Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 22 - 30
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2557
https://journal.fkpt.org/index.php/BIT

Pemodelan Biaya Sewa pada Data Pendidikan Internasional Menggunakan Pendekatan *Machine Learning* dan CRISP-DM

Arif Hamied Nababan[1,*], Rezeki Nauli Lumban Gaol[2], Fauziah Rahmadhani[3]

[1] Teknik Komputer dan Informatika, Teknologi Rekayasa Perangkat Lunak, Politeknik Negeri Medan, Medan, Indonesia
Email: [1,*]arifhamied@polmed.ac.id, [2]renaulilumbangaol@gmail.com, [3]fauziahramadhani366@gmail.com

Abstrak- Perkembangan machine learning mendorong pemanfaatannya dalam analisis data pendidikan yang kompleks. Dalam konteks pendidikan internasional, biaya sewa tempat tinggal (Rent_USD) merupakan komponen biaya hidup yang menunjukkan variasi signifikan antar wilayah. Variasi ini dipengaruhi faktor geografis, ekonomi lokal, serta karakteristik lingkungan pendidikan, sehingga memerlukan pemodelan data yang sistematis. Penelitian ini bertujuan memodelkan variabel Rent_USD menggunakan kerangka kerja CRISP-DM (Business Understanding, Data Understanding, Data Preparation, Modeling, dan Evaluation). Tiga algoritma digunakan dalam penelitian ini: Decision Tree sebagai model dasar, Random Forest sebagai pembanding, dan XGBoost sebagai model utama. Untuk mengoptimalkan kinerja, dilakukan penyesuaian hyperparameter melalui GridSearchCV. Evaluasi model diukur menggunakan metrik Mean Absolute Error (MAE), Root Mean Square Error (RMSE), dan koefisien determinasi (R^2). Hasil eksperimen menunjukkan bahwa algoritma XGBoost memberikan kinerja paling unggul dengan nilai RMSE terendah sebesar 93.27 USD dan R^2mencapai 0.96. Capaian tersebut melampaui performa Random Forest yang mencatat RMSE: 114.87 dan R^2: 0.94, serta Decision Tree dengan RMSE: 157.16 dan R^2: 0.89. Selain itu, analisis feature importance mengungkapkan temuan krusial bahwa indikator biaya hidup (Living Cost Index) dan biaya kuliah (Tuition Fee) memiliki pengaruh paling dominan terhadap variasi Rent_USD, dengan kontribusi masing-masing sebesar **58.32%** dan **32.94%**. Penelitian ini memberikan gambaran empiris mengenai penerapan machine learning dalam pemodelan biaya pendidikan internasional serta menjadi referensi penting bagi studi masa depan terkait manajemen data pendidikan.

Kata Kunci: Machine Learning, Rent_USD, Decision Tree, Random Forest, XGBoost, CRISP-DM.

Abstract- Advances in machine learning drive its application in analyzing complex educational data. In international education, housing rent (Rent_USD) is a critical cost-of-living component showing significant variation across regions. These variations are influenced by geography, local economics, and educational environments, requiring systematic data modeling. This study aims to model Rent_USD using the CRISP-DM framework: Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation. Three algorithms were employed: Decision Tree as the baseline, Random Forest as a comparison, and XGBoost as the primary model. To enhance performance, hyperparameter tuning was conducted via GridSearchCV. Model evaluation utilized Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2). The experimental results demonstrate that the XGBoost algorithm delivers the most superior performance, achieving the lowest RMSE of 93.27 USD and an R^2 of 0.96. This performance outperforms Random Forest (RMSE: 114.87, R^2: 0.94) and Decision Tree (RMSE: 157.16, R^2: 0.89). Furthermore, feature importance analysis revealed crucial findings: the Living Cost Index and Tuition Fee are the most dominant factors influencing Rent_USD variations, contributing **58.32%** and **32.94%** respectively. This research provides an empirical overview of machine learning applications in modeling international education costs and serves as a vital reference for future studies regarding educational data management and predictive analytics in global student mobility.

Keywords: Machine Learning, Rent_USD, Decision Tree, Random Forest, XGBoost, CRISP-DM.

1. PENDAHULUAN

Era globalisasi saat ini telah membawa dampak transformasi yang mendalam pada sektor pendidikan tinggi, yang ditandai dengan lonjakan mobilitas mahasiswa internasional secara masif. Fenomena ini tidak hanya mencerminkan pertukaran akademik, tetapi juga perpindahan ekonomi dan demografi yang signifikan antar negara. Mahasiswa yang memutuskan untuk melanjutkan studi ke luar negeri dihadapkan pada ekosistem pengambilan keputusan yang kompleks. Mereka tidak hanya mempertimbangkan reputasi universitas atau kurikulum akademik, tetapi juga dipaksa untuk memperhitungkan berbagai aspek non-akademik yang krusial. Di antara berbagai variabel non-akademik tersebut, biaya hidup (living cost) muncul sebagai faktor determinan utama yang sering kali menjadi penentu kelayakan studi seseorang. Dalam struktur biaya hidup mahasiswa internasional, komponen biaya sewa tempat tinggal (Rent_USD) secara konsisten menempati proporsi pengeluaran terbesar, sering kali melampaui biaya makan dan transportasi. Tidak seperti biaya kuliah (tuition fee) yang umumnya bersifat tetap (fixed cost) dan dapat diprediksi sejak awal penerimaan, biaya sewa bersifat sangat fluktuatif dan bervariasi (variable cost). Variasi ini dipengaruhi oleh interaksi kompleks antara faktor makro-ekonomi seperti inflasi dan nilai tukar mata uang, serta faktor mikro-geografis seperti jarak ke kampus, akses transportasi publik, dan fasilitas keamanan lingkungan. Ketidakpastian dalam memperkirakan biaya sewa ini berpotensi fatal; kesalahan dalam perencanaan finansial dapat menyebabkan stres finansial, gangguan fokus belajar, hingga risiko putus studi (dropout) karena ketidaksanggupan ekonomi.

