Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.20241118074
Vol. 11, No. 1, Februari 2024, hlm. 215-224 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

PREPROCESSING DATA DAN KLASIFIKASI UNTUK PREDIKSI KINERJA AKADEMIK SISWA

Takhamo Gori[*1], Andi Sunyoto[2], Hanif Al Fatta[3]

[1, 2, 3]Universitas Amikom Yogyakarta
Email: [1]takhamo.gori@students.amikom.ac.id, [2]andi@amikom.ac.id, [3]hanif.a@amikom.ac.id
[*]Penulis Korespondensi

(Naskah masuk: 20 November 2023, diterima untuk diterbitkan: 12 Februari 2024)

Abstrak

Pendidikan merupakan aspek penting dalam kehidupan masyarakat dan memiliki peran yang sangat vital untuk menciptakan sumber daya manusia yang handal dan berkualitas dalam menghadapi berbagai tantangan pada era modernisasi. Namun, putus sekolah dan retensi siswa menjadi tantangan serius bagi perkembangan pendidikan saat ini. Salah satu faktor pemicu putus sekolah adalah kinerja akademik siswa yang rendah, mendorong perlunya tindakan pencegahan yang efektif untuk mengurangi tingkat kegagalan pendidikan. Penelitian ini bertujuan untuk memprediksi kinerja akademik siswa dengan mengintegrasikan metode *Correlation-Based Feature Selection* (CFS) dan Algoritma *Naïve Nayes* pada gabungan dataset pelajaran Matematika dan Bahasa Portugis dua sekolah menengah di Portugal. Proses preprocessing data melibatkan integrasi data, pelabelan data, transformasi data, dan pembersihan data diterapkan pada tahap awal penelitian. Hasil penelitian menunjukkan bahwa atribut signifikan yang mempengaruhi kinerja akademik siswa meliputi *G2, G1, Higher, Medu, Studytime, goout, Absences, dan Failures*. Melalui pemodelan algoritma *Naïve Bayes*, metode CFS terbukti meningkatkan nilai *accuracy*, *recall*, *precision*, dan *f1-score* dalam memprediksi kinerja akademik siswa. Sebelum CFS, model *Naïve Bayes* menunjukkan *accuracy* sebesar 89.27%, dengan *recall, precision,* dan *f1-score* masing-masing sebesar 89.27%, 89.86%, dan 89.47%. Setelah implementasi CFS, evaluasi model prediksi mengalami peningkatan signifikan menjadi 91.22%, 91.22%, 92.24%, dan 91.48%.

Kata kunci: Prediksi, Correlation-Based Feature Selection, Naïve Buyes, Kinerja Akademik Siswa

DATA PREPROCESSING AND CLASSIFICATION FOR PREDICTING STUDENT ACADEMIC PERFORMANCE

Abstract

Education is a crucial aspect of community life and plays a highly vital role in creating reliable and high-quality human resources to face various challenges in the era of modernization. However, school dropout and student retention pose serious challenges to the current development of education. One triggering factor for school dropout is low academic performance, necessitating effective preventive measures to reduce the education failure rate. This research aims to predict students' academic performance by integrating the Correlation-Based Feature Selection (CFS) method and the Naïve Bayes Algorithm on the combined dataset of Mathematics and Portuguese language subjects from two secondary schools in Portugal. The data preprocessing process involves data integration, data labeling, data transformation, and data cleaning applied at the early stage of the study. The research results indicate that significant attributes influencing students' academic performance include G2, G1, Higher, Medu, Studytime, goout, Absences, and Failures. Through the Naïve Bayes algorithm modeling, the CFS method has proven to enhance the accuracy, recall, precision, and f1-score values in predicting students' academic performance. Before CFS, the Naïve Bayes model showed an accuracy of 89.27%, with recall, precision, and f1-score at 89.27%, 89.86%, and 89.47%, respectively. After the implementation of CFS, the predictive model evaluation experienced a significant improvement, reaching 91.22%, 91.22%, 92.24%, and 91.48%.

Keywords: *Prediction, Correlation-Based Feature Selection, Naïve Buyes, Student Academic Performance*

1. PENDAHULUAN

Pendidikan merupakan aspek penting dalam kehidupan masyarakat dan memiliki peran yang sangat vital untuk menciptakan sumber daya manusia yang handal dan berkualitas dalam menghadapi berbagai tantangan pada era

modernisasi. Kualitas pendidikan menjadi landasan untuk menciptakan individu yang produktif dan berkontribusi positif dalam masyarakat. Namun, dalam konteks pendidikan saat ini, terdapat tantangan serius terkait dengan putus sekolah dan retensi siswa (Tjandra, Kusumawardani & Ferdiana, 2022). Fenomena ini dapat menghambat perkembangan pendidikan dan menghasilkan dampak negatif terhadap masa depan siswa.

Salah satu faktor yang dapat memicu putus sekolah adalah kinerja akademik siswa yang rendah. Siswa dengan kinerja akademik yang tidak memadai sering kali merasa kesulitan dalam proses belajar dan akhirnya mengambil keputusan untuk meninggalkan pendidikan formal (Gusnina, Wiharto, & Salamah, 2022.). Oleh karena itu, tindakan pencegahan yang efektif perlu dilakukan untuk mengurangi tingkat kegagalan dan putus sekolah pada siswa (Ismanto, Ghani, Saleh, Al, & Gunawan, 2022).

Prediksi kinerja akademik siswa menjadi langkah penting dalam upaya meningkatkan kualitas pendidikan (Saifudin, Ekawati, Yulianti, & Desyani, 2020). Kinerja akademik siswa merupakan ukuran kualitas mampu atau tidaknya siswa dalam pencapaian hasil belajar terhadap materi yang diterimanya (Fikron & Zulian, 2021). Dengan memahami dan memprediksi kinerja akademik siswa, pendidik dapat lebih efektif dalam membantu siswa yang menghadapi kesulitan dalam pembelajaran dan memberikan dukungan tambahan yang sesuai dengan kebutuhan masing-masing siswa (Masangu, Jadhav, & Ajoodha, 2020).

Kinerja akademik siswa memiliki dampak yang signifikan di luar dunia pendidikan. Siswa yang berhasil secara akademik cenderung memiliki peluang yang lebih baik untuk mendapatkan pekerjaan yang baik dalam dunia kerja (Adane, Deku, & Asare, 2023). Oleh karena itu, meningkatkan kinerja akademik siswa bukan hanya berkaitan dengan pendidikan, tetapi juga berpengaruh pada kesuksesan siswa dalam kehidupan setelah lulus.

