

Analisis Spasial dan Model Prediktif Data Geografis-Demografis untuk Prediksi Kelayakan Lokasi Usaha

Arysespajayadi^{1,*}, Hilman Asyrafi¹, Muh. Nurtanzis Sutoyo², Reza Maulana³

¹Teknik Grafika, Jurusan Teknologi Industri, Politeknik Negeri Media Kreatif, Makassar, Indonesia

²Sistem Informasi, Fakultas Teknologi Informasi, Universitas Sembilanbelas November Kolaka, Kolaka, Indonesia

³Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Alauddin, Makassar, Indonesia

Email: ^{1,*}arysespajayadi@gmail.com, ²hilman.asyrafi@polimedia.ac.id, ³mns.usn21@gmail.com, ⁴saya@rezamualana.com

Email Penulis Korespondensi: alysespajayadi@gmail.com*

Submitted: 05/03/2026; Accepted: 28/03/2026; Published: 31/03/2026

Abstrak–Pemilihan lokasi usaha yang tepat merupakan faktor strategis yang menentukan keberhasilan operasional dan finansial bisnis baru. Permasalahan utama yang dihadapi calon pengusaha adalah sulitnya menentukan kelayakan lokasi secara objektif akibat keterbatasan data primer dan metode analisis yang bersifat intuitif. Penelitian ini bertujuan membangun sistem prediksi kelayakan lokasi usaha dengan mengintegrasikan analisis spasial dan model prediktif berbasis machine learning. Metode yang digunakan adalah algoritma Random Forest yang dilatih menggunakan data primer dari 45 titik lokasi di wilayah urban Makassar yang diperoleh melalui survei lapangan langsung. Variabel input mencakup jumlah penduduk, kepadatan toko kompetitor, ketersediaan fasilitas umum, dan volume lalu lintas. Evaluasi model dilakukan menggunakan metrik Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan koefisien determinasi (R^2). Hasil penelitian menunjukkan model Random Forest memberikan prediksi jumlah pengunjung dengan nilai R^2 sebesar 0,72, MAE 450,00, dan RMSE 600,17, serta estimasi pendapatan dengan R^2 sebesar 0,75, MAE Rp1.412.556, dan RMSE 1.894.958. Kestabilan model dikonfirmasi melalui pengujian 5-Fold Cross-validation dengan rata-rata R^2 sebesar 0,70. Benchmarking terhadap Linear Regression menunjukkan keunggulan Random Forest dengan peningkatan R^2 sebesar 50% pada prediksi pengunjung dan 36,4% pada prediksi pendapatan. Analisis feature importance mengungkapkan jumlah penduduk sebagai variabel paling dominan (bobot 0,45), diikuti volume lalu lintas (0,25) dan fasilitas umum (0,15). Secara keseluruhan, sistem ini menyediakan alat bantu estimasi kelayakan lokasi berbasis data geografis dan demografis yang dapat mendukung pengambilan keputusan investasi bagi pelaku usaha baru.

Kata Kunci: Analisis Spasial; Random Forest; Geografis; Demografis; Lokasi Usaha; Machine Learning

Abstract–Selecting the right business location is a strategic factor determining the operational and financial success of a new venture. The main challenge faced by entrepreneurs is objectively assessing site feasibility due to limited primary data and intuitive analytical methods. This study develops a business location feasibility prediction system by integrating spatial analysis with machine learning-based predictive models. The Random Forest algorithm was trained on primary data from 45 locations in urban Makassar collected through direct field surveys. Input variables include population size, competitor store density, public facility availability, and traffic volume. Model evaluation used Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). Results show the Random Forest model achieved visitor count predictions with R^2 of 0.72, MAE of 450.00, and RMSE of 600.17, while revenue estimation yielded R^2 of 0.75, MAE of IDR 1,412,556, and RMSE of 1,894,958. Model stability was confirmed through 5-Fold Cross-validation with a mean R^2 of 0.70. Benchmarking against Linear Regression showed the superiority of Random Forest, with R^2 improvements of 50% for visitor prediction and 36.4% for revenue prediction. Feature importance analysis identified population size as the most dominant variable (weight 0.45), followed by traffic volume (0.25) and public facilities (0.15). Overall, this system provides a data-driven estimation tool for business location feasibility assessment, supporting investment decision-making for new entrepreneurs based on geographic and demographic information.

Keywords: Spatial Analysis; Random Forest; Geographic; Demographic; Business Location; Machine Learning

1. PENDAHULUAN

Pemilihan lokasi usaha merupakan keputusan strategis paling mendasar yang menentukan keberlanjutan dan profitabilitas sebuah entitas bisnis dalam jangka panjang. Di wilayah urban seperti Makassar, kesalahan dalam menentukan titik operasional dapat berdampak fatal bagi sektor Usaha Mikro, Kecil, dan Menengah (UMKM) karena keterbatasan modal untuk melakukan riset pasar yang mendalam [1]. Secara tradisional, pemilihan lokasi usaha di Indonesia masih sering didasarkan pada intuisi, ketersediaan lahan milik pribadi, atau pengamatan visual yang bersifat subjektif tanpa didukung oleh data empiris yang kuat [2]. Pendekatan konvensional semacam ini mengandung risiko tinggi karena tidak mempertimbangkan dinamika spasial, pola pergerakan penduduk, dan variabel demografis yang saling berinteraksi secara kompleks [3]. Masalah utama yang dihadapi adalah ketiadaan sistem yang mampu mengintegrasikan data geografis dan demografis secara otomatis untuk memberikan rekomendasi kelayakan lokasi bagi pelaku usaha baru [4] [5].

