Volume 4 No. 2 (July - December 2024), pp. 127-145 | ISSN 2962-2360 (online) | ISSN 2962-3669 (print)
Arcitech: Journal of Computer Science and Artificial Intelligence
https://journal.iaincurup.ac.id/index.php/arcitech/article/view/12048 | DOI : https://dx.doi.org/10.29240/arcitech.v4i2.12048

Penerapan Predictive Analytics untuk Analisis Faktor-faktor yang Mempengaruhi Performa Akademik Siswa

Yanuarini Nur Sukmaningtyas[1], Ronny Makhfuddin Akbar[2], Gita Rohma Utami Asyafiiyah[3]
[1,2,3] Program Studi Informatika, Universitas Islam Majapahit, Jawa Timur, Indonesia
Email : yanuarini.ft@unim.ac.id, ronnyma.ft@unim.ac.id, gitarohma7@gmail.com

Article Information

Article history

Received December 25, 2024
Revised December 29, 2024
Accepted December 29, 2024
Available December 30, 2024

Keywords

Predictive Analytics
Education
Academic Performance
Machine Learning

Corresponding Author:

Yanuarini Nur Sukmaningtyas,
Universitas Islam Majapahit,
Email : yanuarini.ft@unim.ac.id

Abstract

Education in Indonesia currently faces several challenges, particularly the inequality of educational facilities in rural areas, leading to lower academic achievement compared to urban students. This study differs from previous research that focused solely on machine learning with academic data. Using a data-driven predictive analytics approach, the research aims to analyze factors influencing student academic performance, such as study hours, sleep hours, previous scores, and extracurricular involvement. Several machine learning algorithms including Linear Regression, Support Vector Regression (SVR), Random Forest, K-Nearest Neighbors (KNN), and XGBoost were employed to build the prediction model. The results indicated a significant correlation of 0.92 between previous scores and academic performance. Among the five algorithms, the XGBoost model demonstrated superior performance compared to the others. This highlights the effectiveness of the XGBoost model in predicting factors that affect students' academic performance and its potential as a tool for educators to develop more effective learning strategies, ultimately aiming to enhance students' academic achievements significantly.

Keywords : *Predictive Analytics, Education, Academic Performance, Machine Learning*

Abstrak

Pendidikan di Indonesia saat ini masih menghadapi sejumlah masalah, salah satunya adalah ketidakmerataan fasilitas pendidikan khusunya di daerah pedesaaan yang mengakibatkan prestasi akademik di wilayah ini seringkali lebih rendah dibandingkan dengan siswa di perkotaan. Penelitian ini berbeda dari penelitian sebelumnya yang hanya menggunakan *machine learning* pada data akademik. Melalui pendekatan *predictive analytics* dan berbasis data, tujuan penelitian yaitu menganalisis faktor-faktor yang mempengaruhi performa akademik siswa, seperti jam belajar, jam tidur, skor sebelumnya, dan keterlibatan dalam kegiatan ekstrakurikuler. Beberapa algoritma *machine learning* seperti *Linear Regression, Support Vector Regression (SVR), Random Forest, K-Nearest Neighbors (KNN),* dan *XGBoost*, digunakan dalam membangun model prediksi. Hasil menunjukkan bahwa skor sebelumnya memiliki korelasi yang signifikan sebesar 0,92 terhadap performa akademik. Dari kelima algoritma yang digunakan model *XGBoost* menunjukkan performa terbaik dibanding dengan pemodelan lainnya. Hal ini menunjukkan bahwa model *XGBoost* efektif dalam memprediksi faktor-faktor yang mempengaruhi performa akademik siswa dan dapat digunakan sebagai alat yang membantu para tenaga pendidik dalam menyusun strategi pembelajaran yang lebih baik dan efektif, sehingga dapat meningkatkan prestasi akademik siswa secara signifikan.

Kata Kunci : *Predictive Analytics, Pendidikan, Performa Akademik, Machine Learning*

1. Pendahuluan

Pendidikan memainkan peran penting dalam kehidupan masyarakat untuk menghasilkan sumber daya manusia yang kompeten dan berkualitas tinggi yang dapat menghadapi berbagai tantangan yang muncul di era modernisasi (Gori et al., 2024). Saat ini, pendidikan di Indonesia masing menghadapi sejumlah masalah, seperti ketidakmerataan fasilitas pendidikan, kesenjangan antar daerah dan kualitas guru yang belum merata. Kesenjangan pendidikan antar daerah perkotaan dan pedesaan di Indonesia juga menjadi masalah mendesak. Keterbatasan fasilitas pendidikan yang kurang memadai di daerah pedesaan seringkali mengakibatkan prestasi akademik di wilayah ini lebih rendah dibandingkan dengan siswa di perkotaan (Khusaini, 2020). Data menunjukkan bahwa hanya sebagian kecil penduduk di pedesaaan yang berhasil menamatkan pendidikan menengah dibanding dengan mereka yang tinggal di perkotaan (Khoeriyah, 2024). Kondisi ini menunjukkan bahwa perlu adanya usaha lebih untuk memastikan pemerataan akses dan kualitas pendidikan di seluruh Indonesia.

Peningkatan dan pemerataan kualitas pendidikan memerlukan pendekatan yang terintegrasi dan berbasis data. Pendekatan ini berguna untuk mengidentifikasi berbagai faktor yang memengaruhi kinerja akademis siswa. Salah satu pendekatan yang saat ini semakin banyak diterapkan adalah pemanfaatan teknologi kecerdasan buatan seperti *machine learning* yang telah terbukti efektif dalam menganalisis data pendidikan.

Penelitian sebelumnya mengenai pemanfaatan *machine learning* untuk menganalisis faktor-faktor yang mempengaruhi performa akademik siswa telah menunjukkan bahwa analisis berbasis data dapat digunakan untuk memprediksi hasil prestasi akademik siswa. Namun, pada penelitian sebelumnya masih terdapat beberapa keterbatasan. Data yang digunakan umumnya berskala kecil dan belum sepenuhnya mencakup faktor akademik maupun non-akademik, seperti jam belajar, skor sebelumnya, keterlibatan dalam kegiatan ekstrakurikuler, jam tidur, serta latihan soal. Oleh karena itu, diperlukan pendekatan yang lebih holistik untuk meningkatkan akurasi dalam memprediksi performa akademik siswa. Selain itu, penggunaan berbagai algoritma *machine learning* juga sangat penting untuk menentukan metode prediksi yang paling akurat dan relevan dalam konteks pendidikan.