Faktor internal dan eksternal merupakan aspek yang memiliki dampak signifikan terhadap kinerja akademik siswa. Faktor internal mencakup aspek seperti kesehatan, kemampuan intelektual, bakat dan minat, motivasi, serta kesiapan siswa dalam belajar. Sedangkan faktor eksternal mencakup faktor lingkungan, termasuk lingkungan keluarga, sekolah, dan masyarakat di sekitar siswa (Saputra, 2018). Pemahaman mendalam terhadap faktor-faktor ini memberikan landasan bagi lembaga pendidikan untuk merancang strategi yang lebih efektif dalam meningkatkan kinerja akademik siswa (Yusof, Hashim, Rahman, Yunus, & Fadzillah, 2022).

Dalam era teknologi informasi dan berkembangnya *Big Data*, penerapan *data mining* dan algoritma *machine learning* menjadi solusi yang efektif untuk menganalisis dan memprediksi faktor-faktor yang memengaruhi kinerja akademik siswa. Metode *data mining* telah diterapkan dalam berbagai aspek pengolahan data pendidikan, termasuk retensi siswa, prediksi putus sekolah, analisis data akademik, dan analisis perilaku siswa (Feng, Fan, & Chen, 2022). Data mining adalah proses menemukan relasi baru yang bermakna, pola dan kebiasaan dengan memilah sebagian besar data dengan menggunakan teknologi pengenalan pola seperti statistik dan matematika (Yuli Mardi, 2017). Menerapkan *data mining* pada kumpulan data sangat bermanfaat bagi lembaga pendidikan (Sudais, Safwan, Khalid, & Ahmed, 2022), ini dapat membantu untuk menggali informasi yang berharga tentang faktor-faktor yang mempengaruhi hasil akademik siswa.

Salah satu metode yang digunakan dalam Data Mining adalah *Correlation-Based Feature Selection* (CFS). CFS adalah algoritma filter sederhana yang memeringkat subset fitur menurut fungsi evaluasi heuristik berbasis korelasi. Atribut terpilih adalah atribut yang memiliki korelasi yang tinggi dengan atribut kelasnya namun tidak berkorelasi dengan fitur lainnya (Hall, 1999). Korelasi antar atribut yang tinggi dengan atribut lainnya menunjukkan bahwa atribut tersebut redundan. Atribut yang memiliki korelasi rendah dengan kelas merupakan atribut yang tidak relevan yang kemudian harus dieliminasi (Adi, Pristyanto, & Sunyoto, 2019). Dengan menggunakan CFS, fitur-fitur yang memiliki korelasi kuat dengan kinerja akademik siswa dapat dipilih, sehingga mengurangi dimensi data dan meningkatkan kualitas prediksi algoritma machine learning.

Algoritma *machine learning* merupakan cabang dari kecerdasan buatan yang memungkinkan komputer untuk belajar dari data dan membuat prediksi atau keputusan tanpa perlu pemrograman eksplisit (Purba, 2020). Dalam konteks prediksi kinerja akademik siswa, algoritma *machine learning* telah terbukti menjadi alat yang sangat bermanfaat. Dengan menganalisis data dan pola yang ada, algoritma *machine learning* dapat memprediksi kinerja akademik siswa berdasarkan faktor-faktor yang telah diidentifikasi.

Beberapa penelitian terdahulu telah dilakukan untuk prediksi kinerja akademik siswa, namun, terdapat sejumlah aspek yang harus diperhatikan, seperti *preprocessing* data yang optimal (Chandra & Kumar, 2022), pemilihan fitur yang relevan (Padilha, Lumacad, & Catrambone, 2021) dan algoritma *machine learning* yang paling sesuai untuk memodelkan kinerja akademik siswa.

Penelitian yang dilakukan oleh Musiliu (2020) membandingkan dua teknik pemilihan fitur yaitu *Information Gain Attribute Evaluator* dan *Correlation-Based Features Selection* (CFS) dalam mengidentifikasi faktor-faktor utama yang mempengaruhi akademik siswa sehingga dapat memberikan prediksi yang akurat. Hasil penelitian

disimpulkan bahwa CFS merupakan metode yang paling baik dalam menjaga atribut yang optimal dan memiliki dampak yang sangat besar pada hasil akurasi prediksi. Hasil penelitian juga menunjukkan bahwa tingkat mental dari hubungan, angkat tangan, mengunjungi sumber daya, partisipasi dalam diskusi kelompok, survei orangtua, dan kehadiran dalam kelas sangat berpengaruh terhadap kinerja akademik siswa.

Penelitian lainnya yang dilakukan oleh Adane, Deku, & Asare (2023), memprediksi kinerja akademik siswa menggunakan empat algoritma klasifikasi populer yaitu algoritma C4.5 *Decision tree* (CDT), *Multilayer Perceptron* (MLP), *Naïve Bayes* (NB) dan *Random Forest* (RF) dengan menerapkan metode *feature selection* menggunakan *Information Gain* (IG) dan mengevaluasi model prediksi menggunakan *confusion matrix* dengan rasio pelatihan dan pengujian 80:20, 70:30 dan *10-fold cross validation*. Hasil penelitian ini menemukan bahwa penerapan metode *feature selection* secara keseluruhan dapat meningkatkan performa model prediksi, sedangkan Algoritma NB secara signifikan menunjukan performa terbaik pada keseluruhan rasio pelatihan dan pengujian, dengan performa lebih unggul pada rasio 80:20.

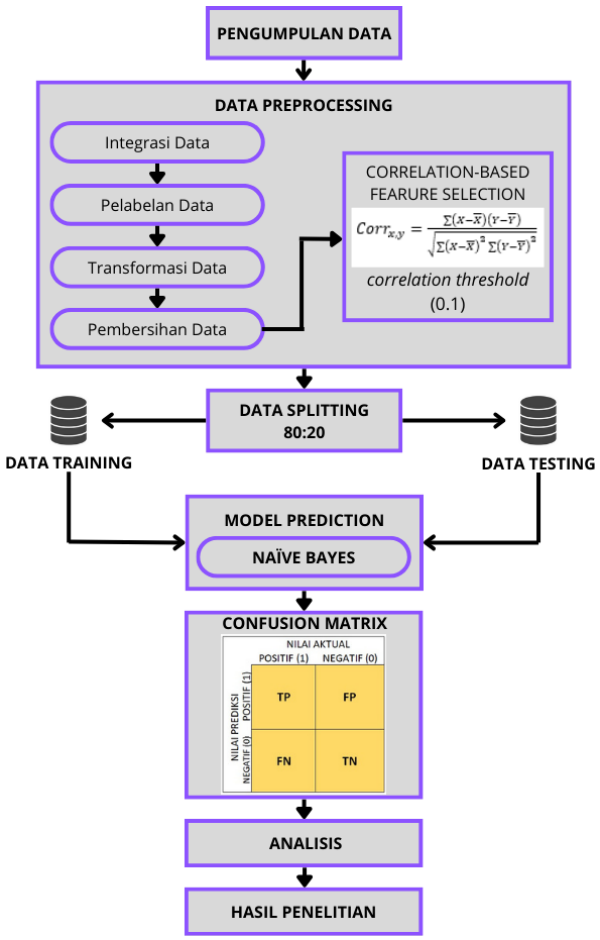
Penelitian selanjutnya yang dilakukan oleh Kumar, Chetan, Shamneesh, Nidhi, & Nazrul (2022), memprediksi performa akademik siswa menggunakan metode *feature selection (filters and wrappers*) dan algoritma machine learning yaitu *Decision Tree*, *JRip*, *Naive Bayes*, *Multilayer Perceptron*, dan *Random Forest*. Penelitian ini menggunakan dua dataset dari dua sekolah menengah Portugal pada pelajaran Matematika dan Bahasa Portugis. Kedua dataset diprediksi berdasarkan nilai akhir siswa dengan metode *Binary Grade System* (BGS) dan *Multi Grade System* (MGS). Hasil penelitian menunjukkan bahwa pemilihan fitur menggunakan metode *wrapper* secara konsisten mampu meningkatkan akurasi model prediksi yang digunakan, sedangkan prediksi dengan teknik BGS menghasilkan performa lebih akurat, baik pada dataset pelajaran Matematika maupun Bahasa Portugis.