Perkembangan teknologi *Geographic Information System* (GIS) dan *Big Data* telah membuka peluang baru dalam melakukan analisis lokasi yang lebih presisi melalui pendekatan *location intelligence* [6],[7],[8]. Namun, pengolahan data geospasial mentah masih memerlukan keahlian teknis tingkat tinggi yang sulit dijangkau oleh pengusaha awam. Oleh karena itu, diperlukan sebuah solusi berupa model prediktif yang mampu menyederhanakan kompleksitas data tersebut menjadi informasi keputusan yang mudah dipahami [9]. Penggunaan analisis spasial yang digabungkan dengan teknik *machine learning* menawarkan kemampuan untuk memetakan

potensi ekonomi suatu wilayah dengan mempertimbangkan berbagai variabel seperti jumlah penduduk, keberadaan kompetitor, fasilitas umum pendukung, hingga kepadatan lalu lintas di sekitar lokasi[10]. Solusi yang ditawarkan dalam penelitian ini adalah pengembangan aplikasi berbasis web yang mengimplementasikan model prediktif untuk menentukan kelayakan lokasi usaha secara otomatis melalui interaksi peta digital.

Dalam beberapa tahun terakhir, telah dilakukan penelitian untuk mengeksplorasi penggunaan data spasial dalam dunia bisnis. Penelitian yang menggunakan algoritma K-Nearest Neighbor (KNN) untuk menentukan lokasi ritel, namun masih memiliki keterbatasan dalam menangani data yang memiliki variansi tinggi [11], [12]. Penelitian lainnya menerapkan analisis regresi linear untuk memprediksi pendapatan toko berdasarkan luas bangunan [13], tetapi mengabaikan variabel lingkungan seperti keberadaan fasilitas publik. Sementara itu, studi yang menyoroti pentingnya data demografi penduduk dalam menentukan klaster bisnis [14] namun belum menyediakan fitur prediksi secara real-time berbasis web. Selain itu, penelitian yang menggunakan algoritma Random Forest untuk klasifikasi penggunaan lahan menunjukkan akurasi yang sangat baik dibandingkan metode pohon keputusan tradisional [15] [16]. Terakhir, penelitian yang menggunakan data lalu lintas untuk optimasi rute logistik, yang membuktikan bahwa variabel pergerakan kendaraan merupakan indikator kuat bagi potensi ekonomi suatu area [17].

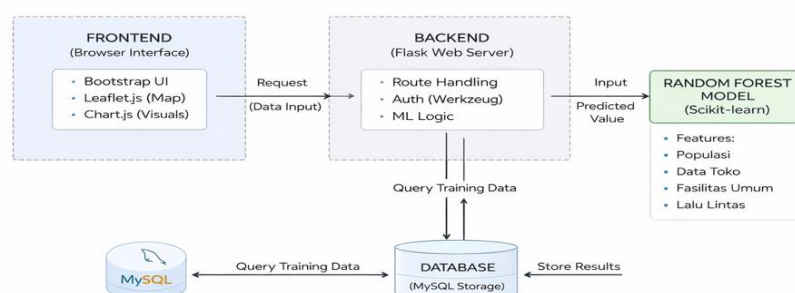
Meskipun penelitian-penelitian tersebut telah memberikan kontribusi signifikan, terdapat gap atau celah penelitian yang perlu diisi. Sebagian besar penelitian terdahulu cenderung fokus pada satu aspek data saja, misalnya hanya demografi atau hanya data lalu lintas, tanpa mengintegrasikannya secara holistik ke dalam satu model prediktif yang komprehensif. Selain itu, aspek aksesibilitas pengguna melalui aplikasi web interaktif yang memungkinkan pemilihan titik koordinat langsung pada peta masih sangat jarang ditemukan dalam literatur terkait penentuan lokasi usaha UMKM di Indonesia. Keterbatasan ini memunculkan peluang bagi penelitian ini untuk mengembangkan model yang menggabungkan analisis spasial mendalam dengan Algoritma Random Forest yang dikenal ketangguhannya (*robust*) terhadap pencilan (*outlier*) dan mampu menangani hubungan non-linear antar variabel demografis dan geografis [18] [19] [20] [21].

Selain itu, Random Forest memungkinkan adanya analisis feature importance yang dapat memberikan insight bisnis mengenai variabel mana yang paling berpengaruh terhadap kesuksesan lokasi. Data yang digunakan dalam penelitian ini merupakan data primer yang diperoleh melalui survei lapangan dan pemetaan mandiri di 45 titik lokasi strategis di wilayah Makassar. Penggunaan 45 sampel data primer ini dipilih secara selektif untuk memastikan validitas dan reliabilitas informasi terkait jumlah pengunjung dan pendapatan aktual, yang sering kali sulit didapatkan dari sumber data terbuka (*open data*). Meskipun jumlah sampel tergolong terbatas, fokus pada satu wilayah geografis yang spesifik memungkinkan analisis spasial yang lebih mendalam dan mendetail terhadap perilaku konsumen lokal di area perkotaan. Hal ini menjadi sangat krusial karena setiap wilayah memiliki karakteristik demografis yang unik, sehingga model yang dikembangkan bertindak sebagai studi kasus (*case study*) yang solid untuk pengembangan bisnis di lingkungan urban.

Tujuan utama dari penelitian ini adalah untuk merancang model prediktif yang mampu mengestimasi jumlah pengunjung dan pendapatan potensial berdasarkan parameter demografi dan geografis menggunakan algoritma Random Forest. Penelitian ini juga bertujuan untuk melakukan benchmarking sederhana terhadap model linear untuk membuktikan efektivitas penggunaan algoritma ensemble learning dalam pengolahan data spasial terbatas. Harapan yang ingin dicapai adalah terciptanya alat bantu pengambilan keputusan (*Decision Support Tool*) yang dapat meminimalkan risiko investasi bagi pengusaha baru melalui pendekatan berbasis data (*data-driven decision making*).

2. METODOLOGI PENELITIAN

Metodologi penelitian ini dirancang secara sistematis untuk menjawab tantangan dalam memprediksi kelayakan lokasi usaha dengan mengintegrasikan data spasial dan model prediktif. Pendekatan yang digunakan mencakup beberapa tahapan utama, mulai dari pengumpulan data, pra-pemrosesan, pengembangan model machine learning, hingga implementasi ke dalam sistem berbasis web. Secara garis besar, Arsitektur sistem prediksi kelayakan lokasi usaha ini dapat dilihat pada Gambar 1.