Pendidikan di Indonesia masih menghadapi tantangan signifikan, khususnya ketidakmerataan fasilitas pendidikan di daerah pedesaan, yang berdampak langsung pada performa akademik siswa. Penelitian sebelumnya telah menunjukkan bahwa faktor akademis dan non-akademis, seperti skor sebelumnya, jam belajar, serta motivasi, memiliki pengaruh signifikan terhadap hasil belajar. Namun, sebagian besar studi berfokus pada wilayah dengan fasilitas memadai, sehingga konteks daerah pedesaan kurang terwakili (Ibrahim et al., 2024). Selain itu, algoritma machine learning terbukti efektif dalam memprediksi hasil akademik, tetapi pemilihan algoritma yang optimal untuk menangkap hubungan non-linear antar faktor belum banyak dieksplorasi

(Beckham et al., 2022). Penelitian lain menyoroti bahwa faktor seperti durasi studi dan performa sebelumnya di pendidikan menengah menjadi prediktor kuat, tetapi sering mengabaikan variabel kompleks seperti keterlibatan ekstrakurikuler atau kondisi sosial ekonomi (Al-Alawi et al., 2023).

Penelitian ini bertujuan untuk mengembangkan model prediktif berbasis machine learning yang tidak hanya mengidentifikasi faktor-faktor utama, tetapi juga membandingkan berbagai algoritma, seperti *Linear Regression*, *Support Vector Regression*, *Random Forest*, KNN, dan *XGBoost*, untuk menentukan algoritma dengan akurasi terbaik dalam memprediksi hasil belajar siswa berdasarkan data akademis dan non-akademis. Model ini diharapkan dapat memberikan manfaat praktis bagi para guru dan tenaga pendidik dalam memahami kebutuhan siswa secara lebih mendalam, merancang strategi pembelajaran yang lebih efektif dan merekomendasikan intervensi yang lebih tepat.

2. Kajian Terdahulu

Pada penelitian yang dilakukan oleh (Chitti et al., 2020) menunjukkan bahwa analisis berbasis data dapat mengidentifikasi faktor-faktor yang mempengaruhi kinerja akademis siswa seperti jam belajar, skor sebelumnya, dan keterlibatan dalam kegiatan ekstrakulikuler. Namun, penelitian tersebut terbatas pada konteks variabel spesifik yang dianalisis secara linear tanpa mengeksplorasi hubungan non-linear antar faktor, yang dapat memberikan wawasan lebih mendalam.

Pada penelitian yang dilakukan oleh (Afandi et al., 2024) menunjukkan bahwa analisis berbasis data dapat mengidentifikasi faktor-faktor yang mempengaruhi kinerja akademis siswa seperti usia, pendidikan orang tua, kota asal dan kesulitan dalam belajar. Meskipun hasil penelitian ini memberikan wawasan penting, terdapat beberapa keterbatasan yang perlu diperhatikan, terutama pada penggunaan dataset yang relatif kecil, yaitu hanya melibatkan 22 responden. Penggunaan dataset yang terbatas ini menyebabkan akurasi dari algoritma machine learning yang digunakan menjadi seragam dan tidak menunjukkan perbedaan yang signifikan, sehingga membatasi generalisasi hasil penelitian pada populasi yang lebih luas.

Penelitian serupa yang dilakukan oleh (Herman & Yefta Christian, 2022) menggunakan algoritma *Distributed Random Forest, Naïve Bayes, Generalized Linear Model*, dan *Gradient Boosting Machine* untuk menganalisis faktor-faktor yang memengaruhi performa akademik mahasiswa, juga menunjukkan bahwa faktor seperti nilai ujian akhir (UAS), nilai ujian tengah semester (UTS), jumlah tugas dan kehadiran memiliki pengaruh yang signifikan terhadap hasil akademik siswa. Dengan akurasi model mencapai 99,83% menggunakan *Distributed Random Forest*, penelitian ini membuktikan bahwa algortima *machine learning* efektif dalam memprediksi performa akademik siswa. Pada penelitian tersebut fokusnya terbatas pada data mahasiswa di perguruan tinggi.

Perbedaan dari penelitian – penelitian sebelumnya adalah penelitian ini memberikan kontribusi orisinal dengan memperluas cakupan analisis terhadap kombinasi variabel akademis dan non-akademis, seperti jam tidur, keterlibatan ekstrakurikuler, dan skor sebelumnya, serta membandingkan lima algoritma *machine learning*, yakni *Linear Regression*, SVR, *Random Forest*, KNN, dan *XGBoost*. Persamaan dengan penelitian sebelumnya terletak pada penggunaan machine learning pada data akademik siswa. Kontribusi ini memberikan alat yang relevan bagi tenaga pendidik untuk merancang strategi pembelajaran yang lebih efektif, berbasis pada analisis data yang mencakup berbagai aspek penting dalam lingkungan pendidikan.

Predictive Analytics

Analisis prediktif mengacu pada pendekatan yang digunakan dalam memproses data pendidikan yang besar dan komplek guna mengidentifikasi atribut penting yang dianggap memiliki pengaruh pada performa akademik siswa (Gori et al., 2024). Analisis ini melibatkan penerapan algoritma machine learning untuk memodelkan hubungan antara data akademik seperti nilai, jumlah kehadiran dan aktivitas kursus dengan keberhasilan akademik siswa (Alyahyan & Düştegör, 2020) .

Linear Regression

Regresi linear merupakan salah satu pemodelan yang digunakan untuk mengetahui hubungan antara dua variabel atau lebih. Variabel adalah faktor atau aspek yang dapat diukur, diamati, atau dimanipulasi. Pada regresi linear, variabel tersebut terbagi menjadi dua jenis yaitu variabel independen (X) yang dianalogikan sebagai sebab atau pemberi pengaruh dan variabel dependen (Y) atau variabel terpengaruh yang dianalogikan sebagai akibat (Sinaga et al., 2022).

SVR (*Support Vector Regression*)

Support Vector Regression (SVR) merupakan pengembangan dari *Support Vector Machines* (SVM) yang dapat digunakan untuk memecahkan masalah regresi dengan keluaran berupa bilangan rill (Isnaeni, Sudarmin, 2022). Metode ini banyak digunakan dalam penelitian karena mampu menangani data yang komples dan meminimalkan nilai error. Metode ini juga mampu menyelesaikan masalah overfitting, sehingga output dihasilkan memiliki nilai akurasi yang baik (Isnaeni, Sudarmin, 2022).

Random Forest

Random Forest adalah algoritma *machine learning* yang terdiri dari sekumpukan *decision tree* yang digunakan untuk mengkategorikan data ke dalam kelas tertentu. Untuk menentukan hasil akhir, pohon keputusan dibuat dengan menentukan node akar dengan beberapa node daun (Alita & Rahman, 2020). Metode ini banyak digunakan karena

mampu meningkatkan akurasi model dan mencegah masalah overfitting (Suci Amaliah et al., 2022).

