbagai bidang, termasuk ilmu pengetahuan, penelitian, dan pengambilan keputusan [8]. Data yang digunakan berupa dataset PGK yang diperoleh dari UCI *Machine Learning Repository*. Dataset ini berisi 400 sampel pasien, terdiri dari dua class yaitu ("ckd" dan "not-ckd"), 24 atribut seperti Usia, Tekanan darah, *Gravitas (Specific Gravity), Albumin, Kreatinin (Serum Creatinine), Natrium (Sodium), Kalium (Potassium), Hemoglobin, Volume sel darah merah (PCV - Packed Cell Volume),* Warna sel darah putih (WCC - *White Blood Cell Count*), Warna sel darah merah (RCC - *Red Blood Cell Count),* Bilirubin, dan Glukosa (*Blood Glucose Random*) [3].

2.2 *Pre-processing* Data

Preprocessing data adalah proses mempersiapkan data mentah agar siap digunakan dalam analisis atau model machine learning. Proses ini mencakup pembersihan, transformasi, dan integrasi data untuk memastikan kualitas, keakuratan, dan format yang sesuai untuk analisis lebih lanjut [9]. Kegiatan pada tahap ini meliputi (a) Penanganan nilai hilang *(missing values)* dengan metode imputasi (b) Normalisasi atau transformasi data bila diperlukan. (c) Encoding variabel kategorikal menjadi numerik.

2.3 Klasifikasi

Klasifikasi adalah proses pengelompokkan atau penggolongan sesuatu berdasarkan ciri-ciri atau karakteristik yang sama, sehingga dapat diidentifikasi, dibandingkan, dan dipelajari dengan lebih mudah [10]. Dalam konteks data, klasifikasi adalah proses memberikan label atau kategori pada data untuk memudahkan analisis dan pemodelan [11]. Tiga algoritma akan diuji yaitu Decision Tree, Naive Bayes, dan SVM. Selanjutnya Proses dilakukan pada dua platform: WEKA dan RapidMiner

2.4 Cross-validation

Merupakan teknik statistik yang digunakan dalam ML dan pemodelan prediktif lainnya untuk menilai kinerja dan kemampuan generalisasi suatu model [12]. Berdasarkan geeksforgeeks.org, pada *cross-validation*, data yang tersedia akan

dibagi ke dalam subset *(fold),* supaya dapat dilakukan pelatihan dan pengujian model berkali-kali. Teknik ini memberikan estimasi performa model yang lebih akurat pada data yang tidak terlihat [13]. Manfaat penting lainnya dari *cross validation*, yakni membantu data analis mengatasi masalah *overfitting* atau kondisi saat model terlalu spesifik pada data pelatihan sehingga kurang baik dalam menganalisis data baru [14]. Skema *fold cross-validation* digunakan untuk mengevaluasi performa model [12].

2.5 Evaluasi

Evaluasi dilakukan untuk mengetahui kinerja algoritma ML. Kami mengevaluasi kinerja algoritma ML menggunakan Confusion Matrix (CM) berupa nilai *Accuracy, Precission*, dan *Recall*. CM adalah tabel yang digunakan untuk mengevaluasi kinerja suatu model klasifikasi dalam ML [15]. Matriks ini memvisualisasikan perbandingan antara hasil prediksi model dengan nilai aktual, membantu memahami kesalahan dan kelemahan model secara detail [16]. Perbandingan hasil dilakukan untuk melihat keunggulan relatif antar platform dan algoritma. Analisis visual dilakukan dengan grafik perbandingan hasil.

