Prosiding Seminar Nasional Teknologi dan Sistem Informasi (SITASI) 2025
Surabaya, 11 November 2025

ANALISIS FAKTOR PERFORMA AKADEMIK MATEMATIKA SISWA DENGAN PERBANDINGAN ALGORITMA MACHINE LEARNING

ANALYSIS OF STUDENT MATHEMATICS PERFORMANCE FACTORS THROUGH MACHINE LEARNING ALGORITHM COMPARISON

Tasbi Khatuz Zuhriya[1*], Erna Daniati[1]
*E-mail: ernadaniati@unpkediri.ac.id

[1]Sistem Informasi, Fakultas Teknik dan Ilmu Komputer, Universitas Nusantara PGRI Kediri

Abstrak

Matematika merupakan salah satu ilmu yang sering digunakan manusia dalam kehidupan sehari-hari, karena dengan adanya ilmu matematika dapat membantu manusia untuk membentuk pola pikir kemampuan penalaran manusia. Meski begitu tidak semua siswa menyukai mata pelajaran matematika, namun bagi sebagian siswa yang menganggap matematika merupakan ilmu yang menyenangkan, hal tersebut bisa dipengaruhi oleh beberapa faktor. Dalam penelitian ini akan melakukan analisis mengenai faktor performa akademik matematika siswa yang meliputi latar belakang pendidikan orang tua, gender, status ekonomi, les / kursus tambahan, serta kemampuan membaca dan menulis dengan menggunakan perbandingan model algoritma *machine learning* seperti *Regresi Linier, Random Forest,* dan *Gradient Boosting*. Dari ketiga model algoritma tersebut, *Regresi Linier* dengan nilai RMSE = 5.37, MAE = 4.23, serta nilai R^2 tertinggi dengan nilai 0.88. Dan kesimpulan dari penelitian ini faktor yang paling berpengaruh dengan performa akademik matematika siswa adalah gender, latar belakang pendidikan orang tua, serta kemampuan membaca dan menulis.

Kata kunci**:** *Performa Akademik, Matematika, Machine Learning* .

Abstract

Mathematics is one of the sciences that is often used by humans in everyday life, because with the existence of mathematics, it can help humans to form a mindset of human reasoning abilities. Even so, not all students like mathematics, but for some students who consider mathematics to be a fun science, this can be influenced by several factors. In this study, an analysis will be conducted on the factors of students' mathematical academic performance which include parental educational background, gender, economic status, tutoring / additional courses, and reading and writing skills by using a comparison of machine learning algorithm models such as Linear Regression, Random Forest, and Gradient Boosting. Of the three algorithm models, Linear Regression with an RMSE value of 5.37, MAE = 4.23, and the highest R2 value with a value of 0.88. And the conclusion of this study is that the factors that most influence students' mathematical academic performance are gender, parental educational background, and reading and writing skills.

Keywords*: Academic Performance, Mathematics, Machine Learning.*

1. PENDAHULUAN

Matematika merupakan salah satu ilmu yang sering digunakan manusia dalam kehidupan sehari-hari, karena dengan adanya ilmu matematika dapat membantu manusia untuk membentuk pola

pikir kemampuan penalaran manusia [1]. Namun sebagian siswa menganggap mata pelajaran matematika menjadi sumber kecemasan dan ketakutan atau sering disebut dengan *math anxiety. Math anxiety* bisa muncul saat siswa dihadapkan pada situasi yang melibatkan matematika khususnya di lingkungan akademik atau sekolah [2]. Meski begitu, tidak semua siswa menjadikan mata pelajaran matematika sebagai sumber kecemasan dan ketakutan, sebagian dari mereka ada yang menganggap mata pelajaran matematika sebagai tantangan yang perlu dipecahkan.

