

Perbandingan Akurasi Algoritma C4.5 Dan Naïve Bayes Terhadap Klasifikasi Kelulusan Berdasarkan Data Mahasiswa Universitas Lancang Kuning

Ahmad Zamsuri¹, Vebby², Dwi Muladi³

^{1,2,3} Program Studi Teknik Informatika Fakultas Ilmu Komputer
Universitas Lancang Kuning

^{1,2,3} Jl.Yos Sudarso KM. 8 Rumbai, Pekanbaru, Riau, Telp.0811 753 2015

email: ahmadzamsuri@unilak.ac.id¹, vebby@unilak.ac.id², dwimuladi49@gmail.com³

Abstrak

Dalam era teknologi informasi, penggunaan database untuk data besar semakin umum. Namun, data sering diabaikan dan dihapus, padahal bisa memberikan nilai. Data mining digunakan untuk menggali pengetahuan dari data besar. Penelitian ini membandingkan algoritma C4.5 dan Naïve Bayes untuk memprediksi kelulusan mahasiswa. Hasilnya, C4.5 memiliki akurasi 86,7%, sedangkan Naïve Bayes 76%. Ini menunjukkan C4.5 lebih baik dalam memprediksi kelulusan. Hasil ini relevan untuk meningkatkan kualitas pendidikan di Universitas Lancang Kuning.

Kata Kunci: Data Mining, Naïve Bayes, C4.5, Rapid Miner, Klasifikasi

Abstract

In the era of information and advanced technology, the use of databases for large-scale data has become widespread. However, data is often neglected and deleted, even though it could provide value. Data mining is used to extract knowledge from large datasets. This research compares the C4.5 and Naïve Bayes algorithms for predicting student graduation. The results show that C4.5 has an accuracy of 86.7%, while Naïve Bayes has 76%. This indicates that C4.5 is better at predicting graduation. These findings are relevant for improving the quality of education at Lancang Kuning University.

Keywords: Data Mining, Naïve Bayes, C4.5, Rapid Miner, Classification

1. PENDAHULUAN

Pada masa informasi dan era teknologi canggih seperti beberapa tahun terakhir ini dapat kita lihat bahwa sudah banyak sistem komputer yang dibangun dengan menggunakan desain database untuk data berskala besar. Apalagi didukung dengan tempat penyimpanan data yang sangat besar pula sehingga mendukung manusia untuk leluasa menyimpan banyaknya data. Semakin banyak data tersimpan maka semakin banyak pula penumpukan yang terjadi pada data sangat banyak pada tempat penyimpanan data ke dalam satu atau beberapa kelas yang sudah didefinisikan sebelumnya. Kondisi data yang bertumpuk terus menerus ini akan sia-sia jika tidak dimanfaatkan kembali untuk kebutuhan informasi dimana kebutuhan informasi dari tahun ke tahun terus

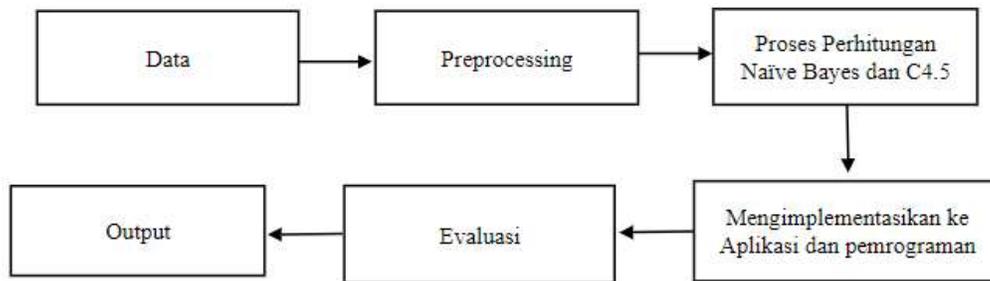
meningkat.(Saputra & Primadasa, 2018). Database yang tersimpan di media penyimpanan jarang sekali dimanfaatkan oleh sebagian besar penggunaanya dan bahkan dalam jangka waktu tertentu data-data tersebut dihapus karena dianggap sampah dan hanya memenuhi media penyimpanan saja. Anggapan tersebut tidak sepenuhnya benar, karena sesungguhnya database dalam ukuran yang besar dapat memberikan informasi yang dibutuhkan untuk berbagai kepentingan, baik untuk kepentingan bisnis dalam mengambil keputusan maupun untuk ilmu pengetahuan dan penelitian. (Mardi, 2017)