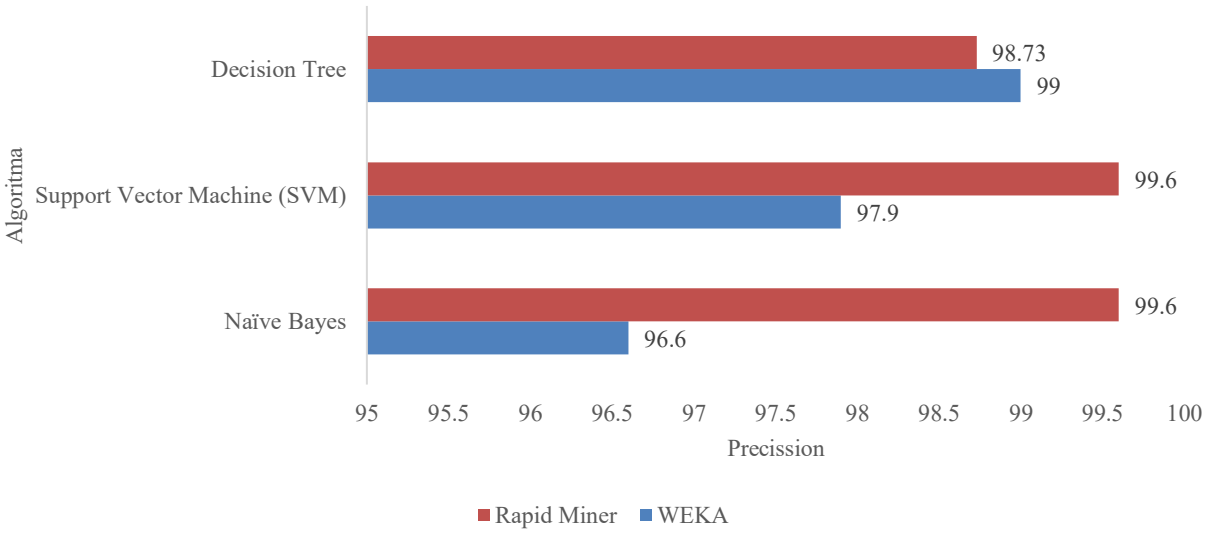
3 Hasil dan pembahasan

Kami melakukan perbandingan kinerja algoritma klasifikasi ML menggunakan tools WEKA versi 3.9.6 dan RapidMiner 10.0.0. pada dataset PGK. Hasil kinerja masing-masing algoritma ML menggunakan tools WEKA dan RapidMiner ditampilkan dalam Tabel 1.

Tabel 1. Perbandingan *accuracy tools* WEKA dan RapidMiner

No.	Algoritma	WEKA	RapidMiner
1	Naïve Bayes	95	**99.5**
2	Support Vector Machine (SVM)	97.75	98.75
3	Decision Tree	99	95
Rata-rata		**97.25**	**97.75**

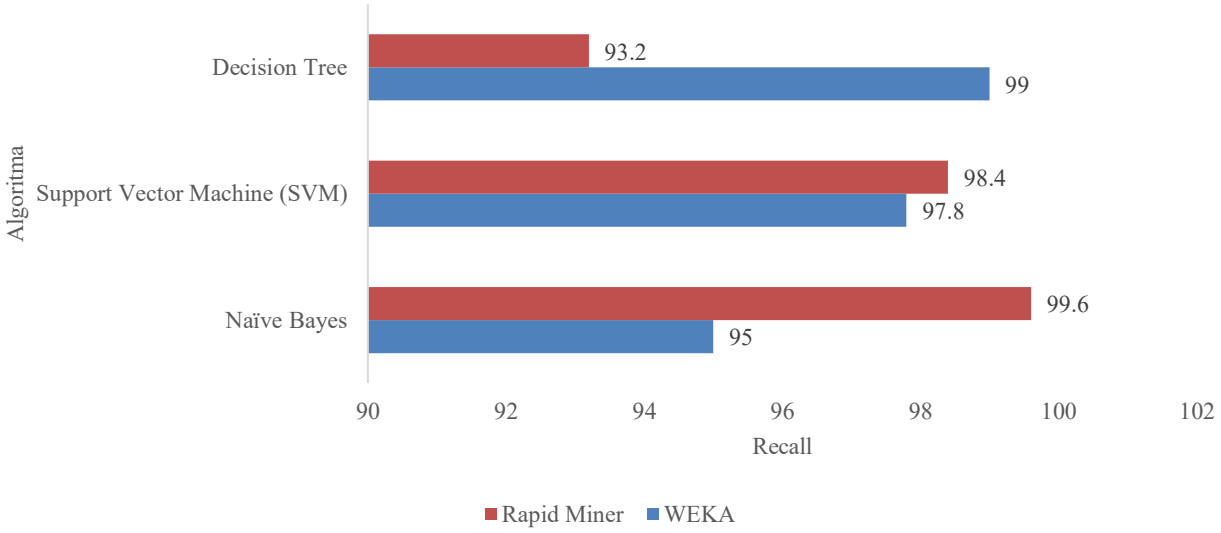
Berdasarkan Tabel 1, algoritma Naive Bayes mampu melakukan klasifikasi dengan akurasi tertinggi. Sedangkan hasil perbandingan kinerja tools, secara umum RapidMiner memiki akurasi yang lebih baik dari WEKA dengan gap 0.5%. Kami menemukan bahwa kombinasi algoritma Naive Bayes dan tools RapidMiner memiliki kinerja terbaik. Selain itu, kami melakukan perbandingan precission di yang ditampilkan pada Gambar 2.