K-Nearest Neighbors (KNN)

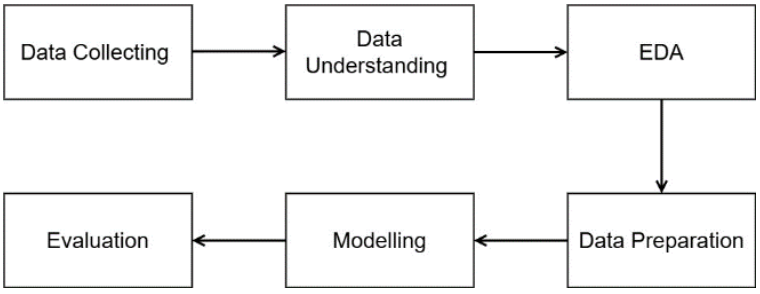
KNN (*K-Nearest Neighbors*) adalah metode klasifikasi yang didasarkan pada nilai kedekatan jarak antara objek dengan data latih atau data uji. Metode ini bekerja dengan mengklasifikasikan data baru berdasarkan jarak ke sejumlah data terdekat atau tetangga terdekatnya. Meskipun sederhana, KNN mampu memberikan performa dan akurasi yang tinggi, menjadikannya algoritma yang mudah digunakan namun tetap efektif. Pendekatan KNN mirip dengan metode penggolongan, di mana data baru dikelompokkan ke dalam kelas tertentu berdasarkan jarak terdekat dengan data yang sudah ada (Hendriyanto & Betha Nurina Sari, 2022).

XGBoost

eXtreme Gradient Boosting (XGBoost) adalah algoritma *machine learning* berbasis metode peningkatan (*boosting)* yang menggunakan prinsip *gradient boosting*. Dengan tujuan meningkatkan akurasi prediksi, algoritma ini berkonsentrasi pada contoh-contoh yang salah diklasifikasikan oleh model sebelumnya (Jan Melvin Ayu Soraya Dachi & Pardomuan Sitompul, 2023). XGBoost dirancang untuk mengoptimalkan kemampuan komputasi dan menghindari *overfitting*. Selain itu, model yang digunakan XGBoost untuk membangun struktur pohon regresi lebih teratur sehingga memungkinkan peningkatkan kinerja dan pengurangan kompleksitas pada model. Hal ini menjadikan XGBoost efektif dalam menyelesaikan masalah regresi, klasifikasi, dan ranking dengan sangat baik dan mampu menemukan solusi optimal untuk berbagai situasi (Herni Yulianti et al., 2022).

3. Metodologi Penelitian

Metode yang digunakan pada penelitian ini dilakukan melalui enam tahapan utama, yaitu *data collecting, data understanding, data preparation, exploratory data analysisis (EDA), modelling,* dan *evaluation*. Gambaran lengkap dari tahapan penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Tahapan Penelitian

a. Data *Collecting*

Pada tahap data *collecting*, proses pengumpulan data dilakukan dengan mengakses berbagai sumber yang relevan untuk memperoleh data yang dibutuhkan dalam penelitian. Data yang digunakan dalam penelitian ini berasal dari dataset publik berjudul *Student Performance* yang diunduh melalui *website Kaggle*. Dataset ini menyajikan informasi yang relevan untuk menganalisis faktor-faktor yang mempengaruhi performa akademik siswa.

b. Data *Undertanding*

Pada tahap *data understanding*, analisis dilakukan untuk memahami isi dan struktur dari dataset *Student Performance* yang telah diperoleh sebelumnya melalui website *kaggle*. Pemahaman terhadap dataset mencakup ekplorasi awal, seperti analisis deskriptif terhadap setiap fitur, serta memastikan kualitas data dengan melakukan pemeriksaan terhadap potensi permasalahan seperti, *missing values, duplicate value*, dan data *outliers*, yang dapat mempengaruhi hasil analisis. Langkah ini dilakukan untuk menjaga kualitas data dan mendapatkan pemahaman awal yang diperlukan untuk proses pengolahan data berikutnya.

c. *Exploratory Data Analysis* (EDA)

Tahapan *Exploratory Data Analysis* (EDA) dilakukan untuk mengeksplorasi data secara mendalam guna memahami pola, hubungan dan karakteristik antar variabel. Terdapat dua teknik EDA yang digunakan pada penelitan ini, yaitu *Univariate Analysis* dan *Multivariate Analysisis*. Analisis *univariate* dilakukan untuk melihat distribusi dan karakteristik dari masing-masing variabel, baik numerik maupun kategorik. Sedangkan analisis multivariate dilakukan untuk mengeksplorasi hubungan antara dua atau lebih variabel numerik maupun kategorik.

d. Data *Preparation*

Proses yang dilakukan pada tahapan ini hampir sama seperti pada tahap data *cleaning*, yaitu meliputi penghapusan data duplikat, melakukan fitur *encoding* untuk mengubah data kategorik menjadi bentuk numerik, dan normalisasi data untuk mentransformasikan data ke dalam skala yang sama. Selain itu, pada tahap ini juga dilakukan data *splitting* untuk membagi data menjadi dua subset, yaitu data latih (*training*) dan data uji (*testing*).

e. Modelling

Penelitian ini menggunakan lima algoritma *machine learning*, yaitu *Linear Regression*, *Support Vector Regression* (SVR), *Random Forest*, *K-Nearest Neighbors* (KNN), dan *XGBoost*, dengan setiap algoritma memiliki asumsi dasar dan kekuatan spesifik yang memengaruhi

hasil serta interpretasinya. *Linear Regression* mengasumsikan hubungan linear antara variabel independen dan dependen, yang memungkinkan interpretasi yang jelas terhadap hubungan antar variabel, tetapi dapat menghasilkan hasil yang kurang akurat jika terdapat pola non-linear dalam data. *Support Vector Regression* (SVR), di sisi lain, mengasumsikan bahwa data dapat dipisahkan dengan margin maksimal dalam ruang fitur, sehingga cocok untuk menangani data dengan pola non-linear yang kompleks, meskipun membutuhkan tuning parameter seperti kernel dan C untuk hasil optimal. *Random Forest* tidak memiliki asumsi khusus tentang distribusi data, menjadikannya ideal untuk dataset besar yang tidak terstruktur dengan baik; model ini menggunakan prinsip *ensemble* untuk meningkatkan akurasi dan mengurangi *overfitting*, namun kehilangan interpretasi langsung karena sifatnya yang berbasis banyak pohon keputusan. KNN mengasumsikan bahwa data dengan fitur serupa memiliki hasil yang serupa, membuatnya fleksibel tanpa memerlukan asumsi distribusi data, tetapi performanya sangat tergantung pada nilai parameter k dan skala fitur. *XGBoost* mengasumsikan bahwa boosting iteratif dapat secara progresif memperbaiki kesalahan model sebelumnya, sehingga efektif untuk menangani data besar dan kompleks, meskipun membutuhkan *tuning hyperparameter* yang lebih mendalam untuk performa maksimal.