Untuk mengatasi supaya tidak terjadi penumpukan data berskala besar, digunakanlah suatu teknik penggalian informasi dari data yang sudah bertumpuk tersebut. Teknik yang disebut dengan *Data mining*. *Data mining* merupakan teknik yang sering digunakan untuk menggali informasi yang tersembunyi dalam data yang besar. Sehingga *Data mining* bisa menjadi solusi untuk mengatasi tumpukan data yang ada pada tempat penyimpanan,dengan menggunakan teknik data mining tersebut kita dapat menemukan informasi yang berupa pola, ciri, dan aturan atau dikenal sebagai istilah *knowledge*. Salah satu misi Universitas Lancang Kuning ialah Meningkatkan Kualitas Pendidikan untuk Menghasilkan Lulusan Kompetitif Berskala Nasional. Kualitas pendidikan disini terkait langsung dengan kemampuan mahasiswa yang merupakan aspek penting dalam menentukan keberhasilan penyelenggaraan program studi pada suatu perguruan tinggi. Salah satu indikator dari keberhasilan suatu perguruan tinggi dapat dilihat dari banyaknya tingkat kelulusan mahasiswa setiap tahunnya. Pada peneliti sebelumnya sudah melakukan penelitian untuk mengukur tingkat kelulusan mahasiswa Fakultas Ilmu Komputer Universitas Lancang Kuning melalui klasifikasi lulus tepat waktu atau lulus tidak tepat waktu menggunakan algoritma *Naïve Bayes Classifier* yang dilakukan oleh (N. Nasution et al., 2015) berdasarkan evaluasi kinerja akademik di tingkat 3 atau semester 6 dengan hasil dari penelitian dengan pembangunan model Naïve Bayes terbaik berada pada akurasi 76%, kemudian model tersebut diterapkan untuk prediksi kelulusan mahasiswa pada tingkat 3, dengan perolehan hasil 48 mahasiswa lulus tepat waktu dan 13 mahasiswa lulus tidak tepat waktu. Pada penelitian ini atribut yang digunakan adalah jenis kelamin; program studi; asal sekolah; daerah asal; tempat tanggal lahir; indeks prestasi semester 1 sampai dengan 5; indeks prestasi kumulatif; dan target kelulusan.

Dari beberapa penelitian di atas, dapat disimpulkan bahwa atribut yang signifikan dalam penentuan kelulusan secara umum adalah Indeks Prestasi Kumulatif (IPK) dengan menggunakan algoritma *Naïve Bayes*, *k-nearest neighbor*, *Decision Tree*, *neural network*. Sehingga pada penelitian ini akan dilakukan uji coba perbandingan algoritma klasifikasi *C4.5*, *Naïve Bayes classifier* dengan beberapa atribut terpilih seperti NIM, Nama, Fakultas, Atribut nilai dan Indeks Prestasi Kumulatif (IPK) yang dapat meningkatkan akurasi. Efek untuk lancang kuning adalah dapat melihat gambaran tingkat kelulusan mahasiswa di universitas lancang kuning.

2. METODE PENELITIAN

Penelitian ini dilakukan dengan beberapa tahap, berikut tahapan penelitian yang dilakukan seperti pada Gambar 1.



Gambar 1. Tahapan Penelitian

2.1. Data Mining

Data mining yaitu istilah yang digunakan untuk menemukan pengetahuan yang tersembunyi di dalam database. Data mining merupakan proses semi otomatis yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengidentifikasi informasi pengetahuan potensial berguna yang bermanfaat yang tersimpan dalam database besar. (Fatma Ayu Rahman et al., 2020)

Knowledge Discovery In Databases (KDD) Adalah proses untuk mencari dan mengidentifikasi pola (pattam) dalam data. yang mana pola yang ditemukan mudah dipahami. (Saputra & Primadasa, 2018)

Adapun tahapan *Knowledge Discovery In Databases* (KDD) adalah :