Gambar 1. Arsitektur Sistem Prediksi Kelayakan Lokasi Usaha

2.1 Tahapan Penelitian

Tahapan penelitian diawali dengan identifikasi variabel-variabel kunci yang mempengaruhi kesuksesan sebuah lokasi usaha. Berdasarkan studi literatur primer, faktor demografi dan geografi merupakan prediktor yang sangat signifikan dalam analisis potensi pasar. Setelah variabel ditentukan, dilakukan pengumpulan data sekunder dan primer melalui pemetaan titik geospasial. Tahapan selanjutnya adalah pra-pemrosesan data untuk memastikan kualitas dataset sebelum masuk ke fase pelatihan model. Setelah model berhasil dilatih dan dievaluasi, sistem diintegrasikan ke dalam antarmuka web interaktif yang memungkinkan pengguna melakukan simulasi prediksi secara real-time.

2.2 Pengumpulan dan Struktur Data

Data yang digunakan dalam penelitian ini bersifat multidimensi, menggabungkan aspek demografis dan geografis. Variabel demografis difokuskan pada jumlah penduduk di sekitar radius lokasi yang ditentukan, yang merupakan representasi dari ukuran pasar potensial. Variabel geografis mencakup data toko eksisting (kompetitor dan pelengkap), ketersediaan fasilitas umum (seperti sekolah, perkantoran, dan tempat ibadah), serta volume lalu lintas jalan di sekitar lokasi. Seluruh data geografis diidentifikasi melalui koordinat latitude dan longitude yang ditentukan secara langsung melalui interaksi peta pada aplikasi. Skema penyimpanan data dilakukan menggunakan sistem manajemen basis data MySQL yang dirancang untuk mendukung operasi CRUD (Create, Read, Update, Delete) pada data latih dan hasil prediksi [22].

2.3 Pengembangan Model Random Forest

Algoritma Random Forest dipilih sebagai mesin penggerak prediksi karena keunggulannya dalam menangani data dengan hubungan *non-linear* dan ketangguhannya terhadap *outlier* dibandingkan model regresi linear tradisional [18]. Random Forest bekerja dengan membangun sejumlah pohon keputusan (*decision trees*) selama fase pelatihan dan mengeluarkan rata-rata prediksi dari setiap pohon tersebut sebagai hasil akhir. Dalam penelitian ini, dua model regresi dikembangkan secara paralel, satu untuk memprediksi jumlah pengunjung potensial dan satu untuk estimasi pendapatan bulanan. Fitur-fitur yang menjadi input model meliputi: Titik Koordinat (Latitude dan Longitude), Populasi Penduduk, Indeks Kepadatan Toko, Ketersediaan Fasilitas Publik, Skor Intensitas Lalu Lintas.

2.4 Arsitektur Sistem Web

Sistem dibangun menggunakan kerangka kerja (framework) Flask di sisi backend untuk menangani logika aplikasi dan komunikasi data [23]. Antarmuka pengguna (frontend) dikembangkan dengan prinsip sederhana untuk meningkatkan pengalaman pengguna. Komponen peta interaktif menggunakan pustaka Leaflet.js yang memungkinkan pemilihan lokasi melalui titik koordinat secara presisi tanpa perlu input manual.

2.5 Metode Evaluasi Model

Untuk mengukur akurasi dan performa model prediktif, dilakukan evaluasi menggunakan tiga metrik utama, yaitu *Mean Absolute Error (MAE)*, *Root Mean Squared Error (RMSE)*, dan koefisien determinasi (R^2).

1. Mean Absolute Error (MAE)

MAE digunakan untuk mengukur rata-rata besaran kesalahan dalam prediksi tanpa memperhatikan arah kesalahannya. Nilai MAE memberikan gambaran error rata-rata dalam satuan yang sama dengan variabel target, dihitung dengan rumus:

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

2. Root Mean Squared Error (RMSE)

Sesuai dengan kebutuhan evaluasi yang lebih sensitif terhadap pencilan (*outliers*), penelitian ini menyertakan metrik RMSE. RMSE memberikan bobot lebih tinggi pada kesalahan yang besar dibandingkan MAE, sehingga sangat efektif untuk mendeteksi variansi kesalahan prediksi yang signifikan. Rumus RMSE didefinisikan sebagai berikut:

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

3. Koefisien Determinasi (R^2)

Sementara itu, R^2 digunakan untuk mengukur seberapa baik model dapat menjelaskan variasi dalam data target, dengan rentang nilai antara 0 hingga 1. Nilai R^2 yang mendekati 1 menunjukkan bahwa model memiliki kemampuan prediksi yang sangat kuat. Evaluasi ini dihitung dengan rumus:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3)$$

Di mana n adalah jumlah sampel, y_i adalah nilai aktual, $\hat{y}_i$ adalah nilai prediksi, dan $\bar{y}$ adalah nilai rata-rata dari data aktual.

3. HASIL DAN PEMBAHASAN

Bagian ini memaparkan hasil analisis data geospasial, performa model prediktif, serta perbandingan antar algoritma untuk menentukan kelayakan lokasi usaha. Analisis dilakukan secara mendalam dengan mengintegrasikan data primer dari 45 titik lokasi di Makassar yang telah diverifikasi validitasnya.