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 22 - 30
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2557
https://journal.fkpt.org/index.php/BIT

Secara konvensional, calon mahasiswa melakukan estimasi biaya sewa melalui metode heuristik sederhana, seperti pencarian manual di portal properti, bertanya pada forum alumni, atau menggunakan nilai rata-rata nasional yang disediakan oleh brosur universitas. Namun, pendekatan-pendekatan ini memiliki kelemahan fundamental: mereka bersifat statis dan gagal menangkap heterogenitas data. Nilai rata-rata sering kali bias dan tidak mencerminkan realitas harga di lokasi spesifik, terutama di kota-kota metropolitan pendidikan di mana disparitas harga antar distrik bisa sangat ekstrem. Oleh karena itu, diperlukan sebuah pendekatan analitik yang mampu memodelkan hubungan non-linear antar variabel secara sistematis untuk menghasilkan prediksi yang presisi.

Di sinilah teknologi Machine Learning (ML) menawarkan solusi yang relevan. Berbeda dengan model statistik tradisional yang sering kali terikat pada asumsi linieritas dan normalitas data, algoritma ML modern memiliki kemampuan untuk mempelajari pola-pola tersembunyi (latent patterns) dari dataset yang besar, berisik (noisy), dan multivariat. Dengan melatih model pada data historis yang mencakup variabel negara, kota, reputasi universitas, dan indeks ekonomi, ML dapat menghasilkan estimasi biaya sewa yang dipersonalisasi dan lebih akurat.

Sejumlah penelitian terdahulu telah meletakkan dasar bagi penerapan ML di sektor pendidikan, meskipun fokusnya masih terfragmentasi. Dalam penelitiannya berhasil mendemonstrasikan kekuatan algoritma ML untuk memprediksi kinerja akademik mahasiswa magister [1]. Studi ini penting karena membuktikan bahwa data pendidikan yang kompleks dapat dimodelkan dengan baik, namun sayangnya, variabel target mereka terbatas pada nilai akademik (GPA) dan tidak menyentuh aspek finansial mahasiswa.

Di sisi lain, melakukan studi komparatif yang komprehensif menggunakan Decision Tree, Random Forest, SVM, dan Neural Network untuk menganalisis faktor kepuasan mahasiswa [2]. Penelitian ini memberikan wawasan berharga mengenai keunggulan metode ensemble dalam menangani data survei yang heterogen. Namun, dalam studi tersebut, variabel finansial hanya diperlakukan sebagai salah satu fitur prediktor (variabel independen), bukan sebagai target utama (variabel dependen) yang harus diprediksi.

Kesenjangan serupa juga ditemukan pada penelitian yang menerapkan algoritma Decision Tree untuk sistem rekomendasi jurusan kuliah [3]. Meskipun penelitian ini berhasil menonjolkan aspek interpretabilitas model pohon keputusan—di mana aturan keputusan dapat dijelaskan dengan logika "jika-maka"—fokusnya adalah pada masalah klasifikasi (memilih kategori jurusan), bukan masalah regresi (memprediksi nilai numerik biaya).

Sebaliknya, literatur di domain ekonomi properti (real estate) telah sangat maju dalam memprediksi harga sewa. [4],[5] telah membuktikan bahwa algoritma seperti Random Forest dan XGBoost sangat superior dalam memprediksi harga sewa ruko dan rumah tinggal. Temuan ini didukung oleh serangkaian penelitian global lainnya oleh Tauryawati [6], [7], [8], [9], [10] yang secara konsisten menunjukkan bahwa metode ensemble learning mampu meminimalkan kesalahan prediksi (error rates) pada platform seperti Airbnb. Juga memperkuat argumen ini dengan menunjukkan bahwa XGBoost memberikan kinerja terbaik dalam memprediksi angka harapan hidup, sebuah variabel numerik yang memiliki karakteristik distribusi data mirip dengan biaya hidup [11]. Turut menyimpulkan bahwa algoritma berbasis pohon (tree-based) lebih tahan terhadap outlier dibandingkan regresi linier klasik [12].

Berdasarkan tinjauan literatur mendalam tersebut, teridentifikasi sebuah celah penelitian (research gap) yang signifikan. Terdapat dikotomi yang jelas: penelitian pendidikan cenderung mengabaikan prediksi finansial mendalam [1], [3], sementara penelitian properti cenderung menggunakan data pasar umum yang tidak memperhitungkan konteks spesifik mahasiswa (seperti lokasi kampus atau jenis visa) [4], [12]. Data properti umum sering kali memasukkan variabel seperti "kemewahan interior" atau "luas tanah" yang mungkin tidak relevan atau tidak tersedia dalam konteks data pendidikan internasional.

