Jurnal Elektronika dan Teknik Informatika Terapan (JENTIK)
Volume 3 Nomor 4 Desember 2025
e-ISSN : 2988-0874; p-ISSN : 2988-0866, Hal 65-85
DOI: https://doi.org/10.59061/jentik.v3i4.1550
Tersedia: https://e-journal.poltek-kampar.ac.id/index.php/JENTIK

Prediksi Jumlah Penduduk Malaysia Berdasarkan Faktor Demografi dan Etnis Menggunakan Metode Random Forest

Nurkholis[1*], Andri Nofiar. Am[2], Hayatul Khairul Rahmat[3], Rahmad Akbar[4]
[1]Program Studi Teknologi Rekayasa Logistik, Politeknik Kampar, Indonesia.
[2]Program Studi Bisnis Digital, Politeknik Negeri Bengkalis, Indonesia.
[3,4]Program Studi Administrasi Bisnis Internasional, Politeknik Kampar, Indonesia.
Korespondensi penulis: nurkholis11151996@gmail.com

Abstract. *Accurate population prediction is essential for development planning, particularly in multiethnic countries such as Malaysia, yet conventional statistical approaches generally do not explicitly incorporate ethnic composition. This study aims to develop a population prediction model for Malaysia using a Random Forest Regressor that accounts for demographic factors and the composition of six ethnic groups. Historical data from 16 Malaysian states spanning 1981–2025 (673 state-year observations) were processed through feature engineering, including population lag variables, growth rate, and ethnic proportions, yielding 21 independent variables. The model was trained using a time-aware train-test split (80% training, 20% testing) and validated with 5-fold time series cross-validation. Results on the test set (2020–2025) showed strong performance, with R^2 = 0.9769, RMSE = 259 thousand people, MAE = 97 thousand people, and sMAPE = 3.17%. Feature importance analysis revealed that the previous year's population (pop_lag1) and gender composition were the most dominant predictors. Projections indicate Malaysia's population will reach approximately 38.5 million by 2030. These findings demonstrate that Random Forest with demographic-ethnic features can produce more accurate population predictions, providing a data-driven policy support tool for the Malaysian government.*
Keywords: *demographics, ethnicity, machine learning, population prediction, random forest.*

Abstrak. Prediksi jumlah penduduk yang akurat sangat penting bagi perencanaan pembangunan, terutama di negara multietnis seperti Malaysia, namun pendekatan statistik konvensional umumnya belum mengintegrasikan faktor etnisitas secara eksplisit. Penelitian ini bertujuan mengembangkan model prediksi jumlah penduduk Malaysia menggunakan Random Forest Regressor dengan mempertimbangkan faktor demografi dan komposisi enam kelompok etnis. Data historis 16 negeri bagian Malaysia periode 1981–2025 (673 observasi state-year) diolah melalui rekayasa fitur meliputi variabel lag populasi, laju pertumbuhan, dan proporsi etnis, menghasilkan 21 variabel independen. Model dilatih menggunakan pembagian data time-aware (80% latih, 20% uji) dan divalidasi dengan 5-fold time series cross-validation. Hasil pada data uji (2020–2025) menunjukkan performa sangat baik dengan R^2 = 0,9769, RMSE = 259 ribu jiwa, MAE = 97 ribu jiwa, dan sMAPE = 3,17%. Analisis feature importance mengungkapkan bahwa populasi tahun sebelumnya (pop_lag1) serta komposisi jenis kelamin merupakan prediktor paling dominan. Proyeksi menunjukkan populasi Malaysia diperkirakan mencapai sekitar 38,5 juta jiwa pada tahun 2030. Temuan ini membuktikan bahwa Random Forest dengan fitur demografi-etnis mampu menghasilkan prediksi populasi yang lebih akurat, sekaligus menyediakan alat pendukung kebijakan kependudukan berbasis data bagi pemerintah Malaysia.
Kata kunci: demografi, etnisitas, machine learning, prediksi penduduk, random forest.