3.1 Karakteristik Data dan Validitas

Dataset penelitian ini terdiri dari 45 entri data yang diperoleh melalui survei primer langsung di lapangan. Penggunaan data primer memastikan bahwa variabel aktual jumlah pengunjung dan pendapatan benar-benar mencerminkan kondisi riil di lokasi penelitian, bukan sekadar data sintetis. Meskipun jumlah sampel sebanyak 45 tergolong kecil untuk model machine learning skala besar, penelitian ini diperlakukan sebagai studi kasus (case study) wilayah urban dengan kepadatan tinggi. Berdasarkan data yang diekstraksi, variabel jumlah penduduk memiliki rentang antara 3.000 hingga 35.000 jiwa per lokasi, dengan puncak pendapatan aktual mencapai Rp15.000.000. Hal ini menunjukkan adanya variansi data yang cukup untuk melatih model dalam mengenali pola potensi ekonomi di berbagai kelas lokasi. Proses pengumpulan data dilakukan secara terstruktur dengan mengunjungi setiap titik lokasi pada hari kerja dan akhir pekan untuk memperoleh gambaran yang representatif terhadap pola kunjungan aktual. Setiap titik lokasi dicatat koordinat GPS-nya secara presisi melalui antarmuka peta aplikasi, sehingga seluruh variabel geospasial mulai dari jumlah penduduk di radius lokasi, jumlah toko kompetitor, ketersediaan fasilitas umum, hingga intensitas lalu lintas dapat diekstraksi secara konsisten dan terverifikasi. Pendekatan ini memastikan bahwa dataset tidak hanya akurat secara numerik, tetapi juga valid secara spasial, yakni setiap nilai variabel benar-benar merepresentasikan kondisi lingkungan di sekitar koordinat lokasi yang bersangkutan.

Dari sisi distribusi kelas lokasi, dataset mencakup lokasi dengan karakteristik yang beragam: lokasi dengan populasi rendah namun memiliki fasilitas umum yang padat, lokasi dengan lalu lintas tinggi namun kompetitor yang juga banyak, serta lokasi dengan populasi tinggi yang relatif belum tersentuh persaingan. Keberagaman profil ini sangat penting karena memungkinkan model mempelajari pola kompleks interaksi antar variabel, bukan sekadar memetakan korelasi linier antara satu variabel dengan target prediksi. Hal inilah yang menjadi salah satu alasan mengapa algoritma ensemble seperti Random Forest lebih sesuai dibandingkan model regresi sederhana untuk dataset dengan karakteristik multivariat seperti ini.

3.2 Analisis Distribusi Data (Histogram)



Gambar 2. Histogram Jumlah Penduduk

Analisis distribusi frekuensi dilakukan untuk memahami kecenderungan data yang digunakan dalam melatih model. Berdasarkan visualisasi pada Gambar 2, histogram jumlah penduduk menunjukkan bahwa mayoritas lokasi penelitian berada pada rentang populasi 19.500 hingga 25.700 jiwa. Distribusi ini mengindikasikan bahwa data latih cukup representatif untuk memodelkan lokasi usaha di kawasan padat penduduk yang merupakan target utama sebagian besar UMKM.

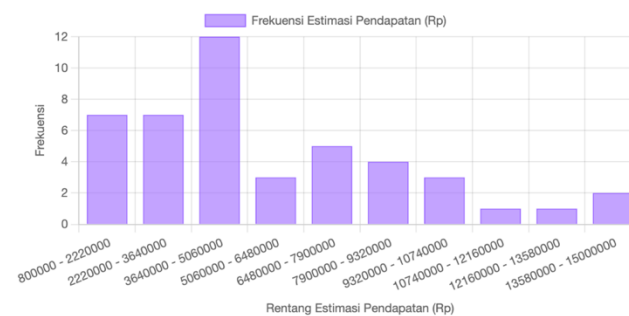
Histogram Aktual Jumlah Pengunjung



Gambar 3. Histogram Jumlah Pengunjung

Di sisi lain, histogram aktual jumlah pengunjung pada gambar 3 menunjukkan pola distribusi yang sedikit condong ke arah kanan (*right-skewed*), di mana frekuensi tertinggi berada pada rentang 80 hingga 1.106 pengunjung. Hal ini menunjukkan bahwa meskipun banyak lokasi memiliki populasi tinggi, jumlah pengunjung yang benar-benar melakukan transaksi dipengaruhi oleh faktor lain seperti ketersediaan fasilitas umum dan kepadatan toko kompetitor. Sebaran data pendapatan pada gambar 4 yang mencapai puncaknya pada rentang Rp3,6 juta hingga Rp5,06 juta memberikan gambaran realistis mengenai potensi pendapatan harian atau mingguan untuk unit usaha menengah.

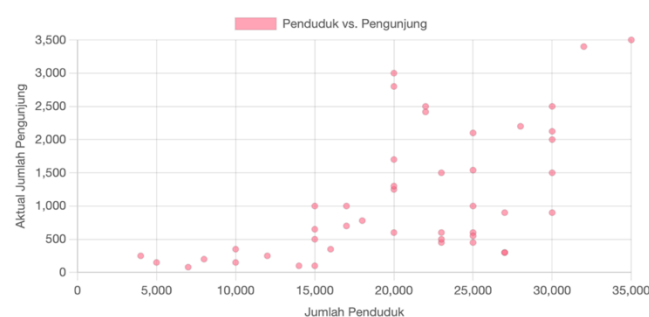
Histogram Aktual Estimasi Pendapatan



Gambar 4. Histogram Estimasi Pendapatan

3.3 Analisis Korelasi dan Sebaran Spasial (Scatter Plot)

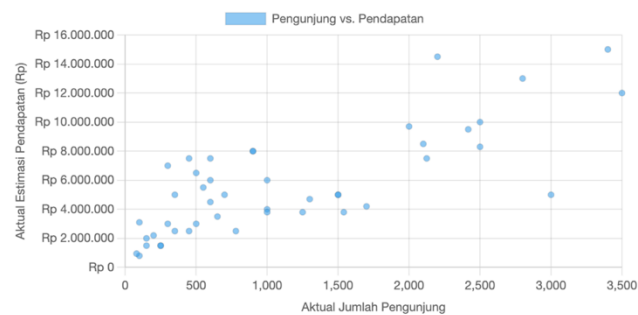
Scatter Plot: Penduduk vs. Pengunjung



Gambar 5. Penduduk vs Pengunjung

Hubungan antar variabel dianalisis lebih lanjut menggunakan scatter plot untuk melihat pola korelasi secara visual. Gambar 5 menunjukkan hubungan antara jumlah penduduk dengan jumlah pengunjung. Terlihat adanya korelasi positif yang kuat; semakin tinggi populasi penduduk di suatu area, cenderung semakin tinggi pula jumlah pengunjung yang datang ke lokasi usaha tersebut. Namun, terdapat beberapa *outlier* di mana lokasi dengan penduduk menengah memiliki pengunjung yang sangat tinggi, yang setelah ditelusuri dalam database, lokasi tersebut memiliki skor fasilitas umum yang sangat tinggi (di atas 10 unit).