Berdasarkan penelitian yang dilakukan oleh [3] faktor yang mempengaruhi personal siswa dalam belajar dibagi menjadi dua faktor yaitu faktor internal dan faktor eksternal. Faktor internal diantaranya motivasi, kepercayaan diri, rasa ingin tahu, IQ dan lain sebagainya. Sedangkan dalam faktor eksternal diantaranya adalah dukungan keluarga khususnya orang tua.

Berdasarkan hasil penelitian yang dilakukan oleh [4] melakukan analisis prestasi siswa faktor fasilitas belajar, kebiasaan belajar, partisipasi dalam kursus tambahan, motivasi diri, dan dukungan orang tua. Dari hasil penelitian tersebut menunjukan bahwa dukungan orang tua memiliki pengaruh terbesar, diikuti oleh motivasi diri, kebiasaan belajar dan mengikuti kursus tambahan. Namun nilai dari R^2 sebesar -0,34 yang mengindikasikan bahwa model belum optimal dalam menjelaskan variabilitas data target.

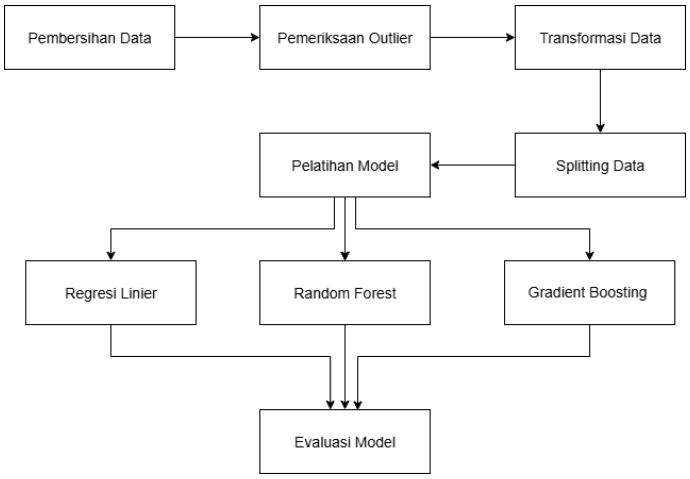
Dalam penelitian ini, menggunakan algoritma *Regresi Linier. Regresi Linier* merupakan salah satu metode yang digunakan untuk mengetahui hubungan antara variabel independen (bebas) dan hubungan garis lurus dengan variabel dependen (terikat). Selain itu *Regresi Linier* juga merupakan metode statistik yang digunakan untuk melakukan prediksi mengenai karakteristik kualitas maupun kuantitas [5].

Selain itu dalam penelitian ini juga menggunakan perbandingan *Machine Learning* yang lain seperti *Random Forest* dan *Gradient Boosting*. *Random Forest* merupakan algoritma yang bekerja dengan menggunakan metode pemisahan biner rekursif untuk mencapai node akhir dalam struktur pohon yang berdasar pada pohon klasifikasi dan regresi [6]. Sedangkan *Gradient Boosting* merupakan algoritma *Machine Learning* yang digunakan untuk menyelesaikan permasalahan regresi dan klasifikasi, model prediksi yang dihasilkan dalam algoritma ini terdiri dari ansambel model prediksi lemah, yang biasanya berupa pohon keputusan [7].

Penelitian ini bertujuan untuk mengetahui faktor-faktor yang mempengaruhi akademik matematika siswa dengan menggunakan perbandingan algoritma *Regresi Linier, Random Forest,* dan *Gradient Boosting*. Dengan memahami pengaruh ini, diharapkan dapat memberikan wawasan baru mengenai faktor-faktor sosial dan ekonomi yang berkontribusi terhadap prestasi akademik siswa. Hasil penelitian ini juga diharapkan dapat menjadi acuan bagi pendidik dan pembuat kebijakan dalam merancang strategi pembelajaran yang lebih efektif serta dukungan pendidikan yang sesuai dengan latar belakang keluarga siswa.