1. LATAR BELAKANG

Prediksi jumlah penduduk merupakan salah satu kebutuhan mendasar dalam perencanaan pembangunan nasional yang efektif. Pemerintah di seluruh dunia memerlukan proyeksi populasi yang akurat untuk mengalokasikan sumber daya secara optimal di bidang infrastruktur, layanan kesehatan, pendidikan, perumahan, dan

Naskah Masuk: 01 September 2025; Revisi: 15 Oktober 2025; Diterima: 25 November 2025; Terbit: 30 Desember 2025

kebijakan sosial. Tanpa proyeksi yang andal, perencanaan jangka menengah dan panjang berisiko mengalami ketidaksesuaian antara kapasitas layanan publik dan kebutuhan masyarakat yang terus berkembang (Wilson et al., 2022). Tantangan ini menjadi semakin kompleks pada negara-negara multietnik, di mana dinamika pertumbuhan penduduk tidak bersifat homogen, melainkan bervariasi secara signifikan berdasarkan komposisi etnis, distribusi jenis kelamin, dan pola migrasi regional. Malaysia merupakan salah satu negara Asia Tenggara dengan komposisi demografis yang paling beragam. Berdasarkan data Jabatan Perangkaan Malaysia (DOSM, 2024), jumlah penduduk Malaysia pada tahun 2024 diperkirakan mencapai 34,1 juta jiwa, meningkat dari 33,4 juta jiwa pada tahun 2023 dengan laju pertumbuhan 1,9 persen per tahun. Namun, pertumbuhan ini tidak merata; empat negeri bagian Sabah (4,0%), W.P. Kuala Lumpur (3,0%), Selangor (2,1%), dan Johor (1,9%) mencatat laju pertumbuhan di atas rata-rata nasional. Dari sisi komposisi etnis, Bumiputera Melayu dan kelompok Bumiputera lainnya mencakup sekitar 69,9% dari populasi warga negara pada tahun 2023, sementara etnis Cina sebesar 22,8% dan India sebesar 6,6% (Department of Statistics Malaysia, 2024). Lebih jauh, proyeksi DOSM mengindikasikan populasi Malaysia akan mencapai puncaknya pada 42,37 juta jiwa di tahun 2059 sebelum mengalami penurunan, disertai perubahan signifikan pada lanskap etnis di mana proporsi etnis Cina diproyeksikan turun dari 21,1% 2030 menjadi hanya 14,8% pada tahun 2060 (Department of Statistics Malaysia, 2025). Dinamika multidimensional ini menegaskan bahwa proyeksi kependudukan Malaysia tidak dapat dipisahkan dari faktor etnis.

Metode statistik tradisional seperti ARIMA (Autoregressive Integrated Moving Average) dan model kohort-komponen telah lama digunakan dalam proyeksi demografis. Meskipun metode-metode tersebut memiliki fondasi matematis yang kuat, penelitian terkini menunjukkan keterbatasan substansial dalam menangkap hubungan non-linier yang kompleks antara variabel demografis dan sosial-ekonomi (Kontopoulou et al., 2023). Kwarteng & Andreevich (2024) dalam perbandingan metode ARIMA, SARIMA, dan Prophet menemukan bahwa model konvensional cenderung menghasilkan bias signifikan ketika dihadapkan pada tren demografi yang fluktuatif dan non-stasioner. Lebih lanjut, Wilson et al. (2022) menegaskan bahwa pendekatan machine learning belum banyak dieksplorasi untuk prediksi populasi di wilayah multietnik kecil, meskipun

potensinya sangat besar. Di Malaysia secara khusus, sebagian besar studi demografis masih mengandalkan pendekatan statistik konvensional yang tidak mengintegrasikan variabel etnisitas secara eksplisit sebagai fitur prediktif, sehingga menghasilkan proyeksi yang kurang mencerminkan realitas keragaman demografis negara tersebut.