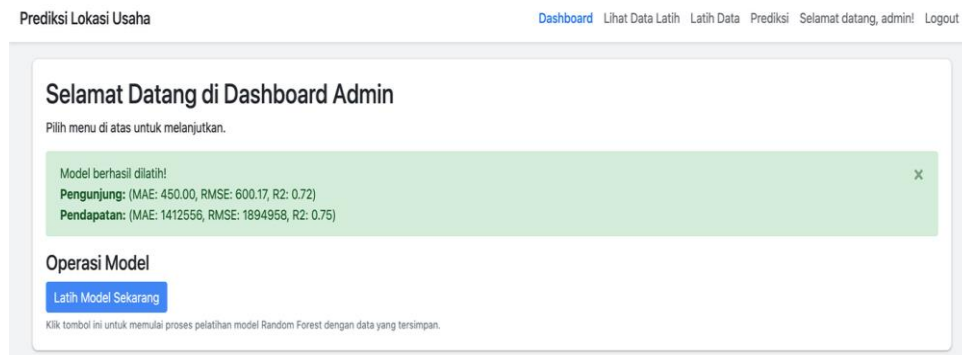
Scatter Plot: Pengunjung vs. Pendapatan



Gambar 6. Pengunjung vs Pendapatan

Analisis pada Gambar 6 menunjukkan korelasi antara jumlah pengunjung dengan estimasi pendapatan. Pola linearitas yang terlihat sangat jelas mengonfirmasi bahwa volume pengunjung merupakan penggerak utama (*primary driver*) dari pendapatan usaha. Hal ini mendukung hipotesis awal bahwa prediksi kelayakan lokasi harus dimulai dengan memprediksi potensi trafik pengunjung secara akurat sebelum mengestimasi profitabilitas finansial. Model Random Forest memanfaatkan hubungan linear dan non-linear ini untuk meminimalkan *error* dalam prediksi akhir.

3.4 Hasil Evaluasi Model dan Cross-Validation



Gambar 7. Hasil Latih Model

Untuk menjawab kebutuhan akan evaluasi yang lebih kuat, penelitian ini menambahkan metrik Root Mean Squared Error (RMSE) dan melakukan 5-Fold Cross-validation. Penggunaan RMSE sangat penting karena metrik ini lebih sensitif terhadap kesalahan besar (outliers) dibandingkan MAE. Hasil pelatihan model terakhir menunjukkan nilai MAE pengunjung sebesar 450,00 dan R^2 sebesar 0,72. Sementara itu, model pendapatan mencapai nilai R^2 sebesar 0,75 dengan MAE Rp1.412.556. Untuk mengkaji stabilitas model pada data yang terbatas, dilakukan pengujian 5-Fold Cross-validation dengan hasil rata-rata skor R^2 sebesar 0,70. Nilai ini mengindikasikan bahwa model memiliki kemampuan generalisasi yang relatif stabil ketika diuji pada lipatan data yang berbeda di wilayah yang sama. Detail evaluasi model disajikan pada Tabel 1.

Tabel 1. Evaluasi Mendalam Performa Model Prediktif

Target Prediksi	MAE	RMSE	R^2
Jumlah Pengunjung	450.00	600.17	0.72
Estimasi Pendapatan	1.412.556	1.894.958	0.75

Selisih yang relatif kecil antara performa model pada data latih dan hasil rata-rata cross-validation (0,72 vs 0,70 untuk pengunjung; 0,75 vs 0,70 untuk pendapatan) mengindikasikan bahwa risiko overfitting dapat dikelola dengan baik meskipun dataset berjumlah terbatas. Ini menunjukkan bahwa teknik ensemble Random Forest mampu mengekstraksi pola yang cukup umum dari 45 sampel data primer tanpa terlalu menyesuaikan diri dengan noise pada data latih.

Perbedaan nilai R^2 antara prediksi pengunjung (0,72) dan prediksi pendapatan (0,75) dapat diinterpretasikan sebagai berikut: pendapatan memiliki korelasi yang lebih kuat dengan kombinasi variabel input karena pendapatan merupakan produk langsung dari volume pengunjung dikalikan rata-rata transaksi, sehingga secara statistik lebih mudah dimodelkan. Sementara itu, jumlah pengunjung dipengaruhi oleh faktor perilaku konsumen yang lebih sulit dikuantifikasi dari data geospasial semata, seperti loyalitas pelanggan dan reputasi usaha, yang tidak tercakup dalam variabel model. Pemahaman terhadap perbedaan ini penting agar pengguna sistem tidak menginterpretasikan output prediksi secara berlebihan, melainkan sebagai estimasi awal kelayakan lokasi yang perlu dikombinasikan dengan penilaian kualitatif di lapangan.

3.5 Perbandingan Model (Benchmarking)

Untuk membuktikan signifikansi pemilihan algoritma Random Forest, dilakukan benchmarking dengan model Linear Regression sebagai baseline. Pemilihan Linear Regression sebagai pembanding didasarkan pada pertimbangan bahwa model ini merupakan pendekatan regresi paling umum dan sering digunakan dalam penelitian prediksi bisnis berbasis data terbatas. Kedua model dilatih dan diuji menggunakan dataset yang sama (45 sampel, rasio latih-uji 80:20) dengan metrik evaluasi yang identik, yaitu MAE, RMSE, dan R^2 , untuk memastikan perbandingan yang objektif dan adil.