Proses seleksi fitur dilakukan dengan mempertimbangkan variabel-variabel seperti skor sebelumnya, jam belajar, jam tidur, jumlah soal latihan yang dikerjakan siswa, dan keterlibatan dalam kegiatan ekstrakurikuler, berdasarkan literatur yang menunjukkan relevansinya terhadap performa akademik. Pemilihan algoritma ini didasarkan pada kebutuhan untuk mengeksplorasi pendekatan yang beragam, baik yang bersifat sederhana seperti *Linear Regression* dan KNN, maupun yang lebih canggih seperti *Random Forest* dan *XGBoost*, guna memastikan bahwa model terbaik dapat dipilih sesuai karakteristik dataset. Selain itu, faktor seperti kemampuan menangani data besar (*XGBoost*), fleksibilitas (KNN), dan kepraktisan implementasi (*Linear Regression*) menjadi pertimbangan utama. Pendekatan ini memungkinkan evaluasi kinerja model dari berbagai perspektif, sehingga memberikan hasil yang lebih komprehensif dan akurat dalam memprediksi performa akademik siswa.

f. Evaluation

Tahapan evaluasi dilakukan menggunakan dua metriks utama yaitu *Mean Absolute Error (MAE)* dan *Root Mean Squared Error (RMSE). MAE* digunakan untuk mengukur rata-rata kesalahan absolut antara nilai aktual dan prediksi, serta memberikan gambaran tentang seberapa besar kesalahan prediksi model dengan nilai sebenarnya (Suryanto, 2019). Rumus untuk menghitung *MAE* dapat dilihat pada persamaan 1.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad (1)$$

Keterangan :
n = jumlah sampel dalam data
y_i = nilai aktual
$\hat{y}_i$ = nilai prediksi

Sedangkan *RMSE* digunakan untuk memberikan bobot yang lebih besar pada kesalahan yang signifikan, menjadikannya lebih sensitif terhadap adanya outlier (Sihombing et al., 2023). Metrik ini efektif untuk mengevaluasi model yang memiliki kesalahan prediksi cukup besar. Rumus untuk menghitung RMSE dapat dilihat pada persamaan 2.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (2)$$

Keterangan :
n = jumlah sampel dalam data
y_i = nilai aktual
$\hat{y}_i$ = nilai prediksi

Tujuan dari penggunaan kedua metrik ini adalah untuk memberikan gambaran kinerja model yang lebih jelas. Sementara MAE secara umum digunakan untuk mengukur kesalahan rata-rata, RMSE digunakan untuk memberikan penilaian yang lebih mendalam dengan mempertimbangkan kesalahan yang lebih besar, sehingga dapat menunjukkan seberapa baik model menangani data dengan variasi atau outlier yang lebih signifikan.

4. Hasil dan Pembahasan

Pada bagian ini, peneliti akan menguraikan analisis yang komprehensif mengenai penerapan predictive analytics untuk mengevaluasi variabel yang mempengaruhi performa akademik siswa. Fokus dari pembahasan ini terletak pada validasi data terkumpul dan pengujian efektivitas model-model prediktif yang telah diimplementasikan.

Melalui analisis ini, peneliti menyajikan bagaimana variabel tertentu mempengaruhi hasil pendidikan dan menjelaskan potensi aplikasi praktis dari hasil penelitian ini dalam konteks pendidikan. Hasil dan pembahasan ini diharapkan dapat memperdalam pemahaman mengenai faktor-faktor yang mempengaruhi performa akademik serta mengidentifikasi metodologi yang paling efektif untuk mendukung kebijakan pendidikan yang didukung oleh data yang kuat.

a. Data Collecting

Data yang digunakan dalam penelitian ini berasal dari dataset publik berjudul *Student Performance* yang diunduh melalui website *Kaggle*. Dataset ini terdiri dari 6 kolom dan 10.000 baris data yang mencakup sejumlah variabel penting, yaitu, *Hours Studied* (jumlah total jam belajar), *Previous Scores* (skor yang diperoleh pada tes sebelumnya), *Extracurricular Activities* (keterlibatan dalam kegiatan ekstrakurikuler, dengan kategori Ya/Tidak), *Sleep Hours* (rata-rata jumlah jam tidur siswa per hari), *Sample Question Papers Practiced* (jumlah soal latihan yang dikerjakan siswa), dan *Performance Index* (ukuran performa setiap siswa), seperti yang dapat dilihat pada Gambar 2.

	Hours Studied	Previous Scores	Extracurricular Activities	Sleep Hours	Sample Question Papers Practiced	Performance Index
0	7	99	Yes	9	1	91.0
1	4	82	No	4	2	65.0
2	8	51	Yes	7	2	45.0
3	5	52	Yes	5	2	36.0
4	7	75	No	8	5	66.0
...	...	...	...	...	...	...
9995	1	49	Yes	4	2	23.0
9996	7	64	Yes	8	5	58.0
9997	6	83	Yes	8	5	74.0
9998	9	97	Yes	7	0	95.0
9999	7	74	No	8	1	64.0

10000 rows × 6 columns

Gambar 2. Dataset *Student Performance*

b. Data Understanding

Pada Gambar 3, hasil analisis awal menunjukkan bahwa dataset yang digunakan telah memenuhi kualitas data yang baik, dengan tidak adanya *missing value*.

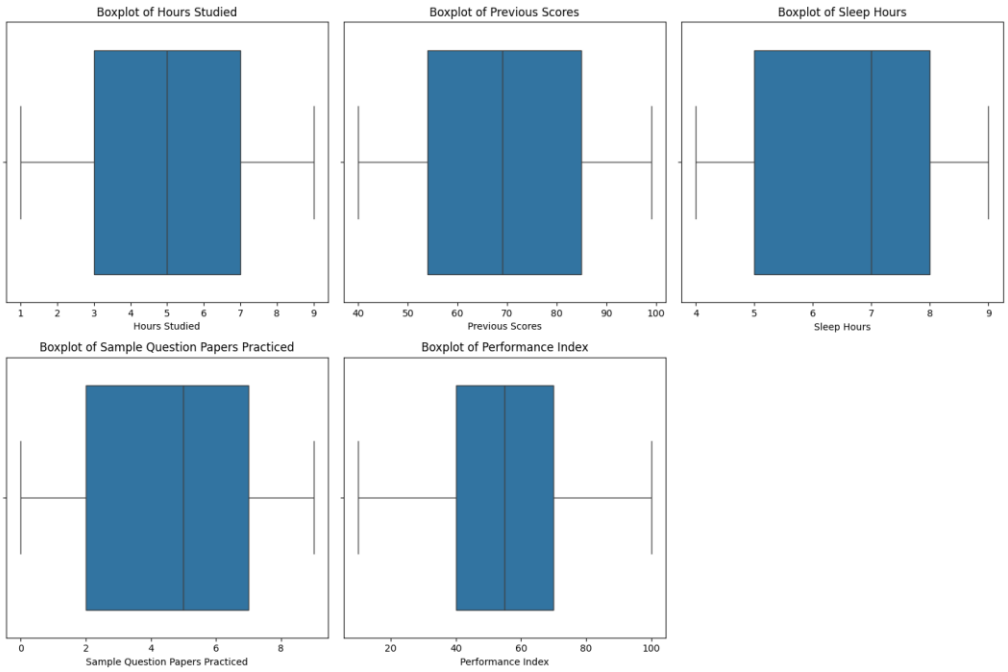
```
#Missing value
df.isnull().sum()
```