Berbagai penelitian mutakhir telah membuktikan keunggulan algoritma ensemble berbasis pohon keputusan dalam prediksi populasi. Baker, Swanson & Tayman (2023) menunjukkan bahwa Boosted Regression Trees menghasilkan akurasi yang lebih tinggi dan outlier yang lebih sedikit dibandingkan metode Hamilton-Perry dalam prediksi populasi tingkat sensus AS, Penelitian Zhao et al. (2025) membuktikan bahwa Random Forest dengan feature selection berbasis MDA mampu mencapai $R^2 = 0{,}913$ dan MAPE = 0,174 dalam spesialisasi populasi spasial. Dalam Tanmoy (2024) dalam studi komparatif menemukan bahwa Random Forest secara konsisten mengungguli Decision Tree, Linear Regression, dan K-Nearest Neighbors untuk prediksi populasi. Namun demikian, penelitian yang secara spesifik mengaplikasikan Random Forest dengan mengintegrasikan fitur demografi jenis kelamin, komposisi absolut enam kelompok etnis Malaysia, proporsi etnis, dan fitur lag populasi historis dalam satu kerangka prediktif belum pernah dilaporkan dalam literatur. Kebaruan (novelty) penelitian ini terletak pada integrasi komprehensif 21 fitur multidimensi yang mencakup aspek demografi dan etnis secara bersamaan, dengan validasi berbasis waktu (time-aware split) yang mencegah kebocoran data, serta analisis feature importance untuk mengidentifikasi kontribusi relatif setiap kelompok etnis terhadap dinamika populasi.

Bertolak dari kesenjangan penelitian di atas, studi ini bertujuan untuk: (1) membangun model prediksi jumlah penduduk Malaysia menggunakan Random Forest yang mengintegrasikan faktor demografi (populasi jenis kelamin, rasio jenis kelamin, lag populasi 1–3 tahun, tingkat pertumbuhan) dan faktor etnis (populasi absolut dan proporsi enam kelompok etnis: Bumiputera Melayu, Bumiputera lainnya, Cina, India, WN lain, dan Non-WN) pada 16 negeri bagian Malaysia untuk periode 1981–2025; (2) mengevaluasi performa model secara komprehensif menggunakan metrik R^2, RMSE, MAE, dan sMAPE dengan validasi 5-fold cross-validation; dan (3) mengidentifikasi faktor-faktor paling berpengaruh terhadap dinamika kependudukan Malaysia melalui

analisis Gini Feature Importance, yang dapat menjadi landasan empiris bagi perumusan kebijakan kependudukan yang lebih berbasis data.

2. KAJIAN TEORITIS

2.1 Dinamika Kependudukan Malaysia

Malaysia merupakan negara multikultural dengan tiga kelompok etnis utama: Bumiputera (Melayu dan suku asli Sabah-Sarawak), Cina, dan India. Komposisi etnis ini bukan sekadar karakteristik demografis, melainkan merupakan faktor struktural yang membentuk pola fertilitas, migrasi, dan urbanisasi berbeda antar negeri bagian (Elsawahli, 2016). Bumiputera Melayu mencatat tingkat kelahiran alami yang lebih tinggi dibandingkan etnis Cina dan India, sementara populasi etnis Cina mengalami penurunan proporsi secara konsisten sejak kemerdekaan 1957 akibat tingkat fertilitas yang lebih rendah (Department of Statistics Malaysia, 2024). Pola divergen ini mengindikasikan bahwa proyeksi populasi Malaysia yang akurat harus mempertimbangkan komposisi etnis secara eksplisit, bukan sekadar agregat total penduduk. Distribusi spasial penduduk Malaysia juga sangat heterogen. Selangor, sebagai negeri bagian paling padat, menyumbang lebih dari 21% total populasi nasional pada tahun 2024, sementara W.P. Labuan dan W.P. Putrajaya masing-masing hanya dihuni sekitar 100 ribu jiwa (DOSM, 2024). Ketimpangan skala ini menimbulkan tantangan metodologis tersendiri dalam pemodelan prediksi, karena model yang dilatih pada data heterogen rentan terhadap bias pada wilayah berskala kecil (Baker, Swanson & Tayman, 2023). Proyeksi DOSM untuk periode 2030–2060 mengindikasikan percepatan ketidakseimbangan gender dan penurunan tajam proporsi etnis Cina, yang semakin memperkuat urgensi model prediktif yang mampu menangkap dinamika multivariat ini.

2.2 Algoritma Random Forest

Random Forest adalah algoritma ensemble learning yang diusulkan oleh Breiman (2001) dan telah menjadi salah satu metode machine learning paling banyak digunakan dalam berbagai domain prediksi. Algoritma ini membangun sejumlah besar pohon keputusan secara paralel menggunakan teknik bootstrap aggregating (bagging) dan pemilihan fitur secara acak (random subspace method). Setiap pohon dilatih pada subset data yang berbeda-beda, dan prediksi akhir diperoleh melalui rata-rata (untuk regresi) dari

seluruh pohon. Secara matematis, prediksi Random Forest dapat dinyatakan sebagai: $\hat{y}(x) = (1/B) \Sigma^{b}_{=1}{}^{B} f^{b}(x)$, di mana B adalah jumlah pohon dan $f^{b}(x)$ adalah prediksi pohon ke-b untuk input x.