Penelitian ini bertujuan untuk menjembatani kesenjangan tersebut dengan memodelkan variabel Rent_USD secara spesifik pada dataset pendidikan internasional. Penelitian ini mengadopsi kerangka kerja standar industri CRISP-DM untuk menjamin ketigakakuan metodologis. Tiga algoritma state-of-the-art akan dibandingkan: Decision Tree (DT) sebagai model baseline yang menawarkan transparansi, Random Forest (RF) sebagai representasi metode bagging yang stabil, dan XGBoost sebagai representasi metode boosting yang akurat. Melalui pendekatan ini, diharapkan dapat dihasilkan model yang tidak hanya akurat secara statistik, tetapi juga relevan secara praktis untuk membantu calon mahasiswa dalam perencanaan studi.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menggunakan pendekatan machine learning dengan kerangka kerja CRISP-DM (Cross-Industry Standard Process for Data Mining). CRISP-DM dipilih karena menyediakan tahapan penelitian yang sistematis dan terstruktur, sehingga sesuai untuk menggambarkan alur penerapan metode dan pengujian model dalam penelitian berbasis

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 22 - 30
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2557
https://journal.fkpt.org/index.php/BIT

data. Penggunaan kerangka kerja CRISP-DM juga telah banyak diadopsi pada penelitian machine learning [13] dan penelitian data mining dibidang pendidikan karena mampu memberikan panduan yang jelas mulai dari pemahaman masalah hingga evaluasi model [14]

Secara umum, tahapan penelitian yang dilakukan terdiri atas lima tahap utama, yaitu Business Understanding, Data Understanding, Data Preparation, Modeling, dan Evaluation. Alur tahapan penelitian ini ditunjukkan pada Gambar 2.1.

Gambar 1.Alur Tahapan Penelitian Menggunakan CRISP-DM

Tahapan tersebut digunakan untuk memastikan bahwa proses penelitian dilakukan secara sistematis, mulai dari pemahaman masalah hingga pengujian hasil model.

2.2 Business Understanding

Tahap pertama, Business Understanding, adalah fondasi dari seluruh penelitian. Pada tahap ini, fokus utamanya adalah menerjemahkan masalah dunia nyata (kesulitan mahasiswa mengestimasi biaya) menjadi masalah teknis data mining (regresi).

a. Definisi Masalah: Tingginya variabilitas biaya sewa tempat tinggal (Rent_USD) menciptakan asimetri informasi bagi calon mahasiswa internasional. Informasi yang tersedia sering kali kadaluwarsa atau terlalu umum (agregat nasional), sehingga tidak dapat dijadikan dasar perencanaan anggaran yang valid.
b. Tujuan Penelitian:
   - Membangun model prediktif yang dapat mengestimasi Rent_USD dengan tingkat kesalahan seminimal mungkin.
   - Mengevaluasi dan membandingkan kinerja tiga algoritma berbeda untuk mengetahui mana yang paling cocok untuk karakteristik data pendidikan.
c. Kriteria Kesuksesan: Model dianggap berhasil jika mampu memberikan prediksi dengan nilai R-Squared (R^2) yang tinggi (mendekati 1.0) dan nilai Root Mean Square Error (RMSE) yang rendah, serta serta model dinilai berdasarkan akurasi prediksi untuk optimalisasi stok [15] dan juga memberikan wawasan mengenai fitur mana yang paling berpengaruh.

2.3 Data Understanding

Tahap Data Understanding mencakup pengumpulan data awal, deskripsi data, eksplorasi data, dan verifikasi kualitas data. Proses eksplorasi dan analisis statistik deskriptif diperlukan untuk memahami pola data serta mendeteksi potensi anomali sejak awal [16]. Dataset yang digunakan dalam penelitian ini adalah data sekunder yang merepresentasikan profil biaya pendidikan internasional. Data ini bersifat heterogen karena mengandung campuran antara data demografis, geografis, dan ekonomi.

Langkah-langkah eksplorasi yang dilakukan meliputi:

a. Identifikasi Struktur Data: Memeriksa dimensi data (jumlah baris dan kolom) serta tipe data dasar (integer, float, object).

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 22 - 30
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2557
https://journal.fkpt.org/index.php/BIT

b. Analisis Statistik Deskriptif: Melihat distribusi data melalui nilai mean, median, standar deviasi, serta nilai minimum dan maksimum untuk mendeteksi adanya anomali awal.
c. Deteksi Kualitas Data: Memeriksa keberadaan missing values (data yang hilang) dan data duplikat yang dapat mendistorsi hasil pelatihan model.

Tabel 1. Artribut Dataset

No	Nama Atribut	Tipe Data	Deskripsi
1	Country	Kategorikal	Negara lokasi institusi
2	City	Kategorikal	Kota lokasi institusi
3	University	Kategorikal	Nama institusi
4	Living_Cost_Index	Numerik	Indikator biaya hidup
5	Rent_USD	Numerik	Biaya sewa tempat tinggal

Tahap ini bertujuan untuk memastikan bahwa dataset dipahami dengan baik sebelum dilakukan pemrosesan lebih lanjut.

2.4 Data Preparation

Tahap Data Preparation sering kali memakan waktu 60-80% dari total waktu proyek data mining. Tahap ini bertujuan mengubah data mentah menjadi format yang "bersih" dan siap dikonsumsi oleh algoritma matematika. Proses yang dilakukan meliputi:

a. Seleksi Atribut (Feature Selection):
Tidak semua atribut dalam dataset mentah relevan untuk prediksi. Atribut yang tidak memiliki korelasi logis atau statistik dengan target Rent_USD dibuang untuk mengurangi dimensi data dan mencegah noise. Fitur utama yang dipertahankan adalah fitur lokasi (Country, City) dan ekonomi (Living_Cost_Index).
b. Penanganan Data Kategorikal (Encoding):
Algoritma machine learning secara matematis hanya dapat memproses angka, bukan teks. Oleh karena itu, kolom kategorikal seperti 'Country' dan 'City' harus dikonversi. Teknik yang digunakan adalah Label Encoding atau One-Hot Encoding. Label Encoding mengubah setiap kategori menjadi angka unik (misal: USA=0, UK=1). Teknik ini dipilih untuk menjaga efisiensi memori, terutama pada variabel dengan kardinalitas tinggi seperti 'City'.
c. Pembagian Data (Splitting):
Untuk menguji kemampuan generalisasi model, dataset dibagi menjadi dua bagian terpisah: data latih (training set) dan data uji (testing set). Proporsi yang digunakan adalah 80:20 (80% untuk pelatihan, 20% untuk pengujian). Pembagian ini dilakukan secara acak namun terkontrol menggunakan parameter random_state untuk memastikan hasil eksperimen konsisten dan dapat direproduksi (reproducible).