2. METODOLOGI

Pada penelitian ini, data yang digunakan dalam penelitian merupakan dataset “Students Performance | Clean Dataset” yang didapatkan dari platform Kaggle dengan 1000 baris dan 10 kolom. Berikut gambar mengenai tahapan analisis yang dilakukan dalam penelitian :



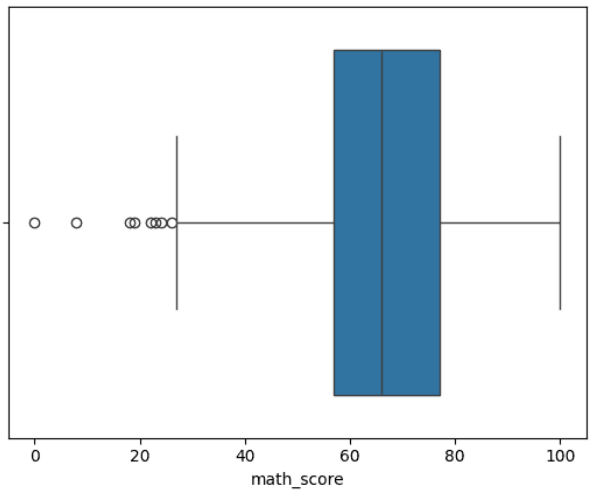
Gambar 1. Alur Diagram Penelitian

2.1 Pembersihan Data

Langkah pertama yang dilakukan untuk analisis data adalah pembersihan data. Pembersihan data merupakan suatu tahapan yang melibatkan identifikasi, perbaikan, serta penghapusan ketidakakuratan data untuk meningkatkan kualitas data [8]. Dalam penelitian ini pembersihan data dilakukan dengan menghapus kolom yang tidak relevan, seperti kolom ”race_ethnicity”. Selanjutnya dilakukan pengecekan *duplicate value* dan *missing value.*

2.2 Pemeriksaan Outlier

Setelah melakukan pembersihan data, langkah selanjutnya adalah melakukan pemeriksaan *Outlier*. *Outlier* merupakan suatu objek data yang tidak mengikuti perilaku umum dari data [9]. Seperti yang terlihat pada Gambar 2. terdapat *Outlier* dalam data, *Outlier* tersebut tidak dihilangkan karena setiap siswa memiliki nilai matematika yang berbeda-beda. Hal ini sangat penting untuk mempertahankan berbagai macam aneka data dan memastikan *Regresi Linier, Random Forest,* dan *Gradient Boosting* dapat menangkap variasi nilai matematika siswa.



Gambar 2. Diagram Outlier

2.3 Transformasi Data

Transformasi data dilakukan secara bertahap untuk menyesuaikan format dan mempermudah analisis. Dengan mengubah data latar belakang pendidikan orang tua, gender, status ekonomi, les / kursus tambahan menjadi bentuk indeks dengan rentang tertentu [10]. Pada tahap transformasi data dilakukan proses *one-hot-encoding. One-Hot-Encoding* merupakan metode yang digunakan untuk melakukan cluster pada data bertipe campuran, metode ini mengkonversi setiap atribut kategori menjadi *dummy* dimana nilai 1 diberikan untuk kategori tertentu sedangkan kategori yang lain bernilai 0 [11]. Hal ini dilakukan karena *Regresi Linier, Random Forest,* dan *Gradient Boosting* membutuhkan data numerik untuk melakukan perhitungan dan analisis. Dalam penelitian dilakukan *encoding* untuk kolom "parental_level_of_education", "gender", "lunch", dan "test_preparation_course".

2.4 Splitting Data

Setelah proses *encoding* selesai, langkah selanjutnya adalah *splitting data. Splitting data* merupakan suatu proses yang dilakukan untuk membagi data menjadi data training dan data test [12]. Dalam penelitian ini *splitting* dilakukan dengan data pelatihan dengan rasio 0.8 dan data pengujian dengan rasio 0.2.

