Jurnal Cakrawala dan Integrasi Pendidikan
Vol. 2, No. 2, Juli 2026, pp. 76-88
DOI: -
E-ISSN 3109-6328

Analisis Komparatif Metode Regresi dan Machine Learning untuk Prediksi Performa Akademik Mahasiswa

Comparative Analysis of Regression and Machine Learning Methods for Predicting Student Academic Performance

Nurkhalik Wahdanial Asbara[1,*]; Muhammad Rijal[2]; Irmawati[3]
[1,2] *Institut Teknologi dan Bisnis Nobel Indonesia, Makassar 90221, Indonesia*
[3] *IRMEX Digital Akademika, Makassar 90155, Indonesia*
[1] khalikwahdanial@nobel.ac.id; [2] rijal@nobel.ac.id; [3] irmawati@irmexdigika.com
** Corresponding author*

Abstrak

Penelitian ini membandingkan performa Linear Regression, Ridge Regression, Lasso Regression, dan Random Forest Regression dalam memprediksi performa akademik mahasiswa menggunakan Higher Education Students Performance Evaluation Dataset dari UCI Machine Learning Repository dengan 145 data dan GRADE sebagai target. Meskipun berbagai penelitian telah membandingkan algoritma prediksi performa akademik, masih terbatas penelitian yang mengintegrasikan analisis korelasi, uji multikolinearitas VIF, seleksi fitur Lasso, dan feature importance Random Forest dalam satu kerangka komparatif yang utuh pada dataset dengan multikolinearitas tinggi. Pemilihan keempat algoritma didasarkan pada karakteristik dataset yang memiliki multikolinearitas tinggi (22 dari 30 fitur dengan VIF > 10), sehingga diperlukan perbandingan antara regresi konvensional, regularisasi (Ridge dan Lasso), dan ensemble learning (Random Forest). Metodologi mencakup pra-pemrosesan data, analisis korelasi Pearson, uji multikolinearitas VIF, pembangunan model, dan evaluasi menggunakan R^2, RMSE, dan MAE. Hasil analisis korelasi menunjukkan hanya 8 dari 30 fitur yang berkorelasi signifikan dengan GRADE (r tertinggi = 0,3355). Uji VIF mengidentifikasi 22 fitur dengan VIF > 10, mengindikasikan multikolinearitas sangat tinggi. Lasso Regression menjadi model terbaik dengan R^2 testing 0,1535, RMSE 1,9735, dan MAE 1,5905, mengungguli Ridge Regression (R^2 = 0,1476), Random Forest (R^2 = 0,1346), dan Linear Regression (R^2 = -0,0902). Meskipun Lasso Regression memberikan performa terbaik, nilai R^2 testing yang hanya 0,1535 mengindikasikan bahwa sebagian besar variasi nilai akademik belum dapat dijelaskan oleh model, yang disebabkan oleh keterbatasan jumlah data (n=145), lemahnya korelasi linier antar fitur dengan GRADE (r tertinggi = 0,3355), serta kemungkinan adanya faktor-faktor lain di luar dataset yang memengaruhi performa akademik. Ridge menggunakan alpha 100,0 dan Lasso menggunakan alpha 0,1 dengan eliminasi 10 fitur. Random Forest menunjukkan overfitting dengan rata-rata R^2 cross-validation negatif (-0,0108). Fitur 29 (Kebiasaan Pendidikan) dan fitur 11 (Pertanyaan Keluarga) teridentifikasi sebagai prediktor paling konsisten, dengan fitur 29 memiliki korelasi positif signifikan terhadap GRADE (r = 0,3155; p-value = 0,0001). Penelitian menyimpulkan bahwa Lasso Regression lebih efektif pada dataset dengan multikolinearitas tinggi, serta kebiasaan pendidikan merupakan faktor dominan dalam prediksi performa akademik.

Kata Kunci: Prediksi Performa Akademik, Regresi Linier, Ridge Lasso, Random Forest, Multikolinearitas, Educational Data Mining

Abstract

This study compares the performance of Linear Regression, Ridge Regression, Lasso Regression, and Random Forest Regression in predicting student academic performance using the Higher Education Students Performance Evaluation Dataset from the UCI Machine Learning Repository, consisting of 145 data points with GRADE as the target variable. Although various studies have compared prediction algorithms for academic performance, research integrating correlation analysis, VIF multicollinearity testing, Lasso feature selection, and Random Forest feature importance within a unified comparative framework on high-multicollinearity datasets remains limited. The selection of these four algorithms was based on the dataset's high multicollinearity characteristics (22 out of 30 features with VIF > 10), necessitating a comparison between conventional regression, regularization methods (Ridge and Lasso), and ensemble learning (Random Forest). The methodology includes data preprocessing, Pearson correlation analysis, VIF multicollinearity testing, model development, and evaluation using R^2, RMSE, and MAE. Correlation analysis showed that only 8 out of 30 features had a significant correlation with GRADE (highest r = 0.3355). VIF testing identified 22 features with VIF > 10, indicating very high multicollinearity. Lasso Regression was the best model with R^2 testing of 0.1535, RMSE of 1.9735, and MAE of 1.5905, outperforming Ridge Regression (R^2 = 0.1476), Random Forest (R^2 = 0.1346), and Linear Regression (R^2 = -0.0902). Although Lasso Regression provided the best performance, the R^2 testing value of only 0.1535 indicates that most of the variation in academic performance cannot be explained by the model, due to limited data (n=145), weak linear correlation between features and GRADE (highest r = 0.3355), and the possibility of other factors beyond the dataset affecting academic performance. Ridge used alpha 100.0 and Lasso used alpha 0.1 with elimination of 10 features. Random Forest showed overfitting with a negative average cross-validation R^2 (-0.0108). Feature 29 (Educational Habits) and feature 11 (Family Questions) were identified as the most consistent predictors, with feature 29 having a significant positive correlation with GRADE (r = 0.3155; p-value = 0.0001). The study concludes that Lasso

Regression is more effective on datasets with high multicollinearity, and educational habits are the dominant factor in predicting academic performance.

Keywords: *Academic Performance Prediction, Linear Regression, Ridge Lasso, Random Forest, Multicollinearity, Educational Data Mining*

Pendahuluan

Sistematika Perkembangan teknologi informasi telah mendorong transformasi pada berbagai bidang, termasuk sektor pendidikan. Salah satu dampak perkembangan tersebut adalah meningkatnya pemanfaatan data akademik untuk mendukung proses pengambilan keputusan melalui pendekatan Educational Data Mining (EDM). EDM merupakan cabang ilmu yang memanfaatkan teknik statistik, data mining, dan machine learning untuk menemukan pola serta menghasilkan informasi yang dapat digunakan dalam meningkatkan kualitas pembelajaran dan pengelolaan pendidikan [1]. Salah satu penerapan EDM yang banyak dikembangkan adalah prediksi performa akademik mahasiswa (Student Performance Prediction), yaitu proses memprediksi hasil belajar mahasiswa berdasarkan karakteristik, aktivitas pembelajaran, maupun faktor-faktor lain yang memengaruhi pencapaian akademik [2].

Prediksi performa mahasiswa menjadi salah satu topik penting dalam pendidikan tinggi karena mampu membantu institusi mengidentifikasi mahasiswa yang berpotensi mengalami penurunan prestasi sejak dini. Hasil prediksi tersebut dapat digunakan sebagai dasar penyusunan strategi pembelajaran, pemberian pendampingan akademik, maupun evaluasi kurikulum sehingga dapat meningkatkan keberhasilan studi mahasiswa [3]. Selain itu, meningkatnya jumlah data akademik yang tersimpan pada sistem informasi perguruan tinggi menjadikan penerapan metode machine learning sebagai solusi yang efektif dalam menghasilkan model prediksi yang lebih cepat, objektif, dan akurat dibandingkan pendekatan konvensional [4].

Berbagai metode machine learning telah diterapkan untuk memprediksi performa akademik mahasiswa. Algoritma yang sering digunakan antara lain Decision Tree, Support Vector Machine, Artificial Neural Network, Random Forest, Naïve Bayes, Linear Regression, Ridge Regression, dan Lasso Regression [5]. Masing-masing algoritma memiliki keunggulan dan keterbatasan yang dipengaruhi oleh karakteristik data, jumlah atribut, distribusi target, serta hubungan antarvariabel. Oleh karena itu, pemilihan model prediksi yang sesuai masih menjadi tantangan dalam penelitian di bidang Educational Data Mining [6].

Muhammad Bin Roslan dan Chwen Chen melakukan systematic literature review terhadap 58 artikel yang diterbitkan pada periode 2015–2021 untuk mengidentifikasi tren penelitian, faktor-faktor yang paling banyak dikaji, serta metode yang digunakan dalam prediksi performa akademik mahasiswa [7]. Hasil kajian menunjukkan bahwa penelitian pada bidang ini didominasi oleh identifikasi faktor yang memengaruhi performa mahasiswa, evaluasi kinerja algoritma data mining, serta penerapan data mining pada sistem e-learning. Selain itu, data akademik dan karakteristik demografis mahasiswa merupakan faktor yang paling sering digunakan dalam proses prediksi, sedangkan pendekatan klasifikasi dengan algoritma Decision Tree menjadi metode yang paling banyak diterapkan [7]. Alice Villar dan Carolina Robledo Velini de Andrade membandingkan beberapa algoritma machine learning untuk memprediksi student dropout dan keberhasilan akademik mahasiswa. Hasil penelitian menunjukkan bahwa algoritma boosting, terutama LightGBM dan CatBoost yang dioptimalkan menggunakan Optuna, memberikan performa prediksi yang lebih baik dibandingkan algoritma klasifikasi tradisional seperti Decision Tree, Support Vector Machine, dan Random Forest [8].