2.5 Modeling

Tahap ini adalah inti dari eksperimen komputasi, di mana tiga algoritma dengan paradigma berbeda diterapkan:

a. Decision Tree (DT):
Sebagai model baseline, Decision Tree bekerja dengan memecah data menjadi himpunan bagian yang lebih kecil berdasarkan aturan keputusan "jika-maka". Algoritma ini memilih fitur pemisah terbaik berdasarkan kriteria Information Gain atau Gini Impurity. Keunggulan utamanya adalah "White Box Nature"—model ini sangat mudah diinterpretasikan dan divisualisasikan. Namun, DT tunggal memiliki kelemahan fatal yaitu kecenderungan untuk overfitting (terlalu menghafal data latih) sehingga buruk saat memprediksi data baru. Decision Tree sering digunakan sebagai model dasar karena mudah diinterpretasikan meskipun memiliki kecenderungan overfitting [17].
b. Random Forest (RF):
Untuk mengatasi kelemahan DT, digunakan algoritma Random Forest. Ini adalah metode ensemble tipe Bagging (Bootstrap Aggregating). RF membangun ratusan pohon keputusan secara paralel. Setiap pohon dilatih menggunakan sampel acak dari data (bootstrap), dan prediksi akhir diambil dari rata-rata prediksi seluruh pohon. Pendekatan "kebijaksanaan orang banyak" (wisdom of crowds) ini terbukti secara teoritis dan empiris mampu menurunkan varians model secara drastis dan meningkatkan stabilitas prediksi, sebagaimana dijelaskan oleh Breiman.
c. XGBoost (Extreme Gradient Boosting):
Algoritma ketiga dan model utama dalam penelitian ini adalah XGBoost. Berbeda dengan RF yang membangun pohon secara paralel/independen, XGBoost menggunakan pendekatan Boosting. Pohon-pohon dibangun secara sekuensial (berurutan); pohon kedua bertugas memperbaiki kesalahan (residual error) dari pohon pertama, pohon ketiga memperbaiki pohon kedua, dan seterusnya. Algoritma ini menggunakan teknik optimasi Gradient Descent untuk meminimalkan fungsi kerugian (loss function). Dikembangkan oleh Chen dan Guestrin, XGBoost dikenal

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 22 - 30
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2557
https://journal.fkpt.org/index.php/BIT

memiliki eksekusi yang sangat cepat, skalabilitas tinggi, dan fitur regularisasi bawaan yang mencegah overfitting pada data tabular yang kompleks.

Agar kinerja model maksimal, parameter bawaan (default) sering kali tidak cukup. Penelitian ini menerapkan teknik GridSearchCV, yaitu pencarian grid secara sistematis untuk menemukan kombinasi parameter terbaik. Parameter yang dioptimasi tercantum pada Tabel 2.2.

Tabel 2. Daftar Hyperparameter yang Dioptimasi

Algoritma	Hyperparameter
Decision Tree	max_depth, min_samples_split
Random Forest	n_estimators, max_depth
XGBoost	n_estimators, learning_rate, max_depth

2.6 Evaluation

Tahap terakhir adalah evaluasi model menggunakan data uji (unseen data). Karena ini adalah permasalahan regresi (prediksi nilai kontinu), metrik akurasi klasifikasi (seperti presisi/recall) tidak dapat digunakan. Tiga metrik regresi standar industri dipilih:

a. Mean Absolute Error (MAE): Menghitung rata-rata selisih absolut antara nilai prediksi dan nilai aktual. MAE memberikan gambaran kesalahan yang mudah dipahami secara intuitif (misal: "rata-rata prediksi meleset sebesar $100").
b. Root Mean Square Error (RMSE): Menghitung akar kuadrat dari rata-rata kuadrat kesalahan. Berbeda dengan MAE, RMSE memberikan "hukuman" lebih besar pada kesalahan prediksi yang bernilai besar (large errors). Ini penting dalam konteks biaya, karena kesalahan prediksi yang sangat jauh bisa berakibat fatal bagi perencanaan keuangan mahasiswa.
c. Koefisien Determinasi (R^2 Score): Mengukur seberapa baik variasi dalam variabel dependen (Rent_USD) dapat dijelaskan oleh model. Nilai R^2 berkisar antara 0 hingga 1. Nilai mendekati 1 menunjukkan bahwa model sangat akurat dan mampu menangkap pola data dengan sangat baik.

Penggunaan kombinasi metrik Mean Absolute Error(MAE),Root Mean Square Error(RMSE), dan koefisien determinasi R^2 banyak digunakan dalam penelitian regresi untuk memberikan evaluasi performa model secara komprehensif [18].