	0
Hours Studied	0
Previous Scores	0
Extracurricular Activities	0
Sleep Hours	0
Sample Question Papers Practiced	0
Performance Index	0

dtype: int64

Gambar 3. Cek *Missing Value*

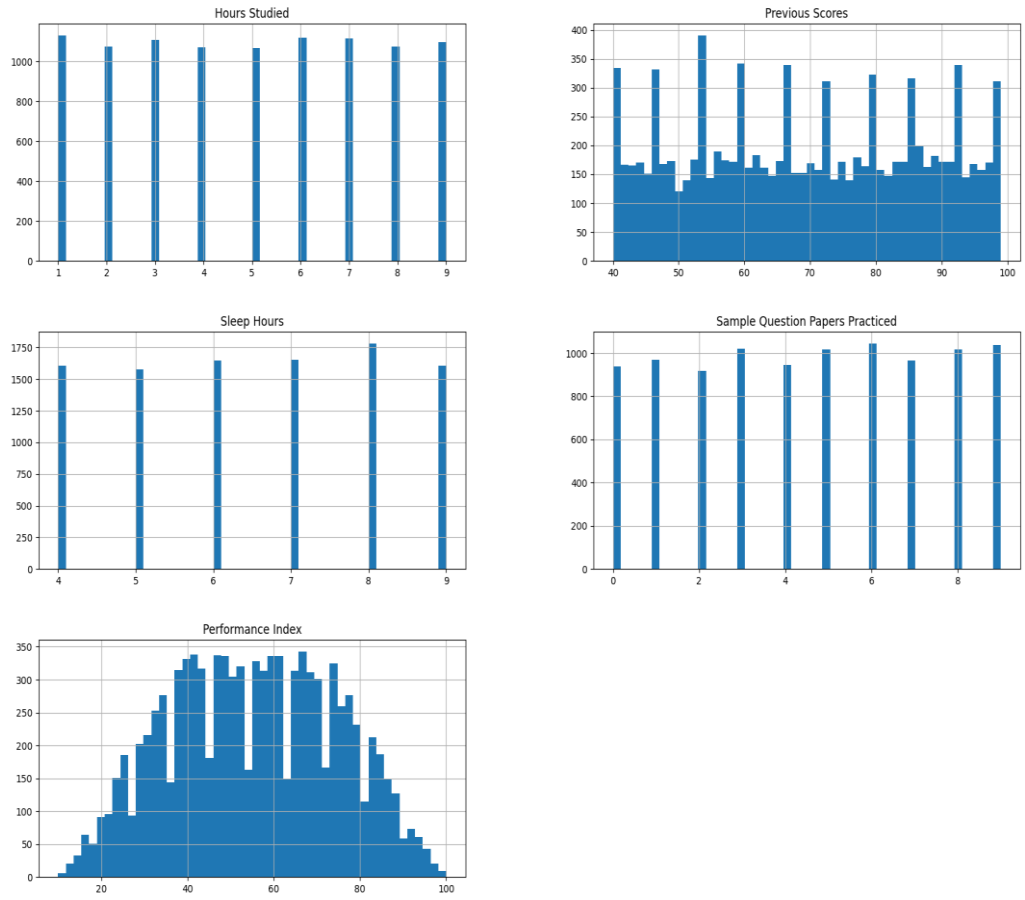
Selain itu, hasil pada Gambar 4 juga menunjukkan bahwa data yang digunakan tidak memiliki data *outlier*, sehingga data dapat dianggap cukup konsisten untuk dilanjutkan pada proses analisis lebih lanjut. Namun, ditemukan sebanyak 127 data duplikat dalam dataset yang perlu diatasi melalui proses data *cleaning* pada tahap data *preparation*.



Gambar 4. Data *Outlier*

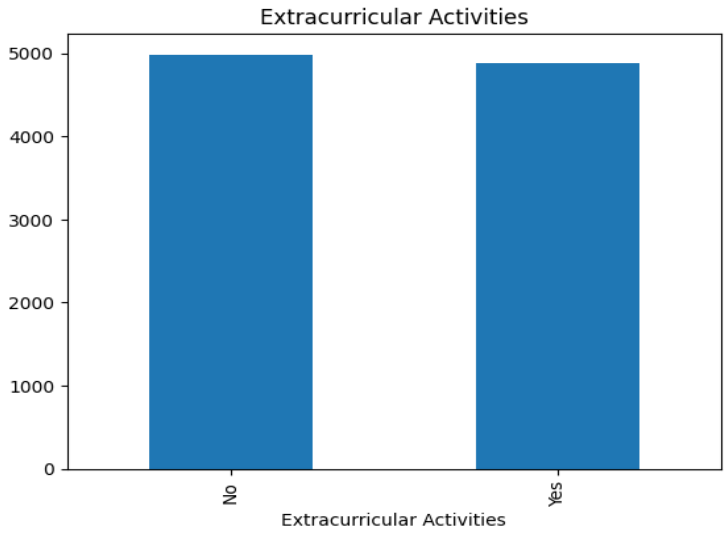
c. *Exploratory Data Analysis* (EDA)

Analisis awal dilakukan untuk mengetahui distribusi data dan memahami bagaimana variabel target dan variabel prediktor berhubungan satu sama lain. Proses pada tahapan ini mencakup analisis distribusi univariat dan multivariat yang dilakukan melalui visualisasi grafik serta penggunaan matriks korelasi untuk mengidentifikasi pola-pola penting dalam dataset. Seperti yang terlihat pada Gambar 5, visualisasi hasil analisis Univariat data numerik menunjukkan distribusi berbagai variabel yang berpengaruh terhadap performa akademik siswa.



