p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 897-907
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1094

Analisis Komparatif Linear Regression dan Decision Tree untuk Prediksi Skor QS World University Rankings 2025

Dyah Puspita Sari[1*], Hafiyyan Putra Pratama[2]

[1,2]Program Studi Sistem Telekomunikasi, Universitas Pendidikan Indonesia Kampus UPI di Purwakarta, Purwakarta, Indonesia
Email: [1*]dyahps01@upi.edu, [2]hafiyyan@upi.edu
(*Email Corresponding Author: dyahps01@upi.edu)
Received: June 2,2026 | Revision: June 5,2026 | Accepted: June 19,2026

Abstrak

Sistem perangkingan universitas dunia telah menjadi tolak ukur global yang krusial dalam mengukur kualitas institusi pendidikan tinggi, produktivitas riset, dan keunggulan akademik. Dataset QS World University Rankings 2025 menyediakan seperangkat indikator evaluasi yang komprehensif, mencakup reputasi akademik, reputasi pemberi kerja, rasio dosen-mahasiswa, sitasi per fakultas, serta berbagai indikator internasionalisasi. Penelitian ini melakukan studi komparatif regresi machine learning untuk memprediksi Overall Score universitas berdasarkan indikator-indikator tersebut. Dua model supervised learning diterapkan, yaitu Regresi Linear dan Decision Tree Regressor. Dataset yang terdiri dari 1.503 entri dan 28 kolom diproses melalui tahapan preprocessing menyeluruh, meliputi penanganan nilai hilang dengan imputasi median, deteksi outlier menggunakan metode IQR, pengkodean variabel kategorikal dengan LabelEncoder, dan normalisasi fitur menggunakan StandardScaler. Data dibagi dengan rasio 80:20 untuk pelatihan dan pengujian. Metrik evaluasi yang digunakan mencakup Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), dan koefisien determinasi (R^2). Hasil penelitian menunjukkan bahwa Regresi Linear secara signifikan mengungguli Decision Tree, dengan capaian R^2 sebesar 0,9985, MAE sebesar 0,3662, dan RMSE sebesar 0,7427. Validasi silang 5-fold mengonfirmasi stabilitas model Regresi Linear dengan R^2 rata-rata 0,9374 ± 0,0668. Analisis feature importance mengidentifikasi Academic Reputation Score sebagai prediktor paling berpengaruh terhadap Overall Score, konsisten dengan temuan analisis korelasi (r = 0,90).

Kata Kunci: Decision Tree, Machine Learning, QS Rankings, Regresi Linear, Supervised Learning

Abstract

University ranking systems have become a critical global benchmark for measuring institutional quality, research productivity, and academic excellence. The QS World University Rankings 2025 dataset provides a comprehensive set of evaluation indicators, including academic reputation, employer reputation, faculty-to-student ratio, citations per faculty, and internationalization metrics. This study conducts a comparative machine learning regression analysis to predict universities' Overall Score based on these indicators. Two supervised learning models were employed: Linear Regression and Decision Tree Regressor. The dataset, comprising 1,503 entries and 28 columns, underwent thorough preprocessing including median imputation, IQR-based outlier detection, label encoding, and StandardScaler normalization. An 80:20 train-test split was applied. Evaluation metrics include MAE, MSE, RMSE, and R^2. Results show that Linear Regression significantly outperformed Decision Tree, achieving $R^2 = 0.9985$, MAE = 0.3662, and RMSE = 0.7427. Five-fold cross-validation confirmed model stability with a mean R^2 of 0.9374 ± 0.0668. Feature importance analysis identified Academic Reputation Score as the dominant predictor (r = 0.90).

Keywords: *Decision Tree, Linear Regression, Machine Learning, QS Rankings, Supervised Learning*

1. PENDAHULUAN

Pendidikan tinggi memainkan peran sentral dalam pembangunan sumber daya manusia dan daya saing bangsa di era globalisasi. Institusi pendidikan tinggi dituntut untuk tidak hanya unggul secara lokal, tetapi juga diakui dalam kancah internasional melalui berbagai sistem penilaian global. Salah satu instrumen paling berpengaruh dalam mengukur kualitas universitas di tingkat dunia adalah sistem perangkingan universitas, di mana QS World University Rankings yang diterbitkan oleh Quacquarelli Symonds (QS) merupakan salah satu yang paling bergengsi dan banyak dirujuk oleh institusi, calon mahasiswa, pembuat kebijakan, dan mitra industri [1].

Edisi QS World University Rankings 2025 mengevaluasi lebih dari 1.500 universitas dari berbagai penjuru dunia berdasarkan serangkaian indikator kinerja yang terstruktur dan terstandarisasi. Edisi ini memperkenalkan indikator baru seperti International Research Network, Employment Outcomes, dan Sustainability, yang merefleksikan perhatian global terhadap relevansi karir lulusan dan keberlanjutan institusi [2]. Setiap indikator memiliki bobot tertentu yang berkontribusi terhadap Overall Score universitas, dengan Academic Reputation memiliki bobot tertinggi sebesar 30%.