Dalam penelitian ini, kami mengkombinasikan dua dataset siswa dari dua sekolah menengah di Portugal pada pelajaran Matematika dan Bahasa Portugis untuk memprediksi keberhasilan siswa pada akhir semester menggunakan metode CFS dan model prediksi *Naïve Bayes*. CFS digunakan untuk mengidentifikasi atribut-atribut yang paling relevan terhadap kinerja akademik siswa sedangkan algoritma *Naive Bayes* digunakan untuk membangun model prediksi kinerja akademik siswa berdasarkan atribut-atribut yang telah dipilih. Kami melakukan evaluasi dan validasi kinerja model prediksi menggunakan *Confusion Matrix* dengan rasio data pelatihan dan data pengujian 80:20.

Tujuan penelitian ini adalah untuk memprediksi kinerja akademik siswa berdasarkan nilai akhir siswa guna mendukung para pendidik dalam mengambil tindakan pencegahan terhadap siswa yang berisiko. Sejumlah proses *preprocessing* data diterapkan untuk meningkatkan tingkat akurasi model prediksi. Dengan mencapai tujuan-tujuan ini, penelitian ini diharapkan dapat memberikan panduan yang berguna bagi lembaga pendidikan, guru, dan pemangku kepentingan lainnya dalam upaya untuk meningkatkan kinerja akademik siswa dan efektivitas sistem pendidikan.

2. METODE PENELITIAN

Tahapan metode penelitian yang dilakukan dalam penelitian ini dimulai dari pengumpulan data, *preprocessing* data yang meliputi: integrasi data, pelabelan data, transformasi data, pembersihan data, dan pemilihan fitur menggunakan metode *Correlation-Based Feature Selection* (CFS). Setelah itu, membangun model prediksi menggunakan algoritma Naïve Bayes, mengevaluasi model prediksi dengan *Confusion Matrix*, analisis hasil, dan kesimpulan hasil. Tahapan penelitian ditunjukkan pada gambar 1.



Gambar 1. Tahapan Penelitian

2.1. Pengumpulan data

Data yang digunakan dalam penelitian ini adalah dua data sekunder pelajaran Bahasa Portugis dan Matematika dari dua sekolah menengah Portugal (Cortez & Silva, 2008). Jumlah data dari kedua dataset masing-masing 396 pada dataset pelajaran Matematika dan 649 pada dataset pelajaran Bahasa Portugis. Dataset ini dapat diakses pada wabsite Kaggle melalui link https://www.kaggle.com/datasets/impapan/student-performance-data-set. Kedua dataset terdiri dari 33 atribut mencakup demografi siswa, sosial, dan akademik. Informasi atribut dataset dapat dilihat pada Tabel 1.

Tabel 1. Informasi Atribut Dataset

Atribut	Deskripsi
school	Sekolah siswa (biner: Gabriel Pereira atau Mousinho da Silveira)
sex	Jenis kelamin siswa (biner: perempuan atau laki-laki)
age	Usia siswa (numerik: 15 - 22)
address	Jenis alamat rumah siswa (biner: perkotaan atau pedesaan)
famsize	Ukuran keluarga (biner: <= 3 atau > 3).
pstatus	Status hidup bersama orang tua (biner: tinggal bersama atau terpisah)
medu	Pendidikan ibu (numerik: 0 - 4)
fedu	Pendidikan ayah (numerik: 0 - 4)
mjob	Pekerjaan ibu (nominal)
Fjob	Pekerjaan ayah (nominal)
Reason	Alasan memilih sekolah ini (nominal: dekat dengan rumah, reputasi sekolah, preferensi kursus atau lainnya)
guardian	Wali siswa (nominal: ibu, ayah atau lainnya)
traveltime	Waktu perjalanan dari rumah ke sekolah (numerik: 1 – 4)
studytime	Jumlah belajar mingguan (numerik: 1 - 4)
failures	Jumlah kegagalan kelas di masa lalu (numerik: n jika 1 <= n < 3, lainnya 4)
schoolsup	Dukungan pendidikan tambahan (biner: ya atau tidak)
famsup	Dukungan pendidikan dari keluarga (biner: ya atau tidak)
paid	Kelas berbayar tambahan (biner: ya atau tidak)
activities	Kegiatan ekstrakurikuler (biner: ya atau tidak)
nursery	Bersekolah di taman kanak-kanak (biner: ya atau tidak)
higher	Keinginan mengambil pendidikan tinggi (biner: ya atau tidak)
internet	Akses internet di rumah (biner: ya atau tidak)
romantic	Menjalin hubungan romantis (biner: ya atau tidak)
famrel	Kualitas hubungan keluarga (numerik: 1 - 5)
freetime	Waktu luang sepulang sekolah (numerik: 1 - 5)
goout	Pergi bersama teman (numerik: 1 - 5)
dalc	Konsumsi alkohol di hari kerja (numerik: 1 - 5)
walc	Konsumsi alkohol akhir pekan (numerik: 1 - 5)
health	Status kesehatan (numerik: dari 1 - 5)
absensces	Jumlah ketidakhadiran di sekolah (numerik: 0 - 93)
G1	Nilai periode pertama (numerik: 0 - 20)
G2	Nilai periode kedua (numerik: 0 - 20)
G3	Nilai akhir (numerik: 0 - 20, variabel target)

2.2. Integrasi Data

Integrasi data merupakan proses manyatukan data dari berbagai sumber ke dalam satu database baru. Proses penggabungan ini dapat terjadi pada atribut yang sama atau melibatkan penambahan atribut, yang hasilnya adalah peningkatan informasi yang lebih komprehensif (Kharis & Zili, 2022). Pada tahap ini, dataset dari dua sekolah negara Portugal yang masing-masing dataset pelajaran Matematika dan Bahasa Portugis akan digabungkan menjadi satu dataset tunggal.