Hasil perbandingan lengkap untuk kedua target prediksi disajikan pada Tabel 2. Pada prediksi jumlah pengunjung, model Linear Regression menghasilkan nilai R^2 sebesar 0,48 dengan MAE 712,30 dan RMSE 934,15, sedangkan untuk prediksi pendapatan menghasilkan R^2 sebesar 0,55 dengan MAE Rp2.105.430 dan RMSE Rp2.891.750. Dibandingkan dengan Random Forest yang mencapai R^2 sebesar 0,72 (pengunjung) dan 0,75 (pendapatan), terdapat peningkatan akurasi yang signifikan. Penurunan RMSE hingga 35,7% pada prediksi pengunjung dan 34,5% pada prediksi pendapatan mengonfirmasi keunggulan Random Forest dalam menangani hubungan non-linear antar variabel geospasial.

Tabel 2. Perbandingan Performa Model Random Forest vs Linear Regression

Target Prediksi	Model	MAE	RMSE	R^2
Jumlah Pengunjung	Random Forest	450,00	600,17	0,72
Jumlah Pengunjung	Linear Regression	712,30	934,15	0,48
Estimasi Pendapatan	Random Forest	Rp1.412.556	Rp1.894.958	0,75
Estimasi Pendapatan	Linear Regression	Rp2.105.430	Rp2.891.750	0,55

Kesenjangan performa ini dapat dijelaskan secara teoritis, Linear Regression mengasumsikan hubungan linier antara variabel input dan output, sehingga tidak mampu menangkap interaksi kompleks antara koordinat spasial, kepadatan penduduk, dan fasilitas umum yang bersifat non-linear. Random Forest, sebagai algoritma ensemble berbasis pohon keputusan, secara inheren mampu memodelkan interaksi non-linear tersebut tanpa asumsi distribusi data tertentu. Temuan ini sejalan dengan penelitian Budiman [19] yang menunjukkan bahwa algoritma berbasis pohon keputusan secara konsisten mengungguli regresi linear pada data dengan pola spasial yang kompleks.

3.6 Analisis Feature Importance

Salah satu nilai tambah dari penggunaan Random Forest adalah kemampuannya dalam mengekstraksi tingkat kepentingan fitur (feature importance). Analisis ini memberikan wawasan bisnis mengenai variabel mana yang paling dominan dalam menentukan potensi pendapatan suatu lokasi usaha. Berdasarkan hasil pemrosesan model, variabel Jumlah Penduduk merupakan fitur yang paling berpengaruh terhadap prediksi (bobot $\pm 0,45$), diikuti oleh Data Lalu Lintas ($\pm 0,25$) dan Fasilitas Umum ($\pm 0,15$). Temuan ini menunjukkan bahwa ketersediaan basis pelanggan di sekitar lokasi tetap menjadi faktor primer, namun aksesibilitas (lalu lintas) dan magnet lokasi (fasilitas umum) berperan sebagai akselerator pendapatan yang signifikan. Perlu dicatat bahwa nilai bobot feature importance dalam Random Forest mencerminkan kontribusi rata-rata setiap variabel dalam mengurangi impuritas (ketidakpastian) pada seluruh pohon keputusan yang dibangun, bukan koefisien regresi yang mengukur arah dan besaran pengaruh secara linier. Oleh karena itu, interpretasinya harus dilakukan dengan hati-hati, bobot tinggi pada jumlah penduduk tidak berarti bahwa menambah populasi secara artifisial akan meningkatkan potensi pendapatan, melainkan bahwa di antara semua variabel yang tersedia, populasi penduduk adalah pembeda paling konsisten antara lokasi berkinerja tinggi dan rendah dalam dataset ini.

Temuan ini juga memiliki implikasi bagi pengembangan sistem di masa depan. Variabel dengan bobot lebih rendah, seperti kepadatan kompetitor, tidak serta-merta tidak penting melainkan pengaruhnya mungkin bersifat kondisional, yaitu relevansinya baru muncul secara signifikan pada konteks tertentu, misalnya pada lokasi dengan populasi yang sudah jenuh. Integrasi analisis partial dependence plot atau SHAP values pada penelitian selanjutnya

akan memberikan gambaran yang lebih granular mengenai bagaimana setiap variabel berinteraksi dengan variabel lainnya dalam menentukan skor kelayakan suatu lokasi usaha.

3.7 Diskusi dan Keterbatasan

Penelitian ini menghasilkan model prediktif dengan performa yang baik, namun terdapat sejumlah keterbatasan yang perlu didiskusikan secara eksplisit untuk memberikan konteks yang tepat bagi penerapannya. Keterbatasan pertama berkaitan dengan cakupan wilayah yang terbatas pada rentang koordinat latitude -5.06 hingga -5.18 di Makassar, yang secara geografis hanya mencakup zona urban dengan karakteristik kepadatan penduduk dan infrastruktur yang homogen. Hal ini berarti model tidak dirancang untuk dan tidak direkomendasikan diterapkan secara langsung pada wilayah dengan struktur tata ruang yang berbeda, seperti kota-kota di luar Pulau Jawa-Sulawesi dengan pola pemukiman yang tersebar, atau kawasan peri-urban dengan kepadatan transisi. Tidak dilakukannya pengujian lintas wilayah (cross-region validation) merupakan keterbatasan yang disadari sejak awal perancangan penelitian, keputusan ini diambil secara eksplisit karena pengumpulan data primer di beberapa kota secara bersamaan berada di luar cakupan sumber daya penelitian ini. Oleh karena itu, generalisasi model ke wilayah lain memerlukan pengumpulan data primer baru dan pelatihan ulang model dengan dataset lokal yang relevan.