Perkembangan pesat dalam bidang kecerdasan buatan dan machine learning telah membuka peluang baru dalam analisis data pendidikan. Machine learning memungkinkan ekstraksi pola tersembunyi dan pembangunan model prediktif

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 897-907
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1094

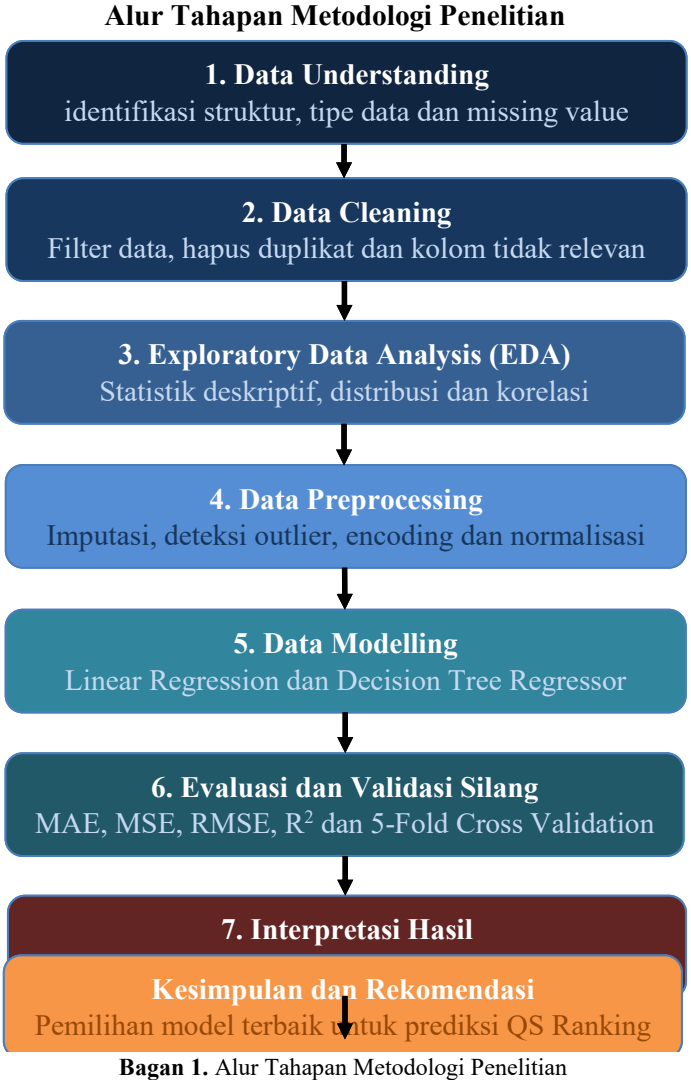
dari dataset berdimensi tinggi, sehingga memberikan interpretasi yang lebih mendalam dari data yang kompleks [3]. Penerapan teknik machine learning dalam konteks pendidikan tinggi telah berkembang pesat, mulai dari prediksi performa mahasiswa, deteksi risiko putus kuliah, hingga prediksi peringkat universitas secara global [4], [5].

Terdapat kesenjangan penelitian mengenai faktor-faktor penentu skor QS dari perspektif machine learning. Sebagian besar studi berfokus pada analisis deskriptif atau statistik konvensional, tanpa mengeksplorasi kapabilitas prediktif model machine learning modern. Penelitian terbaru oleh Basireddy et al. [6] dan Navia-Gamero et al. [7] telah menunjukkan bahwa model machine learning mampu memprediksi skor QS dengan akurasi tinggi; namun, perbandingan sistematis antara model parametrik dan non-parametrik masih relatif terbatas dalam literatur.

Penelitian ini bertujuan mengisi kesenjangan tersebut melalui studi komparatif regresi machine learning yang komprehensif menggunakan dataset QS World University Rankings 2025. Dua algoritma supervised learning yang representatif dipilih: Regresi Linear sebagai model parametrik berbasis asumsi linearitas, dan Decision Tree Regressor sebagai model non-parametrik yang mampu menangkap pola non-linear. Secara spesifik, penelitian ini mencakup: (1) membangun dan mengevaluasi kedua model untuk memprediksi Overall Score; (2) menilai generalisasi model melalui validasi silang 5-fold; (3) melakukan tuning hiperparameter pada Decision Tree; dan (4) menganalisis feature importance untuk mengidentifikasi prediktor paling berpengaruh.

2. METODOLOGI PENELITIAN

Penelitian ini mengadopsi pendekatan kuantitatif komputasional yang terstruktur dalam tujuh tahapan utama: pemahaman data (data understanding), pembersihan data (data cleaning), eksplorasi data (Exploratory Data Analysis/EDA), pra-pemrosesan data sebelum pemodelan (data preprocessing before modelling), pelatihan dan pemodelan (data modelling), evaluasi dan validasi silang, serta interpretasi hasil. Seluruh implementasi dilakukan menggunakan Python 3.x dengan pustaka pandas, numpy, matplotlib, seaborn, dan scikit-learn [8].



Bagan 1. Alur Tahapan Metodologi Penelitian

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 897-907
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1094

Bagan 1. mengilustrasikan alur tujuh tahapan metodologi penelitian ini secara berurutan. Tahap pertama, Data Understanding, bertujuan memperoleh gambaran awal struktur dan kualitas dataset. Dilanjutkan dengan Data Cleaning untuk memfilter dan membersihkan data tidak valid. Pada tahap ketiga, Exploratory Data Analysis (EDA) dilakukan untuk mengungkap distribusi dan korelasi antar variabel. Selanjutnya, Data Preprocessing mempersiapkan data melalui imputasi, deteksi outlier, encoding, dan normalisasi sebelum masuk ke tahap Data Modelling, di mana dua algoritma Regresi Linear dan Decision Tree Regressor dilatih dan diuji. Tahap keenam, Evaluasi dan Validasi Silang, mengukur performa model menggunakan MAE, MSE, RMSE, R^2, serta 5-fold cross-validation. Proses diakhiri dengan Interpretasi Hasil yang menganalisis feature importance dan implikasi praktis temuan penelitian.

2.1 Dataset

Dataset yang digunakan adalah QS World University Rankings 2025 yang diperoleh dari sumber data publik. Dataset terdiri dari 1.503 baris yang merepresentasikan institusi universitas dari berbagai negara, dan 28 kolom mencakup atribut institusional serta indikator evaluasi QS. Karena variabel target (Overall_Score) hanya memiliki 601 nilai non-null, hanya baris dengan Overall_Score yang lengkap yang digunakan dalam pemodelan, menghasilkan dataset kerja final sebanyak 601 sampel.