Gambar 2. Perbandingan nilai prescission algoritma ML menggunakan tools WEKA dan RapidMiner

Berdasarkan perbandingan kinerja recall pada Gambar 2, terlihat bahwa pada pengunaan tools RapidMiner memiliki nilai precission lebih tinggi dibandingkan menggunakan WEKA. Pada tools RapidMiner, Algoritma SVM dan Naive Bayes memiliki nilai precission terbaik, sedangkan algoritma Decision Tree memiliki kinerja terburuk. Pada penggunaan tools

WEKA, menghasilkan precission yang berbeda. Justru algoritma Decision Tree memiliki precissoin terbaik, diikuti SVM dan Naive Bayes. Secara umum kinerja precission menggunakan tools RapidMiner memiliki nilai lebih baik dengan selisih (gap) sebesar 1.47%. Pada evaluasi nilai precission, kombinasi algoritma SVM, dan Naive Bayes dengan tools RapidMiner memiliki kombinasi terbaik. Selain itu, kami melakukan perbandingan nilai recall. Hasil perbandingan di ditampilkan pada Gambar 3.



Gambar 3. Perbandingan nilai recall algoritma ML menggunakan tools WEKA dan RapidMiner

Berdasarkan perbandingan kinerja recall pada Gambar 3, penggunanaan tools RapidMiner dan WEKA menggunakan nilasi recall yang berbeda. Hasil pada tools RapidMiner menempatkan NaiveBayes memiliki recall terbaik, diikuti SVM dan Decision Tree. Sedang pada tools WEKA, Decision Tree memiliki kinerja terbaik, diikuti SVM, dan Niave Bayes. Secara umum kinerja recall menggunakan tools WEKA memiliki nilai lebih baik dengan selisih (gap) sebesar 0.2%. Pada evaluasi nilai recall, kombinasi algoritma Naive Bayes dengan tools Rapid Miner memiliki kombinasi terbaik.

4 Kesimpulan

Berdasarkan hasil penelitian dan analisis terhadap performa algoritma klasifikasi pada platform WEKA dan RapidMiner, dapat disimpulkan bahwa kinerja algoritma berbeda antar platform: di tools WEKA, algoritma Decision Tree menunjukkan performa terbaik dengan akurasi 99%, diikuti oleh SVM dan Naive Bayes. Di RapidMiner, Naive Bayes memberikan hasil paling akurat dengan nilai akurasi mencapai 99,5%, sedangkan SVM menyusul di bawahnya. Secara umum tools RapidMiner memiliki kinerja akurasi yang lebih baik. Namun, setiap platform memiliki keunggulan unik. WEKA unggul dari segi fleksibilitas eksperimen dan dukungan terhadap *tuning* parameter yang lebih teknis. RapidMiner lebih ramah pengguna berkat antarmuka grafis yang intuitif dan kemampuan visualisasi proses yang baik, cocok untuk pengguna non-programmer. Pemilihan platform sebaiknya disesuaikan dengan kebutuhan pengguna

Untuk eksperimen lanjutan yang kompleks dan teknis, WEKA lebih ringan dalam eksekusi model karena tidak membutuhkan memori yang besar; sedangkan untuk kebutuhan pengembangan sistem prototipe klinis yang cepat dan *user-friendly*, RapidMiner menjadi pilihan yang lebih efisien. Dengan demikian, kedua platform memiliki potensi yang besar dalam pengembangan sistem pendukung keputusan berbasis ML untuk diagnosis PGK. Penelitian lanjutan disarankan untuk menguji kombinasi algoritma *ensemble*, serta integrasi data klinis *real-time* untuk mendukung penerapan di lingkungan klinis yang sesungguhnya

Daftar Pustaka

[1] Z. Chen, Z. Zhang, R. Zhu, Y. Xiang, and P. B. Harrington, “Diagnosis of patients with chronic kidney disease by using two fuzzy classifiers,” *Chemom. Intell. Lab. Syst.*, vol. 153, pp. 140–145, 2016, doi: 10.1016/j.chemolab.2016.03.004.