Penelitian lain dilakukan oleh Wanyonyi dan Masinde yang menganalisis pengaruh teknik prapemrosesan data terhadap performa berbagai model machine learning pada beberapa dataset publik [9]. Penelitian tersebut mengevaluasi kinerja beberapa algoritma, antara lain Random Forest, Logistic Regression, K-Nearest Neighbor, dan Long Short-Term Memory (LSTM), menggunakan metrik accuracy, precision, recall, F1-score, dan waktu pelatihan. Hasil penelitian menunjukkan bahwa proses prapemrosesan data memberikan pengaruh yang signifikan terhadap performa model machine learning, sehingga tahapan tersebut menjadi bagian penting dalam pembangunan model prediksi [9]. Amjad Abu Saa melakukan penelitian mengenai penerapan Educational Data Mining (EDM) untuk menganalisis performa mahasiswa pada pendidikan tinggi [10]. Penelitian tersebut mengeksplorasi berbagai faktor personal dan sosial yang secara teoritis dapat memengaruhi performa akademik mahasiswa. Hasil penelitian menunjukkan bahwa faktor-faktor tersebut dapat digunakan untuk membangun model klasifikasi dan prediksi performa mahasiswa, sehingga EDM dapat menjadi pendekatan yang efektif dalam menemukan pola serta informasi penting dari data pendidikan [10].

Meskipun berbagai penelitian telah dilakukan, masih terdapat beberapa keterbatasan yang menjadi celah penelitian (research gap). Pertama, sebagian besar penelitian hanya membandingkan satu atau dua algoritma prediksi sehingga belum memberikan gambaran mengenai performa beberapa metode regresi pada dataset yang sama [1], [10], sebagian penelitian lebih berfokus pada peningkatan akurasi prediksi tanpa melakukan analisis korelasi antarfitur maupun pengujian multikolinearitas yang berpotensi memengaruhi kestabilan model regresi [5], penelitian terdahulu umumnya belum mengintegrasikan analisis korelasi, pengujian Variance Inflation Factor (VIF), seleksi fitur menggunakan Lasso Regression, serta analisis feature importance menggunakan Random Forest dalam satu kerangka

analisis yang utuh [8], [3]. Keterbatasan-keterbatasan ini menunjukkan bahwa masih diperlukan penelitian yang mengintegrasikan pendekatan statistik dan machine learning secara komprehensif untuk menghasilkan pemahaman yang lebih mendalam tentang faktor-faktor yang memengaruhi performa akademik dan metode prediksi yang paling sesuai untuk dataset dengan karakteristik tertentu

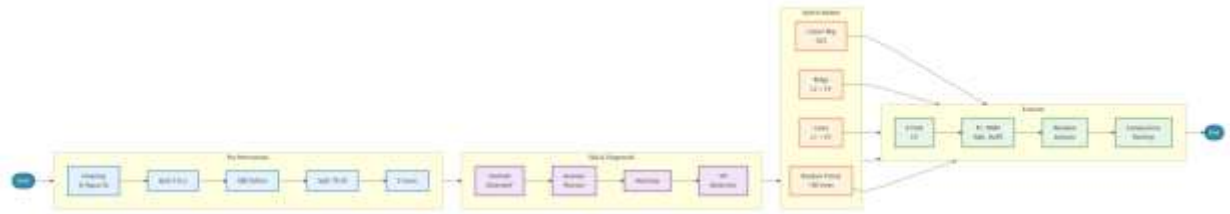
Kebaruan (state of the art) penelitian ini terletak pada integrasi analisis statistik dan machine learning dalam satu kerangka eksperimen yang komprehensif. Berbeda dengan penelitian sebelumnya yang umumnya hanya berfokus pada pembangunan model prediksi, penelitian ini mengombinasikan analisis korelasi Pearson, pengujian multikolinearitas menggunakan Variance Inflation Factor (VIF), seleksi fitur menggunakan Lasso Regression, interpretasi variabel melalui Random Forest Feature Importance, serta perbandingan empat algoritma regresi menggunakan dataset yang sama. Kontribusi baru yang dihasilkan mencakup: (1) kontribusi metodologis berupa kerangka evaluasi komparatif yang terintegrasi, (2) kontribusi empiris berupa identifikasi fitur 29 (Kebiasaan Pendidikan) dan fitur 11 (Pertanyaan Keluarga) sebagai prediktor paling konsisten melalui triangulasi dari berbagai metode, serta (3) kontribusi praktis berupa rekomendasi untuk sistem peringatan dini berbasis kebiasaan pendidikan mahasiswa. Pendekatan tersebut tidak hanya menghasilkan model prediksi, tetapi juga memberikan informasi mengenai faktor-faktor yang paling memengaruhi performa akademik mahasiswa.

Pada penelitian ini digunakan dataset Student Performance yang terdiri atas 145 data mahasiswa dengan 30 variabel aktivitas pembelajaran sebagai variabel independen dan GRADE sebagai variabel dependen. Berdasarkan hasil eksplorasi data, seluruh atribut tidak memiliki missing value, namun ditemukan 22 dari 30 variabel memiliki nilai VIF lebih dari 10, yang menunjukkan adanya tingkat multikolinearitas yang tinggi. Kondisi tersebut mengindikasikan bahwa metode regularisasi seperti Ridge Regression dan Lasso Regression berpotensi memberikan hasil yang lebih stabil dibandingkan regresi linier konvensional. Untuk menguji hal tersebut, penelitian ini membangun model prediksi menggunakan empat algoritma, yaitu Linear Regression, Ridge Regression, Lasso Regression, dan Random Forest Regression, dengan evaluasi performa menggunakan Coefficient of Determination (R^2), Adjusted R^2, Root Mean Square Error (RMSE), dan Mean Absolute Error (MAE).

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk menganalisis dan membandingkan performa Linear Regression, Ridge Regression, Lasso Regression, dan Random Forest Regression dalam memprediksi performa mahasiswa berdasarkan 30 variabel aktivitas pembelajaran, mengidentifikasi faktor-faktor yang paling berpengaruh terhadap capaian akademik mahasiswa, serta menentukan model terbaik berdasarkan nilai R^2, RMSE, dan MAE sebagai dasar pengembangan sistem pendukung keputusan pada bidang pendidikan tinggi.

Metode

Data mining adalah proses penggunaan statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstrak dan merangkum informasi untuk penggunaan praktis [11]. Data mining didefinisikan sebagai proses menemukan pola dalam data. Berdasarkan fungsinya, data mining dapat dikategorikan menjadi deskripsi, estimasi, prediksi, klasifikasi, klasterisasi, dan asosiasi [12]. Algoritma data mining membantu mengeksplorasi data dan memfasilitasi penemuan informasi [13]. Penelitian ini terdiri dari empat tahapan: pra-pemrosesan data, analisis data eksplorasi dan diagnostik multikolinearitas, pembangunan model menggunakan pendekatan hybrid regresi dan machine learning, serta pengukuran performa sistem. Diagram alir proses diilustrasikan pada Gambar 1.



Gambar 1. Diagram Alir Penelitian

A. Dataset

Setiap Penelitian ini menggunakan Higher Education Students Performance Evaluation Dataset yang diperoleh dari UCI Machine Learning Repository pada tautan ***https://archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation***. Dataset ini berisi data yang dikumpulkan dari mahasiswa program sarjana di Turki, terdiri dari beberapa atribut dengan satu atribut yang berfungsi sebagai label. Atribut-atribut tersebut mencakup berbagai metrik aktivitas belajar seperti frekuensi mengangkat tangan di kelas, kehadiran, jam belajar, partisipasi grup diskusi, posting forum, frekuensi mengunjungi pengumuman mata kuliah, dan aktivitas persiapan ujian. Label pada dataset ini adalah GRADE, yang merupakan nilai akhir mahasiswa pada mata kuliah tertentu dengan skala ordinal mulai dari 0 (Gagal/FF) hingga 7 (Lulus dengan predikat tertinggi/AA). Dataset ini juga mencakup kolom Student ID dan Course ID yang dihapus selama pra-pemrosesan karena merupakan pengenal unik yang tidak memiliki nilai prediktif terhadap target.