2.5 Pelatihan Model

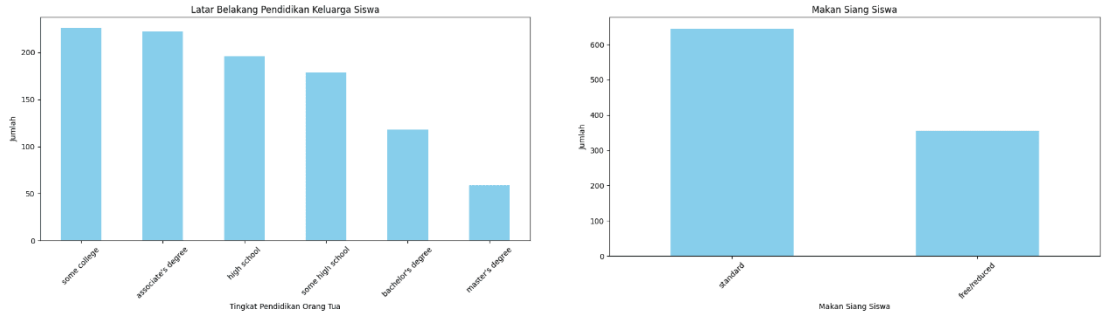
Keberhasilan penggunaan *Regresi Linier, Random Forest,* dan *Gradient Boosting* dapat dilihat dari kinerja model terhadap data uji. Dalam penelitian ini *Regresi Linier* menggunakan package *statsmodels, Random Forest* menggunakan package *scikit-learn* dengan memanfaatkan algoritma *ensemble learning* berbasis pohon keputusan, sedangkan *Gradient Boosting* menggunakan package *sklearn* dengan parameter utama seperti jumlah estimator (n_estimators = 200), learning rate sebesar 0.1, serta kedalaman maksimum pohon (max_depth = 5). Penetapan nilai random_state = 42 juga dilakukan guna menjaga konsistensi hasil pada setiap proses pelatihan.

2.6 Evaluasi Model

Setelah melakukan pelatihan pada model dilakukan evaluasi menggunakan data pengujian. Selanjutnya melakukan perhitungan metrik kinerja seperti RMSE (*Root Mean Square Error),* MAE (*Mean Absolute Error*), dan R^2 (koefisien determinasi) untuk menilai seberapa baik model dalam membuat prediksi.

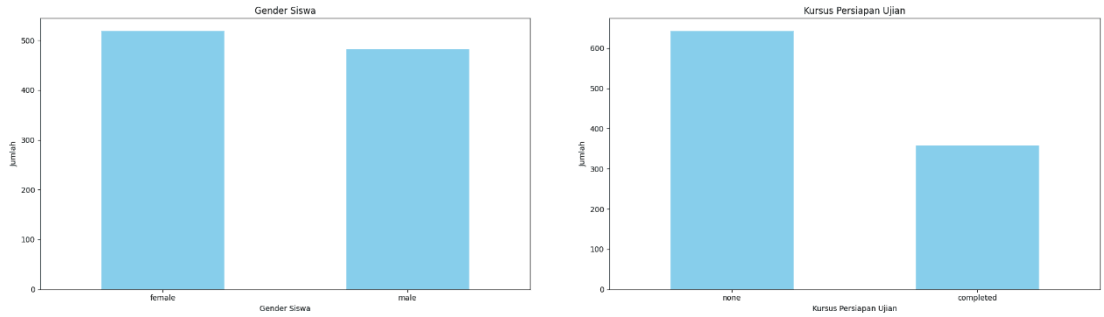
3. HASIL DAN PEMBAHASAN

Berdasarkan 1000 data siswa yang telah dianalisis diketahui bahwa sebagian besar siswa memiliki keluarga dengan latar belakang pendidikan *some college* kemudian diikuti dengan latar belakang pendidikan *associate's degree, high school, some high school, bachelor's degree,* serta *master's degree*. Selain itu juga diketahui bahwa banyak siswa yang tidak mendapatkan makan siang gratis, hal tersebut menunjukan bahwa banyak siswa yang mempunyai tingkat sosial ekonomi menengah ke atas. Hal ini dapat dibuktikan melalui gambar dibawah ini.