Tabel 1. Deskripsi Variabel Dataset QS World University Rankings 2025

Fitur	Tipe Data	Keterangan
Academic_Reputation_Score	Numerik	Skor reputasi akademik berdasarkan survei global
Employer_Reputation_Score	Numerik	Skor reputasi dari perspektif pemberi kerja
Faculty_Student_Score	Numerik	Rasio antara jumlah dosen dan mahasiswa
Citations_per_Faculty_Score	Numerik	Jumlah sitasi ilmiah per anggota fakultas
International_Faculty_Score	Numerik	Proporsi dosen dari luar negeri
International_Students_Score	Numerik	Proporsi mahasiswa internasional
International_Research_Network_Score	Numerik	Jaringan riset internasional
Employment_Outcomes_Score	Numerik	Tingkat keberhasilan karir lulusan
Sustainability_Score	Numerik	Skor keberlanjutan lingkungan dan sosial
Region	Kategorikal	Wilayah geografis universitas
SIZE	Kategorikal	Ukuran institusi (S/M/L/XL)
FOCUS	Kategorikal	Fokus akademik institusi
Overall_Score	Numerik (Target)	Skor keseluruhan QS (variabel target)

2.2 Data Understanding

Tahap data understanding dilakukan untuk memperoleh gambaran awal terhadap struktur dan kualitas dataset. Dataset QS World University Rankings 2025 memuat 1.503 baris dan 28 kolom [9]. Proses ini mencakup identifikasi tipe data setiap kolom: kolom bertipe numerik kontinu (float64) seperti Academic_Reputation_Score dan Employer_Reputation_Score; kolom bertipe integer (int64) seperti RANK_2025 dan RANK_2024; serta kolom bertipe kategorikal (object/string) meliputi Institution_Name, Location, Region, SIZE, FOCUS, RES., dan STATUS.

Pengecekan missing value mengungkapkan bahwa Overall_Score hanya memiliki 601 nilai valid dari 1.503 baris, sehingga hanya baris tersebut yang digunakan dalam pemodelan. Variabel kategorikal mencakup Region (6 kategori), SIZE (4 kategori), dan FOCUS (4 kategori).

2.3 Data Cleaning

Data cleaning dilakukan dalam beberapa langkah sistematis [10]. Langkah pertama adalah pemilihan subset data: karena kolom Overall_Score hanya memiliki 601 nilai non-null dari total 1.503 baris, dilakukan filtering untuk mempertahankan baris dengan Overall_Score yang tersedia. Nilai nol pada kolom skor yang tidak bermakna diganti dengan NaN. Kolom tidak relevan (rank, identitas, STATUS) dieliminasi, menghasilkan 13 kolom: 9 fitur numerik, 3 fitur kategorikal, dan 1 target. Tidak ditemukan baris duplikat dalam dataset.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 897-907
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1094

2.4 Exploratory Data Analysis (EDA)

EDA dilakukan untuk memperoleh gambaran distribusi data dan hubungan antar variabel [11]. Analisis mencakup empat jenis utama: (1) statistik deskriptif menggunakan describe() dari pandas; (2) histogram dan density plot (KDE) untuk memvisualisasikan distribusi fitur, termasuk Overall_Score yang berpola right-skewed; (3) boxplot untuk mendeteksi outlier, di mana International_Research_Network_Score menunjukkan nilai ekstrem paling signifikan; dan (4) matriks korelasi Pearson yang divisualisasikan sebagai heatmap. Hasil menunjukkan Academic_Reputation_Score berkorelasi paling kuat dengan Overall_Score (r = 0,90), diikuti Employer_Reputation_Score (r = 0,78) [11].

2.5 Data Preprocessing Before Modelling

Preprocessing data mencakup lima langkah berurutan [12]. Pertama, imputasi median diterapkan untuk mengisi nilai hilang (NaN) karena bersifat robust terhadap outlier. Kedua, deteksi outlier menggunakan metode IQR (Q1 − 1,5×IQR hingga Q3 + 1,5×IQR); ditemukan 20 outlier pada International_Research_Network_Score yang diganti dengan median [13]. Ketiga, variabel kategorikal (Region, SIZE, FOCUS) dikonversi ke numerik menggunakan LabelEncoder dari scikit-learn [8]. Keempat, StandardScaler diterapkan untuk menstandarisasi fitur numerik (mean = 0, std = 1) [8]. Kelima, dataset dibagi dengan rasio 80:20 (random_state=42), menghasilkan 480 sampel pelatihan dan 120 sampel pengujian [8].

2.6 Data Modelling

Dua algoritma supervised regression diimplementasikan dan dibandingkan dalam memprediksi Overall_Score. Model pertama, Regresi Linear, diimplementasikan menggunakan LinearRegression (sklearn) dengan metode OLS dan parameter default. Model kedua, Decision Tree Regressor (sklearn), diuji dalam dua konfigurasi: default (max_depth=None) dan tuned (max_depth=6, min_samples_leaf=5) untuk mengatasi overfitting [14], [15]. Evaluasi menggunakan MAE, MSE, RMSE, dan R^2 pada 120 sampel pengujian [16], dilengkapi analisis feature importance [8].

2.7 Validasi Silang (Cross-Validation)

Validasi silang k-fold diterapkan untuk menilai generalisasi model secara lebih andal. Penelitian ini menggunakan k=5 (5-fold cross-validation) yang diimplementasikan menggunakan fungsi cross_val_score dari scikit-learn dengan parameter cv=5 dan scoring='r2' [8]. Nilai k=5 dipilih sebagai keseimbangan antara bias dan varians estimasi performa. Validasi silang juga diterapkan pada Decision Tree default untuk membandingkan stabilitas kedua model, mengingat Decision Tree rentan menghasilkan hasil yang sangat berbeda akibat varians yang tinggi [14], [17].