Menurut Hu & Szymczak (2023) dalam tinjauan komprehensif tentang Random Forest untuk analisis data longitudinal menegaskan bahwa RF merupakan alternatif superior terhadap metode statistik standar, khususnya dalam konteks data berdimensi tinggi dengan struktur korelasi temporal yang kompleks. Keunggulan ini sangat relevan untuk prediksi populasi yang bersifat time-series dengan banyak fitur demografis yang saling berkorelasi. Zhao et al. (2025) lebih lanjut membuktikan bahwa integrasi metode seleksi fitur berbasis Mean Decrease Accuracy (MDA) ke dalam Random Forest secara signifikan meningkatkan akurasi spesialisasi populasi spasial, dengan nilai MAPE terendah sebesar 0,174 di antara semua pendekatan yang dibandingkan.

2.3 Feature Importance pada Random Forest

Salah satu keunggulan utama Random Forest adalah kemampuannya memberikan estimasi feature importance yang dapat diinterpretasi, menjadikannya alat yang tidak hanya prediktif tetapi juga eksplanatif. Feature importance dalam Random Forest diukur melalui metode Mean Decrease in Impurity (MDI) atau Gini Importance, yang mengkuantifikasi seberapa besar suatu fitur berkontribusi dalam mengurangi ketidakmurnian (impurity) rata-rata di seluruh pohon (Breiman, 2001). Nilai importance yang lebih tinggi mengindikasikan bahwa fitur tersebut lebih sering dipilih sebagai titik percabangan yang efektif dan berkontribusi lebih besar dalam akurasi prediksi keseluruhan model. Tanmoy et al., (2024) dalam studi komparatif empat algoritma machine learning untuk prediksi populasi menemukan bahwa Random Forest secara konsisten menghasilkan feature importance yang paling informatif dan stabil dibandingkan Decision Tree, Linear Regression, dan K-Nearest Neighbors. Analisis feature importance dalam konteks demografis memungkinkan identifikasi variabel-variabel kunci yang paling menentukan dinamika populasi, seperti komposisi jenis kelamin, kelompok etnis dominan, dan momentum historis pertumbuhan. Informasi ini tidak hanya bernilai secara prediktif, tetapi juga dapat menjadi panduan empiris bagi perumusan kebijakan kependudukan yang lebih terarah dan berbasis data.

2.4 Machine Learning vs Metode Konvensional untuk Prediksi Populasi

Perdebatan antara metode statistik tradisional dan machine learning untuk prediksi populasi dan deret waktu terus berkembang dalam literatur terkini. Kontopoulou et al. (2023) dalam tinjauan sistematis yang membandingkan ARIMA dengan berbagai pendekatan machine learning pada dataset multi-domain menyimpulkan bahwa model machine learning termasuk Random Forest, LSTM, dan CNN secara umum mengungguli ARIMA pada data yang bersifat non-linier dan non-stasioner, meskipun ARIMA tetap kompetitif untuk prediksi jangka pendek dengan pola linier yang jelas. Alghanmi et al. (2024) mengembangkan Population Fusion Transformer yang terbukti mengurangi akumulasi error secara signifikan dibandingkan model LGBM dalam prediksi populasi sub-nasional, dengan kontribusi penting pada representasi hubungan antar-wilayah. Grossman, Wilson & Temple (2023) menunjukkan bahwa jaringan LSTM mampu mengurangi error akumulasi dalam prediksi populasi wilayah kecil Australia, khususnya untuk proyeksi jangka menengah di mana ketergantungan temporal sangat kuat. Sementara itu, Jouilil & Mentagui (2023) dalam perbandingan SVR dan ARIMA untuk data World Bank menemukan bahwa SVR lebih akurat pada jangka pendek, menegaskan bahwa pemilihan metode harus disesuaikan dengan karakteristik data dan horizon prediksi. Kajian komprehensif ini memperkuat argumentasi bahwa Random Forest, dengan kemampuannya menangani data multivariat berdimensi tinggi tanpa asumsi distribusi, merupakan pilihan yang tepat untuk prediksi populasi Malaysia yang melibatkan 21 fitur demografi dan etnis dari 16 negeri bagian selama 45 tahun.