[2] "Global, regional, and national burden of chronic kidney disease, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017.," *Lancet (London, England)*, vol. 395, no. 10225, pp. 709–733, Feb. 2020, doi: 10.1016/S0140-6736(20)30045-3.

[3] J. Norouzi, A. Yadollahpour, S. A. Mirbagheri, M. M. Mazdeh, and S. A. Hosseini, "Predicting Renal Failure Progression in Chronic Kidney Disease Using Integrated Intelligent Fuzzy Expert System," *Comput. Math. Methods Med.*, vol. 2016, 2016, doi: 10.1155/2016/6080814.

[4] M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms.," *J. Pathol. Inform.*, vol. 14, p. 100189, 2023, doi: 10.1016/j.jpi.2023.100189.

[5] M. A. Islam, M. Z. H. Majumder, and M. A. Hussein, "Chronic kidney disease prediction based on machine learning algorithms," *J. Pathol. Inform.*, vol. 14, 2023, doi: 10.1016/j.jpi.2023.100189.

[6] B. Klopper *et al.*, "Defining software architectures for big data enabled operator support systems," *IEEE Int. Conf. Ind. Informatics*, vol. 0, pp. 1288–1292, 2016, doi: 10.1109/INDIN.2016.7819366.

[7] G. Ramesh, T. V Rajini kanth, and D. Vasumathi, "A Comparative Study of Data Mining Tools and Techniques for Business Intelligence," in *Performance Management of Integrated Systems and its Applications in Software Engineering*, 2020, pp. 163–173. doi: 10.1007/978-981-13-8253-6_15.

[8] A. Tejawati, H. S. Pakpahan, and W. Susantini, "Drugs Diagnose Level using Simple Multi-Attribute Rating Technique (SMART)," *Proc. - 2nd East Indones. Conf. Comput. Inf. Technol. Internet Things Ind. EIConCIT 2018*, no. Double L, pp. 357–362, 2018, doi: 10.1109/EIConCIT.2018.8878564.

[9] C. Karima and W. Anggraeni, "Performance Analysis of the Ada-Boost Algorithm For Classification of Hypertension Risk With Clinical Imbalanced Dataset," *Procedia Comput. Sci.*, vol. 234, pp. 645–653, 2024, doi: https://doi.org/10.1016/j.procs.2024.03.050.

[10] M. Jasri, A. Wijaya, and R. Sunggara, "Penerapan Data Mining untuk Klasifikasi Penyakit Demam Berdarah Dengue (DBD) Dengan Metode Naïve Bayes (Studi Kasus Puskesmas Taman Krocok)," *SMARTICS J.*, vol. 8, no. 1, pp. 35–42, 2022, [Online]. Available: https://doi.org/10.21067/smartics.v8i1.7375

[11] M. F. Ijaz, G. Alfian, M. Syafrudin, and J. Rhee, "Hybrid Prediction Model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, Synthetic Minority Over Sampling Technique (SMOTE), and random forest," *Appl. Sci.*, vol. 8, no. 8, 2018, doi: 10.3390/app8081325.

[12] W. Thungrut and N. Wattanapongsakorn, "Diabetes classification with fuzzy genetic algorithm," *Adv. Intell. Syst. Comput.*, vol. 769, pp. 107–114, 2019, doi: 10.1007/978-3-319-93692-5_11.

[13] H. Sulistiani, A. Syarif, K. Muludi, and Warsito, "Performance evaluation of feature selections on some ML approaches for diagnosing the narcissistic personality disorder," *Bull. Electr. Eng. Informatics*, vol. 13, no. 2, pp. 1383–1391, 2024, doi: 10.11591/eei.v13i2.6717.

[14] X. He *et al.*, "Sample-efficient deep learning for COVID-19 diagnosis based on CT scans," *IEEE Trans. Med. Imaging*, vol. XX, no. Xx, 2020, doi: 10.1101/2020.04.13.20063941.

[15] M. Ohsaki, P. Wang, K. Matsuda, S. Katagiri, H. Watanabe, and A. Ralescu, "Confusion-matrix-based kernel logistic regression for imbalanced data classification," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 9, pp. 1806–1819, 2017, doi: 10.1109/TKDE.2017.2682249.

[16] I. Düntsch and G. Gediga, "Confusion Matrices and Rough Set Data Analysis," *J. Phys. Conf. Ser.*, vol. 1229, no. 1, 2019, doi: 10.1088/1742-6596/1229/1/012055.