2.3. Pelabelan Data

Pelabelan data merupakan proses yang melibatkan pemberian label, kategori, atau kode tertentu pada setiap entitas atau elemen data untuk mengidentifikasi atau mengklasifikasikan data berdasarkan karakteristik atau atribut tertentu. Pada tahap ini, pelabelan dilakukan dengan membuat atribut baru pada dataset yang telah diintegrasikan yang diberi nama "*performance"*. Atribut "*performance"* digunakan untuk menggambarkan kategori hasil akademik siswa, yaitu "*Fail"* atau "*Pass"*. Proses pembuatan label ini didasarkan pada nilai akhir yang diperoleh siswa pada atribut G3. Jika nilai kurang dari 10, siswa akan dilabeli "*Fail"* yang mengindikasikan bahwa mereka gagal. Sebaliknya, jika nilai mencapai 10 sampai dengan 20, siswa akan dilabeli "*Pass"* yang menunjukkan bahwa mereka telah berhasil seperti yang ditunjukkan pada Tabel 2.

Tabel 2. Pelabelan Data

Atribut	Nilai	Label
Performance	≥10	*Pass*
	<10	*Fail*

2.4. Transformasi Data

Transformasi data merupakan suatu tahapan penting dalam mengubah representasi data sesuai kebutuhan dalam metode data mining (Azizah, Bachtiar, & Adinugroho, 2022). Setelah proses pelabelan data, data dengan tipe kategorik diubah menjadi data dengan tipe numerik menggunakan proses *Label Encoding*. *Label Encoding* adalah metode dalam pengolahan data yang mengubah data kategori atau data ordinal menjadi data numerik dengan memberikan label atau kode numerik pada setiap kategori atau tingkatan data tersebut. Setiap tabel yang berisi data dengan tipe string atau teks akan mengalami transformasi ke bentuk numerik. Dengan pendekatan ini, setiap entitas string atau teks akan diwakili oleh deretan angka yang mencerminkan variasi atau tingkat teks tersebut (Prasetyo, Mercifia, Averina, Sunyoto, & Budiarjo, 2022), sehingga data tersebut dapat memenuhi asumsi analisis yang sesuai untuk diproses dalam pembuatan model prediksi.

Sebelum melibatkan proses *Label Encoding* pada tahap transformasi data, atribut *G1, G2,* dan Absences akan dikategorikan seperti yang ditunjukkan dalam Tabel 3. Kategori pada atribut *G1* dan *G2* disesuaikan dengan level sistem penilaian Erasmus di negara Portugal.

Tabel 3. Kategori Atribut *G1*, *G2* & *Absences*

Atribut	Kategori	Transformasi
G1 & *G2*	≥16	A
	≥14	B
	≥12	C
	≥10	D
	<10	F
Absences	≤10	Rendah
	≤30	Sedang
	>30	Tinggi

2.5. Pembersihan Data

Proses pembersihan data, atau yang dikenal sebagai "*Data cleaning*", merupakan serangkaian langkah untuk membersihkan data, seperti mengidentifikasi dan mengisi nilai yang hilang (*missing values*) pada data, menghapus data yang tidak konsisten (*noisy/outlier*), mencari dan menghapus data yang duplikat, serta menangani data yang tidak lengkap (Putra & Putri, 2022). Selain itu, menghilangkan atribut yang tidak berkontribusi terhadap hasil analisis yang ingin dicapai adalah salah satu tahapan yang dilakukan dalam *Data Cleaning* (Agusriandi, Elihami, Syarif, & Samad, 2022). Pada tahap ini, atribut yang dianggap tidak dibutuhkan dalam penelitian ini akan dieliminasi, diantaranya: atribut *G3, school,* dan *age*.

2.6. Correlation-Based Feature Selection

Seleksi fitur merupakan tahap penting dalam analisis data, melibatkan pemilihan subset atribut yang paling relevan dan informatif dari sekumpulan atribut data yang tersedia untuk digunakan dalam pembuatan model. Penelitian ini mengusulkan pendekatan *Correlation-Based* untuk mengurangi dimensi data yang tinggi, mengoptimalkan waktu komputasi, dan memilih kombinasi fitur terbaik guna meningkatkan kinerja dalam proses pelatihan dan evaluasi (Alomari, Nuiaa, Alyasseri, Mohammed, Sani, Esa, & Musawi, 2023). Pengukuran korelasi merujuk pada perbandingan antara dua variabel atau fitur yang berbeda dalam suatu dataset. Jika dua fitur tidak menunjukkan hubungan yang signifikan, korelasinya mendekati nol; sebaliknya, jika terdapat hubungan, nilai korelasi mendekati ±1 (Zulfiqar, Huang, Lv, Sun, Dao, & Lin, 2022). Pengukuran ini penting karena membantu meningkatkan pemahaman tentang hubungan antara variabel yang dapat mengidentifikasi pola yang mungkin mempengaruhi hasil analisis. Korelasi antara atribut (x) dan target (y) dihitung menggunakan Persamaan 1.

$$Corr_{x,y} = \frac{\sum(Xi-\overline{X})(Yi-\overline{Y})}{\sqrt{\sum(Xi-\overline{X})^2 \sum(Yi-\overline{Y})^2}} \quad (1)$$

Dimana $Corr_{x,y}$ merupakan korelasi antara variabel X dan Y, $\overline{X}$ dan $\overline{Y}$ adalah nilai rata-rata (*mean*) dari masing - masing variabel X dan Y.

2.7. Splitting Data

Dalam konteks *machine learning*, *splitting data* umumnya digunakan untuk membagi dataset menjadi dua komponen utama yaitu data pelatihan (*Data Training)* dan data pengujian (*Data Testing*). *Data training* adalah kumpulan data yang digunakan untuk melatih model berdasarkan informasi yang benar sebelumnya, sementara *data testing* digunakan untuk menguji sejauh mana model tersebut berhasil mengklasifikasikan data dengan benar (Azmi, Hermawan, & Avianto, 2023). Pembagian jumlah data antara *data training* dan *data testing* merupakan faktor penting dalam menentukan tingkat akurasi, sehingga kesalahan dalam menentukan proporsi kedua jenis data ini akan berdampak pada nilai akurasi yang diperoleh (Musu, Ibrahim & Heriadi, 2021). Dalam penelitian ini, presentase data yang digunakan dalam pembagian data yaitu 80% *data training* dan 20% *data testing*.