B. Preprocessing

Dataset yang diperoleh harus diproses terlebih dahulu untuk memastikan keakuratan informasi. Dalam penelitian ini, empat tahapan preprocessing dilakukan: pembersihan data, pemisahan fitur dan target, deteksi outlier, dan standarisasi fitur. Tahap pertama adalah pembersihan data, yang sangat penting karena dataset mungkin mengandung data yang tidak lengkap, tidak relevan, atau tidak akurat. Preprocessing juga membantu menghilangkan potensi masalah seperti missing values [14]. Dalam penelitian ini, pembersihan data dilakukan dengan mengidentifikasi nilai yang hilang menggunakan fungsi isnull().sum() dan menghapus data yang tidak diperlukan. Kolom Student ID dan Course ID dihapus karena merupakan pengenal unik yang tidak berkontribusi dalam memprediksi variabel target. Tahap kedua adalah pemisahan fitur dan target. Dataset dipisahkan menjadi variabel independen (X) yang terdiri dari semua kolom kecuali STUDENT ID, COURSE ID, dan GRADE, serta variabel dependen (y) yaitu kolom GRADE. Tahap ketiga adalah deteksi outlier. Outlier dideteksi menggunakan metode Interquartile Range (IQR). Metode ini menghitung kuartil pertama (Q1) dan kuartil ketiga (Q3), kemudian menghitung IQR = Q3 - Q1. Batas bawah ditentukan sebagai Q1 - 1,5 × IQR, dan batas atas ditentukan sebagai Q3 + 1,5 × IQR. Titik data yang berada di luar rentang [batas bawah, batas atas] diidentifikasi sebagai outlier. Deteksi outlier diterapkan pada variabel target (GRADE) dan seluruh fitur independen (X) menggunakan fungsi detect_outliers_iqr(). Jumlah outlier yang terdeteksi pada setiap fitur dicatat dan dilaporkan untuk analisis lebih lanjut. Tahap keempat adalah pembagian data. Dataset dibagi menjadi data training (70%) dan data testing (30%) menggunakan Stratified Shuffle Split dengan parameter n_splits=1, test_size=0.3, dan random_state=42. Stratifikasi dilakukan dengan mengkategorikan GRADE ke dalam empat kelompok menggunakan fungsi pd.cut() dengan batas kategori (bins) = [-1, 2, 4, 6, 8] dan label = [0, 1, 2, 3] untuk memastikan distribusi variabel target yang seimbang di kedua set data. Hasil pembagian data berupa X_train, X_test, y_train, dan y_test. Tahap kelima adalah standarisasi fitur. Tujuan dari standarisasi adalah untuk mentransformasi data sehingga setiap fitur memiliki rata-rata 0 dan standar deviasi 1. Tahap ini penting untuk model yang sensitif terhadap skala data, seperti Regresi Linier, Ridge Regression, dan Lasso Regression. Standarisasi dilakukan menggunakan StandardScaler yang menerapkan normalisasi Z-score [14]. Standardisasi di-fit pada data training menggunakan fit_transform() dan kemudian diterapkan untuk mentransformasi data testing menggunakan transform() guna mencegah kebocoran data. Hasil standarisasi adalah X_train_scaled dan X_test_scaled. Random Forest tidak memerlukan standarisasi karena algoritma berbasis decision tree tidak sensitif terhadap skala data.

Hasil verifikasi menunjukkan bahwa seluruh 145 data tidak memiliki missing value pada semua fitur, sehingga tidak diperlukan imputasi data. Variabel target GRADE memiliki distribusi yang cukup bervariasi dengan 8 kategori nilai (0 hingga 7), dimana frekuensi masing-masing kategori bervariasi dari 8 (5,52%) untuk nilai 0 hingga 35 (24,14%) untuk nilai 1. Meskipun terdapat variasi frekuensi, ketidakseimbangan data tidak menjadi isu kritis karena GRADE bersifat ordinal dan tidak biner, serta penggunaan Stratified Shuffle Split (dengan pengelompokan GRADE ke dalam 4 kategori untuk keperluan stratifikasi) telah menjaga proporsi kelas GRADE tetap seimbang di kedua set data (training dan testing).

C. Analisis Data Eksplorasi dan Diagnostik Multikolinearitas

Sebelum membangun model prediksi, analisis data eksplorasi dilakukan untuk memahami hubungan antar variabel dan mengidentifikasi potensi masalah yang dapat mempengaruhi performa model.

Analisis statistik deskriptif dilakukan pada seluruh fitur dan variabel target untuk memahami karakteristik data. Analisis ini mencakup informasi umum dataset menggunakan df.info(), ukuran pemusatan dan penyebaran data menggunakan df.describe(), serta pengecekan nilai yang hilang menggunakan df.isnull().sum(). Distribusi frekuensi nilai GRADE dihitung menggunakan value_counts() dan divisualisasikan dalam bentuk count plot menggunakan sns.countplot(). Label jumlah ditambahkan pada setiap batang untuk memudahkan pembacaan.

Analisis korelasi Pearson digunakan untuk mengukur kekuatan dan arah hubungan linier antara setiap fitur dengan variabel target GRADE. Koefisien korelasi (r) dan p-value dihitung menggunakan fungsi pearsonr() dari SciPy untuk setiap kolom pada data X terhadap y. Fitur dengan p-value < 0,05 dianggap memiliki korelasi yang signifikan secara statistik dengan GRADE. Hasil analisis disajikan dalam bentuk tabel yang diurutkan dari korelasi tertinggi ke terendah dan divisualisasikan menggunakan bar plot horizontal dengan sns.barplot(). Garis referensi pada koefisien korelasi 0 dan ±0,2 ditambahkan untuk menunjukkan batas korelasi lemah.

Heatmap korelasi dibuat menggunakan 15 fitur dengan korelasi tertinggi terhadap GRADE. Matriks korelasi dihitung menggunakan X[top_features].corr() dan divisualisasikan menggunakan sns.heatmap() dengan matriks segitiga atas (np.triu) untuk menghindari duplikasi informasi. Nilai korelasi yang tinggi antar fitur independen (mendekati 1 atau -1) mengindikasikan adanya multikolinearitas.

Multikolinearitas antar fitur independen dideteksi menggunakan Variance Inflation Factor (VIF). VIF mengukur seberapa besar varians dari koefisien regresi meningkat akibat adanya korelasi antar fitur independen. VIF untuk fitur ke-i dihitung sebagai $VIF = 1 / (1 - R_i^2)$, dimana R_i^2 adalah koefisien determinasi dari regresi fitur ke-i terhadap seluruh fitur independen lainnya. Dalam penelitian ini, VIF dihitung pada seluruh data X (data asli sebelum pembagian data training dan testing) menggunakan fungsi calculate_vif() yang memanfaatkan variance_inflation_factor dari statsmodels. Fungsi ini menghitung VIF untuk setiap fitur dan mengurutkannya dari nilai tertinggi ke terendah. Nilai

VIF sebesar 1 menunjukkan tidak ada korelasi dengan fitur lain, nilai antara 1 dan 5 menunjukkan multikolinearitas moderat, nilai antara 5 dan 10 menunjukkan multikolinearitas tinggi, dan nilai lebih dari 10 menunjukkan multikolinearitas sangat tinggi. Fitur dengan VIF > 10 diidentifikasi dan dilaporkan sebagai fitur dengan multikolinearitas tinggi yang dapat mempengaruhi stabilitas koefisien regresi linier. Hasil VIF digunakan sebagai dasar pemilihan metode regularisasi (Ridge dan Lasso) yang dapat menangani masalah multikolinearitas.

D. Pendekatan Hybrid: Regresi dan Machine Learning

Penelitian ini menggunakan pendekatan hybrid yang menggabungkan metode regresi dan *machine learning* untuk membangun model prediksi performa akademik mahasiswa. Empat model dibangun: Regresi Linier Berganda, Ridge Regression, Lasso Regression, dan Random Forest Regressor.

1. Regresi Linier Berganda: Regresi linier berganda adalah metode statistik yang digunakan untuk memodelkan hubungan linier antara satu variabel dependen dengan dua atau lebih variabel independen [15]. Model ini digunakan sebagai baseline untuk mengukur seberapa baik variasi GRADE dapat dijelaskan oleh kombinasi linier dari fitur-fitur aktivitas belajar. Model dinyatakan sebagai $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + ... + \beta_nX_n + \varepsilon$, dimana Y adalah variabel target (GRADE), β_0 adalah intersep, β_1 hingga β_n adalah koefisien regresi untuk setiap fitur, X_1 hingga X_n adalah fitur-fitur aktivitas belajar, dan ε adalah error term. Model dilatih menggunakan data training yang telah distandarisasi (X_train_scaled) dengan metode Ordinary Least Squares (OLS) melalui LinearRegression() dari scikit-learn. Prediksi dilakukan pada data training dan testing yang telah distandarisasi. Koefisien regresi yang dihasilkan menunjukkan besaran dan arah pengaruh setiap fitur terhadap GRADE. Koefisien diurutkan dari nilai tertinggi ke terendah dan 10 koefisien teratas ditampilkan.
2. Ridge Regression (Regularisasi L2): Ridge Regression memperluas regresi linier dengan menambahkan penalti L2 pada fungsi loss untuk mengatasi multikolinearitas dan mencegah overfitting [16]. Fungsi objektif meminimalkan jumlah kuadrat residual ditambah α kali jumlah kuadrat koefisien, dimana α (alpha) adalah parameter regularisasi yang mengontrol kekuatan penalti. Ketika α = 0, model setara dengan regresi linier biasa. Seiring meningkatnya α, semua koefisien menyusut mendekati nol tetapi tidak pernah menjadi tepat nol. Ridge mempertahankan semua fitur dalam model sambil mengecilkan fitur yang kurang penting. Nilai α optimal dipilih menggunakan RidgeCV dengan 5-fold cross-validation (cv=5) pada rentang kandidat alpha [0,001, 0,01, 0,1, 1,0, 10,0, 100,0] dengan R^2 sebagai metrik scoring. Alpha dengan rata-rata R^2 cross-validation tertinggi dipilih sebagai parameter optimal. Model Ridge dibangun menggunakan Ridge(alpha=best_alpha_ridge) dan dilatih pada data training yang telah distandarisasi.
3. Lasso Regression (Regularisasi L1): Lasso (Least Absolute Shrinkage and Selection Operator) Regression menggunakan penalti L1 yang memiliki kemampuan untuk melakukan seleksi fitur otomatis dengan mereduksi koefisien fitur yang tidak signifikan menjadi tepat nol [17]. Fungsi objektif meminimalkan jumlah kuadrat residual ditambah α kali jumlah nilai absolut koefisien. Semakin besar α, semakin banyak koefisien yang menjadi tepat nol. Fitur dengan koefisien nol secara otomatis tereliminasi dari model, menghasilkan model yang sparse dan lebih mudah diinterpretasikan. Nilai α optimal dipilih menggunakan LassoCV dengan 5-fold cross-validation (cv=5) pada rentang kandidat alpha [0,0001, 0,001, 0,01, 0,1, 1,0]. Parameter max_iter=10000 digunakan untuk memastikan konvergensi, dan random_state=42 untuk reprodusibilitas. Model Lasso dibangun menggunakan Lasso(alpha=best_alpha_lasso, max_iter=10000, random_state=42) dan dilatih pada data training yang telah distandarisasi. Fitur dengan koefisien tidak nol diidentifikasi sebagai fitur yang dipilih Lasso dan dilaporkan jumlah serta daftarnya.
4. Random Forest Regressor: Random Forest adalah algoritma ensemble learning berbasis bagging yang membangun sejumlah besar decision tree dan menggabungkan prediksinya untuk menghasilkan output yang lebih akurat dan stabil [18]. Algoritma ini bekerja dengan mengambil sampel bootstrap dari data training dan memilih subset acak dari fitur pada setiap node split. Jumlah fitur yang dipertimbangkan untuk setiap split ditentukan oleh parameter max_features='sqrt'. Setiap decision tree dibangun hingga mencapai kriteria penghentian yang ditentukan.