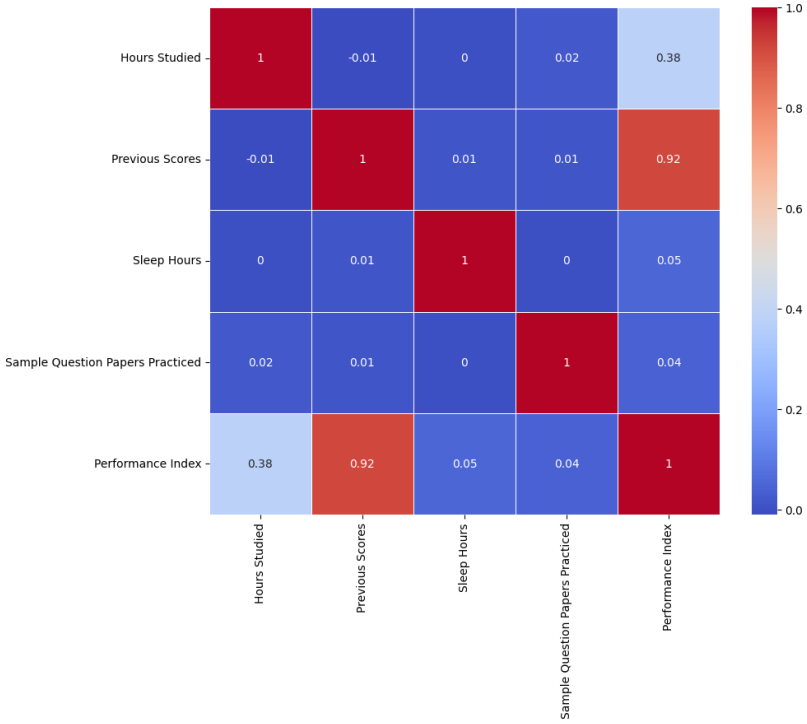
Gambar 5. Analisis Univariat Data Numerik

Dari gambar diatas diketahui bahwa sebagian besar siswa menghabiskan waktu 1 hingga 9 jam per hari untuk belajar. Grafik sebaran skor sebelumnya (*previous score*) menunjukkan bahwa jumlah siswa yang memperoleh nilai 50 cukup tinggi, mencapai sekitar 350 orang. Selain itu, sebagian besar siswa memiliki pola tidur yang konsisten, berkisar antara 6 hingga 8 jam setiap hari. Dalam hal latihan soal, sebagian besar siswa melakukan latihan sebanyak 3, 5, 6, 8, atau 9 kali. Grafik indeks kinerja juga menunjukkan bahwa sebagian besar siswa memiliki indeks kinerja antara 38 dan 75. Selain itu, hasil analisis univariat pada data kategorik yang ditampilkan pada Gambar 6 menunjukkan bahwa jumlah siswa yang mengikuti dan tidak mengikuti kegiatan ekstrakurikuler hampir seimbang, masing-masing kelompok sekitar 4.900 siswa.



Gambar 6. Analisis Univariat Data Kategorik

Setelah melakukan analisis univariat, matriks korelasi digunakan sebagai bagian dari analisis multivariat untuk mengeksplorasi hubungan antar variabel numerik.



Gambar 7. Matriks Korelasi

Seperti yang dapat dilihat pada Gambar 7, hasil matriks korelasi menunjukkan bahwa :

1. *Hour studied* dan *Performance index* memiliki korelasi positif moderat sebesar 0,38, yang menunjukkan bahwa semakin banyak jam belajar yang dilakukan oleh siswa, maka performa akademik siswa juga cenderung meningkat. Namun, karena nilai korelasinya tidak cukup kuat, dapat disimpulkan bahwa jam belajar bukanlah satu-satunya faktor yang paling mempengaruhi performa akademik siswa. Masih ada faktor lain yang juga mempengaruhi indeks kinerja (*Performance index*) selain jam belajar.
2. Ada korelasi positif yang signifikan antara *Previous scores* dan *Performance index* dengan skor 0,92. Hal ini menunjukkan bahwa nilai akademik siswa di masa lalu sangat memengaruhi kinerja mereka di masa mendatang. Siswa yang memiliki nilai tinggi di masa lalu cenderung mempertahankan atau bahkan meningkatkan kinerjanya.
3. *Sleep hours* dan *Performance index* memiliki korelasi positif yang sangat lemah yaitu 0,05, yang menunjukkan bahwa jam tidur tidak siswa tidak memiliki pengaruh yang signifikan terhadap performa akademik mereka.
4. *Sample question paper practiced* dan *performance indeks* menunjukkan korelasi positif yang sangat lemah yakni sebesar 0,04. Hasil ini menunjukkan bahwa latihan soal memiliki pengaruh yang sangat kecil terhadap performa akademik siswa. Meskipun latihan soal membantu meningkatkan pemahaman siswa, temuan ini menunjukkan bahwa masih ada faktor lain yang lebih berpengaruh terhadap performa akademik siswa.

Secara keseluruhan, matriks korelasi ini menunjukkan bahwa nilai sebelumnya (Previous Scores) merupakan prediktor terkuat bagi indeks kinerja siswa, sementara jam belajar, jam tidur, dan latihan soal memiliki pengaruh yang lebih moderat atau lemah.

d. *Data Preparation*

Data *preparation* merupakan tahap untuk mempersiapkan data sebelum masuk ke tahap pembuatan model *Machine Learning*. Tahapan yang dilakukan meliputi penghapusan data duplikat sebanyak 127 entri, sehingga mengurangi data yang sebelumnya berjumlah 10.000 menjadi 9.873 data. Teknik encoding dilakukan untuk mengubah data kategorik seperti E*xtracurricular Activities* menjadi bentuk numerik, sebagaimana yang terlihat pada Gambar 8 dan Gambar 9.

	Hours Studied	Previous Scores	Sleep Hours	Sample Question Papers Practiced	Performance Index	Extracurricular Activities_No	Extracurricular Activities_Yes
0	7	99	9	1	91.0	False	True
1	4	82	4	2	65.0	True	False
2	8	51	7	2	45.0	False	True
3	5	52	5	2	36.0	False	True
4	7	75	8	5	66.0	True	False

Gambar 8. Sebelum dilakukan Fitur *Encoding*

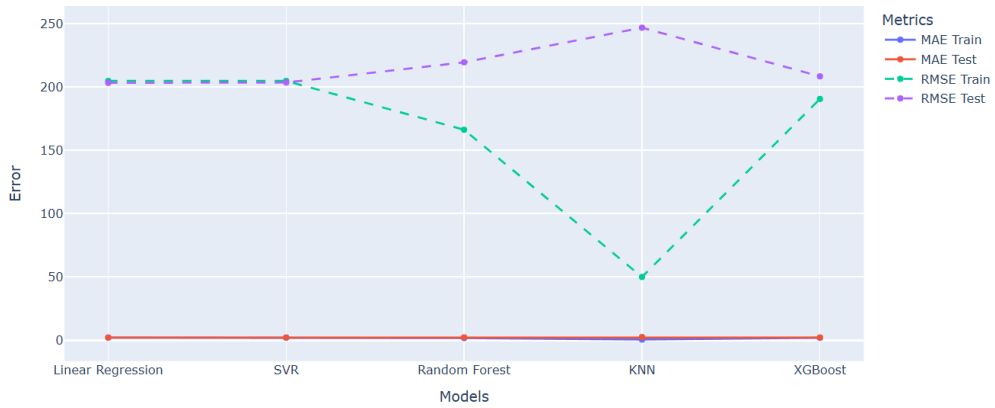
	Hours Studied	Previous Scores	Sleep Hours	Sample Question Papers Practiced	Performance Index	Extracurricular Activities_No	Extracurricular Activities_Yes
0	7	99	9	1	91.0	0	1
1	4	82	4	2	65.0	1	0
2	8	51	7	2	45.0	0	1
3	5	52	5	2	36.0	0	1
4	7	75	8	5	66.0	1	0