2.8. Model Prediksi

Pada tahapan ini, algoritma *Naïve Bayes* digunakan untuk membangun model prediksi. *Naïve Bayes* merupakan sebuah algoritma klasifikasi yang memproyeksikan peluang di masa depan berdasarkan pengalaman masa lalu melalui penerapan metode probabilitas dan statistik yang dikenal sebagai Teorema Bayes. Algoritma ini bekerja dengan mengasumsikan bahwa semua atribut bersifat independen atau tidak saling tergantung berdasarkan nilai-nilai yang diberikan dalam variabel kelas (Yudana, Suyanto, & Nasiri, 2023). Umumnya *Naïve Bayes* menggunakan Teorema Bayes dengan rumus ditunjukkan pada Persamaan 2 (Kurniawan, Cahyono, Nofiyati, Maryanto, Fadli, & Indraswari, 2020).

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)} \quad (2)$$

Keterangan:

P(H|E): probabilitas kondisional akhir saat hipotesis H terjadi setelah diberikan bukti E.

P(E|H): probabilitas bahwa bukti E akan memengaruhi hipotesis H.

P(H): probabilitas awal bahwa hipotesis H terjadi tanpa memperhitungkan bukti apa pun.

P(E): probabilitas awal bahwa bukti E terjadi tanpa memperhitungkan hipotesis atau bukti lainnya.

Pada tipe data numerik/kontinu, algoritma Naïve Bayes dihitung dengan menggunakan distribusi *Gaussian* atau *Gaussian Density* sebagaimana ditunjukkan dalam Persamaan 3 (Fahrudy, & 'Uyun, 2022).

$$P(X_i = x_i|Y = y_i) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} exp^{-\frac{(x_i-\mu_{ij})^2}{2\sigma^2 ij}} \quad (3)$$

dimana P adalah probabilitas; Xi adalah atribut ke-i; xi adalah nilai atribut ke-i; Y adalah kelas yang dicari, yj adalah subkelas Y yang dicari; μ adalah rata-rata dari semua atribut (mean), σ adalah varians dari semua atribut (standar deviasi).

2.9. Evaluasi Model

Evaluasi model pada tahap penelitian ini dilakukan untuk mengukur seberapa baik model yang telah dibangun berdasarkan nilai *accuracy*, *precission*, *recall*, dan *F1-Score* menggunakan *Confusion Matrix*. *Confusion Matrix* merupakan sebuah matrik dua dimensi yang menggambarkan perbandingan antara hasil prediksi dengan kelas data sebenarnya, kemudian menghitung jumlah prediksi yang benar dan yang salah untuk setiap kategori kelas (Riska, Purnawansyah, Darwis, & Astuti, 2023). Terdapat empat variabel yang sangat berperan dalam *Confusion Matrix* seperti ditunjukkan pada Tabel 4.

Tabel 4. Confusion Matrix

		Nilai Aktual	
		Positif (1)	Negatif (0)
Nilai Prediksi	Positif (1)	*TP*	*FP*
	Negatif (0)	*FN*	*TN*

Keterangan:
True Positive (TP) adalah nilai prediksi benar dan nilai sebenarnya benar.
True Negative (TN) adalah nilai prediksi salah dan nilai sebenarnya salah.
False Positive (FP) adalah nilai prediksi benar dan nilai sebenarnya salah.
False Negative (FN) adalah nilai prediksi salah dan nilai sebenarnya benar.

Model prediksi yang dibangun akan diuji dengan perhitungan menggunakan persamaan 4-7 (Adane, Deku, & Asare, 2023).

$$Accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (4)$$

$$Recall = \frac{(TP)}{(TP+FN)} \quad (5)$$

$$Precission = \frac{TP}{TP+FP} \quad (6)$$

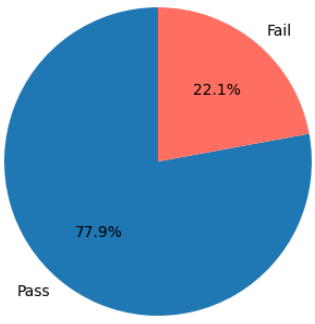
$$F1 - Score = \frac{2*Precission*Recall}{(Precission+Recall)} \quad (7)$$

3. HASIL DAN PEMBAHASAN

Analisis data dalam penelitian ini dilakukan menggunakan bahasa pemrograman *Python* dan *platform Google Colab*. Proses tersebut melibatkan berbagai tahap, seperti preprocessing data, pemilihan fitur, pembuatan model prediksi, hingga evaluasi model. Selama proses ini, penelitian memanfaatkan berbagai alat (*libraries* dan *frameworks*) yang disediakan oleh *Python* dan *Google Colab* untuk mendukung analisis dan pengolahan data.