Gambar 3. Diagram Latar Belakang Pendidikan Keluarga Dan Tingkat Sosial Ekonomi Siswa

Selanjutnya dari hasil analisis data tersebut juga diketahui banyak jumlah siswa perempuan daripada siswa laki-laki. Dan dari hasil analisis data juga diketahui bahwa lebih banyak siswa yang tidak mengikuti les / kursus tambahan daripada siswa yang mengikuti les / kursus tambahan. Hal ini dapat dibuktikan melalui gambar berikut.



Gambar 4. Diagram Gender Dan Les / Kursus Tambahan Siswa

3.1 *Regresi Linier*

Setelah melakukan analisis pada data, langkah selanjutnya adalah melakukan uji model Regresi Linier. Dari hasil uji untuk model Regresi Linier diketahui nilai dari RMSE 5.3 dan MAE 4.23 yang menjelaskan bahwa rata-rata kesalahan prediksi relatif kecil. Kemudian untuk nilai R2 sebesar 0.88, ini menjelaskan bahwa model mampu menjelaskan sekitar 88% variasi data terhadap variabel data target yaitu nilai matematika.

Selain itu pada analisis model Regresi Linier juga diketahui hasil nilai koefisien dari setiap kategori, seperti yang bisa dilihat dari tabel berikut:

Table 1. Nilai Koefisien Latar Belakang Pendidikan Keluarga Siswa

	Koefisien
Associate's Degree	0.0050
Bachelor's Degree	-0.681
High School	0.658
Master's Degree	-1.693
Some College	0.850
Some High School	0.470

Dari tabel diatas diketahui bahwa latar belakang pendidikan *master's degree* memberikan koefisien negatif yaitu dengan koefisien -1.693, yang artinya orang tua berpendidikan master cenderung memiliki skor lebih rendah. Sebaliknya latar belakang pendidikan seperti *some college* dengan nilai koefisien +0.850 dan *high school* dengan nilai koefisien +0.658 cenderung meningkatkan skor.

Table 2. Nilai Koefisien Gender Siswa

	Koefisien
Perempuan	-6.724
Laki-laki	6.333

Dari tabel diatas diketahui bahwa siswa dengan gender perempuan diprediksi memiliki skor lebih rendah dengan nilai koefisien -6.724 sedangkan siswa dengan gender Laki-laki 6.333 diprediksi lebih tinggi. Hal ini terjadi adanya indikasi perbedaan performa akademik matematika siswa berdasarkan gender.

Table 3. Nilai Koefisien Makan Siang Gratis

	Koefisien
Free / Reduced	-1.989
Standard	1.598

Dari tabel diatas diketahui bahwa siswa yang mendapatkan makan siang gratis dengan nilai koefisien -1.989 sedikit lebih rendah dibandingkan dengan siswa yang tidak mendapatkan makan siang gratis dengan nilai koefisien 1.598.

Table 4. Nilai Koefisien Les / Kursus Tambahan

	Koefisien
Completed	-1.662
None	1.271

Dari tabel diatas diketahui bahwa siswa yang mengikuti les / kursus tambahan memiliki nilai koefisien negatif sebesar -1.662 sedangkan siswa yang tidak mengikuti les / kursus tambahan memiliki nilai koefisien +1.271. hal ini terjadi kemungkinan siswa yang mengikuti les / kursus tambahan memiliki kemampuan matematika lebih rendah.