2.7 Deployment

Dalam siklus CRISP-DM standar, tahap Deployment melibatkan penerapan model ke lingkungan produksi (misalnya, membuat aplikasi web). Namun, dalam lingkup penelitian ini, deployment dibatasi pada penyajian hasil analisis, rekomendasi model terbaik, dan implikasi manajerial bagi pemangku kepentingan (mahasiswa dan konsultan pendidikan). Tujuan utamanya adalah memperoleh pengetahuan (knowledge discovery) untuk menjawab permasalahan penelitian yang telah dirumuskan.

3. HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil dan pembahasan dari penerapan metode machine learning untuk memodelkan variabel biaya sewa tempat tinggal (Rent_USD) pada data pendidikan internasional. Proses penelitian dilakukan berdasarkan tahapan CRISP-DM yang telah dijelaskan pada Bab II, sehingga hasil yang diperoleh merupakan keluaran langsung dari tahapan Modeling dan Evaluation.

Tujuan utama pada tahap ini adalah untuk mengevaluasi kinerja beberapa algoritma machine learning, yaitu Decision Tree, Random Forest, dan XGBoost, dalam memodelkan variabel Rent_USD. Selain itu, dilakukan analisis terhadap faktor-faktor yang memengaruhi variasi nilai Rent_USD berdasarkan model terbaik. Hasil penelitian disajikan dalam bentuk tabel dan gambar untuk memudahkan interpretasi dan pembahasan.

3.1 Hasil Penerapan Tahapan CRISP-DM

Penerapan tahapan CRISP-DM dalam penelitian ini memberikan kerangka kerja yang sistematis dalam proses pemodelan. Pada tahap Business Understanding, permasalahan penelitian difokuskan pada pemodelan variasi biaya sewa dalam konteks pendidikan internasional. Tahap ini menghasilkan tujuan yang jelas, yaitu membangun dan mengevaluasi model machine learning untuk variabel Rent_USD.

Tahap Data Understanding menghasilkan pemahaman terhadap karakteristik dataset yang digunakan, termasuk jenis atribut, distribusi data, serta potensi permasalahan seperti nilai kosong dan variasi data antar wilayah. Tahap Data Preparation memastikan bahwa data telah siap digunakan melalui proses seleksi atribut, encoding data kategorikal, dan

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 22 - 30
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2557
https://journal.fkpt.org/index.php/BIT

pembagian data menjadi data latih dan data uji. Tahap Modeling dan Evaluation menjadi inti dari penelitian ini. Pada tahap tersebut, model dilatih menggunakan data latih dan diuji menggunakan data uji untuk memperoleh nilai evaluasi kinerja. Hasil dari tahapan ini dibahas secara rinci pada subbagian berikutnya.

3.2 Hasil Pelatihan Model Machine Learning

Pada tahap pelatihan model, tiga algoritma machine learning diterapkan pada dataset yang sama, yaitu Decision Tree, Random Forest, dan XGBoost. Seluruh model dilatih menggunakan data latih yang identik agar perbandingan kinerja dapat dilakukan secara adil. Untuk meningkatkan kinerja model, dilakukan penyesuaian hyperparameter menggunakan GridSearchCV sebagaimana dijelaskan pada Bab II. Proses ini bertujuan untuk memperoleh konfigurasi parameter yang sesuai tanpa mengubah fokus utama penelitian. Penyesuaian hyperparameter dilakukan secara terbatas, sehingga tetap menjaga kesederhanaan dan keterulangan penelitian.
Hasil pelatihan menunjukkan bahwa setiap algoritma memiliki karakteristik yang berbeda dalam mempelajari pola data. Perbedaan tersebut juga ditentukan pada studi perbandingan SVM, Random Forest, dan XGBoost [19]. Decision Tree sebagai model dasar menunjukkan kemampuan yang cukup baik dalam menangkap pola sederhana, namun memiliki keterbatasan dalam menangani kompleksitas data. Random Forest dan XGBoost, sebagai metode ensemble, menunjukkan kemampuan yang lebih baik dalam memodelkan hubungan non-linear antar variabel.

3.3 Hasil Evaluasi Kinerja Model

Evaluasi kinerja model dilakukan menggunakan tiga metrik regresi, yaitu Mean Absolute Error (MAE), Root Mean Square Error (RMSE), dan koefisien determinasi (R^2). Metrik tersebut umum digunakan dalam evaluasi model regresi berbasis machine learning [20]. Ketiga metrik ini dipilih karena mampu memberikan gambaran menyeluruh mengenai tingkat kesalahan prediksi dan kemampuan model dalam menjelaskan variasi data Rent_USD.

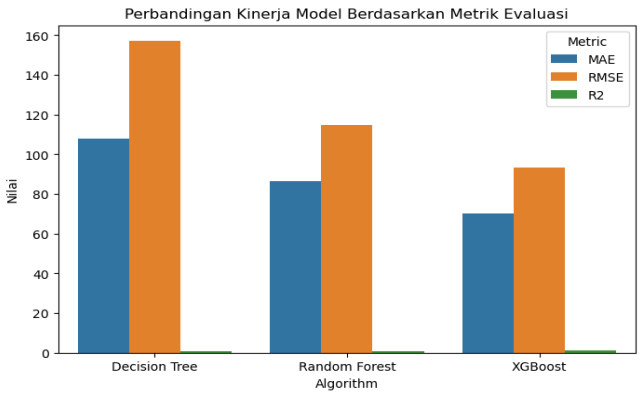
Tabel 3. Daftar Hyperparameter yang Dioptimasi

Algoritma	MAE	RMSE	R^2
Decision Tree	108.023357	157.157180	0.885934
Random Forest	86.585884	114.865381	0.939065
XGBoost	69.987701	93.267545	0.959826