Hasil akhir prediksi adalah rata-rata dari prediksi seluruh pohon. Parameter yang digunakan dalam penelitian ini adalah n_estimators=150, max_depth=12, min_samples_split=5, min_samples_leaf=2, max_features='sqrt', rando m_state=42, dan n_jobs=-1. Random Forest tidak memerlukan standarisasi data karena algoritma decision tree membuat keputusan berdasarkan threshold pada nilai fitur yang tidak terpengaruh oleh skala. Oleh karena itu, model dilatih menggunakan data asli (X_train, y_train) tanpa standarisasi. Random Forest menghasilkan feature importance yang mengukur kontribusi relatif setiap fitur terhadap prediksi. Metode yang digunakan adalah Mean Decrease in Impurity (MDI), yang menghitung rata-rata penurunan impurity yang dihasilkan oleh setiap fitur di seluruh pohon dalam ensemble. Nilai feature importance diambil dari atribut feature_importances_ dari model Random Forest yang telah dilatih, kemudian diurutkan dari nilai tertinggi ke terendah. Sepuluh fitur dengan nilai importance tertinggi ditampilkan, dan 15 fitur teratas divisualisasikan dalam bentuk bar plot horizontal.

E. Validasi Model dengan Cross-Validation

Untuk memastikan model Random Forest memiliki performa yang stabil dan tidak overfitting dilakukan validasi menggunakan 5-fold Cross-Validation pada data training. Data training (X_train, y_train) dibagi menjadi 5 bagian

yang sama besar secara acak. Model dilatih pada 4 bagian dan divalidasi pada 1 bagian yang tersisa. Proses ini diulang sebanyak 5 kali sehingga setiap bagian menjadi data validasi tepat satu kali. Rata-rata dan standar deviasi dari skor R^2 dihitung dari kelima iterasi. Cross-validation dilakukan menggunakan fungsi cross_val_score() dengan parameter cv=5 dan scoring='r2'. Hasil cross-validation berupa rata-rata R^2, standar deviasi R^2, dan skor individual untuk setiap fold ditampilkan.

F. Pengukuran Performa Sistem

Pengujian dalam penelitian ini dilakukan menggunakan perhitungan yang berasal dari metrik evaluasi untuk model regresi, yang mengevaluasi performa model prediksi [24]. Metrik evaluasi digunakan untuk menghitung nilai R^2, Adjusted R^2, RMSE, MAE, dan MAPE [25],[26],[27] dan memberikan wawasan tentang jenis kesalahan yang dibuat oleh model prediksi [28]. Fungsi evaluate_model() dibuat untuk menghitung seluruh metrik evaluasi secara otomatis dengan parameter y_true, y_pred, model_name, dataset_type, dan n_features.

R^2 atau Koefisien Determinasi mengukur proporsi varians variabel target yang dapat dijelaskan oleh model. Semakin mendekati 1, semakin baik performa model [31]. R^2 dihitung menggunakan (1).

$$R2 = 1 - \sum i = 1n(yi - y^-)2\sum i = 1n(yi - y\hat{}i)2 \quad (1)$$

Adjusted R^2 merupakan modifikasi dari R^2 yang memberikan penalti untuk penambahan fitur yang tidak signifikan, sehingga lebih adil untuk membandingkan model dengan jumlah fitur berbeda [32]. Adjusted R^2 dihitung menggunakan (2)

$$Adjusted\ R2 = 1 - n - k - 1(1 - R2)(n - 1) \quad (2)$$

dimana n adalah jumlah sampel dan k adalah jumlah fitur. Jika parameter n_features tidak diberikan, jumlah fitur default diambil dari X.shape[1].

RMSE (Root Mean Squared Error) mengukur besarnya rata-rata kesalahan prediksi dalam satuan yang sama dengan variabel target. RMSE memberikan penalti lebih besar pada kesalahan besar karena menggunakan kuadrat error [33]. RMSE dihitung menggunakan (3).

$$RMSE = n1\sum i = 1n(yi - y^i)2 \quad (3)$$

RMSE diperoleh dengan menghitung MSE menggunakan mean_squared_error() terlebih dahulu, kemudian mengambil akar kuadratnya menggunakan np.sqrt().

MAE (Mean Absolute Error) mengukur rata-rata nilai absolut dari kesalahan prediksi. MAE memberikan interpretasi langsung dalam satuan yang sama dengan variabel target. MAE dihitung menggunakan (4).

$$MAE = n1\sum i = 1n \mid yi - y\hat{}i \mid \quad (4)$$

MAE dihitung menggunakan fungsi mean_absolute_error().

MAPE (Mean Absolute Percentage Error) mengukur akurasi prediksi dalam bentuk persentase. MAPE dihitung menggunakan (5).

$$MAPE = n1\sum i = 1nyiyi - y\hat{}i \times 100\% \quad (5)$$

MAPE hanya dapat dihitung jika seluruh nilai aktual tidak sama dengan nol. Karena GRADE memiliki nilai 0 (Fail/FF), perhitungan MAPE dalam penelitian ini dilakukan dengan pengecekan kondisi menggunakan (y_true != 0).all(). Jika terdapat nilai aktual sama dengan 0, MAPE tidak dihitung dan dikembalikan sebagai NaN (Not a Number) untuk menghindari pembagian dengan nol.

Evaluasi performa model dilakukan pada data training dan data testing untuk setiap model. Metrik pada data testing digunakan untuk menilai performa model sebenarnya pada data yang belum pernah dilihat, sedangkan metrik pada data training digunakan untuk mendeteksi adanya overfitting. Overfitting terdeteksi jika model memiliki performa sangat baik pada data training namun performa jauh menurun pada data testing.

G. Analisis Diagnostik Regresi

Analisis diagnostik regresi dilakukan pada model Regresi Linier Berganda menggunakan data testing. Residual dihitung sebagai selisih antara nilai aktual (y_test) dan nilai prediksi (y_test_pred_lr).

Residual vs Predicted Plot dibuat menggunakan scatter() dengan sumbu X berupa nilai prediksi dan sumbu Y berupa nilai residual. Garis horizontal merah pada y=0 ditambahkan sebagai referensi. Plot ini digunakan untuk memeriksa asumsi linearitas dan homoskedastisitas. Residual yang tersebar secara acak di sekitar garis nol tanpa pola tertentu mengindikasikan asumsi terpenuhi. Q-Q Plot dibuat menggunakan stats.probplot() dengan parameter

dist="norm" untuk membandingkan distribusi residual dengan distribusi normal teoritis. Titik-titik yang mengikuti garis diagonal menunjukkan residual terdistribusi normal.

Histogram residual dibuat menggunakan hist() dengan 15 bin untuk menampilkan distribusi frekuensi residual. Garis vertikal merah pada x=0 ditambahkan sebagai referensi. Bentuk yang mendekati lonceng simetris di sekitar nol mengindikasikan normalitas residual.

Scatter plot antara nilai aktual (y_test) dan nilai prediksi (y_test_pred_lr) dibuat untuk mengevaluasi akurasi model secara visual. Garis diagonal merah dari nilai minimum ke maksimum menunjukkan prediksi sempurna ($y = \hat{y}$). Plot ini juga menampilkan nilai R^2 pada judul. Plot prediksi vs aktual juga dibuat untuk model Random Forest menggunakan y_test dan y_test_pred_rf.

H. Perbandingan Komparatif Model (Hybrid Comparison)

Setelah keempat model (Regresi Linier, Ridge, Lasso, Random Forest) dibangun dan dievaluasi, dilakukan perbandingan komparatif untuk menentukan model terbaik. Seluruh metrik (R^2, RMSE, dan MAE untuk Training dan Testing) dikumpulkan dalam satu DataFrame metrics df yang berisi metode dan nilai-nilai metriknya.

Visualisasi komparatif dibuat dalam satu figure dengan tiga subplot berdampingan menggunakan plt.subplots(1, 3). Ketiga subplot menampilkan diagram batang berkelompok (grouped bar chart) yang membandingkan:

1. R^2 Score (Training vs Testing) untuk setiap model pada subplot pertama
2. RMSE (Training vs Testing) untuk setiap model pada subplot kedua
3. MAE (Training vs Testing) untuk setiap model pada subplot ketiga

Setiap batang menampilkan label nilai di atasnya. Sumbu X menampilkan nama metode dengan rotasi 45 derajat. Warna biru (royalblue) digunakan untuk data training dan oranye (darkorange) untuk data testing.

Model dengan R^2 Testing tertinggi ditetapkan sebagai model terbaik. Seluruh model diurutkan berdasarkan performa R^2 Testing dari tertinggi ke terendah dan ditampilkan sebagai ranking.