Table 5. Nilai Koefisien Reading Dan Writing

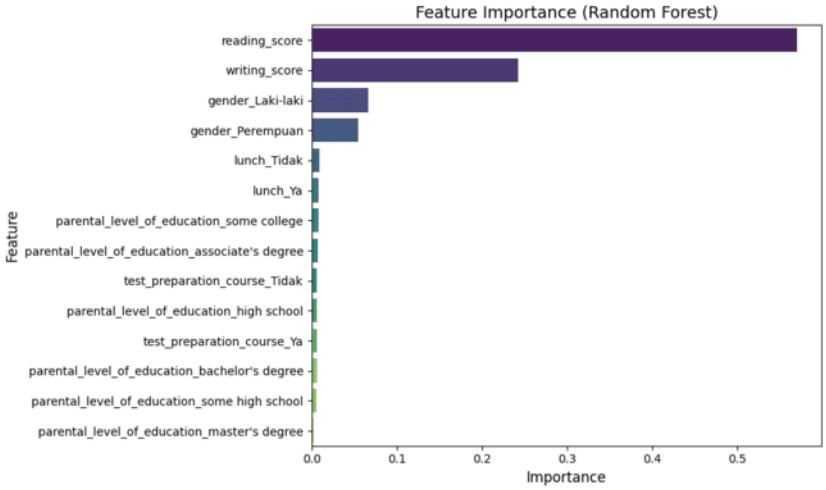
	Koefisien
Reading	0.295
Writing	0.673

Dari tabel diatas diketahui bahwa *reading score* memiliki nilai koefisien +0.295 sedangkan *writing score* memiliki nilai koefisien +0.673. ini menjelaskan bahwa siswa yang memiliki nilai tinggi dalam membaca dan menulis cenderung memiliki nilai matematika yang lebih tinggi.
Dari seluruh penjelasan diatas dapat diketahui bahwa model *regresi linier* memiliki performa prediksi yang cukup baik dengan nilai $R^2 = 0.882$ dan nilai RMSE yang rendah. Faktor yang sangat mempengaruhi tingkat akademik matematika siswa adalah gender, latar pendidikan keluarga, serta kemampuan membaca dan menulis.

3.2 *Random Forest*

Setelah melakukan analisis dengan model Regresi Linier, selanjutnya adalah melakukan analisis dengan model Random Forest. Dari hasil analisis menggunakan model Random Forest diketahui bahwa nilai dari R2 = 0.839 yang menunjukan bahwa model dapat menjelaskan sekitar 83,9% variasi data. Dan nilai RMSE = 6.25 dan MAE = 4.95 yang menjelaskan bahwa tingkat kesalahan prediksi masih cukup kecil.

Selain itu hasil analisis model random forest juga diketahui nilai prediksi tertinggi berada pada reading dengan skor 0.57 dan writing 0.24 yang menjadi faktor dominan prediksi, kedua faktor tersebut jauh lebih berpengaruh dibanding faktor lain. Hal tersebut dapat dibuktikan melalui gambar diagram bawah ini.

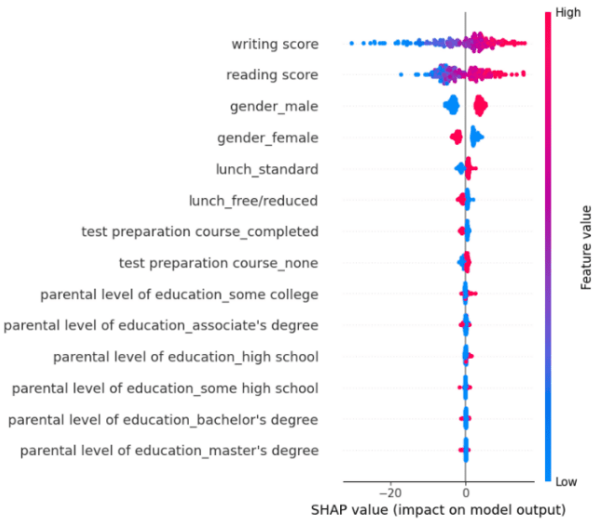


Gambar 6. Diagram Featur Importance

Dari hasil analisis model *random forest* yang telah dilakukan dapat disimpulkan bahwa model ini mampu melakukan prediksi yang cukup baik dengan akurasi tinggi. Meskipun performa dari model ini sedikit lebih rendah dibanding *regresi linier.* Dimana faktor utama yang memengaruhi model ini adalah kemampuan membaca dan menulis siswa.