Berdasarkan Tabel 3 dapat diamati bahwa Decision Tree menghasilkan nilai MAE dan RMSE yang lebih tinggi dibandingkan algoritma lainnya, serta nilai R^2 yang lebih rendah. Hal ini menunjukkan bahwa kemampuan Decision Tree dalam memodelkan variasi Rent_USD masih terbatas, terutama pada data yang bersifat heterogen. Random Forest menunjukkan peningkatan kinerja dibandingkan Decision Tree, yang ditunjukkan oleh nilai MAE dan RMSE yang lebih rendah serta nilai R^2 yang lebih tinggi. Peningkatan ini disebabkan oleh mekanisme ensemble yang menggabungkan banyak pohon keputusan sehingga mampu mengurangi varians model. XGBoost memberikan kinerja terbaik di antara ketiga algoritma yang diuji. Nilai MAE dan RMSE yang dihasilkan oleh XGBoost paling rendah, sementara nilai R^2 paling tinggi. Hal ini menunjukkan bahwa XGBoost mampu memodelkan hubungan antar variabel dengan lebih efektif dibandingkan algoritma lainnya.

3.4 Perbandingan Kinerja Antar Algoritma

Gambar yang dihasilkan membuktikan secara visual bahwa model XGBoost Anda memiliki performa yang baik karena titik-titik datanya cenderung mengikuti garis diagonal (linear), yang sejalan dengan nilai R^2 yang tinggi (0.96) yang Anda dapatkan sebelumnya.



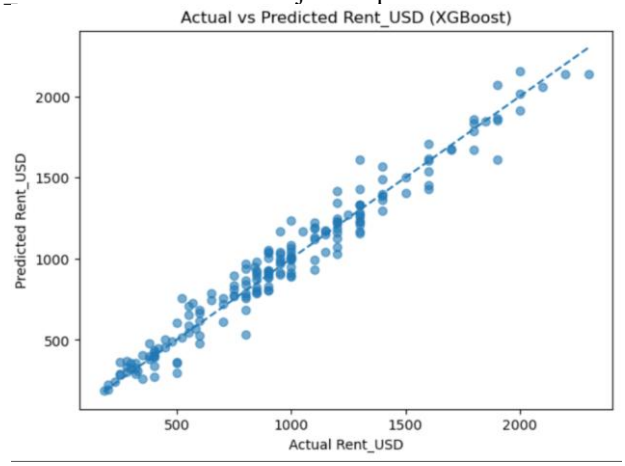
Gambar 2. Actual vs Predicted Rent_USD(XGBoost)

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 22 - 30
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2557
https://journal.fkpt.org/index.php/BIT

Perbandingan kinerja antar algoritma menunjukkan bahwa pemilihan algoritma memiliki pengaruh signifikan terhadap hasil pemodelan Rent_USD. Decision Tree sebagai model dasar memiliki keunggulan dalam hal interpretabilitas, namun kurang optimal dalam menangani data dengan kompleksitas tinggi.
Random Forest menunjukkan kinerja yang lebih stabil dibandingkan Decision Tree, terutama dalam mengurangi kesalahan prediksi. Penggunaan banyak pohon keputusan memungkinkan model ini menangkap variasi data dengan lebih baik. Namun demikian, Random Forest masih memiliki keterbatasan dalam mengoptimalkan kesalahan prediksi dibandingkan XGBoost. XGBoost sebagai model utama menunjukkan keunggulan dalam meminimalkan kesalahan prediksi dan meningkatkan kemampuan penjelasan model. Mekanisme gradient boosting memungkinkan model melakukan optimasi secara bertahap, sehingga kesalahan pada iterasi sebelumnya dapat diperbaiki pada iterasi berikutnya. Hal ini menjadikan XGBoost lebih efektif dalam memodelkan variasi Rent_USD pada data pendidikan internasional. Dari ini di ketahui bahwa ke konsistenan penelitian yang menunjukan keunggulan XGBoost dibandingkan Decision Tree [21].

3.5 Visualisasi Hasil Prediksi

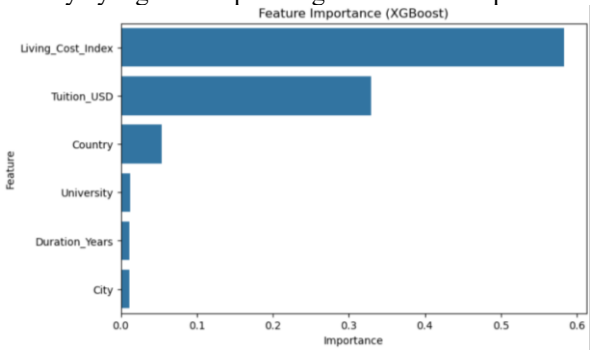
Untuk memberikan gambaran yang lebih jelas mengenai kinerja model, dilakukan visualisasi perbandingan antara nilai aktual dan nilai prediksi Rent_USD. Visualisasi ini ditunjukkan pada Gambar 4.



Gambar 4. Perbandingan Nilai Aktual dan Prediksi Rent_USD Menggunakan Model XGBoost
(Gambar berupa grafik scatter atau garis yang menunjukkan hubungan antara nilai aktual dan nilai prediksi Rent_USD)
Berdasarkan Gambar 4, dapat diamati bahwa sebagian besar titik prediksi berada di sekitar garis diagonal, yang menunjukkan bahwa model XGBoost mampu menghasilkan prediksi yang mendekati nilai aktual. Meskipun masih terdapat beberapa penyimpangan, secara umum pola prediksi menunjukkan kesesuaian yang baik dengan data aktual. Algoritma XGBoost sangat direkomendasikan untuk digunakan dalam sistem estimasi biaya studi internasional karena presisi, stabil, dan mampu menjelaskan faktor ekonomi yang memengaruhi harga sewa.