Analisis fitur penting dilakukan dengan membandingkan lima fitur teratas dari Random Forest (berdasarkan feature importance) dan lima fitur teratas dari Regresi Linier (berdasarkan nilai absolut koefisien). Interseksi antara kedua daftar fitur tersebut dihitung untuk mengidentifikasi fitur-fitur yang konsisten penting di kedua pendekatan.

Hasil dan Diskusi

A. Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini terdiri dari 145 data mahasiswa dengan 33 kolom awal. Setelah dilakukan pra-pemrosesan dan penghapusan kolom STUDENT ID dan COURSE ID, diperoleh 30 fitur sebagai variabel independen dan GRADE sebagai variabel target. Berdasarkan informasi tambahan, fitur-fitur tersebut dikelompokkan menjadi tiga kategori: fitur 1-10 merupakan pertanyaan pribadi, fitur 11-16 mencakup pertanyaan keluarga, dan fitur 17-30 mencakup kebiasaan pendidikan. Seluruh data tidak memiliki missing value dan seluruh fitur bertipe numerik (integer).

Variabel target GRADE memiliki distribusi sebagai berikut: nilai 0 (Fail) sebanyak 8 mahasiswa (5,52%), nilai 1 (DD) sebanyak 35 mahasiswa (24,14%), nilai 2 (DC) sebanyak 24 mahasiswa (16,55%), nilai 3 (CC) sebanyak 21 mahasiswa (14,48%), nilai 4 (CB) sebanyak 10 mahasiswa (6,90%), nilai 5 (BB) sebanyak 17 mahasiswa (11,72%), nilai 6 (BA) sebanyak 13 mahasiswa (8,97%), dan nilai 7 (AA) sebanyak 17 mahasiswa (11,72%). Distribusi ini menunjukkan bahwa mayoritas mahasiswa memperoleh nilai pada rentang 1-3 (55,17%), sedangkan nilai tertinggi (AA) dan terendah (Fail) memiliki frekuensi yang relatif lebih kecil.

Hasil pengecekan *outlier* pada variabel target GRADE menunjukkan tidak terdapat outlier dengan batas bawah -5,0 dan batas atas 11,0. Namun, deteksi *outlier* pada fitur independen menunjukkan adanya *outlier* pada 16 fitur, dengan jumlah *outlier* tertinggi pada fitur 3 (42 outlier), fitur 15 (42 outlier), fitur 18 (46 outlier), fitur 19 (42 outlier), fitur 20 (31 outlier), fitur 22 (35 outlier), dan fitur 24 (22 outlier).

B. Hasil Analisis Korelasi

Analisis korelasi Pearson antara setiap fitur dengan GRADE menunjukkan bahwa dari 30 fitur yang dianalisis, hanya 8 fitur yang memiliki korelasi signifikan (p-value < 0,05) dengan GRADE. Fitur dengan korelasi tertinggi terhadap GRADE adalah:

1. Fitur 2 (Pertanyaan Pribadi): r = 0,3355; p-value = 0,0000
2. Fitur 29 (Kebiasaan Pendidikan): r = 0,3155; p-value = 0,0001
3. Fitur 30 (Kebiasaan Pendidikan): r = 0,2486; p-value = 0,0026
4. Fitur 18 (Kebiasaan Pendidikan): r = 0,1956; p-value = 0,0184
5. Fitur 5 (Pertanyaan Pribadi): r = 0,1674; p-value = 0,0441

Koefisien korelasi yang diperoleh tergolong lemah ($r < 0,5$) untuk seluruh fitur. Hal ini menunjukkan bahwa secara individual, tidak ada fitur yang memiliki hubungan linier yang kuat dengan GRADE. Dari tiga kategori fitur, kebiasaan pendidikan (fitur 17-30) mendominasi korelasi signifikan dengan 4 fitur, diikuti pertanyaan pribadi (fitur 1-10) dengan 3 fitur, dan pertanyaan keluarga (fitur 11-16) dengan 1 fitur.

C. Hasil Diagnostik Multikolinearitas (VIF)

Hasil perhitungan Variance Inflation Factor (VIF) pada data X (sebelum split) menunjukkan adanya multikolinearitas yang sangat tinggi di antara fitur-fitur independen. Dari 30 fitur yang dianalisis, 22 fitur memiliki nilai VIF > 10, yang merupakan indikator multikolinearitas tinggi. Fitur dengan VIF tertinggi adalah fitur 25 (VIF = 30,9097), fitur 4 (VIF = 29,2781), fitur 30 (VIF = 24,9885), fitur 19 (VIF = 22,6759), dan fitur 27 (VIF = 22,0567). Multikolinearitas tinggi terjadi di semua kategori: pertanyaan pribadi (fitur 4, 5, 6, 7, 1, 2, 3), pertanyaan keluarga (fitur 15, 12, 17, 14, 10), dan kebiasaan pendidikan (fitur 25, 30, 19, 27, 18). Hanya 8 fitur yang memiliki VIF di bawah 10, dengan fitur 9 (VIF = 4,6516) dan fitur 8 (VIF = 4,4274) sebagai fitur dengan multikolinearitas paling rendah. Temuan ini menegaskan perlunya penggunaan metode regularisasi seperti Ridge dan Lasso Regression yang dapat menangani masalah multikolinearitas.

D. Hasil Pemodelan Regresi

1. Regresi Linier Berganda

Model Regresi Linier Berganda yang dilatih pada data *training* yang telah distandarisasi menunjukkan performa yang tidak baik. Pada data training, model menghasilkan R^2 sebesar 0,4808 dengan RMSE 1,5907 dan MAE 1,3125. Adjusted R^2 pada data training hanya sebesar 0,2583. Pada data *testing*, performa model sangat buruk dengan R^2 negatif sebesar -0,0902, yang berarti model lebih buruk daripada prediksi menggunakan rata-rata (baseline). RMSE pada data *testing* sebesar 2,2396 dan MAE sebesar 1,7089. Selisih yang sangat besar antara R^2 training (0,4808) dan R^2 testing (-0,0902) menunjukkan adanya overfitting yang parah. Koefisien Regresi Linier dengan nilai absolut tertinggi adalah fitur 5 (0,5775), fitur 14 (0,5456), fitur 29 (0,5380), fitur 11 (0,5176), dan fitur 3 (0,5021).

2. Ridge Regression

Ridge Regression dengan alpha optimal 100,0 menunjukkan performa yang lebih baik dibandingkan Regresi Linier pada data testing. Pada data *training*, model menghasilkan R^2 sebesar 0,3411 dengan RMSE 1,7920 dan MAE 1,5500. Pada data testing, model menghasilkan R^2 positif sebesar 0,1476 dengan RMSE 1,9803 dan MAE 1,6826. Nilai alpha yang tinggi (100,0) menunjukkan bahwa regularisasi yang kuat diperlukan untuk mengatasi multikolinearitas. Ridge Regression berhasil mengurangi *overfitting*, terlihat dari selisih R^2 training dan testing yang lebih kecil (0,1935).

3. Lasso Regression

Lasso Regression dengan alpha optimal 0,1 menunjukkan performa terbaik di antara semua model pada data *testing*. Pada data training, model menghasilkan R^2 sebesar 0,4119 dengan RMSE 1,6929 dan MAE 1,4446. Pada data testing, model menghasilkan R^2 positif sebesar 0,1535 dengan RMSE 1,9735 dan MAE 1,5905.

Lasso Regression melakukan seleksi fitur otomatis dengan mereduksi koefisien beberapa fitur menjadi nol. Dari 30 fitur, Lasso memilih 20 fitur dengan koefisien tidak nol, yang berarti 10 fitur dieliminasi karena dianggap tidak signifikan. Berdasarkan kategori:

- Pertanyaan Pribadi (fitur 1-10): 8 fitur dipertahankan, 2 fitur dieliminasi
- Pertanyaan Keluarga (fitur 11-16): 4 fitur dipertahankan, 2 fitur dieliminasi
- Kebiasaan Pendidikan (fitur 17-30): 8 fitur dipertahankan, 6 fitur dieliminasi

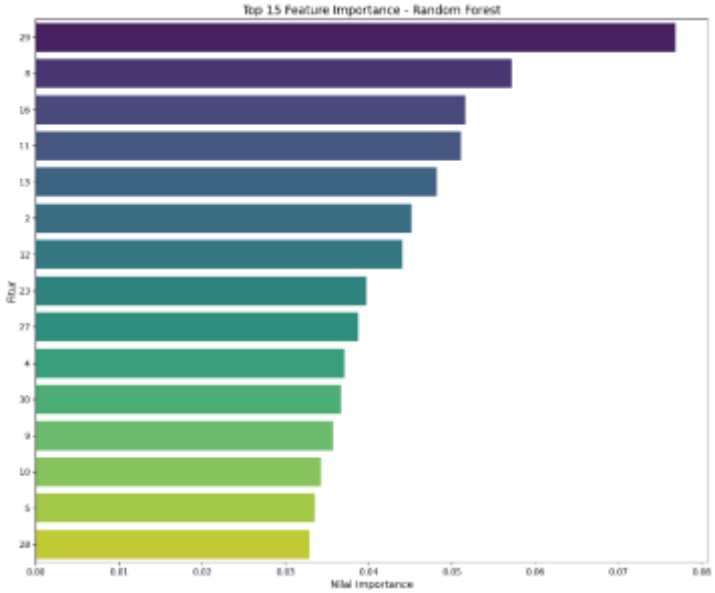
Fitur dengan koefisien positif tertinggi pada Lasso adalah fitur 5 (0,3730), fitur 3 (0,3188), fitur 14 (0,3133), fitur 2 (0,3027), dan fitur 11 (0,2607). Fitur dengan koefisien negatif tertinggi adalah fitur 1 (-0,3422), fitur 8 (-0,3326), fitur 9 (-0,3285), fitur 16 (-0,2565), dan fitur 21 (-0,2438).