3.3 *Gradient Boosting*

Setelah melakukan analisis dengan model *Random Forest,* selanjutnya adalah melakukan analisis dengan model *Gradient Boosting.* Dari hasil analisis menggunakan model *Gradient Boosting* diketahui bahwa nilai dari $R^2 = 0.835$ yang menunjukan bahwa model dapat menjelaskan sekitar 83,5% variasi data. Dan nilai RMSE = 6.33 dan MAE = 5.01 yang menjelaskan bahwa tingkat kesalahan prediksi masih cukup kecil.



Gambar 7. Diagram SHAP *Gradient Boosting*

Dari diagram SHAP *Gradient Boosting* diatas dapat diketahui faktor kemampuan *writing* dan *reading* menjadi pengaruh terbesar terhadap kemampuan akademik matematika siswa. Sedangkan faktor lain seperti gender, status makan siang, dan les / kursus tambahan memiliki pengaruh lebih kecil, sedangkan latar belakang pendidikan orang tua tidak berpengaruh signifikan terhadap kemampuan akademik matematika siswa.

3.4 Evaluasi Model

Setelah melakukan analisis menggunakan model algoritma *Regresi Linier, Random Forest,* dan *Gradient Boosting*, selanjutnya melakukan evaluasi ketiga model yang dapat dilihat dari tabel berikut :

Table 6. Evaluasi Model

	RMSE	MAE	R^2
Regresi Linier	5.37	4.23	0.88
Random Forest	6.25	4.95	0.84
Gradient Boosting	6.33	5.01	0.84

Dari tabel diatas dapat diketahui bahwa model algoritma *Regresi Linier* dapat memberikan performa terbaik dibandingkan dengan model *Random Forest* dan *Gradient Boosting,* ini karena model algoritma *Regresi Linier* memiliki nilai RMSE = 5.37 dan MAE = 4.23, nilai ini merupakan nilai paling kecil dari model algoritma lain. Hal ini menjelaskan bahwa prediksi model algoritma *Regresi Linier* menghasilkan hasil prediksi lebih mendekati dengan nilai sebenarnya.
Serta memiliki nilai R^2 tertinggi di dengan nilai 0.88 dimana nilai tersebut mampu menjelaskan 88% variasi data. Sedangkan *Random Forest* dan *Gradient Boosting* memiliki akurasi lebih rendah di angka 0.84 serta terdapat error lebih besar.

4. KESIMPULAN DAN SARAN

Dari pemaparan hasil dan pembahasan yang telah dijelaskan dapat disimpulkan bahwa faktor performa akademik matematika pada siswa diantara latar belakang pendidikan orang tua, gender, status ekonomi, les / kursus tambahan, serta kemampuan membaca dan menulis, faktor yang paling berpengaruh dengan performa akademik matematika siswa adalah gender, latar belakang pendidikan orang tua, serta kemampuan membaca dan menulis. Sebagian besar siswa yang memiliki performa tinggi dalam akademik matematika adalah siswa perempuan, keluarga dengan latar belakang pendidikan *some college* dan *high school*, serta siswa yang memiliki kemampuan membaca dan menulis
Rekomendasi bagi peneliti selanjutnya adalah peneliti dapat menambahkan variabel lain dalam penelitian dengan menggunakan data primer melalui survei atau menggunakan data akademik sekolah serta menggunakan algoritma *machine learning* yang beragam. Selain itu peneliti selanjutnya melakukan evaluasi model dengan teknik validasi yang lebih lengkap.

5. DAFTAR RUJUKAN

[1] M. Ardiansyah, “Kontribusi Tingkat Pendidikan Orang Tua, Lingkungan, dan Kecerdasan Logis Terhadap Kemampuan Berpikir Kritis Matematis,” *Jurnal Pendidikan Matematika (Kudus)*, pp. 185–200.
[2] L. Nurhidayati, “KECEMASAN MATEMATIKA (MATH ANXIETY) DAN DAMPAKNYA TERHADAP PRESTASI BELAJAR,” *JURNAL ILMIAH IPA DAN MATEMATIKA*, vol. 2, no. 3, pp. 61–66, Aug. 2024, doi: 10.61116/jiim.v2i3.477.