1. *selection*
Pada tahap ini dilakukan pemilihan data yang akan digunakan dalam proses *data mining*.
 2. *Pre-processing*
Dalam tahap ini kehandalan data ditingkatkan, termasuk pembersihan data seperti menangani data yang tidak lengkap, menghilangkan gangguan atau outlier. Tahapan *Pre-Processing Data* perlu dilakukan agar data terhindar dari duplikasi data, data yang tidak konsisten, memperbaiki kesalahan data, atau dilakukan penambahan data untuk menunjang sistem yang dibuat.
 3. *Transformation*
Untuk mengubah data dari bentuk asalnya menjadi data yang siap untuk ditambang. Karena hal tersebut dapat memudahkan dalam proses penggalian data untuk menemukan suatu pengetahuan baru.
 4. *Data mining*
Proses untuk mencari pola atau *knowledge* dari suatu kumpulan data yang selama ini tidak diketahui untuk mendapatkan informasi.
 5. *Evaluasi*
Hasil akhir dari informasi yang di temukan pada pola yang dihasilkan proses *data mining*.
-

2.2. Algoritma C4.5

Algoritma *C4.5* adalah salah satu metode untuk membuat decision tree berdasarkan training data yang telah disediakan. algoritma *C4.5* merupakan pengembangan dari *ID3*. Algoritma klasifikasi *decision tree* menentukan nilai target sampel baru berdasarkan berbagai nilai atribut dari data yang tersedia. Pada gambar 2 menunjukkan bahwa pohon keputusan memiliki 8 daun dan ukuran total pohon keputusan adalah 15 elemen. *Algoritma C4.5* merupakan algoritma yang digunakan untuk membentuk pohon keputusan. Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. (Budiman & Niqotaini, 2021)

$$3.1 \text{ Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * \text{Entropy}(S_i)$$

Keterangan :

- S : Himpunan Kasus
- A : Atribut
- N : Jumlah partisi atribut A
- |S_i| : Jumlah kasus pada partisi ke i
- |S| : Jumlah kasus dalam S

Sedangkan perhitungan nilai *Entropy* dapat dilihat pada persamaan 2:

$$3.1 \text{ entropy}(A) = - \sum_{i=1}^n \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

Keterangan :

- S : Himpunan kasus
- A : Atribut
- N : Jumlah partisi atribut A
- |S_i|: Jumlah kasus pada partisi ke i
- |S|: Jumlah kasus dalam S

2.3. Naïve Bayes

Naïve Bayes merupakan klasifikasi probabilitas sesuai dengan Teori *Bayes*. *Naive Bayes* menganggap bahwa efek dari nilai atribut pada kelas tertentu independen dari nilai-nilai atribut lainnya,

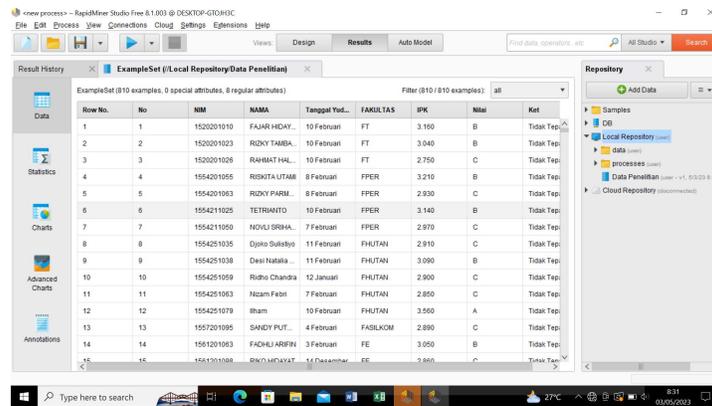
$$4.1 P(H|X) = \frac{P(X|H).P(H)}{P(X)}$$

Di mana:

- X : Data dengan kelas yang belum diketahui
- H : Hipotesa data kelas
- P(H|X) : Probabilitas hipotesa H berdasar kondisi
- X P(H): Probabilitas hipotesa H
- P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis
- H P(X): Probabilitas X.

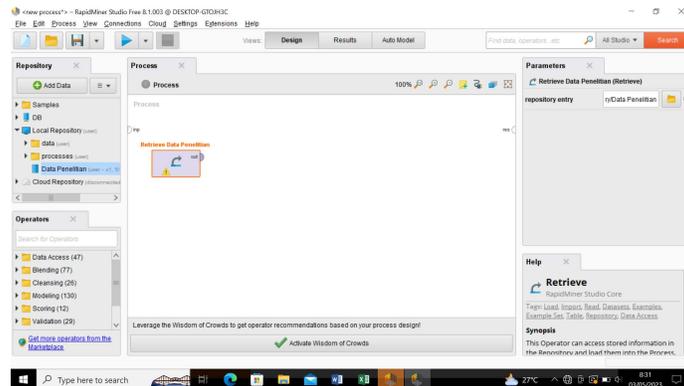
3. HASIL DAN PEMBAHASAN

1. Import data kedalam rapidminer.
Sebelum kita malakuka perhitungan, kita imporkan data yang akan kita gunakan kedalam rapidminer. seperti gambar 5.2 yang ada dibawah ini.