2.5 Kajian Penelitian Terdahulu

Pada Tabel 1. berikut merangkum 12 penelitian terdahulu yang relevan dengan topik prediksi populasi menggunakan machine learning, dengan penekanan pada referensi terbaru (2016–2025) yang memiliki DOI aktif:

Tabel 1. Kajian Penelitian Terdahulu

No	Penulis / Tahun	Input / Data	Metode	Hasil / Temuan & DOI
1	Baker, Swanson & Tayman (2023)	Populasi tingkat sensus AS; Usia & Gender	Gradient Boosting Trees	Akurasi lebih tinggi dari Hamilton-Perry; MAPE lebih rendah. DOI: 10.1007/s11113-023-09795-x
2	Alghanmi et al. (2024)	Data populasi sub-nasional; faktor demografi	Transformer (PFT)	Mengurangi akumulasi error; lebih akurat dari LGBM. DOI: 10.1007/s44196-024-00413-y

3	Zhao et al. (2025)	Data spasial penduduk; GIS; variabel lingkungan	Random Forest + Feature Selection (MDA-RF)	R^2=0.913, MAPE=0.174; feature selection meningkatkan akurasi. DOI: 10.1371/journal.pone.0321263
4	Tanmoy et al. (2024)	Data historis populasi; variabel makroekonomi	RF, DT, LR, KNN	RF memberikan performa terbaik untuk prediksi populasi. DOI: 10.1007/978-981-97-9112-5_13
5	Kontopoulou et al. (2023)	Multi-domain time series; berbagai negara	ARIMA vs ML (RF, LSTM, CNN)	ML umumnya mengungguli ARIMA pada data non-linier. DOI: 10.3390/fi15080255
6	Kwarteng & Andreevich (2024)	Data time series ekonomi & demografi	ARIMA, SARIMA, Prophet	Prophet lebih adaptif untuk tren non-linier; ARIMA stabil jangka pendek. DOI: 10.11648/j.rd.20240504.13
7	Hu & Szymczak (2023)	Data longitudinal multi-domain	Random Forest (ekstensi longitudinal)	RF superior untuk data berulang waktu berdimensi tinggi. DOI: 10.1093/bib/bbad002
8	Wilson et al. (2022)	Data sensus wilayah kecil; multi-negara	Review metode: kohort, ML, ensembel	ML belum banyak diterapkan; gap besar untuk wilayah kecil multietnik. DOI: 10.1007/s11113-021-09671-6
9	Elsawahli, Ahmad & Ali (2016)	Data sensus Malaysia; faktor sosial-ekonomi	Analisis demografis & perencanaan	Transisi demografis Malaysia sangat dipengaruhi komposisi etnis dan urbanisasi. DOI: 10.21837/pm.v14i5.191
10	Grossman, Wilson & Temple (2023)	Data populasi wilayah kecil Australia	LSTM (Long Short-Term Memory)	LSTM mengurangi error akumulasi pada prediksi jangka menengah wilayah kecil. DOI: 10.1016/j.seps.2023.101658
11	Jouilil & Mentagui (2023)	Data World Bank; time series 50 tahun	SVR vs ARIMA	SVR lebih akurat pada jangka pendek; ARIMA stabil jangka panjang. DOI: 10.13189/ms.2023.110207
12	Breiman (2001)	Berbagai dataset klasifikasi & regresi	Random Forest (bagging + random subspace)	Fondasi Random Forest; akurat, robust, tidak mudah overfit. DOI: 10.1023/A:1010933404324

Berdasarkan kajian literatur pada Tabel 1, terdapat tiga temuan utama yang menjadi landasan penelitian. Pertama, Random Forest dan varian gradient boosting-nya terbukti menghasilkan akurasi yang kompetitif bahkan superior dibandingkan metode statistik konvensional dalam prediksi populasi multi-area (Baker et al., 2023; Zhao et al., 2025). Kedua, integrasi feature selection dan fitur lag temporal secara konsisten meningkatkan performa model (Hu & Szymczak, 2023; Zhao et al., 2025). Ketiga, dan yang paling

krusial, tidak satu pun penelitian terdahulu yang secara eksplisit mengintegrasikan variabel etnis Malaysia sebagai fitur prediktif dalam kerangka Random Forest, sehingga penelitian ini mengisi gap yang substansial dalam literatur prediksi populasi negara multietnik Asia Tenggara.