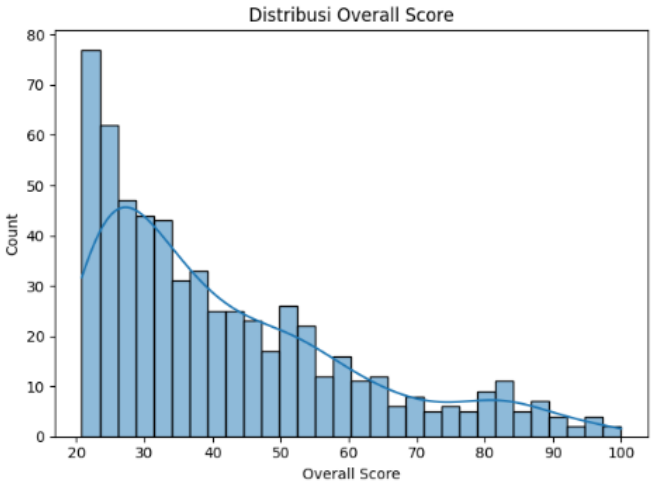
3. HASIL DAN PEMBAHASAN

3.1 Hasil Exploratory Data Analysis

Dataset awal terdiri dari 1.503 institusi dengan 28 atribut. Dari jumlah tersebut, hanya 601 rekaman yang memiliki nilai Overall_Score yang valid untuk pemodelan. Ketidaklengkapan ini terjadi karena universitas di luar ambang batas kualitatif tertentu tidak mendapatkan Overall Score numerik dari QS [2].

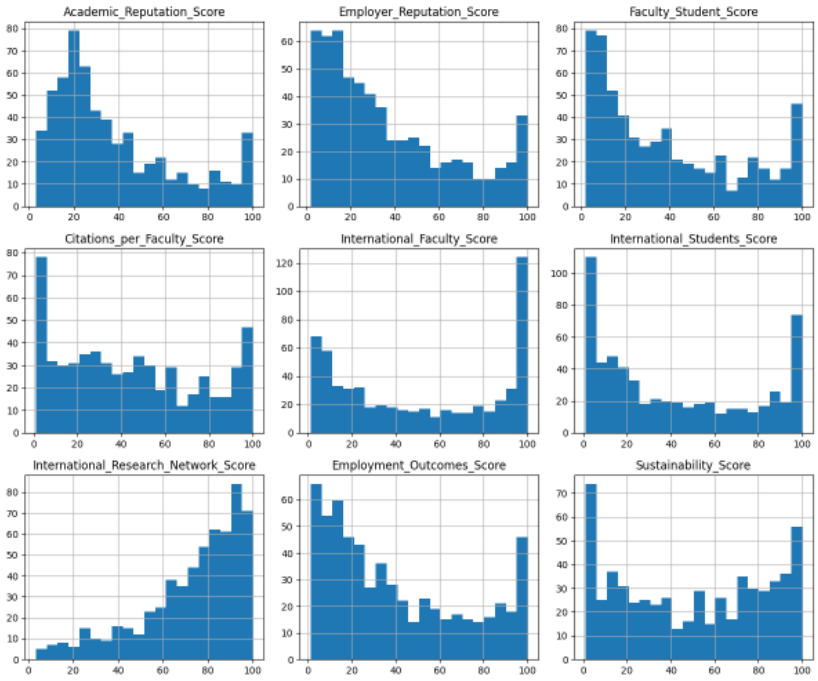
Analisis statistik deskriptif mengungkapkan variasi signifikan antar indikator. Academic_Reputation_Score memiliki rata-rata sekitar 20,79 dengan rentang yang luas, mencerminkan disparitas besar dalam reputasi akademik antar universitas. Employment_Outcomes_Score rata-rata sekitar 51,10, menunjukkan distribusi lebih terpusat.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 897-907
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1094



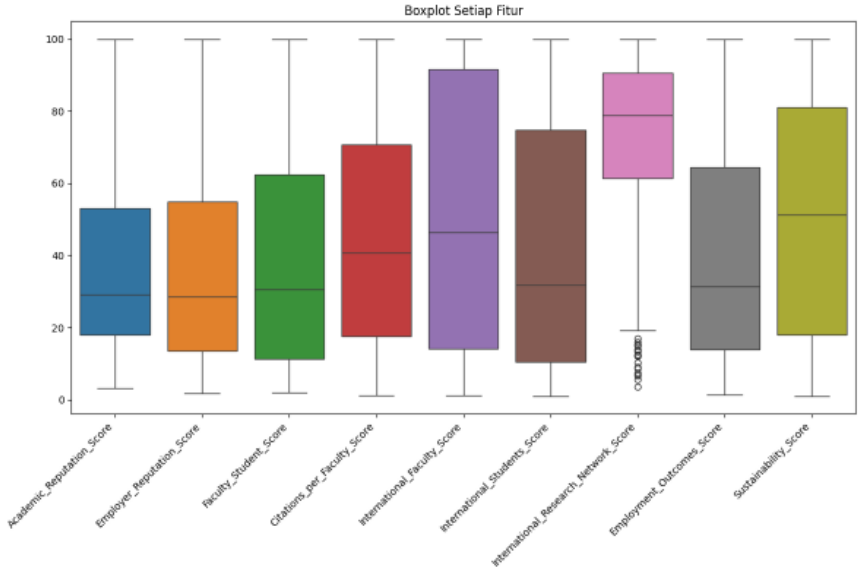
Gambar 1. Distribusi nilai Overall Score pada dataset QS World University Rankings 2025

Berdasarkan Gambar 1, distribusi Overall Score bersifat right-skewed (miring ke kanan), dengan sebagian besar universitas terkonsentrasi pada kisaran skor 20–40. Hanya sedikit institusi, umumnya universitas-universitas elite seperti MIT, Harvard, Oxford, dan Cambridge yang berhasil mencapai skor di atas 80.