E. Hasil Random Forest Regressor

Model Random Forest Regressor dengan parameter n_estimators=150, max_depth=12 menunjukkan overfitting yang signifikan. Pada data training, model menghasilkan R^2 sebesar 0,6611 (tertinggi di antara semua model) dengan RMSE 1,2852 dan MAE 1,1240. Pada data testing, performa menurun drastis dengan R^2 hanya 0,1346, RMSE 1,9954, dan MAE 1,7245. Selisih R^2 training dan testing sebesar 0,5265 menegaskan adanya overfitting.

Hasil 5-fold Cross-Validation mengkonfirmasi ketidakstabilan model dengan rata-rata R^2 sebesar -0,0108 (negatif) dan standar deviasi 0,0980. Skor individual pada setiap fold bervariasi dari 0,1099 hingga -0,1555, menunjukkan performa model sangat bergantung pada subset data yang digunakan.

Feature Importance dari Random Forest menunjukkan fitur-fitur paling berkontribusi terhadap prediksi: fitur 29 (0,0769), fitur 8 (0,0572), fitur 16 (0,0516), fitur 11 (0,0511), dan fitur 13 (0,0482). Berdasarkan kategori, distribusi feature importance adalah: kebiasaan pendidikan (6 fitur di 10 teratas), pertanyaan pribadi (2 fitur), dan pertanyaan keluarga (2 fitur).



Gambar 2. Top 15 Feature Importance Random Forest Regressor

Berdasarkan Gambar 2, fitur dengan importance tertinggi adalah fitur 29 (0,0769), diikuti oleh fitur 8 (0,0572), fitur 16 (0,0516), fitur 11 (0,0511), dan fitur 13 (0,0482). Berdasarkan kategori, distribusi feature importance pada 10 teratas didominasi oleh kebiasaan pendidikan (fitur 29, 27, 23, 12, 4), diikuti oleh pertanyaan pribadi (fitur 8, 2), dan pertanyaan keluarga (fitur 16, 11, 13).

F. Perbandingan Komparatif Model

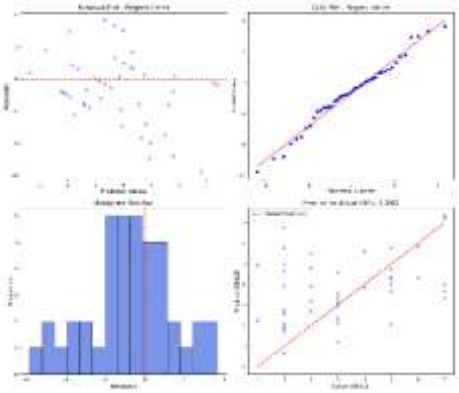
Berdasarkan hasil evaluasi pada data testing, model Lasso Regression menunjukkan performa terbaik dengan R^2 testing sebesar 0,1535, diikuti oleh Ridge Regression (R^2 = 0,1476), Random Forest (R^2 = 0,1346), dan Regresi Linier (R^2 = -0,0902).

Tabel 1. Perbandingan Komparatif Model

Peringkat	*Model*	*R^2 Testing*	*RMSE Testing*	*MAE Testing*
1	Lasso	0,1535	19,735	15,905
2	Ridge	0,1476	19,803	16,826
3	Random Forest	0,1346	19,954	17,245
4	Regresi Linier	-0,0902	22,396	17,089

G. Analisis Fitur Penting

Interseksi antara lima fitur teratas dari Random Forest (feature importance) dan lima fitur teratas dari Regresi Linier (koefisien absolut) menunjukkan bahwa fitur 29 (Kebiasaan Pendidikan) dan fitur 11 (Pertanyaan Keluarga) muncul di kedua daftar. Fitur 29 juga merupakan fitur dengan korelasi Pearson tertinggi kedua terhadap GRADE (r = 0,3155; p-value = 0,0001), semakin menguatkan pentingnya fitur ini. Hasil Analisis Diagnostik Regresi, Analisis diagnostik residual dilakukan pada model Regresi Linier Berganda menggunakan data testing untuk memverifikasi asumsi regresi. Hasil analisis disajikan pada Gambar 2 yang terdiri dari empat plot: (a) Residual vs Predicted Plot, (b) Q-Q Plot, (c) Histogram Residual, dan (d) Scatter Plot Prediksi vs Aktual



Gambar 3. Residual plot diagnostik

Berdasarkan Gambar 2, residual tersebar di sekitar garis horizontal nol tanpa menunjukkan pola sistematis tertentu, yang mengindikasikan bahwa asumsi linearitas dan homoskedastisitas terpenuhi. Gambar 2 menunjukkan Q-Q Plot dimana sebagian besar titik residual mengikuti garis diagonal, meskipun terdapat sedikit penyimpangan pada bagian ekor yang menunjukkan residual mendekati distribusi normal. Gambar 2 menampilkan histogram residual yang mendekati bentuk lonceng dan berpusat di sekitar nol, memperkuat indikasi normalitas residual. Gambar 2 menyajikan scatter plot antara nilai aktual dan prediksi GRADE dengan garis diagonal merah sebagai referensi prediksi sempurna. Titik-titik data tersebar di sekitar garis diagonal, meskipun dengan variasi yang cukup lebar mengingat nilai R^2 testing yang rendah (-0,0902).

H. Pembahasan

Penelitian ini bertujuan untuk memprediksi kinerja siswa pada akhir semester menggunakan teknik pembelajaran mesin dengan memanfaatkan faktor pribadi, keluarga, dan kebiasaan pendidikan. Hasil penelitian menunjukkan bahwa pendekatan hybrid yang mengombinasikan model regresi dengan regularisasi, yaitu Ridge dan Lasso Regression, serta machine learning Random Forest mampu memberikan gambaran yang komprehensif mengenai faktor-faktor yang berkaitan dengan performa akademik. Di antara model yang dibandingkan, Lasso Regression menunjukkan performa terbaik dengan nilai R^2 testing sebesar 0,1535, yang berarti model mampu menjelaskan sekitar 15,35% variasi nilai GRADE pada data testing. Rendahnya performa seluruh model dapat dipengaruhi oleh beberapa kondisi dataset. Dataset hanya terdiri atas 145 mahasiswa dengan 30 fitur, sehingga jumlah data training yang hanya 101 sampel membatasi kemampuan model dalam mempelajari pola yang kompleks. Selain itu, hasil analisis Variance Inflation Factor (VIF) menunjukkan bahwa 22 dari 30 fitur memiliki nilai VIF > 10, yang mengindikasikan adanya multikolinearitas tinggi pada fitur-fitur dalam kategori pribadi, keluarga, maupun kebiasaan pendidikan. Kondisi tersebut dapat menyebabkan ketidakstabilan koefisien pada model regresi standar, sedangkan Ridge dan Lasso mampu mengatasinya melalui mekanisme regularisasi. Korelasi Pearson yang relatif lemah, dengan seluruh nilai $r < 0,5$, juga menunjukkan bahwa hubungan antara faktor-faktor yang diamati dan GRADE tidak sepenuhnya bersifat linier. Selain itu, terdapat indikasi overfitting pada seluruh model, meskipun Ridge menunjukkan selisih R^2 training dan testing yang paling kecil, yaitu 0,1935, sehingga mengindikasikan bahwa regularisasi L2 lebih efektif dalam mengendalikan kompleksitas model pada dataset dengan multikolinearitas tinggi.

Hasil analisis menunjukkan bahwa faktor pribadi, keluarga, dan kebiasaan pendidikan memiliki kontribusi yang berbeda terhadap prediksi GRADE. Pada kelompok faktor pribadi yang terdiri atas fitur 1–10, fitur 2 memiliki korelasi paling tinggi terhadap GRADE dengan nilai $r = 0,3355$ dan signifikan secara statistik. Temuan tersebut diperkuat oleh koefisien positif pada Lasso Regression sebesar 0,3027, yang menunjukkan bahwa peningkatan nilai pada fitur tersebut berkaitan dengan peningkatan GRADE yang diprediksi. Sebaliknya, fitur 1, 8, dan 9 memiliki koefisien negatif pada Lasso masing-masing sebesar -0,3422, -0,3326, dan -0,3285. Hasil tersebut menunjukkan bahwa peningkatan nilai pada karakteristik pribadi yang direpresentasikan oleh fitur-fitur tersebut cenderung berkaitan dengan penurunan GRADE yang diprediksi, meskipun korelasi negatifnya tidak seluruhnya signifikan secara statistik. Pada faktor keluarga yang terdiri atas fitur 11–16, fitur 11 menjadi salah satu fitur yang relatif konsisten karena muncul sebagai fitur penting pada Random Forest dan Regresi Linier serta memiliki koefisien positif pada Lasso sebesar 0,2607. Hal ini menunjukkan bahwa aspek keluarga yang direpresentasikan oleh fitur 11 berpotensi memberikan dukungan terhadap performa akademik. Sebaliknya, fitur 16 memiliki koefisien negatif sebesar -0,2565, yang menunjukkan bahwa peningkatan nilai fitur tersebut berkaitan dengan penurunan GRADE yang diprediksi. Temuan ini memperlihatkan bahwa pengaruh lingkungan keluarga terhadap performa akademik tidak bersifat seragam, karena aspek keluarga tertentu dapat memberikan kontribusi positif, sementara aspek lainnya dapat berkaitan dengan hasil akademik yang lebih rendah.