3.6 Pembahasan Hasil Penelitian

Hasil penelitian ini sejalan dengan temuan penelitian terdahulu yang menyatakan bahwa metode ensemble learning, seperti Random Forest dan XGBoost, cenderung memberikan kinerja yang lebih baik dibandingkan model tunggal seperti Decision Tree dalam permasalahan regresi. Keunggulan XGBoost dalam penelitian ini juga konsisten dengan penelitian-penelitian sebelumnya yang menerapkan algoritma tersebut pada data numerik yang kompleks.



Gambar 5. Feature Importance (XGBoost)

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 22 - 30
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2557
https://journal.fkpt.org/index.php/BIT

Penggunaan GridSearchCV sebagai bagian dari proses pemodelan membantu memperoleh konfigurasi hyperparameter yang lebih sesuai, sehingga meningkatkan kinerja model tanpa menambah kompleksitas yang berlebihan. Hal ini menunjukkan bahwa penyesuaian hyperparameter yang dilakukan secara terbatas sudah cukup untuk meningkatkan performa model dalam konteks penelitian ini. Namun demikian, perlu ditekankan bahwa data yang digunakan dalam penelitian ini merupakan data sekunder yang merepresentasikan estimasi biaya sewa dalam konteks pendidikan internasional, bukan data transaksi sewa aktual. Oleh karena itu, hasil penelitian ini lebih tepat dipahami sebagai pemodelan pola variasi biaya sewa, bukan sebagai prediksi nilai transaksi riil.

4. KESIMPULAN

Penelitian ini bertujuan untuk memodelkan variasi biaya sewa tempat tinggal (Rent_USD) pada data pendidikan internasional menggunakan pendekatan *machine learning* dengan kerangka kerja CRISP-DM. Berdasarkan hasil penelitian, dapat disimpulkan bahwa pendekatan CRISP-DM mampu memberikan alur penelitian yang sistematis dan terstruktur, mulai dari pemahaman masalah hingga evaluasi kinerja model. Penerapan tahapan ini membantu memastikan bahwa proses pemodelan dilakukan secara konsisten dan sesuai dengan tujuan penelitian. Hasil evaluasi menunjukkan adanya perbedaan kinerja yang jelas antara algoritma *Decision Tree*, *Random Forest*, dan XGBoost. *Decision Tree* sebagai model dasar menunjukkan keterbatasan dalam menangani kompleksitas data, sementara *Random Forest* mampu meningkatkan stabilitas prediksi. Namun, algoritma XGBoost memberikan kinerja terbaik berdasarkan metrik MAE, RMSE, dan koefisien determinasi (R^2), yang membuktikan bahwa metode *ensemble learning* berbasis *gradient boosting* lebih efektif dalam menangkap hubungan non-linear pada data pendidikan internasional.

Keberhasilan penelitian ini semakin diperkuat oleh hasil analisis *feature importance* pada model XGBoost yang memberikan gambaran faktor-faktor penentu biaya sewa. Ditemukan bahwa indikator biaya hidup (**Living Cost Index**) dan biaya kuliah (**Tuition Fee**) merupakan faktor yang paling dominan dalam memengaruhi variasi nilai Rent_USD, dengan kontribusi masing-masing sebesar **58.32%** dan **32.94%**. Temuan ini mengindikasikan bahwa variasi biaya sewa dalam konteks pendidikan internasional tidak hanya dipengaruhi oleh lokasi geografis secara makro, tetapi secara signifikan ditentukan oleh karakteristik ekonomi lokal dan lingkungan pendidikan sekitarnya.

Meskipun memberikan hasil yang relevan, terdapat beberapa keterbatasan yang perlu diperhatikan. Data yang digunakan merupakan data sekunder yang bersifat estimasi dan tidak mencerminkan transaksi sewa aktual. Selain itu, penyesuaian *hyperparameter* dilakukan secara terbatas dan penelitian ini hanya mengevaluasi tiga algoritma. Oleh karena itu, penelitian selanjutnya diharapkan dapat menggunakan dataset yang lebih luas, menerapkan metode optimasi yang lebih mendalam, serta mengeksplorasi pendekatan model lain untuk meningkatkan kualitas serta generalisasi hasil penelitian.

REFERENCES

[1] S. Alturki, L. Cohausz, and H. Stuckenschmidt, “Predicting Master’s students’ academic performance: an empirical study in Germany,” *Smart Learning Environments*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40561-022-00220-y.

[2] A. D. Riyanto, A. M. Wahid, and A. A. Pratiwi, “ANALYSIS OF FACTORS DETERMINING STUDENT SATISFACTION USING DECISION TREE, RANDOM FOREST, SVM, AND NEURAL NETWORKS: A COMPARATIVE STUDY,” *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 4, pp. 187–196, Jul. 2024, doi: 10.52436/1.jutif.2024.5.4.2188.

[3] A. Anugerah *et al.*, “SISTEM REKOMENDASI JURUSAN KULIAH BAGI CALON MAHASISWA BARU UNIVERSITAS BSI MARGONDA FAKULTAS TEKNIK DAN INFORMATIKA MENGGUNAKAN ALGORITMA C4.5,” 2025.