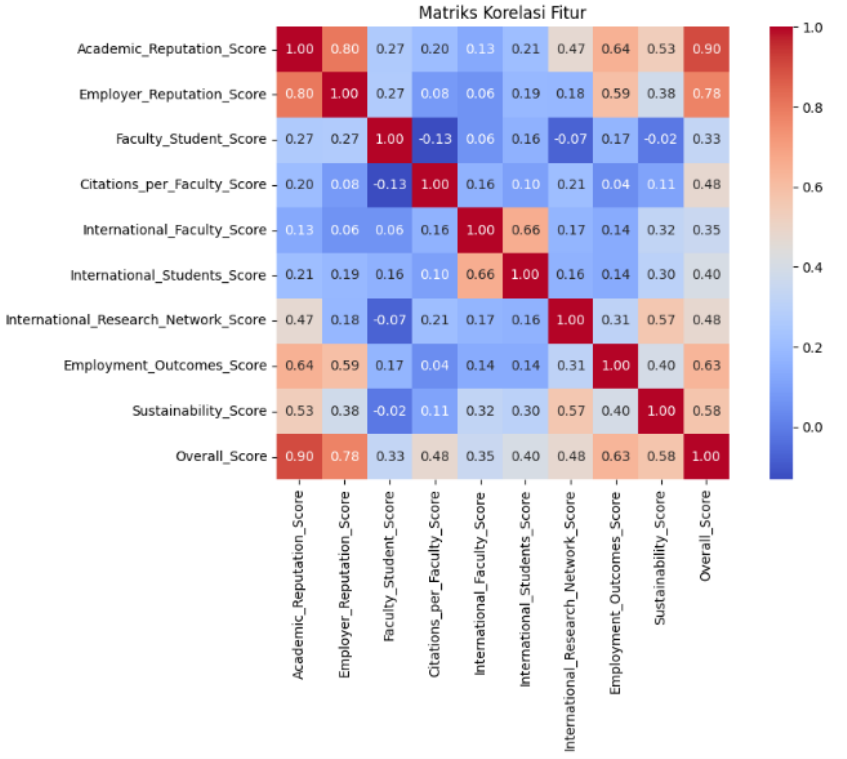
Gambar 2. Distribusi setiap fitur/indikator penilaian QS

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 897-907
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1094



Gambar 3. Boxplot setiap fitur

Berdasarkan Gambar 2 dan 3, distribusi masing-masing fitur menunjukkan pola yang beragam. International_Research_Network_Score menunjukkan sejumlah nilai ekstrem yang ditangani pada tahap preprocessing.



Gambar 4. Heatmap matriks korelasi antar fitur prediktor dan variabel target Overall_Score

Berdasarkan Gambar 4, Academic_Reputation_Score memiliki korelasi tertinggi terhadap Overall_Score (r = 0,90), mengindikasikan hubungan linear positif yang sangat kuat. Employer_Reputation_Score berada di posisi kedua (r = 0,78), diikuti Employment_Outcomes_Score (r = 0,63) dan Sustainability_Score (r = 0,58).

Tabel 2. Nilai Korelasi Fitur terhadap Overall_Score

Fitur	Nilai Korelasi (r)	Kategori Korelasi
Academic_Reputation_Score	0,90	Sangat Kuat
Employer_Reputation_Score	0,78	Kuat
Employment_Outcomes_Score	0,63	Sedang-Kuat

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 897-907
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1094

Sustainability_Score	0,58	Sedang
International_Research_Network_Score	0,48	Sedang
Citations_per_Faculty_Score	0,48	Sedang
International_Students_Score	0,40	Lemah-Sedang
International_Faculty_Score	0,35	Lemah-Sedang
Faculty_Student_Score	0,33	Lemah

3.2 Hasil Preprocessing Data

Inspeksi awal dataset mengonfirmasi tidak ada anomali nilai nol yang signifikan dalam fitur numerik sebelum pelatihan. Imputasi median berhasil mengisi seluruh nilai hilang, menghasilkan matriks fitur yang sepenuhnya lengkap. Analisis outlier menggunakan metode IQR mengidentifikasi 20 nilai ekstrem secara eksklusif pada kolom International_Research_Network_Score, yang kemudian diganti dengan median kolom tersebut. Label encoding berhasil mengonversi tiga variabel kategorikal menjadi representasi numerik ordinal. Standardisasi menggunakan StandardScaler mengubah distribusi seluruh fitur numerik menjadi zero-mean dan unit-variance, memastikan tidak ada fitur yang mendominasi proses pembelajaran akibat perbedaan skala. Dataset final: 480 sampel pelatihan dan 120 sampel pengujian.