Pada kelompok kebiasaan pendidikan yang terdiri atas fitur 17–30, pengaruh yang lebih dominan terlihat dibandingkan kelompok faktor lainnya. Fitur 29 merupakan fitur yang paling konsisten karena masuk dalam lima besar fitur penting pada seluruh model dan memiliki korelasi signifikan dengan GRADE sebesar $r = 0,3155$. Fitur 18

juga menunjukkan korelasi positif yang signifikan dengan GRADE sebesar r = 0,1956 dan memiliki koefisien positif pada Lasso sebesar 0,2390. Hasil tersebut menunjukkan bahwa kebiasaan belajar atau aktivitas pendidikan tertentu memiliki hubungan yang lebih kuat dengan performa akademik dibandingkan sebagian faktor pribadi dan keluarga. Namun demikian, tidak seluruh kebiasaan pendidikan memberikan pengaruh positif. Fitur 21 memiliki koefisien negatif sebesar -0,2438 pada Lasso, sehingga menunjukkan bahwa karakteristik yang direpresentasikan oleh fitur tersebut berpotensi berkaitan dengan penurunan GRADE. Dengan demikian, kategori kebiasaan pendidikan menjadi aspek penting yang perlu diperhatikan karena di dalamnya terdapat faktor yang berpotensi mendukung maupun menghambat performa akademik.

Perbandingan model menunjukkan bahwa Ridge dan Lasso Regression lebih efektif dibandingkan Regresi Linier standar dalam menghadapi kondisi multikolinearitas pada dataset. Ridge menggunakan nilai alpha sebesar 100,0, yang menunjukkan bahwa penalti L2 yang relatif kuat diperlukan untuk menstabilkan koefisien model. Sementara itu, Lasso dengan alpha sebesar 0,1 tidak hanya menghasilkan nilai R^2 testing tertinggi, tetapi juga melakukan seleksi fitur dengan mengeliminasi 10 dari 30 fitur yang tersedia. Hasil seleksi tersebut menunjukkan bahwa Lasso mempertahankan 8 dari 10 fitur pada kategori pribadi, 4 dari 6 fitur pada kategori keluarga, dan 8 dari 14 fitur pada kategori kebiasaan pendidikan. Dengan demikian, meskipun beberapa fitur dieliminasi karena kontribusinya yang relatif rendah atau redundan, ketiga kategori faktor tetap memiliki representasi dalam model akhir. Temuan ini memperlihatkan bahwa regularisasi tidak hanya berperan dalam mengurangi dampak multikolinearitas, tetapi juga membantu menghasilkan model yang lebih sederhana dan lebih mudah diinterpretasikan.

Meskipun Random Forest dikenal sebagai salah satu algoritma yang mampu menangani hubungan non-linier dan interaksi antarfitur, performanya pada penelitian ini belum menunjukkan hasil yang optimal. Model Random Forest menghasilkan rata-rata R^2 cross-validation sebesar -0,0108, yang menunjukkan bahwa kemampuan generalisasi model masih rendah pada dataset yang digunakan. Salah satu penyebab yang mungkin adalah ukuran dataset yang relatif kecil, yaitu hanya 101 sampel untuk data training, sementara parameter yang digunakan berupa n_estimators = 150 dan max_depth = 12 menghasilkan model dengan tingkat kompleksitas yang relatif tinggi dibandingkan jumlah data yang tersedia. Kondisi tersebut berpotensi meningkatkan varians model dan menyebabkan model kesulitan melakukan generalisasi terhadap data yang belum pernah dilihat sebelumnya. Hasil ini menunjukkan bahwa pada dataset berukuran kecil dengan multikolinearitas tinggi, model yang lebih sederhana dan dilengkapi regularisasi, seperti Lasso Regression, dapat memberikan hasil yang lebih stabil dibandingkan model ensemble yang lebih kompleks.

Analisis feature importance dan koefisien model menunjukkan adanya beberapa fitur yang secara konsisten berpotensi menjadi prediktor penting terhadap kinerja akademik. Fitur 29 yang berasal dari kategori kebiasaan pendidikan dan fitur 11 yang berasal dari kategori faktor keluarga secara konsisten muncul dalam lima besar fitur penting pada Random Forest dan Regresi Linier. Fitur 29 juga memiliki korelasi positif yang signifikan terhadap GRADE sebesar r = 0,3155 serta koefisien positif pada Lasso sebesar 0,2382. Hasil tersebut mengindikasikan bahwa semakin baik karakteristik kebiasaan pendidikan yang direpresentasikan oleh fitur 29, semakin tinggi GRADE yang diprediksi. Sementara itu, fitur 11 memiliki koefisien positif pada Lasso sebesar 0,2607, yang menunjukkan bahwa aspek keluarga yang direpresentasikan oleh fitur tersebut berpotensi memberikan kontribusi positif terhadap performa akademik. Sebaliknya, fitur 1 dan fitur 8 yang termasuk dalam kategori faktor pribadi memiliki koefisien negatif yang relatif kuat pada Lasso. Temuan tersebut mengindikasikan bahwa karakteristik pribadi tertentu dapat menjadi faktor yang perlu mendapatkan perhatian dalam upaya meningkatkan performa akademik. Secara praktis, hasil ini dapat dimanfaatkan oleh pendidik maupun institusi pendidikan sebagai dasar untuk mengidentifikasi faktor-faktor yang berpotensi mendukung atau menghambat pencapaian akademik, meskipun interpretasi terhadap masing-masing fitur perlu disesuaikan dengan makna pertanyaan atau indikator yang mendasarinya.

Penelitian ini memiliki beberapa keterbatasan yang sekaligus menjadi peluang untuk pengembangan penelitian selanjutnya. Ukuran dataset yang relatif kecil, yaitu 145 data dengan 101 data training, membatasi kemampuan model dalam menangkap pola yang lebih kompleks dan dapat memengaruhi generalisasi hasil penelitian. Meskipun demikian, dataset tersebut telah memberikan gambaran awal mengenai hubungan faktor pribadi, keluarga, dan kebiasaan pendidikan dengan performa akademik. Selain itu, ditemukannya multikolinearitas tinggi pada 22 dari 30 fitur merupakan karakteristik penting dari dataset yang perlu diperhatikan, meskipun permasalahan tersebut telah ditangani menggunakan Ridge dan Lasso Regression. Distribusi GRADE yang berada pada rentang 0 hingga 7 juga mencerminkan variasi hasil akademik, sedangkan penggunaan Stratified Shuffle Split membantu mempertahankan proporsi data pada proses pembagian training dan testing. Metrik Mean Absolute Percentage Error (MAPE) tidak dapat digunakan karena terdapat nilai GRADE sebesar 0, sehingga evaluasi dilakukan menggunakan R^2, Adjusted R^2, RMSE, dan MAE untuk memberikan gambaran yang lebih komprehensif mengenai performa model. Selain itu, outlier ditemukan pada 16 fitur dan tidak dihilangkan karena dianggap sebagai bagian dari karakteristik data yang tersedia, tetapi keberadaannya dapat menjadi fokus analisis lebih lanjut pada penelitian berikutnya. Pengelompokan fitur ke dalam tiga kategori, yaitu faktor pribadi, keluarga, dan kebiasaan pendidikan, juga telah memberikan kerangka interpretasi yang relevan, meskipun ketersediaan label atau deskripsi spesifik untuk setiap fitur akan memungkinkan analisis yang lebih mendalam. Secara keseluruhan, penelitian ini berhasil membangun dan membandingkan beberapa model prediksi performa akademik serta mengidentifikasi fitur-fitur yang relatif konsisten sebagai prediktor penting.

Hasil tersebut dapat menjadi dasar bagi penelitian lanjutan dengan dataset yang lebih besar, fitur yang lebih terdefinisi, serta pendekatan pemodelan yang lebih kompleks untuk meningkatkan kemampuan prediksi dan generalisasi model.

Kesimpulan

Berdasarkan hasil penelitian, pendekatan hybrid yang menggabungkan metode regresi dan machine learning berhasil dibangun untuk memprediksi performa akademik mahasiswa, dengan Lasso Regression sebagai model terbaik yang menghasilkan R^2 testing 0,1535, RMSE 1,9735, dan MAE 1,5905, mengungguli Ridge Regression (R^2 = 0,1476), Random Forest (R^2 = 0,1346), dan Regresi Linier (R^2 = -0,0902). Teridentifikasi multikolinearitas sangat tinggi pada 22 dari 30 fitur (VIF > 10) yang menyebabkan Regresi Linier standar mengalami overfitting parah, sementara regularisasi Ridge (alpha 100,0) dan Lasso (alpha 0,1) berhasil mengatasi masalah tersebut. Fitur 29 (Kebiasaan Pendidikan) dan fitur 11 (Pertanyaan Keluarga) teridentifikasi sebagai prediktor paling konsisten yang muncul di lima besar pada model Random Forest dan Regresi Linier, dengan fitur 29 memiliki korelasi positif signifikan terhadap GRADE (r = 0,3155), menegaskan bahwa kebiasaan pendidikan merupakan faktor dominan dalam menentukan performa akademik. Dampak kebiasaan pendidikan terhadap performa akademik ditunjukkan oleh dominasi fitur 29 (Kebiasaan Pendidikan) sebagai prediktor paling konsisten dengan korelasi positif signifikan terhadap GRADE (r = 0,3155) dan koefisien positif pada Lasso (0,2382), yang berarti semakin baik kebiasaan pendidikan mahasiswa, semakin tinggi nilai akhir yang diperoleh. Temuan ini menegaskan bahwa faktor kebiasaan pendidikan lebih dominan dibandingkan faktor pribadi dan keluarga dalam menentukan keberhasilan akademik. Dari data ini, beberapa hal yang dapat dibuat dan ditindaklanjuti adalah: (1) sistem peringatan dini (early warning system) untuk mengidentifikasi mahasiswa yang berpotensi memperoleh nilai rendah berdasarkan pola kebiasaan pendidikan mereka, sehingga institusi dapat memberikan intervensi tepat waktu, (2) rekomendasi personalisasi pembelajaran yang disesuaikan dengan profil kebiasaan pendidikan masing-masing mahasiswa, (3) panduan pengembangan diri bagi mahasiswa untuk fokus meningkatkan aspek kebiasaan pendidikan yang terbukti berkorelasi positif dengan prestasi akademik.