Gambar 9. Sesudah dilakukan Fitur *Encoding*

Setelah tahap *encoding* selesai, data dibagi menjadi dua bagian, 80% data *training* (7.898 data) dan 20% data *test* (1.975 data). Selain itu, normalisasi data juga dilakukan menggunakan *MinMaxScaler* untuk mentransformasi data ke dalam skala yang sama, sehingga semua fitur atau atribut memiliki rentang nilai yang sebanding.

e. Modelling

Setelah dilakukan proses data preparation, tahap pemodelan dilakukan menggunakan lima algoritma machine learning yang sudah dipilih sebelumnya, yaitu *Linear Regression, Support Vector Regression (SVR), Random Forest, K-Nearest Neighbors (KNN),* dan *XGBoost*. Performa setiap model yang digunakan dapat dilihat pada Gambar 10.



Gambar 10. Plot Model

Pada pemodelan *Linear Regression* yang dipilih karena kepraktisan dan kemampuannya dalam menginterpretasikan hubungan antar variabel, menunjukkan performa yang stabil dengan nilai *error* rendah pada data training dan testing. Sementara model *Support Vector Regression* (SVR), yang dipilih karena kemampuan untuk menangani pola data yang lebih kompleks menunjukkan hasil yang sama dengan model *Linear Regression*, tetapi tidak secara signifikan lebih baik. Model *Random Forest* yang dianggap mampu menangani data besar dan mengatasi *overfitting* menunjukkan performa yang baik pada data *training* tetapi tidak pada data *testing,* hal ini mengindikasikan adanya *overfitting* pada model. Hal serupa juga terjadi pada pemodelan *K-Nearest Neighbors* (KNN) yang memiliki *error* terendah pada data training, tetapi karena *overfitting*, performa model menjadi sangar buruk pada data *testing*. Di sisi lain, XGBoost yang dipilih karena kemampuannya untuk menangani data yang rumit dan meningkatkan akurasi model melalui peningkatan *boost*, menunjukkan performa yang sangat baik dengan nilai *error* terendah pada data *testing* dan *training* tanpa ada indikasi terjadinya *overfitting* pada model.

f. Evaluasi Model

Setelah dilakukan pemodelan menggunakan lima algoritma yang sudah dipilih, evaluasi dilakukan dengan menggunakan metrik *Mean Absolute Error* (MAE) dan Root *Mean Square Error* (RMSE), untuk memberikan gambaran tentang tingkat kesalahan prediksi yang dihasilkan oleh model.

Tabel 1. Hasil Evaluasi Model

Model	Data	MAE	RMSE
Linear Regression	Train	1.6254	2.0467
Linear Regression	Test	1.6134	2.0315
SVR	Train	1.6239	2.0467
SVR	Test	1.6159	2.0342
Random Forest	Train	1.3237	1.6598
Random Forest	Test	1.7462	2.1932
KNN	Train	0.1395	0.4991
KNN	Test	1.9848	2.4670
XGBoost	Train	1.5077	1.8996
XGBoost	Test	1.6510	2.0858

Berdasarkan Tabel 1, hasil evaluasi menunjukkan bahwa model *XGBoost* menunjukkan performa terbaik dari kelima model yang digunakan. Hal ini ditunjukkan oleh nilai *Mean Absolute Error* (MAE) yang rendah, sebesar 1,5077 pada data pelatihan dan 1,6510 pada data uji, yang menunjukkan nilai MAE terendah dari kelima model yang digunakan. Selain itu, nilai *Root Mean Squared Error* (RMSE) juga terendah, sebesar 1,8996 pada data

pelatihan dan 2,0858 pada data uji. Hasil ini menunjukkan bahwa model *XGBoost* dapat melakukan prediksi yang lebih baik tanpa mengalami *overvitting*. Dengan keseimbangan yang baik antara data uji dan dan data latih serta kemampuannya memberikan generalisasi yang lebih baik dan prediksi yang lebih akurat, XGBoost menjadi pilihan terbaik untuk digunakan dalam penelitian ini.

Namun, meskipun performa XGBoost unggul, terdapat beberapa keterbatasan dalam penelitian ini. Pertama, dataset yang digunakan mungkin tidak sepenuhnya representatif terhadap populasi siswa secara keseluruhan. Kedua, penelitian ini lebih berfokus pada variabel yang bersifat akademis dan beberapa faktor non-akademis yang terbatas, sehingga tidak mengeksplorasi faktor kontekstual yang lebih luas, seperti kondisi sosial ekonomi keluarga atau akses terhadap sumber daya pendidikan. Ketiga, implementasi XGBoost memerlukan *tuning hyperparameter* yang cukup kompleks, yang mungkin membatasi penggunaannya di lingkungan pendidikan dengan sumber daya teknologi yang terbatas.

Untuk penelitian selanjutnya, disarankan untuk menggunakan dataset yang lebih besar dan beragam untuk meningkatkan generalisasi model. Penelitian juga dapat memperluas cakupan variabel dengan memasukkan faktor-faktor kontekstual, seperti pengaruh lingkungan sosial, kondisi psikologis siswa, dan dukungan keluarga, untuk memberikan analisis yang lebih holistik. Selain itu, eksperimen lebih lanjut dengan algoritma lainnya atau metode deep learning dapat dilakukan untuk mengevaluasi apakah model yang lebih kompleks dapat memberikan peningkatan performa signifikan. Terakhir, pengembangan sistem berbasis aplikasi yang mudah digunakan oleh tenaga pendidik dapat menjadi langkah praktis untuk menerapkan model ini di lapangan, dengan tujuan membantu mereka dalam membuat keputusan yang lebih berbasis data untuk meningkatkan kualitas pendidikan.

5. Kesimpulan

Hasil penelitian menunjukkan bahwa skor sebelumnya (*previous scores*) memiliki pengaruh yang signifikan terhadap performa akademik siswa, dengan nilai korelasi sebesar 0,92. Hal ini munjukkan bahwa prestasi akademik siswa di masa mendatang sangat dipengaruhi oleh prestasi mereka di masa lalu. Dari kelima algoritma yang digunakan, model *XGBoost* menunjukkan kinerja prediksi terbaik dibanding dengan model lainnya. Dengan keakuratan prediksi yang tinggi tanpa indikasi *overfitting*, model *XGBoost* terbukti efektif dalam memprediksi faktor-faktor yang mempengaruhi performa akademik siswa.