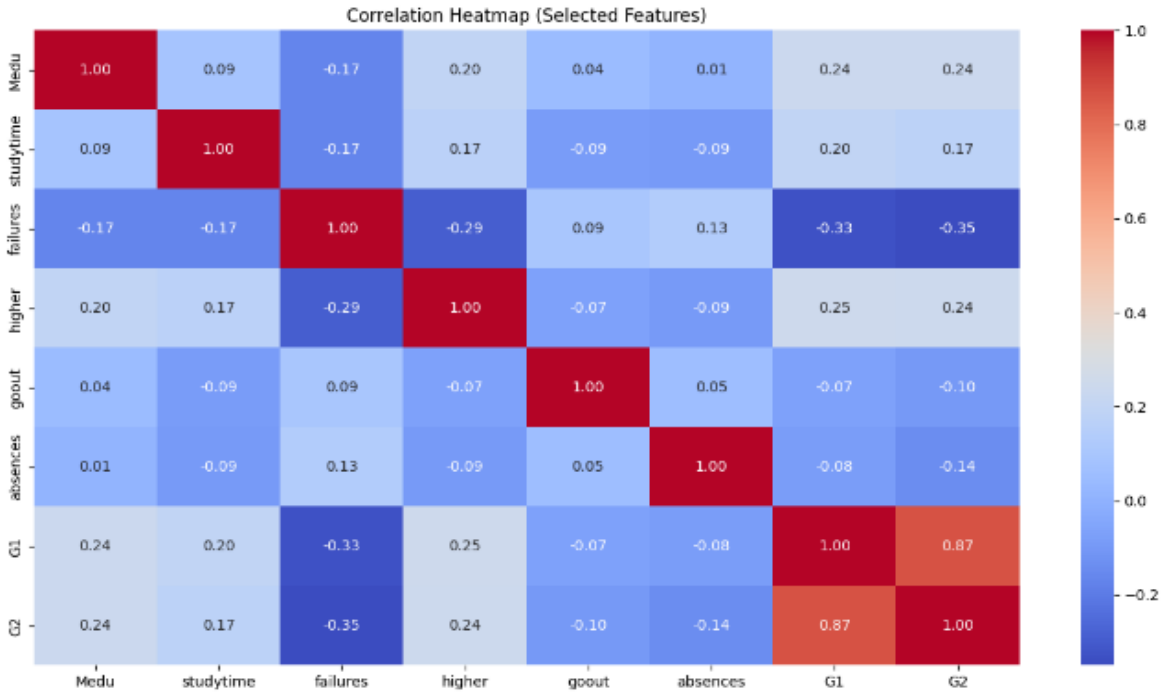
3.1. Preprocessing Data

Pada tahap *preprocessing* data, dataset pelajaran Matematika dan Bahasa Portugis di dua sekolah menengah Portugal digabungkan, masing-masing memiliki 395 entri data untuk pelajaran Matematika dan 649 entri data untuk pelajaran Bahasa Portugis, sehingga total entri mencapai 1044. Tahap selanjutnya melibatkan pelabelan data, di mana 814 entri diberi label 'Pass' dan 230 entri diberi label 'Fail'. Data kemudian ditransformasikan menggunakan proses *Label Encoding* untuk mengkonversi data teks menjadi tipe numerik. Setelah transformasi data, dilakukan pembersihan data dengan menghapus beberapa atribut yang tidak diperlukan, yaitu atribut *G3, school,* dan *age*. Setelah menghapus atribut tersebut, jumlah atribut yang sebelumnya mencapai 34 berkurang menjadi 31. Pada tahap pembersihan data, sebanyak 22 entri data duplikat dihapus, menghasilkan total 796 entri dengan label 1 ('*Pass*') dan 226 entri dengan label 0 ('*Fail*'). Presentase distribusi data dapat dilihat pada Gambar 2.



Gambar 2. Presentase Distribusi Data

Setelah tahap pembersihan data, dilakukan pemilihan fitur menggunakan metode *correlation-based* dengan nilai ambang batas (*correlation threshold*) sebesar 0.1. Hasilnya, dari total 31 atribut data, dipilih 8 atribut yang menunjukkan korelasi di atas ambang batas yang telah ditetapkan terhadap variabel target, sebagaimana ditampilkan dalam Gambar 3 dengan nilai korelasi masing-masing atribut terpilih disajikan dalam Tabel 5.



Gambar 3. Atribut Terpilih

Tabel 5. Nilai Korelasi Atribut Terpilih

Atribut	Nilai Korelasi
G2	0.59
G1	0.55
higher	0.22
Medu	0.11
studytime	0.10
goout	- 0.10
absences	- 0.12
failures	- 0.36

Dalam Tabel 5, atribut *G2* menunjukkan korelasi linear positif tertinggi sebesar 0.59, menandakan hubungan yang sangat kuat dengan variabel target. Atribut *G1* memiliki korelasi linear positif sebesar 0.55, disusul oleh atribut *higher* (0.22), *Medu* (0.11), dan *studytime* (0.10). Sementara itu, atribut *failures* menunjukkan korelasi linear negatif tertinggi sebesar -0.36, mengindikasikan hubungan linear negatif yang paling signifikan dengan variabel target. Atribut *absences* memiliki korelasi linear negatif sebesar -0.12, sedangkan *goout* memiliki korelasi linear negatif sebesar -0.10.

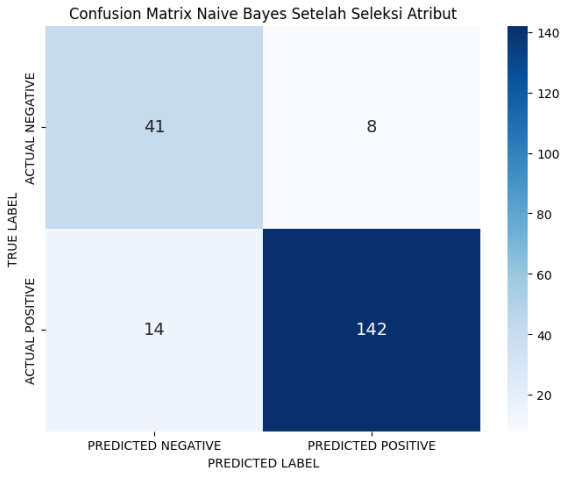
3.2. Prediksi dan Evaluasi Model

Sebelum membangun model prediksi, data dibagi (*Splitting Data*) menjadi 80% *data training* dan 20% *data testing*. Jumlah data masing-masing dapat dilihat pada Tabel 6.

Tabel 6. Jumlah Data Training dan Data Tesing

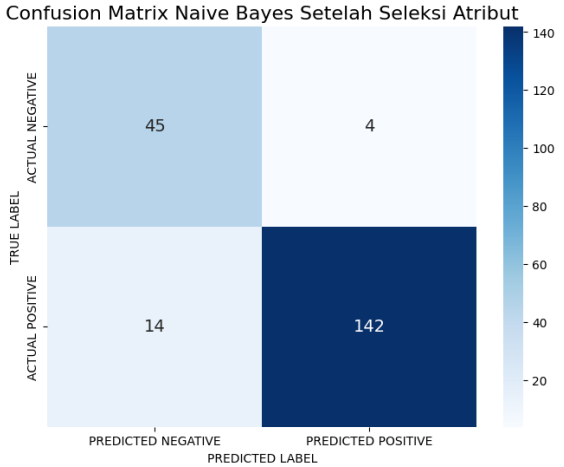
Data	Label 0 (*Fail*)	Label 1 (*Pass*)
Data Training	177	640
Data Testing	49	156

Tabel 6 menampilkan distribusi label pada data training, dengan 177 entri *Fail* dan 640 entri *Pass*, serta data testing dengan 49 entri *Fail* dan 156 entri *Pass*.



Gambar 4. Confusion Matrix Naïve Bayes Sebelum Seleksi Fitur

Gambar 4 menampilkan nilai *confusion matrix* hasil prediksi *Naïve Bayes* sebelum proses *feature selection*. Hasil tersebut menunjukkan bahwa terdapat 142 kasus *True Positive*, 41 kasus *True Negative*, 8 kasus *False Positive*, dan 14 kasus *False Negative*.



Gambar 5. Confusion Matrix Naïve Bayes Setelah Seleksi Fitur

Gambar 5 menampilkan nilai *confusion matrix* hasil prediksi *Naïve Bayes* setelah proses *feature selection*. Gambar ini menunjukkan bahwa terdapat 142 kasus *True Positive*, 45 kasus *True Negative*, 4 kasus *False Positive*, dan 14 kasus *False Negative*.