Deklarasi Kepentingan

Penulis menyatakan tidak terdapat benturan kepentingan (conflict of interest) dalam penelitian dan penulisan artikel ini. Penelitian ini dilakukan secara independen tanpa adanya pengaruh finansial, hubungan personal, atau afiliasi organisasi yang dapat memengaruhi hasil dan interpretasi data yang disajikan.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Institut Teknologi dan Bisnis Nobel Indonesia dan Divisi Penerbitan & Publikasi, PT. Irmex Digital Akademika atas dukungan fasilitas serta akses terhadap sumber daya yang telah memungkinkan terlaksananya penelitian ini. Ucapan terima kasih juga disampaikan kepada tim peneliti dan rekan-rekan yang telah memberikan masukan konstruktif selama proses penulisan artikel ini.

Ketersediaan Data

Dataset yang digunakan dalam penelitian ini adalah Higher Education Students Performance Evaluation Dataset yang tersedia secara publik melalui UCI Machine Learning Repository pada tautan berikut: https://archive.ics.uci.edu/dataset/856/higher+education+students+performance+evaluation. Dataset tersebut digunakan sesuai dengan ketentuan penggunaan yang ditetapkan oleh pemilik dataset. Informasi lengkap mengenai atribut dan struktur data telah dijelaskan pada bagian Metode artikel ini. Penulis tidak melakukan modifikasi terhadap dataset selain proses pra-pemrosesan standar (penghapusan kolom identitas unik, standarisasi, dan pembagian data) yang telah diuraikan secara rinci dalam naskah. Karena dataset ini bersifat publik dan telah disitasi dengan benar, maka tidak diperlukan izin tambahan untuk penggunaannya dalam penelitian ini.

Penggunaan Ai Dan Deklarasi Penggunaan Ai Generatif

Penulis menyatakan bahwa dalam proses penyusunan artikel ini, telah digunakan perangkat berbasis kecerdasan buatan generatif, yaitu DeepSeek, yang dimanfaatkan secara terbatas sebagai alat bantu untuk parafrase teks dalam bahasa Indonesia dan Inggris guna meningkatkan kualitas kebahasaan tanpa mengubah substansi ilmiah, serta untuk perbaikan kode program yang diterapkan pada lingkungan pengembangan PyCharm, meliputi saran optimasi, identifikasi kesalahan sintaks, dan penyesuaian teknis sesuai kebutuhan analisis data. Penulis menegaskan bahwa seluruh keputusan terkait desain penelitian, metodologi, pemilihan algoritma, analisis data, interpretasi hasil, hingga perumusan kesimpulan sepenuhnya merupakan kewenangan dan tanggung jawab penulis, dan peran DeepSeek hanya bersifat sebagai pendukung teknis operasional dan kebahasaan, bukan sebagai pengganti pemikiran kritis manusia. Setelah memanfaatkan alat tersebut, seluruh konten naskah dan kode program telah ditinjau, diverifikasi, dan disunting secara manual oleh penulis untuk memastikan kesesuaian substansi, koherensi logika, akurasi ilmiah, dan orisinalitas karya. Penulis bertanggung jawab penuh atas seluruh isi artikel yang dipublikasikan.

Daftar Pustaka

[1] M. Ardianti, O. D. Nurhayati, and B. Warsito, "Model Prediksi Kinerja Siswa Berdasarkan Data Log LMS Menggunakan Ensemble Machine Learning," *JST (Jurnal Sains dan Teknologi)*, vol. 12, no. 3, pp. 562–571, 2023, doi: 10.23887/jstundiksha.v12i3.59816.

[2] Y. N. Sukmaningtyas, R. M. Akbar, and G. R. U. Asyafiiyah, "Penerapan Predictive Analytics untuk Analisis Faktor-faktor yang Mempengaruhi Performa Akademik Siswa," *Arcitech: Journal of Computer Science and Artificial Intelligence*, vol. 4, no. 2, pp. 127–145, Dec. 2024, doi: 10.29240/arcitech.v4i2.12048.

[3] M. Hooda, C. Rana, O. Dahiya, A. Rizwan, and M. S. Hossain, "Artificial Intelligence for Assessment and Feedback to Enhance Student Success in Higher Education," *Mathematical Problems in Engineering*, vol. 2022, no. 1, p. 5215722, 2022, doi: 10.1155/2022/5215722.

[4] E. N. Nst, Sumijan, and G. W. Nurcahyo, "PENERAPAN MACHINE LEARNING MENGGUNAKAN ALGORITMA DECISION TREE UNTUK PREDIKSI TINGKAT KELULUSAN MAHASISWA," *Jurnal Sistem Informasi dan Informatika (Simika)*, vol. 8, no. 2, pp. 428–439, Aug. 2025, doi: 10.47080/simika.v8i2.3958.

[5] E. Ahmed, "Student Performance Prediction Using Machine Learning Algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2024, no. 1, p. 4067721, 2024, doi: 10.1155/2024/4067721.

[6] "View of Using Machine Learning Algorithms to Predict Students' Performance and Improve Learning Outcome: A Literature Based Review." Accessed: Jul. 29, 2026. [Online]. Available: https://stratfordjournalpublishers.org/journals/index.php/Journal-of-Information-and-Techn/article/view/480/1870

[7] M. B. Roslan and C. Chen, "Educational Data Mining for Student Performance Prediction: A Systematic Literature Review (2015-2021)," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 17, no. 5, pp. 147–179, Mar. 2022.

[8] A. Villar and C. R. V. De Andrade, "Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study," *Discov Artif Intell*, vol. 4, no. 1, p. 2, Jan. 2024, doi: 10.1007/s44163-023-00079-z.

[9] Ijsrcseit and Dr. Harikrishna Jethva, "International Journal of Scientific Research in Computer Science, Engineering and Information Technology," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol*, vol. 1, no. 1, p. 1, Jan. 2017, doi: 10.32628/IJSRCSEIT.

[10] A. A. Saa, "Educational Data Mining & Students' Performance Prediction," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 5, 2016.

[11] J. M. Amigo, "Data Mining, Machine Learning, Deep Learning, Chemometrics. Definitions, common points and Trends (Spoiler Alert: VALIDATE your models!)," *Braz. J. Anal. Chem.*, vol. 8, no. 32, pp. 45–61, May 2021, doi: 10.30744/brjac.2179-3425.AR-38-2021.

[12] F. Gullo, "From Patterns in Data to Knowledge Discovery: What Data Mining Can Do," *Physics Procedia*, vol. 62, pp. 18–22, Jan. 2015, doi: 10.1016/j.phpro.2015.02.005.

[13] X. Shu and Y. Ye, "Knowledge Discovery: Methods from data mining and machine learning," *Social Science Research*, vol. 110, p. 102817, Feb. 2023, doi: 10.1016/j.ssresearch.2022.102817.

[14] A. Widianti and I. Pratama, "PENANGANAN MISSING VALUES DAN PREDIKSI DATA TIMBUNAN SAMPAH BERBASIS MACHINE LEARNING: HANDLING MISSING VALUES AND PREDICTION OF WASTE PILE DATA BASED ON MACHINE LEARNING | Rabit : Jurnal Teknologi dan Sistem Informasi Univrab," Jul. 2024, Accessed: Jul. 29, 2026. [Online]. Available: https://jurnal.univrab.ac.id/index.php/rabit/article/view/4789

[15] S. Sulantari, W. Hariadi, E. D. Putra, and A. Anas, "Analisis Regresi Linier Berganda Untuk Memodelkan Faktor yang Mempengaruhi Nilai Penambahan Utang Tahunan Negara Indonesia," *UJMC (Unisda Journal of Mathematics and Computer Science)*, vol. 10, no. 1, pp. 36–46, Jun. 2024, doi: 10.52166/ujmc.v10i1.6631.

[16] S. R. Jannah, A. Rusyana, and E. Ramadhani, "Perbandingan Metode Regresi OLS, Ridge, dan Lasso dalam Memodelkan Kemiskinan di Indonesia Tahun 2024," *Journal of Data Analysis*, pp. 93–100, Dec. 2025, doi: 10.24815/jda.v8i2.365.

[17] A. F. Marwa, S. A. Setiyawan, Y. T. N. Cahyani, and H. D. Cahyono, "OPTIMIZATION OF STOCK PRICE PREDICTION WITH RIDGE REGRESSION AND HYPERPARAMETER SELECTIONS," *Jurnal Teknik Informatika (Jutif)*, vol. 6, no. 1, pp. 141–148, 2025, doi: 10.52436/1.jutif.2025.6.1.2384.

[18] A. Mosavi, F. Sajedi Hosseini, B. Choubin, M. Goodarzi, A. A. Dineva, and E. Rafiei Sardooi, "Ensemble Boosting and Bagging Based Machine Learning Models for Groundwater Potential Prediction," *Water Resour Manage*, vol. 35, no. 1, pp. 23–37, Jan. 2021, doi: 10.1007/s11269-020-02704-3.