[4] Y. F. Munawar and A. Arisal, “Analisis Prediksi Harga Sewa Ruko Menggunakan Pendekatan Machine Learning,” *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 4, no. 3, pp. 2538–2544, Aug. 2025, doi: 10.31004/riggs.v4i3.2184.

[5] B. Wulan Sari and D. Prabowo, “Analisis Perbandingan Prediksi Harga Rumah Dengan Random Forest, Gradient Boosting, dan XGBoost,” *Intellect : Indonesian Journal of Learning and Technological Innovation*, vol. 4, no. 1, pp. 42–51, Jun. 2025, doi: 10.57255/intellect.v4i1.1385.

[6] W. Mulia, M. Lista, and A. SSi, “Comparison of Random Forest, XGBoost, and LightGBM Methods in Estimating Airbnb Accommodation Rental Prices Based on Customers in New York City,” 2023.

[7] S. Cao, W. Liao, and J. Huang, “Research on Renting Price Prediction Based on Machine Learning,” European Alliance for Innovation n.o., May 2024. doi: 10.4108/eai.8-12-2023.2344718.

[8] A. Neyaz, A. Ahmed, A. Singh, G. Noida, and U. Pradesh, “Machine Learning for Rental Price Prediction: Regression Techniques and Random Forest Model.” [Online]. Available: https://ssrn.com/abstract=4587725

[9] X. Wan, X. Li, L. Xiong, Y. Xu, and J. Tian, “Comparison and Optimization Strategies of Airbnb Rental Prediction Models: An Empirical Study Based on Linear Regression, XGBoost and Random Forest,” *Advances in Economics,*

Bulletin of Information Technology (BIT)
Vol 7, No 1, Maret 2026, Hal. 22 - 30
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2557
https://journal.fkpt.org/index.php/BIT

Management and Political Sciences, vol. 197, no. 1, pp. 148–162, Sep. 2025, doi: 10.54254/2754-1169/2025.lh27240.

[10] A. Karim and A. Ernawati, "Uncovering Smartphone Brand Strategies through Specification-Based Clustering and Classification," *Buletin Ilmiah Informatika Teknologi*, vol. 4, no. 1, pp. 24–32, Oct. 2025, doi: 10.58369/biit.v2i3.167.

[11] G. Chairunisa *et al.*, "Life Expectancy Prediction Using Decision Tree, Random Forest, Gradient Boosting, and XGBoost Regressions," *Jurnal Sintak*, vol. 2, no. 2, 2024.

[12] A. Yavuz Özalp and H. Akinci, "Comparison of tree-based machine learning algorithms in price prediction of residential real estate," *Gumushane Universitesi Fen Bilimleri Dergisi*, vol. 14, no. 1, pp. 116–130, Mar. 2024, doi: 10.17714/gumusfenbil.1363531.

[13] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," 2021. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[14] N. Lebkiri *et al.*, "Using Machine Learning for Prediction Students Failure in Morocco: an Application of the CRISP-DM Methodology," *International Journal of Education and Information Technologies*, vol. 15, pp. 344–352, Oct. 2021, doi: 10.46300/9109.2021.15.36.

[15] R. Winurputra and D. E. Ratnawati, "Peramalan Penjualan Produk Menggunakan Extreme Gradient Boosting (XGBoost) dan Kerangka Kerja CRISP-DM untuk Pengoptimalan Manajemen Persediaan (Studi Kasus: UB Mart)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 12, no. 2, pp. 417–428, Apr. 2025, doi: 10.25126/jtiik.2025129451.

[16] D. Ruswanti, D. Susilo, and R. Riani, "Implementasi CRISP-DM pada Data Mining untuk Melakukan Prediksi Pendapatan dengan Algoritma C.45," *Go Infotech: Jurnal Ilmiah STMIK AUB*, vol. 30, no. 1, pp. 111–121, Jun. 2024, doi: 10.36309/goi.v30i1.266.

[17] Y. A. Singgalen, "Penerapan CRISP-DM dalam Klasifikasi Sentimen dan Analisis Perilaku Pembelian Layanan Akomodasi Hotel Berbasis Algoritma Decision Tree (DT)," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 5, no. 2, p. 237, Dec. 2023, doi: 10.30865/json.v5i2.7081.

[18] S. Arti and E. Suherlan, "Evaluasi Kinerja Machine Learning dalam Memprediksi Kemampuan Adaptasi Mahasiswa pada Lingkungan Pembelajaran Daring," *Jurnal Pustaka AI (Pusat Akses Kajian Teknologi Artificial Intelligence)*, vol. 5, no. 1, pp. 50–57, Apr. 2025, doi: 10.55382/jurnalpustakaai.v5i1.901.

[19] M. R. Givari, R. Mochamad, and Y. U. Sulaeman[2], "Perbandingan Algoritma SVM, Random Forest Dan XGBoost Untuk Penentuan Persetujuan Pengajuan Kredit," vol. 16, no. 1, 2022, [Online]. Available: https://journal.uniku.ac.id/index.php/ilkom

[20] A. Armalia Raidani, H. Manurung, M. Sihombing, S. Informasi, and S. Kaputama, "PERBANDINGAN ALGORITMA XGBOOST DAN RANDOM FOREST DENGAN TEKNIK FEATURE ENGINEERING PADA KLASIFIKASI." [Online]. Available: https://journaledutech.com/index.php/great

[21] H. H. Sinaga and S. Agustian, "Pebandingan Metode Decision Tree dan XGBoost untuk Klasifikasi Sentimen Vaksin Covid-19 di Twitter," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 8, no. 3, pp. 107–114, Dec. 2022, doi: 10.25077/teknosi.v8i3.2022.107-114.