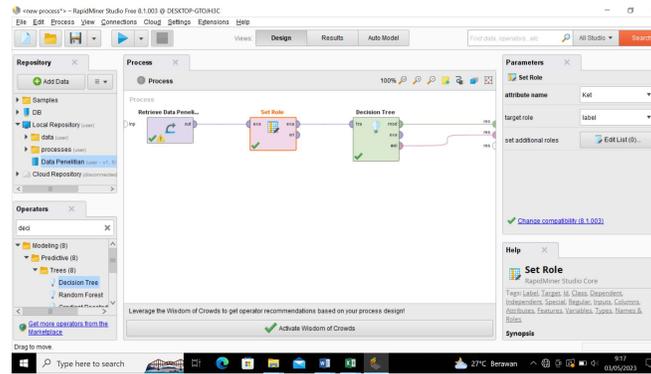


Gambar 5.1. 2. Import data kedalam rapidminer.

2. Jika data sudah selesai melakukan import data, masukan data kedalam proses perhitungan data. Lalu pilih operator yang ingin digunakan untuk melakukan pengujian, hubungkan data dengan operator yang akan digunakan sesuai dengan ketentuan yang ingin dilakukan.

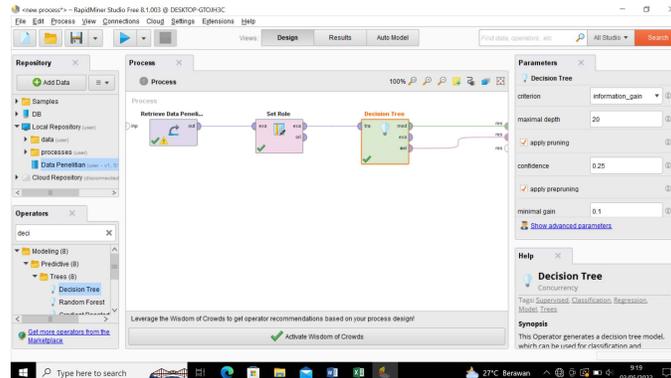


Gambar 5.1. 3. Memasukan data yang kedalam proses rapidminer.



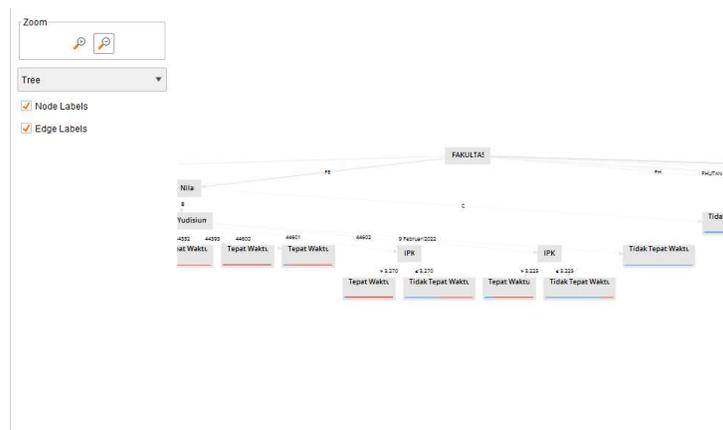
Gambar 5.1. 4. Menghubungkan data dengan algoritma

3. Pada tahap ini kita melakukan seting terhadap algoritma C4.5 nya untuk menentukan atribut mana yang akan di cari perhitungannya serta target role. Dan Seting pada algoritma untuk mencari perhitungan gain yang kita butuhkan.



Gambar 5.1. 6. Seting Algoritma.

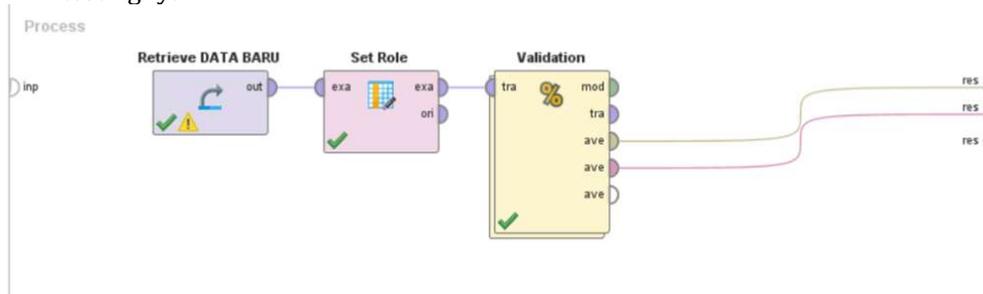
4. Pohon keputusan yang dihasilkan dari rapidminer.