[3] L. F. Tae, Z. Ramdani, and G. A. Shidiq, "Analisis Tematik Faktor-Faktor yang Mempengaruhi Keberhasilan Siswa dalam Pembelajaran Sains," *Indonesian Journal of Educational Assessment*, vol. 2, no. 1, pp. 79–102, Jun. 2019, [Online]. Available: http://ijeajournal.kemdikbud.go.id

[4] W. C. Ferdinan, M. R. Noerfikri, P. A. Panchadri, and F. Ferawati, "Implementasi Algoritma Regresi Linear Berganda untuk Memprediksi Prestasi Siswa," *bit-Tech*, vol. 7, no. 3, pp. 853–864, Apr. 2025, doi: 10.32877/bt.v7i3.2228.

[5] Harsiti, Z. Muttaqin, and E. Srihartini, "PENERAPAN METODE REGRESI LINIER SEDERHANA UNTUK PREDIKSI PERSEDIAAN OBAT JENIS TABLET," *Jurnal Sistem Informasi*, vol. 9, no. 1, pp. 12–16, Mar. 2022, doi: 10.30656/jsii.v9i1.4426.

[6] F. Y. Pamuji and V. P. Ramadhan, "Komparasi Algoritma Random Forest Dan Decision Tree Untuk Memprediksi Keberhasilan Immunotheraphy," *Jurnal Teknologi dan Manajemen Informatika*, vol. 7, no. 1, pp. 46–50, Jun. 2021, doi: 10.26905/jtmi.v7i1.5982.

[7] S. B. Koduri, L. Gunisetti, C. R. Ramesh, and D. Ganesh, "Prediction of crop production using adaboost regression method," *J Phys Conf Ser*, vol. 1228, no. 1, pp. 1–10, 2019.

[8] I. N. Rizki, M. L. Puspita, D. Prayoga, and M. Q. Huda, "IMPLEMENTASI EXPLORATORY DATA ANALYSIS UNTUK ANALISIS DAN VISUALISASI DATA PENDERITA STROKE KALIMANTAN SELATAN MENGGUNAKAN PLATFORM TABLEAU," *Jurnal Informatika dan Teknik Elektro Terapan*, vol. 12, no. 1, Jan. 2024, doi: 10.23960/jitet.v12i1.3856.

[9] A. M. Siregar and A. Puspabhuana, *DATA MINING: Pengolahan Data Menjadi Informasi dengan RapidMiner*. CV Kekata Group.

[10] A. Tholib, M. N. F. Hidayat, S. Supriyono, R. Wulanningrum, and E. Daniati, "Comparison of C4.5 and Naive Bayes for Predicting StudentGraduation Using Machine Learning Algorithms," *International Journal of Engineering and Computer Science Applications*, vol. 2, no. 2, pp. 71–78, Sep. 2023, doi: 10.30812/IJECSA.v2i2.3364.

[11] M. Guntara and F. D. Astuti, "Komparasi Kinerja Label-Encoding dengan One-Hot-Encoding pada Algoritma K-Nearest Neighbor menggunakan Himpunan Data Campuran," *JIKO (Jurnal Informatika dan Komputer)*, vol. 9, no. 2, pp. 352–360, Jun. 2025, doi: 10.26798/jiko.v9i2.1605.

[12] A. Yuniarti, A. Yasin, and Y. A. Nugroho, "Efektifitas Algoritma Data Mining dalam Menentukan Pendonor Darah Potensial," *Syntax: Jurnal Informatika*, vol. 11, no. 01, pp. 12–22, Jun. 2022, doi: 10.35706/syji.v11i01.6595.