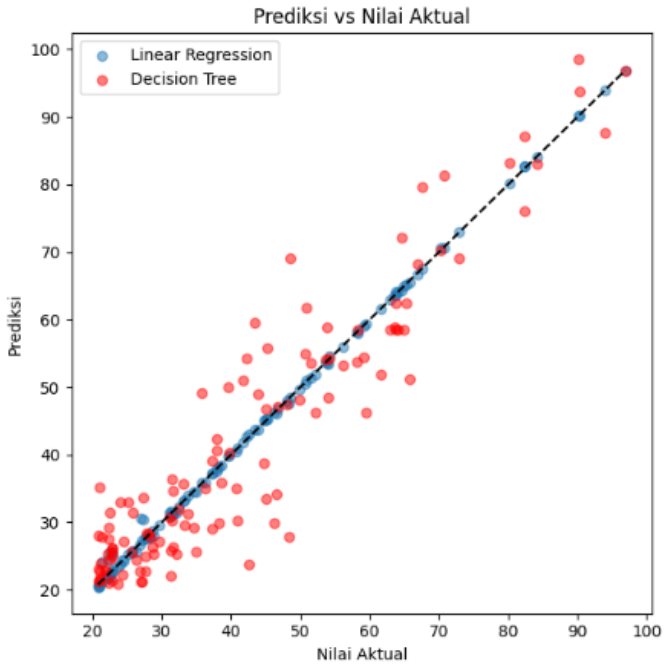
3.3 Evaluasi Model Awal

Kedua model dilatih pada 480 sampel dan dievaluasi pada 120 sampel pengujian. Tabel 3 menyajikan hasil evaluasi komprehensif untuk ketiga konfigurasi model.

Tabel 3. Perbandingan Performa Model pada Set Pengujian

Model	MAE	MSE	RMSE	R^2	Keterangan
Linear Regression	0,3662	0,5516	0,7427	0,9985	Model terbaik
Decision Tree (Default)	5,2250	47,6907	6,9058	0,8675	Overfitting
Decision Tree (Tuned)	5,2456	47,1778	6,8686	0,8690	Peningkatan marginal

Regresi Linear mencapai R^2 sebesar 0,9985, yang mengindikasikan bahwa model ini mampu menjelaskan 99,85% variansi variabel target. Nilai MAE sebesar 0,3662 berarti rata-rata prediksi hanya menyimpang kurang dari 0,37 poin dari nilai aktual. Sebaliknya, Decision Tree default menunjukkan performa yang jauh lebih rendah, dengan MAE sebesar 5,2250 — lebih dari 14 kali lipat dibandingkan Regresi Linear.



Gambar 5. Perbandingan nilai prediksi dan aktual Overall Score antara Regresi Linear dan Decision Tree

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 897-907
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1094

Scatter plot nilai prediksi versus aktual (Gambar 5) memberikan konfirmasi visual terhadap perbedaan performa. Prediksi Regresi Linear terkonsentrasi rapat di sepanjang garis diagonal sempurna (y = x), sedangkan prediksi Decision Tree menunjukkan dispersi yang jauh lebih lebar, terutama untuk universitas dalam kisaran skor menengah (30–70). Hasil ini konsisten dengan temuan Basireddy et al. [6] yang menunjukkan bahwa model dengan struktur linear lebih unggul untuk dataset dengan hubungan linear yang kuat.

3.4 Hasil Validasi Silang

Validasi silang 5-fold diterapkan pada keseluruhan dataset yang telah diproses menggunakan R^2 sebagai metrik scoring. Tabel 4 merangkum hasilnya.

Tabel 4. Hasil Validasi Silang 5-Fold

Model	Mean R^2 (5-Fold CV)	Std Dev
Linear Regression	0,9374	± 0,0668
Decision Tree (Default)	-4,1661	± 1,6075

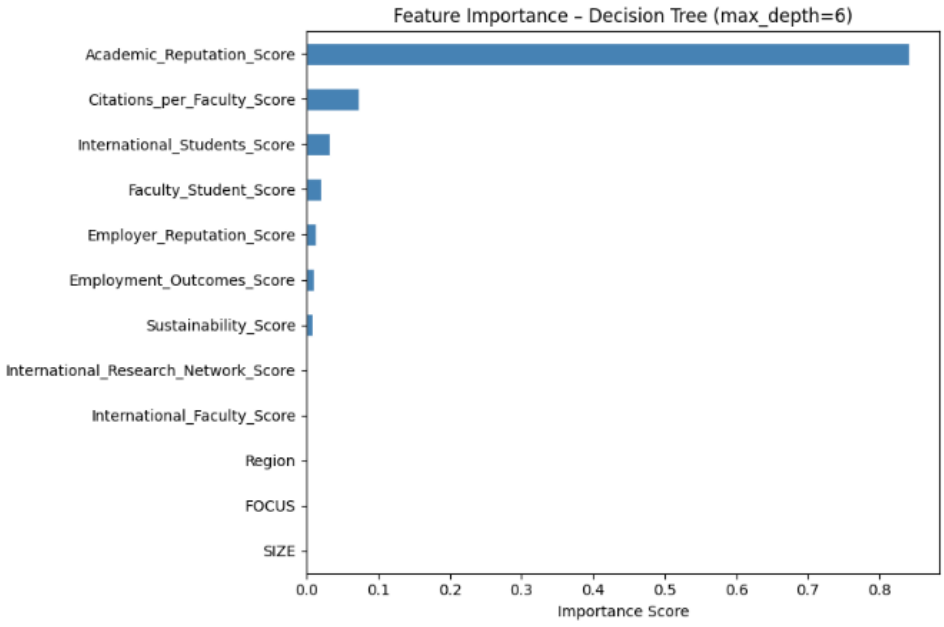
Regresi Linear mempertahankan performa stabil dengan mean R^2 sebesar 0,9374 ± 0,0668 di seluruh lima fold. Decision Tree mengalami penurunan dramatis dengan mean R^2 bernilai negatif (−4,1661 ± 1,6075), yang secara matematis berarti model bahkan lebih buruk daripada memprediksi nilai rata-rata untuk setiap sampel — indikasi overfitting yang sangat parah [18]. Standar deviasi yang sangat besar (1,6075) mengkonfirmasi ketidakstabilan performa Decision Tree terhadap perubahan komposisi data pelatihan.