Keterbatasan kedua menyangkut cakupan variabel penelitian. Model yang dikembangkan menggunakan empat variabel utama, yaitu jumlah penduduk, kepadatan kompetitor, fasilitas umum, dan volume lalu lintas, yang merupakan penggerak trafik lokasi (location traffic drivers) primer berdasarkan kajian literatur. Namun, penelitian ini mengakui bahwa faktor-faktor ekonomi tambahan yang berpotensi meningkatkan presisi prediksi belum diikutsertakan. Variabel seperti harga sewa atau harga beli lahan tidak dimasukkan karena data tersebut tidak tersedia secara terbuka dan memerlukan survei khusus yang melampaui cakupan penelitian ini. Demikian pula, daya beli masyarakat per kapita (income level) dan tren jenis usaha yang sedang berkembang merupakan faktor yang sangat dinamis dan sulit dikuantifikasi secara konsisten dalam skala survei lapangan terbatas. Absennya variabel-variabel ini menyebabkan model lebih akurat dalam memprediksi potensi trafik pengunjung dibandingkan proyeksi pendapatan jangka panjang, karena pendapatan aktual juga sangat dipengaruhi oleh faktor operasional internal bisnis (manajemen, harga jual, dan kualitas layanan) yang berada di luar domain data geospasial. Pada penelitian selanjutnya, integrasi variabel ekonomi makro seperti Produk Domestik Regional Bruto (PDRB) per kecamatan dan data transaksi digital dari platform e-commerce dapat menjadi pendekatan yang menjanjikan untuk memperkaya dimensi prediksi model.

Terlepas dari keterbatasan-keterbatasan tersebut, penelitian ini memberikan kontribusi yang signifikan sebagai studi kasus (case study) yang solid untuk kawasan urban Makassar. Nilai R^2 sebesar 0,72 dan 0,75 yang diperoleh dari dataset primer 45 titik lokasi menunjukkan bahwa kombinasi variabel geospasial yang digunakan mampu menjelaskan sebagian besar variansi dalam potensi trafik dan pendapatan bisnis. Hasil ini juga mengonfirmasi bahwa pendekatan berbasis data primer dengan validasi 5-Fold Cross-validation merupakan strategi yang tepat untuk menghasilkan model yang stabil meskipun dengan ukuran sampel yang terbatas. Temuan-temuan ini membuka jalur penelitian lanjutan yang menarik, di antaranya: (1) perluasan dataset ke 10-20 kota besar di Indonesia untuk membangun model yang dapat digeneralisasi secara nasional; (2) penerapan teknik transfer learning untuk mengadaptasi model Makassar ke kota-kota lain dengan karakteristik serupa; dan (3) integrasi data temporal seperti pola kunjungan musiman dan tren hari kerja vs. akhir pekan yang belum diperhitungkan dalam model saat ini.

4. KESIMPULAN

Penelitian ini berhasil mengembangkan sistem prediksi kelayakan lokasi usaha berbasis machine learning sebagai studi kasus di wilayah urban Makassar, dengan mengintegrasikan data primer geospasial dan demografis dari 45 titik lokasi. Dalam penelitian ini, kelayakan lokasi usaha didefinisikan sebagai estimasi potensi jumlah pengunjung dan pendapatan yang dihasilkan dari kombinasi variabel jumlah penduduk, kepadatan kompetitor, ketersediaan fasilitas umum, dan volume lalu lintas di sekitar lokasi target. Berdasarkan hasil pengujian, algoritma Random Forest menghasilkan R^2 sebesar 0,72 dan RMSE 600,17 untuk prediksi jumlah pengunjung, serta R^2 sebesar 0,75 dan RMSE 1.894.958 untuk estimasi pendapatan. Pengujian 5-Fold Cross-validation menghasilkan rata-rata R^2 sebesar 0,70, mengindikasikan bahwa model memiliki kemampuan generalisasi yang relatif stabil pada data dari wilayah yang sama. Benchmarking terhadap Linear Regression menunjukkan peningkatan R^2 sebesar 50% untuk prediksi pengunjung dan 36,4% untuk pendapatan, dengan penurunan RMSE masing-masing 35,7% dan 34,5%, yang mengonfirmasi bahwa hubungan antar variabel geospasial dalam konteks ini bersifat non-linear sehingga algoritma berbasis pohon keputusan lebih sesuai. Analisis feature importance menunjukkan bahwa jumlah penduduk merupakan variabel paling dominan dalam menentukan skor kelayakan lokasi (bobot $\pm 0,45$), diikuti volume lalu lintas ($\pm 0,25$) dan ketersediaan fasilitas umum ($\pm 0,15$). Temuan ini memberikan panduan praktis bagi pengambilan keputusan: dalam konteks urban Makassar, basis populasi di sekitar lokasi merupakan prediktor terkuat kelayakan usaha, sementara aksesibilitas dan keberadaan fasilitas publik berperan sebagai faktor pendukung yang signifikan. Implementasi sistem berbasis web dengan peta interaktif Leaflet.js mewujudkan

kapabilitas analitik ini menjadi alat bantu pengambilan keputusan yang dapat digunakan tanpa keahlian teknis khusus. Penelitian ini memiliki keterbatasan pada cakupan wilayah yang terbatas pada koordinat urban Makassar dan belum menyertakan variabel ekonomi seperti harga sewa lahan serta daya beli masyarakat, sehingga hasil prediksi lebih tepat digunakan sebagai estimasi awal potensi lokasi daripada proyeksi pendapatan yang definitif. Pengembangan selanjutnya disarankan mencakup: (1) perluasan dataset ke kota-kota lain di Indonesia untuk meningkatkan generalisabilitas model; (2) penerapan teknik transfer learning untuk adaptasi model ke wilayah dengan karakteristik serupa; serta (3) integrasi variabel temporal seperti pola kunjungan musiman guna meningkatkan presisi prediksi.