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan. Dari sisi metodologi, dataset yang digunakan belum sepenuhnya representatif terhadap populasi siswa secara keseluruhan. Selain itu, variabel yang dianalisis, seperti jam belajar dan

keterlibatan ekstrakurikuler, bersifat terbatas dan belum mencakup faktor lain yang mungkin memiliki dampak signifikan, seperti kondisi sosial-ekonomi. Keterbatasan ini dapat memengaruhi generalisasi hasil jika diterapkan di populasi atau konteks pendidikan yang berbeda.

Penerapan teknologi seperti *machine learning* dalam analisis pendidikan, sebagaimana ditunjukkan oleh penelitian ini, memiliki potensi besar untuk merevolusi cara pendidik memahami kebutuhan siswa dan merancang strategi pembelajaran yang lebih efektif. Dengan mengintegrasikan model prediktif yang kuat seperti *XGBoost* ke dalam kebijakan pendidikan, lembaga pendidikan dapat mengambil langkah berbasis data untuk mengidentifikasi dan mengatasi faktor-faktor yang menghambat prestasi siswa. Pendekatan ini tidak hanya membantu meningkatkan hasil akademik tetapi juga membuka peluang untuk menciptakan sistem pendidikan yang lebih adil dan inklusif, khususnya di wilayah dengan keterbatasan fasilitas. Teknologi dapat menjadi alat transformasional dalam menjembatani kesenjangan pendidikan, memberikan dampak positif yang luas pada masa depan generasi penerus.

6. Ucapan Terima Kasih

Kami mengucapkan terima kasih kepada semua pihak yang telah mendukung dan berkontribusi dalam penelitian ini.

7. Pernyataan Penulis

Penulis menyatakan bahwa tidak ada konflik kepentingan terkait publikasi artikel ini. Penulis menyatakan bahwa data dan makalah bebas dari plagiarisme serta penulis bertanggung jawab secara penuh atas keaslian artikel.

Bibliografi

Afandi, K., Arief, M. H., Faizatul Laily, N., & Maulana Nugroho, D. (2024). Analisis Performa Akademik Mahasiswa Menggunakan Social Network Analysis (Studi Kasus: Prodi Bisnis Digital Universitas dr. Soebandi). Journal of Technology and Informatics (JoTI), 5(2), 64–69. https://doi.org/10.37802/joti.v5i2.514

Al-Alawi, L., Al Shaqsi, J., Tarhini, A., & Al-Busaidi, A. S. (2023). Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. *Education and Information Technologies*, *28*(10), 12407–12432. https://doi.org/10.1007/s10639-023-11700-0

Alita, D., & Rahman, A. (2020). Pendeteksian Sarkasme pada Proses Analisis Sentimen Menggunakan Random Forest Classifier. *Jurnal Komputasi*, *8*(2), 50–58. https://doi.org/10.23960/komputasi.v8i2.2615

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in*

Higher Education, *17*(1). https://doi.org/10.1186/s41239-020-0177-7

Beckham, N. R., Akeh, L. J., Mitaart, G. N. P., & Moniaga, J. V. (2022). Determining factors that affect student performance using various machine learning methods. *Procedia Computer Science*, *216*(2022), 597–603. https://doi.org/10.1016/j.procs.2022.12.174

Chitti, M., Chitti, P., & Jayabalan, M. (2020). Need for Interpretable Student Performance Prediction. *Proceedings - International Conference on Developments in ESystems Engineering, DeSE*, *2020-Decem*, 269–272. https://doi.org/10.1109/DeSE51703.2020.9450735

Gori, T., Sunyoto, A., & Al Fatta, H. (2024). Preprocessing Data dan Klasifikasi untuk Prediksi Kinerja Akademik Siswa. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *11*(1), 215–224. https://doi.org/10.25126/jtiik.20241118074

Hendriyanto, M. D., & Betha Nurina Sari. (2022). Penerapan Algoritma K-Nearest Neighbor Dalam Klasifikasi Judul Berita Hoax. *Jurnal Ilmiah Informatika*, *10*(02), 80–84. https://doi.org/10.33884/jif.v10i02.5477

Herman, Y. C. (2022). Analisis Performa Akademik Mahasiswa Menggunakan Distributed Random Forest. *Journal of Applied Informatics and Computing (JAIC)*, *6*(2), 180.

Herni Yulianti, S. E., Oni Soesanto, & Yuana Sukmawaty. (2022). Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. *Journal of Mathematics: Theory and Applications*, *4*(1), 21–26. https://doi.org/10.31605/jomta.v4i1.1792

Ibrahim, I. H., Garba, E. J., & Adejumo, A. (2024). Predictive Model for Identification and Analysis of Factors Impacting Students Academic Performance Using Machine Learning Algorithms. *Kasu Journal of Computer Science*, *1*(September). https://doi.org/10.47514/kjcs/2024.1.3.0013

Isnaeni, Sudarmin, Z. R. (2022). Analisis Support Vector Regression (Svr) Dengan Kernel Radial Basis Function (Rbf) Untuk Memprediksi Laju Inflasi Di Indonesia. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, *4*(1), 30–38. https://doi.org/10.35580/variansiunm13

Jan Melvin Ayu Soraya Dachi, & Pardomuan Sitompul. (2023). Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit. *Jurnal Riset Rumpun Matematika Dan Ilmu Pengetahuan Alam*, *2*(2), 87–103. https://doi.org/10.55606/jurrimipa.v2i2.1470

Khoeriyah, M. (2024). *Menuju Indonesia Emas 2045, tapi Kesenjangan Pendidikan Masih Tinggi?* 29 Oktober.

Khusaini, M. (2020). Prestasi Belajar dan Karakteristik Orang Tua: Studi Perbandingan Sekolah Menengah Atas Perkotaan-Pedesaan. *Jurnal Pendidikan Ekonomi Undiksha*, *12*(2), 296–310.

Sihombing, P. R., Suryadiningrat, S., Sunarjo, D. A., & Yuda, Y. P. A. C. (2023). Identifikasi Data Outlier (Pencilan) dan Kenormalan Data Pada Data Univariat serta Alternatif Penyelesaiannya. *Jurnal Ekonomi Dan Statistik Indonesia*, *2*(3), 307–316. https://doi.org/10.11594/jesi.02.03.07

Sinaga, W. A. L., Sumarno, S., & Sari, I. P. (2022). Penerapan Metode Regresi Linier Berganda Untuk Estimasi Jumlah Penduduk Pada Kecamatan Gunung Malela.

JOMLAI: Journal of Machine Learning and Artificial Intelligence, *1*(1), 55–64. https://doi.org/10.55123/jomlai.v1i1.143

Suci Amaliah, Nusrang, M., & Aswi, A. (2022). Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, *4*(3), 121–127. https://doi.org/10.35580/variansiunm31

Suryanto, A. A. (2019). Penerapan Metode Mean Absolute Error (Mea) Dalam Algoritma Regresi Linear Untuk Prediksi Produksi Padi. *Saintekbu*, *11*(1), 78–83. https://doi.org/10.32764/saintekbu.v11i1.298