3.5 Tuning Hiperparameter Decision Tree

Sebagai upaya mengatasi overfitting pada Decision Tree, tuning hiperparameter dilakukan dengan pembatasan max_depth=6 dan min_samples_leaf=5. Hasil tuning menunjukkan peningkatan marginal: MAE sedikit meningkat dari 5,2250 menjadi 5,2456, MSE menurun dari 47,6907 menjadi 47,1778, dan R^2 meningkat tipis dari 0,8675 menjadi 0,8690. Peningkatan ini jauh dari memadai untuk menjembatani kesenjangan performa yang fundamental antara Decision Tree dan Regresi Linear, mengkonfirmasi bahwa masalah bukan hanya pada kompleksitas berlebihan pohon, tetapi pada ketidaksesuaian mendasar antara asumsi Decision Tree dan struktur linear intrinsik data QS Rankings.

3.6 Analisis Feature Importance

Feature importance diekstraksi dari Decision Tree yang di-tuning untuk mengkuantifikasi kontribusi relatif setiap fitur. Feature importance dihitung berdasarkan total pengurangan impuritas (MSE reduction) yang dapat dikaitkan dengan setiap fitur di seluruh simpul pohon.



Gambar 6. Skor feature importance dari model Decision Tree yang telah di-tuning (max_depth=6)

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 897-907
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1094

Berdasarkan Gambar 6, Academic_Reputation_Score mendominasi feature importance dengan skor sekitar 0,84, bertanggung jawab atas sekitar 84% total pengurangan impuritas. Citations_per_Faculty_Score berada di posisi kedua (≈07%), diikuti International_Students_Score (≈04%). Temuan ini sangat konsisten dengan hasil analisis korelasi (r = 0,90), memberikan validasi lintas-metode yang kuat terhadap dominasi Academic_Reputation_Score. Konsistensi ini juga sejalan dengan temuan penelitian terkini oleh Navia-Gamero et al. [7] yang mengidentifikasi faktor kualitas fakultas sebagai determinan kunci dalam peringkat institusi.

3.7 Keunggulan Regresi Linear: Analisis Struktural

Keunggulan Regresi Linear yang sangat signifikan — dengan R^2 hampir sempurna sebesar 0,9985 — dapat dipahami dalam konteks metodologi perhitungan Overall Score QS. Sebagaimana dideskripsikan oleh QS [2], Overall Score pada dasarnya merupakan kombinasi linear berbobot dari skor-skor indikator konstituennya. Secara matematis, jika Overall_Score = $w_1 \times AR + w_2 \times ER + ... + w_n \times S_n$, maka hubungan antara prediktor dan target secara inheren bersifat linear. Konsekuensinya, Regresi Linear mampu menangkap hubungan fungsional tersebut dengan presisi sangat tinggi. Temuan ini dikonfirmasi oleh studi Li [19] yang juga menemukan bahwa Regresi Linear memberikan performa unggul dalam prediksi skor komprehensif universitas berbasis kombinasi linear berbobot.

3.8 Analisis Kritis Overfitting pada Decision Tree

Fenomena overfitting yang dialami Decision Tree merupakan manifestasi dramatis dari masalah bias-variance tradeoff dalam machine learning [20]. Ketika Decision Tree tanpa batasan dilatih pada 480 sampel, algoritma membangun pohon dengan kedalaman sangat besar, sehingga menghafal seluruh dataset pelatihan. Bukti overfitting paling mencolok adalah disparitas antara performa single test split (R^2 = 0,8675) dan validasi silang (R^2 = −4,1661). Untuk meningkatkan performa Decision Tree secara substansial, diperlukan teknik yang lebih canggih seperti ensemble methods (Random Forest, Gradient Boosting) yang mampu mengatasi varians tinggi melalui agregasi banyak pohon [14], [15].

3.9 Dominasi Academic Reputation: Implikasi Sistemik

Temuan bahwa Academic_Reputation_Score merupakan prediktor paling dominan — baik dalam korelasi (r = 0,90) maupun feature importance (≈84%) — memiliki implikasi sistemik yang signifikan [21]. Dominasi ini mencerminkan desain metodologis QS yang secara eksplisit menempatkan reputasi akademik sebagai komponen berbobot tertinggi (30%) [2]. Dari perspektif kebijakan institusional, temuan ini memberikan sinyal bahwa investasi dalam peningkatan reputasi akademik — melalui output penelitian berkualitas tinggi, kolaborasi internasional, dan peningkatan visibilitas ilmiah — merupakan strategi paling efektif untuk meningkatkan peringkat QS. Namun, Bellantuono et al. [22] menunjukkan adanya bias teritorial yang signifikan dalam perangkingan universitas global, di mana universitas dari negara-negara berbahasa Inggris mendapatkan keuntungan tidak proporsional.

4. KESIMPULAN

Penelitian ini telah melakukan studi komparatif regresi machine learning yang komprehensif pada dataset QS World University Rankings 2025 dan menghasilkan tiga temuan utama. Pertama, Regresi Linear secara konsisten dan substansial mengungguli Decision Tree di seluruh metrik evaluasi, dengan capaian R^2 = 0,9985, MAE = 0,3662, dan RMSE = 0,7427 pada set pengujian, serta R^2 rata-rata = 0,9374 ± 0,0668 pada validasi silang 5-fold, yang secara teoritis dijelaskan oleh kesesuaian antara asumsi linearitas model dan struktur metodologi QS yang merupakan kombinasi linear berbobot dari indikator-indikatornya. Kedua, Decision Tree menunjukkan gejala overfitting yang sangat parah—terbukti dari disparitas besar antara performa single-split (R^2 = 0,8675) dan validasi silang (R^2 = −4,1661)—dan tuning hiperparameter hanya menghasilkan peningkatan marginal yang tidak mampu menjembatani kesenjangan performa fundamental antara kedua model. Ketiga, Academic_Reputation_Score terbukti sebagai prediktor paling dominan dengan feature importance ≈84%, konsisten dengan nilai korelasi Pearson tertinggi (r = 0,90), sehingga institusi pendidikan tinggi yang berupaya meningkatkan peringkat QS disarankan memprioritaskan strategi peningkatan reputasi akademik global; sementara dari sisi metodologis, penelitian ini menegaskan pentingnya pemilihan model yang sesuai dengan karakteristik intrinsik data serta penggunaan validasi silang sebagai instrumen wajib deteksi overfitting, dan penelitian mendatang disarankan mengeksplorasi ensemble methods seperti Random Forest dan Gradient Boosting pada dataset yang lebih besar dan mencakup lebih banyak tahun.