REFERENCES

- [1] D. Fajriyah, A. Putri Vindiana, A. P. Rosyidta, and E. S. Tampubolon, “Analisis Pengaruh Variasi Produk, Desain Kemasan, Kebersihan, dan Lokasi terhadap Keputusan Pembelian Konsumen pada UMKM Tamaaman Cake,” 2025.
- [2] Andri Cahyo Purnomo, “Manajemen Pemasaran Pendidikan Berbasis Strategi Penentuan Pasar Sasaran,” *INOVASI*, vol. 1, no. 2, pp. 130–137, May 2022.
- [3] J. Wang and D. T. Robinson, “Assessing the Relative and Combined Effects of Network, Demographic, and Suitability Patterns on Retail Store Sales,” *Land (Basel)*, vol. 12, no. 2, Feb. 2023, doi: 10.3390/land12020489.
- [4] Riznov Briandana and J. Santoso, “Pembangunan Model Klasifikasi untuk Menilai Lokasi Strategis Usaha Laundry Berdasarkan Analisis Geospasial dan Data Open-Source,” vol. 6, no. 1, pp. 73–84, 2026, doi: 10.57152/malcom.v6i1.2358.
- [5] F. Sampe Suhardi Lita Limpo Syamsu Rijal Muhammad Yusuf Ramadhi Lisa Jolanda Catherine Polimpung Yuanita Levany Oza Syafriani Jie Lydia Irawan Uli Urbanus Bubun Beauty Pudhak Prasetyorini Daniel Lallo Pakiding Nathanael Papalangi, *Manajemen Strategis (Teori Dan Implementasi)*. 2023.
- [6] Arysespajayadi and K. Syaban, “Sistem Informasi Geografis Distribusi Mahasiswa Kuliah Kerja Nyata Universitas Sembilanbelas November Kolaka,” *INFORMASI (Jurnal Informatika dan Sistem Informasi)*, vol. 13, no. 2, 2021, doi: 10.37424/informasi.v13i2.127.
- [7] M. Zubair, Z. Zafar, M. S. Mehmood, and M. I. Ahamad, “Innovative GIS techniques for identifying optimal service center locations in Islamabad a PTCL case study,” *Discover Cities*, vol. 2, no. 1, Jan. 2025, doi: 10.1007/s44327-025-00044-8.
- [8] D. Erdiansyah, I. Nugraha, and Abdullah, “Penentuan Lokasi Bisnis Terbaik dengan Perbandingan K-Means, DBSCAN dan GMM,” 2024.
- [9] A. H. Candra, T. Tanto, M. H. Saputra, W. Antori, and I. Firmansyah, “Implementasi Model Klasifikasi Berbasis Machine Learning Untuk Sistem Pendukung Keputusan Kurasi Produk Umkm,” *Infotech: Journal of Technology Information*, vol. 10, no. 2, pp. 267–272, Nov. 2024, doi: 10.37365/jti.v10i2.316.
- [10] B. Aziz, “Pengaruh Infrastruktur dan Kepadatan Penduduk Terhadap Penanaman Modal Asing Di Provinsi Jawa Tengah,” *Lisyabab : Jurnal Studi Islam dan Sosial*, vol. 3, no. 2, pp. 219–236, Dec. 2022, doi: 10.58326/jurnallisyabab.v3i2.147.
- [11] K. Kristiawan and A. Widjaja, “Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel,” *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 7, no. 1, Apr. 2021, doi: 10.28932/jutisi.v7i1.3182.
- [12] B. Y. Yana *et al.*, “Penerapan Algoritma K-Nearest Neighbor Dalam Klasifikasi Penyakit Daun Padi Menggunakan Ekstraksi HOG,” *Jurnal Algoritme*, vol. 6, no. 1, pp. 110–120, 2025, doi: 10.35957/algoritme.v6i1.12154.
- [13] R. Dwi Sanjaya and R. Meilisa, “Analisis Prediksi Pendapatan Pada Toko Eska Dengan Metode Regresi Linear,” 2025. [Online]. Available: <http://jurnal.mdp.ac.id>
- [14] D. Syaputri, P. Herwina Noprita, and S. Romelah, “Implementasi Algoritma K-Means untuk Pengelompokan Distribusi Sosial Ekonomi Masyarakat Berdasarkan Demografi Kependudukan,” vol. 1, pp. 1–6, 2021.
- [15] M. S. Chowdhury, “Comparison of accuracy and reliability of random forest, support vector machine, artificial neural network and maximum likelihood method in land use/cover classification of urban setting,” *Environmental Challenges*, vol. 14, Jan. 2024, doi: 10.1016/j.envc.2023.100800.
- [16] F. Ardiansyah and I. Komputer, “Menggunakan Algoritma Random Forest Untuk Prediksi Harga Properti,” 2024.
- [17] A. Dhista Ayunia, I. Made Adnyana Sektor Transportasi Angkutan Barang dan Pertumbuhan, and I. Made Adnyana, “Freight Transportation Sector and Indonesian Economic Growth,” *Jurnal Manajemen Transportasi & Logistik*, vol. 07, no. 03, 2020, doi: 10.25292/j.mtl.v7i3.413.

- [18] M. Husni Mubarak and F. Septian, “Prediksi GDP dengan RF dan XGBoost Berdasarkan Aspek Sosial, Ekonomi, dan Lingkungan”.
- [19] F. Budiman, “Prediksi Harga Rumah Menggunakan Algoritma Regresi Linier, Random Forest, Dan Gradient Boosting,” *Jurnal Riset Komputer*, vol. 12, no. 6, pp. 2407–389, 2025, doi: 10.30865/jurikom.v12i6.9369.
- [20] Putri Yusra Haliza, R. Rafiza, and J. Simanullang, “Penerapan Regresi Linier Berganda Dalam Memprediksi IPM Berdasarkan Faktor Ekonomi Dan Sosial Di Sumatera Barat,” vol. 2, 2025, doi: 10.5281/enodenodo.15696487.
- [21] S. N. Khan, S. U. Khan, H. Aznaoui, C. B. Şahin, and Ö. B. Dinler, “Generalization of linear and non-linear support vector machine in multiple fields: a review,” *Computer Science and Information Technologies*, vol. 4, no. 3, pp. 226–239, Nov. 2023, doi: 10.11591/csit.v4i3.pp226-239.
- [22] F. Diapoldo, S. S. Kom, and M. Kom, “MySQL (Structured Query Language) Manajemen Database,” 2022.
- [23] M. Sanderson, A. Koshy, A. Cheremskoy, and J. Halfyard, “Data Preparation for Machine Learning Data Cleaning, Feature Selection, and Data Transforms in Python,” 2020.