Gambar 5.1. 7. Pohon Keputusan

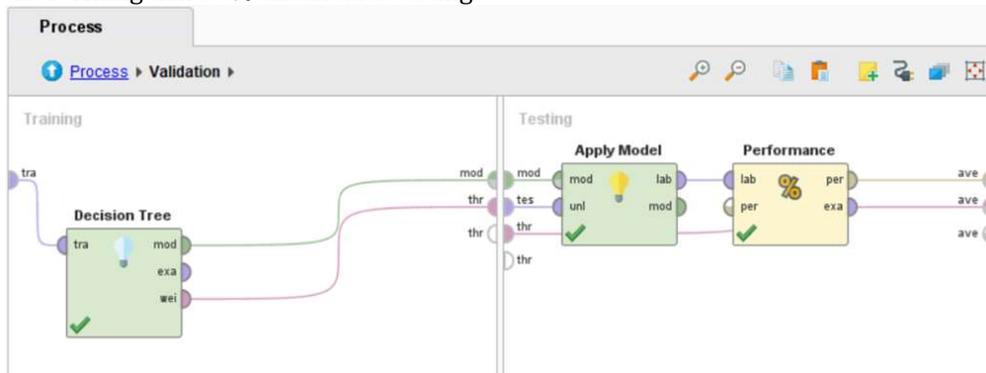
5. Proses Cross Validasi

Pada tahap ini seting pada item validation untuk mengatur algoritma yang ingin kita uji dan berapa performance dari data yang kita gunakan untuk data trining dan data testingnya.



Gambar 5.1. 8. Proses cross validation.

6. Proses di dalam validasi dan seting pada performance berapa persentase data trining dan data testing yang akan di uji. Disini 0.7 split rasio, itu sama dengan 70% data trining dan 30% untuk data testing.



Gambar 5.1. 9. Proses Validation Data

Pengujian pertama dengan performan 40% : 60%, 40% untuk data training dan 60% untuk data testing dari data yang digunakan.

Table View Plot View

accuracy: 100.00%

	true TEPAT WAKTU	true TIDAK TEPAT WAKTU	class precision
pred. TEPAT WAKTU	141	0	100.00%
pred. TIDAK TEPAT WAKTU	0	11	100.00%
class recall	100.00%	100.00%	

Gambar 5.1. 10. Hasil perhitungan Algoritma C4.5

4. KESIMPULAN

Dari hasil pada penjelasan diatas bisa disimpulkan bahwa penggunaan Algoritma C4.5, *Naïve Bayes*, dalam perbandingan akurasi untuk Algoritma C4.5 memiliki nilai 86,7%. Dan untuk Algoritma *Naïve Bayes* memiliki tingkat akurasi sebesar 76%.

Berdasarkan hasil dari pengujian kedua algoritma tersebut, maka dapat disimpulkan bahwa Algoritma C4.5 lebih unggul dibandingkan dengan Algoritma *Naïve Bayes*, dengan tingkat akurasi Algoritma C4.5 sebesar 86,7%

UCAPAN TERIMAKASIH

Saya ingin mengucapkan terima kasih yang sebesar-besarnya atas bantuan, dukungan dan bimbingan dari dosen pembimbing maupun prodi selama proses jurnal kami. Tanpa beliau, pencapaian ini tidak akan mungkin terwujud.