Tabel 7. Perbandingan Nilai Evaluasi Naïve Bayes Sebelum dan Sesudah Pemilihan Fitur

Pemilihan Fitur	**Naïve Bayes**			
	Accuracy	Recall	Precission	F1-Score
Sebelum	89.27%	89.27%	89.86%	89.47%
Setelah	91.22%	91.22%	92.24%	91.48%

Pada Tabel 7, nilai evaluasi model prediksi dibandingkan berdasarkan tingkat *accuracy*, *recall*, *precision*, dan *f1-score* sebelum dan setelah pemilihan fitur. Hasil penelitian menunjukkan bahwa terjadinya peningkatan yang cukup signifikan pada nilai evaluasi model setelah pemilihan fitur. *Accuracy* model meningkat dari 89.27% menjadi 91.22%, sementara nilai *recall*, *precision*, dan *f1-score* juga mengalami peningkatan masing-masing dari 89.27% menjadi 91.22%, dari 89.86% menjadi 92.24%, dan dari 89.47% menjadi 91.48%. Dengan demikian, dapat disimpulkan bahwa *Correlation-Based Feature Selection* (CFS) sangat efektif dalam meningkatkan performa model prediksi.

4. KESIMPULAN DAN SARAN

4.1. Kesimpulan

Berdasarkan hasil dan pembahasan penelitian ini, dapat disimpulkan bahwa penerapan metode *Correlation-Based Feature Selection* (CFS) berhasil mengidentifikasi atribut yang berpengaruh pada variabel target, yang menunjukkan dampaknya terhadap kinerja akademik siswa. Atribut yang signifikan yaitu *G2* (Nilai kelas periode kedua), *G1* (Nilai kelas periode pertama), *Higher* (Keinginan siswa melanjutkan pendidikan tinggi), *Medu* (Pendidikan ibu), *Studytime* (Waktu belajar per minggu), *Goout* (Intensitas keluar dengan teman), *Absences* (Jumlah absen), dan *Failures* (Jumlah gagal pada kelas sebelumnya). Masing-masing atribut memiliki korelasi yang sesuai dengan nilai ambang batas (*correlation threshold*) 0.1, dengan atribut *G2* menunjukkan korelasi linear positif tertinggi, sementara *failures* memiliki korelasi linear negatif tertinggi. Melalui pemodelan algoritma *Naïve Bayes*, terbukti bahwa metode CFS dapat meningkatkan nilai *accuracy, recall, precision,* dan *f1-score* dalam memprediksi kinerja akademik siswa. Sebelum pemilihan fitur dengan CFS, hasil pemodelan Naïve Bayes menunjukkan nilai sebesar 89.27%, 89.27%, 89.86%, dan 89.47% untuk *accuracy, recall, precision,* dan *f1-score*. Setelah pemilihan fitur dengan CFS, evaluasi model prediksi meningkat secara signifikan menjadi 91.22%, 91.22%, 92.24%, dan 91.48%.

4.2. Saran

Saran yang diperoleh dari hasil penelitian ini mencakup strategi penanganan ketidakseimbangan data, terutama saat terdapat perbedaan signifikan dalam jumlah sampel antara kelas mayoritas dan minoritas. Disarankan untuk mempertimbangkan penggunaan strategi oversampling atau undersampling atau metode lainnya guna menyeimbangkan proporsi sampel antar kelas. Selain itu, direkomendasikan menggunakan metode pemilihan fitur lainnya, seperti *Chi-square*, *ANOVA*, *Forward Selection*, dan metode lainnya, untuk menemukan metode yang paling efektif dalam meningkatkan hasil evaluasi model prediksi *Naïve Bayes*. Kemudian, disarankan untuk melakukan perbandingan antara beberapa model prediksi untuk menentukan model yang paling sesuai untuk prediksi kinerja akademik siswa.

DAFTAR PUSTAKA

ADANE, M. D., DEKU, J. K. dan ASARE, E. K., 2023. Performance Analysis of Machine Learning Algorithms in Prediction of Student Academic Performance. *Journal of Advances in Mathematics and Computer Science*, *38*(5), 74–86. doi:10.9734/jamcs/2023/v38i51762.

ADI, S., PRISTYANTO, Y. dan SUNYOTO, A., 2019. The best features selection method and relevance variable for web phishing classification. 2019 International Conference on Information and Communications Technology, ICOIACT 2019. https://doi.org/10.1109/ICOIACT46704.2019.8938566.

AGUSRIANDI, ELIHAMI, SYARIF, I. dan SAMAD, I. S., 2022. Model Analisis Aktivitas Tutor Dalam Learning Management System Berdasarkan Data Log Menggunakan K-Means dan Deteksi

Outlier. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 9(4), 709–716. doi: 10.25126/jtiik.202294764.

ALOMARI, E. S., NUIAA, R. R., ALYASSERI, Z. A., MOHAMMED, H. J., SANI, N. S., ESA, M. I. dan MUSAWI, B. A., 2023. Malware detection using Deep Learning and correlation-based feature selection. *Symmetry*, *15*(1), 123. doi:10.3390/sym15010123.

ANWAR, S., SEPTIAN, F. dan SEPTIANA, R. D., 2019. Klasifikasi Anomali intrusion detection system (IDS) Menggunakan algoritma naïve Bayes classifier Dan Correlation-based feature selection. Jurnal Teknologi Sistem Informasi dan Aplikasi, 2(4), 135. doi:10.32493/jtsi.v2i4.3453.

AZIZAH, R. A., BACHTIAR, F. A. dan ADINUGROHO, S., 2022. Klasifikasi Kinerja Akademik Siswa Menggunakan Neighbor Weighted K-Nearest Neighbor Dengan Seleksi Fitur Information Gain. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 9, 605–614. doi:10.25126/jtiik.202295751.

AZMI, B. N., HERMAWAN, A. dan AVIANTO, D., 2023. Analisis Pengaruh komposisi data training dan data testing Pada penggunaan PCA Dan Algoritma decision tree untuk KLASIFIKASI Penderita Penyakit liver. JTIM: Jurnal Teknologi Informasi dan Multimedia, 4(4), 281–290. doi:10.35746/jtim.v4i4.298.

CHANDRA S. K. dan KUMAR, K. S., 2022. Data Preprocessing and Visualizations Using Machine Learning for Student Placement Prediction. Proceedings of International Conference on Technological Advancements in Computational Sciences, ICTACS 2022. https://doi.org/10.1109/ICTACS56270.2022.9988247.

CORTEZ, P. dan SILVA, A., 2008. Using data mining to predict secondary school student performance. European Concurrent Engineering Conference 2008, Future Business Technology Conference, FUBUTEC 2008.