REFERENCES

[1] C. S. Basireddy, V. K. G. Cheruku, S. Rajagopal, and R. Soangra, “Hybrid prediction models for assessing the Higher Education Institutions Performance in QS World Institution Rankings,” F1000Research, vol. 13, p. 1529, Dec. 2024. https://doi.org/10.12688/f1000research.155847.1

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 897-907
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1094

[2] QS Quacquarelli Symonds, "QS World University Rankings Methodology," QS Top Universities, 2024. [Online]. Available: https://www.topuniversities.com/world-university-rankings/methodology.

[3] M. Javaid, A. Haleem, R. P. Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," International Journal of Intelligent Networks, vol. 3, pp. 58–73, 2022. https://doi.org/10.1016/j.ijin.2022.05.002

[4] S. C. Matz et al., "Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics," Scientific Reports, vol. 13, no. 5608, Apr. 2023. https://doi.org/10.1038/s41598-023-32484-w

[5] H. M. Ahmed et al., "Student performance prediction using machine learning algorithms," Applied Computational Intelligence and Soft Computing, vol. 2024, Art. no. 4067721, 2024. https://doi.org/10.1155/2024/4067721

[6] I. Ullah et al., "Evaluating factors influencing university ranking based on QS ranking 2023–2024 using machine learning algorithms," in Proc. IEEE International Conference, 2024. https://doi.org/10.1109/11013585

[7] U. Navia-Gamero, A. Portilla-Flores, P. Vega-Leal, and M. Pozo-Guerrero, "A data analytics approach for university competitiveness: The QS world university rankings," International Journal on Interactive Design and Manufacturing (IJIDeM), vol. 16, pp. 1803–1812, Jul. 2022. https://doi.org/10.1007/s12008-022-00966-2

[8] S. Raschka, Y. H. Liu, and V. Mirjalili, Machine Learning with PyTorch and Scikit-Learn. Birmingham, UK: Packt Publishing, 2022. https://doi.org/10.17226/26580

[9] T. Nkosi, M. Dlamini, and S. Sibanda, "Towards a data quality framework: Preprocessing and cleaning practices in machine learning pipelines," IEEE Access, vol. 12, pp. 18340–18358, 2024. https://doi.org/10.1109/ACCESS.2024.3360152

[10] T. Nkosi, M. Dlamini, and S. Sibanda, "A systematic review of data cleaning and preprocessing methods for machine learning applications," IEEE Access, vol. 11, pp. 65320–65338, 2023. https://doi.org/10.1109/ACCESS.2023.3288456

[11] W. McKinney, Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022. [Online]. Available: https://www.oreilly.com/library/view/python-for-data/9781098104023/

[12] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022, pp. 64–95. [Online]. Available: https://www.oreilly.com/library/view/hands-on-machine-learning/9781098125967/

[13] S. Boukerche, N. Zhong, and P. Hung, "Outlier detection: Methods, models, and classification," ACM Computing Surveys, vol. 53, no. 3, Art. no. 55, 2022. https://doi.org/10.1145/3381028

[14] P. Schober and T. R. Vetter, "Decision trees in clinical research: Tree structure, overfitting, and cross-validation," Anesthesia & Analgesia, vol. 134, no. 2, pp. 275–278, Feb. 2022. https://doi.org/10.1213/ANE.0000000000005857

[15] R. Loh and W. Y. Loh, "Classification and regression tree methods," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 14, no. 3, Art. no. e1547, 2022. https://doi.org/10.1002/wics.1547

[16] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not," Geoscientific Model Development, vol. 15, no. 14, pp. 5481–5487, Jul. 2022. https://doi.org/10.5194/gmd-15-5481-2022

[17] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, An Introduction to Statistical Learning with Applications in Python. New York: Springer, 2023. https://doi.org/10.1007/978-3-031-38747-0

[18] C. Botchkarev, "Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology," Interdisciplinary Journal of Information, Knowledge, and Management, vol. 14, pp. 45–79, 2023. https://doi.org/10.28945/4184

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 5 No 1 Juni 2026
Hal: 897-907
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v5i1.1094

[19] Y. Li, "Prediction of university comprehensive score based on regression analysis," in Proc. 2022 International Conference on Science and Technology Ethics and Human Future (STEHF 2022), Atlantis Press, Jul. 2022. https://doi.org/10.2991/978-2-494069-79-6_183

[20] Y. A. Alsariera et al., "Assessment and evaluation of different machine learning algorithms for predicting student performance," Computational Intelligence and Neuroscience, vol. 2022, Art. no. 4151487, 2022. https://doi.org/10.1155/2022/4151487

[21] I. D. Stanciu and N. Nistor, "Doctoral capstone theories as indicators of university rankings: Insights from a machine learning approach," Computers in Human Behavior, vol. 164, Art. no. 108504, Mar. 2025. https://doi.org/10.1016/j.chb.2024.108504

[22] L. Bellantuono et al., "Territorial bias in university rankings: A complex network approach," Scientific Reports, vol. 12, Art. no. 4995, 2022. https://doi.org/10.1038/s41598-022-08859-w