DAFTAR PUSTAKA

- [1] Anam, C., ENERGY, H. S., & 2018, undefined. (2018). Perbandingan Kinerja Algoritma C4. 5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa. *Ejournal.Upm.Ac.Id*, 8(1), 2088-4591.
 - [2] Anisa, A., & Mesran, M. (2018). Analisa Pola Pekerjaan Lulusan Stmik Budi Darma Menerapkan Metode C4.5. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, 2(1), 446-449.
 - [3] Annur, H. (2018). Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes. *ILKOM Jurnal Ilmiah*, 10(2), 160-165.
 - [4] Budiman, & Niqotaini, Z. (2021). Perbandingan algoritma klasifikasi data mining untuk Abstrak. *15(3)*, 1-5.
 - [5] C, P. A., & Covid-, U. K. K. (2022). Penerapan algoritme c4.5 untuk klasifikasi kasus covid-19. III.
 - [6] Faid, M., Jasri, M., & Rahmawati, T. (2019). Perbandingan Kinerja Tool Data Mining Weka dan Rapidminer Dalam Algoritma Klasifikasi. *Teknika*, 8(1), 11-16. <https://doi.org/10.34148/teknika.v8i1.95>
 - [7] Fatma Ayu Rahman, A., Wartulas, S., Raya Pagojengan, J. K., & Brebes, P. (2020). Prediksi Kelulusan Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus Di Universitas Peradaban). *Ade Fatma Ayu Rahman IJIR*, 1(2), 70-77.
 - [8] Fay, D. L. (2018). Klasifikasi Dalam Data Mining. *Angewandte Chemie International Edition*, 6(11), 951-952., 6-16.
 - [9] Hendrian, S. (2018). Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan. *Faktor Exacta*, 11(3), 266-274.
 - [10] Hidayanti, A., Siregar, A. M., Lestari, S. A. P., & Cahyana, Y. C. (2021). Model Analisis Kasus Covid-19 Di Indonesia Menggunakan Algoritma Regresi Linier Dan Random Forest. *Petir*, 15(1), 91-101.
 - [11] Mardi, Y. (2017). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. *Edik Informatika*, 2(2), 213-219.
 - [12] Nasution, D. A., Khotimah, H. H., & Chamidah, N. (2019). Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN. *Computer Engineering, Science and System Journal*, 4(1), 78.
-

- [13] Nasution, N., Djahara, K., & Zamsuri, A. (2015). Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes (Studi Kasus : Unilak). Jurnal Fakultas Ilmu Komputer, 1(1), 1-11.
- [14] Novianto, Mohammad Akmaluddin, and Sirojul Munir. 2022. "Jurnal Informatika Terpadu." Jurnal Informatika Terpadu 8(2): 47-61.
- [15] Prihatmono, M. W., & Watratan, A. F. (2019). Implementasi Algoritma C4.5 Menggunakan Python Untuk Klasifikasi Kepuasan Konsumen. Progres, 49-55.
- [16] Retnoningsih, E., & Pramudita, R. (2020). Mengenal Machine Learning Dengan Teknik Supervised Dan Unsupervised Learning Menggunakan Python. Bina Insani Ict Journal, 7(2), 156.
- [17] Sa'adillah Maylawati, D., Mudyawati, M. N., Wahisyam, M. H., & Maulana, R. A. (2021). Perbandingan Algoritma Klasifikasi untuk Analisis Sentimen terhadap Komentar Film (Comparison of Classification Algorithms for Sentiment Analysis on Movie Comments). Gunung Djati Conference Series, 3(April), 71-77.
- [18] Saputra, A. Y., & Primadasa, Y. (2018). Penerapan Teknik Klasifikasi Untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. Techno.Com, 17(4), 395-403.
- [19] Sari, Y. R., Sudewa, A., Lestari, D. A., & Jaya, T. I. (2020). Penerapan Algoritma K-Means Untuk Clustering Data Kemiskinan Provinsi Banten Menggunakan Rapidminer. CESS (Journal of Computer Engineering, System and Science), 5(2), 192. <https://doi.org/10.24114/cess.v5i2.18519>.
- [20] Sibuea, M. L., & Safta, A. (2017). Pemetaan Siswa Berprestasi Menggunakan Metode K-Means Clustering. Jurteksi, 4(1), 85-92.
- [21] Sugianto, C. A. (2015). Analisis Komparasi Algoritma Klasifikasi Untuk Menangani Data Tidak Seimbang Pada Data Kebakaran Hutan. Techno.Com, 14(4), 336-342.
- Triyanto, S., Sunyoto, A., & Arief, M. R. (2021). Analisis Klasifikasi Bencana Banjir Berdasarkan Curah Hujan Menggunakan Algoritma Naïve Bayes. JOISIE (Journal Of Information Systems And Informatics Engineering), 5(2), 109-117.



Prosiding- SEMASTER: Seminar Nasional Teknologi Informasi & Ilmu Komputer is licensed under a [Creative Commons Attribution International \(CC BY-SA 4.0\)](https://creativecommons.org/licenses/by-sa/4.0/)