FAHRUDY, D. dan 'UYUN, S., 2022. Classification of Student Graduation by Naïve Bayes Method by Comparing between Random Oversampling and Feature Selections of Information Gain and Forward Selection. International Journal on Informatics Visualization, 6(4). https://doi.org/10.30630/joiv.6.4.982.

FENG, G., FAN, M. dan CHEN, Y., 2022. Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining. IEEE Access, 10. https://doi.org/10.1109/ACCESS.2022.3151652.

GUSNINA, M., WIHARTO dan SALAMAH, U., 2022. Student Performance Prediction in Sebelas Maret University Based on the Random Forest Algorithm. Ingenierie Des Systemes d'Information, 27(3). https://doi.org/10.18280/isi.270317.

HALL, M.A., 1999. Correlation-Based Feature Selection for Machine Learning. PhD Thesis, University of Waikato, Hamilton.

ISMANTO, E., GHANI, H. A., SALEH, N. I. M., AL AMIEN, J. dan GUNAWAN, R., 2022. Recent systematic review on student performance prediction using backpropagation algorithms. Telkomnika (Telecommunication Computing Electronics and Control), 20(3). https://doi.org/10.12928/TELKOMNIKA.v20i3.21963.

KHARIS, S.A. dan ZILI, A.H., 2022. Learning Analytics Dan Educational Data Mining Pada Data Pendidikan. Jurnal Riset Pembelajaran Matematika Sekolah, 6(1), pp. 12–20. doi:10.21009/jrpms.061.02.

KUMAR, M., CHETAN, S., SHAMNEESH, S., NIDHI dan NAZRUL, I., 2022. Analysis of Feature Selection and Data Mining Techniques to Predict Student Academic Performance. International Conference on Decision Aid Sciences and Applications (DASA). doi:10.1109/dasa54658.2022.9765236.

KURNIAWAN, Y. I., CAHYONO, T., NOFIYATI, MARYANTO, E., FADLI, A. dan INDRASWARI, N. R., 2020. Preprocessing Using Correlation Based Features Selection on Naive Bayes Classification. IOP Conference Series: Materials Science and Engineering, 982(1), doi:10.1088/1757-899x/982/1/012012.

KUSUMA, D. P., 2020, Machine Learning: Teori, Program, dan Studi Kasus, Deepublish, Yogyakarta.

MASANGU, L., JADHAV, A., dan AJOODHA, R., 2021. Predicting student academic performance using data mining techniques. Advances in Science, Technology and Engineering Systems, 6(1). https://doi.org/10.25046/aj060117.

MUSILIU, B., 2020. Comparison of Feature Selection Techniques for Predicting Student's Academic Performance. In International Journal of Research and Scientific Innovation (IJRSI): Vol. VII.

MUSU W., IBRAHIM A. dan HERIADI, 2021. Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4.5. in Seminar Sistem Informasi dan Teknologi Informasi (SISITI), 2021, pp. 186–195.

PADILHA, T.P.P., LUMACAD, G. dan CATRAMBONE, R., 2021. Predicting student performance using feature selection algorithms for Deep Learning Models', 2021 XVI Latin American Conference on Learning Technologies (LACLO) [Preprint]. doi:10.1109/laclo54177.2021.00009.

PRASETYO, V. R., MERCIFIA, M., AVERINA, A., SUNYOTO, L. dan BUDIARJO, 2022. Prediksi Rating Film Pada Website IMDB Menggunakan Metode Neural Network. NERO, 7. doi: http://dx.doi.org/10.21107/nero.v7i1.268.

PUTRA, M. Y. dan PUTRI, D. I., 2022. Pemanfaatan Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Klasifikasi Jurusan Siswa Kelas XI. *Tekno Kompak*, *16(2)*, 176–187.

RISKA, A., PURNAWANSYAH, DARWIS, H. dan ASTUTI, W., 2023. Studi Perbandingan Kombinasi GMI, HSV, KNN, dan CNN Pada Klasifikasi Daun Herbal. Indonesian Journal of Computer Science, 12(3). doi:10.33022/ijcs.v12i3.3210.

SAIFUDIN, A., EKAWATI, YULIANTI dan DESYANI, T., 2020. Forward Selection Technique to Choose the Best Features in Prediction of Student Academic Performance Based on Naïve Bayes. Journal of Physics: Conference Series, 1477(2). https://doi.org/10.1088/1742-6596/1477/3/032007.

SAPUTRA, H. D., ISMET, F. dan ANDRIZAL, A., 2018. Pengaruh Motivasi Terhadap Hasil Belajar Siswa SMK. INVOTEK: Jurnal Inovasi Vokasional Dan Teknologi, 18(1). https://doi.org/10.24036/invotek.v18i1.168.

SUDAIS M., SAFWAN M., KHALID A. M. dan AHMED S., 2022, Students' Academic Performance Prediction Model Using Machine Learning, PREPRINT (Version 1) available at Research Square, https://doi.org/10.21203/rs.3.rs-1296035/v1.

TJANDRA, E., KUSUMAWARDANI, S. S. dan FERDIANA, R., 2022. Student performance prediction in higher education: A comprehensive review. AIP Conference Proceedings, 2470. https://doi.org/10.1063/5.0080187.

YUDANA, F.R., SUYANTO, M. dan NASIRI, A., 2023. Model Klasifikasi Untuk Menentukan Kesiapan Kerja Mahasiswa Dan Kelulusan Tepat Waktu Dengan Metode Machine Learning. IJITECH: Indonesian Journal of Information Technology, 1(1), pp. 1–22.

YULI MARDI., 2019. Data Mining : Klasifikasi Menggunakan Algoritma C4.5. Jurnal Edik Informatika, 2(2).

YUSOF, R., HASHIM, N., RAHMAN, N. A., YUNUS, S. Y. M. dan FADZILLAH, N. A. A., 2022. Academic Performance Prediction Model Using Classification Algorithms: Exploring the Potential Factors. International Journal of Academic Research in Progressive Education and Development, 11(3). https://doi.org/10.6007/ijarped/v11-i3/14753.

ZHANG, Y., YUN, Y., AN, R., CUI, J., DAI, H. dan SHANG, X., 2021. Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis. Frontiers in Psychology, 12. doi:10.3389/fpsyg.2021.698490

ZULFIQAR, H., HUANG, Q.-L., LV, H., SUN, Z.-J., DAO, F. Y. dan LIN, H., 2022. Deep-4mCGP: A Deep Learning Approach to Predict 4MC Sites in Geobacter Pickeringii by Using Correlation-Based Feature Selection Technique. International Journal of Molecular Sciences, 23(3), 1251. doi:10.3390/ijms23031251.