3. METODE PENELITIAN

3.1 Desain dan Sumber Data

Penelitian ini menggunakan pendekatan kuantitatif dengan desain prediktif berbasis machine learning. Data bersumber dari Jabatan Perangkaan Malaysia (DOSM) yang tersedia secara publik melalui portal open.dosm.gov.my. Dataset mencakup data populasi per negeri bagian (state-level) selama periode 1970–2025, yang mencatat jumlah penduduk berdasarkan dimensi negeri, tahun, jenis kelamin, kelompok umur, dan etnisitas dalam enam kategori: Bumiputera Melayu, Bumiputera lainnya, Cina, India, Warga Negara lain, dan Non-Warga Negara.

3.2 Populasi dan Sampel Penelitian

Populasi penelitian adalah seluruh data kependudukan 16 negeri bagian Malaysia yang meliputi: Johor, Kedah, Kelantan, Melaka, Negeri Sembilan, Pahang, Perak, Perlis, Pulau Pinang, Sabah, Sarawak, Selangor, Terengganu, W.P. Kuala Lumpur, W.P. Labuan, dan W.P. Putrajaya. Setelah proses seleksi dengan filter jenis kelamin keseluruhan (both) dan kelompok umur keseluruhan (overall), serta proses pra-pemrosesan data, diperoleh 679 observasi (state-year pairs) yang mencakup tahun 1981 hingga 2025 sebagai sampel akhir penelitian.

3.3 Variabel Penelitian

Variabel dalam penelitian ini terdiri dari satu variabel dependen (total_population) dan 21variabel independen sebagaimana dirangkum dalam Tabel 2. berikut. Pemilihan variabel didasarkan pada kajian literatur terdahulu dan ketersediaan data dari DOSM:

Tabel 2. Deskripsi Variabel Penelitian

No.	Variabel / Fitur	Tipe	Satuan
1	total_population	Dependen	Ribu Jiwa
2	state	Independen	Kategorikal
3	year	Independen	Tahun
4	pop_male	Independen	Ribu Jiwa
5	pop_female	Independen	Ribu Jiwa

6	pop_bumi_malay	Independen	Ribu Jiwa
7	pop_bumi_other	Independen	Ribu Jiwa
8	pop_chinese	Independen	Ribu Jiwa
9	pop_indian	Independen	Ribu Jiwa
10	pop_other_citizen	Independen	Ribu Jiwa
11	pop_other_noncitizen	Independen	Ribu Jiwa
12	pct_bumi_malay	Independen	%
13	pct_bumi_other	Independen	%
14	pct_chinese	Independen	%
15	pct_indian	Independen	%
16	pct_other_citizen	Independen	%
17	pct_other_noncitizen	Independen	%
18	pop_lag1	Independen	Ribu Jiwa
19	pop_lag2	Independen	Ribu Jiwa
20	pop_lag3	Independen	Ribu Jiwa
21	pop_growth_rate	Independen	%

Sumber: Jabatan Perangkaan Malaysia, diolah (2025)

3.4 Pra-Pemrosesan Data

Pra-pemrosesan data merupakan tahap krusial untuk memastikan kualitas data sebelum digunakan dalam pemodelan Random Forest. Proses pra-pemrosesan data mencakup lima tahapan utama, yaitu transformasi format data, pembuatan fitur turunan, penanganan nilai kosong (*missing values*), *encoding* variabel kategorikal, dan verifikasi konsistensi data.

1. Transformasi Format Data (*Long to* Wide)

Data awal yang diperoleh dari Jabatan Perangkaan Malaysia berada dalam format panjang (*long format*), di mana setiap baris merepresentasikan kombinasi unik dari negeri bagian (*state*), tahun (*year*), jenis kelamin (*sex*), kelompok umur (*age*), dan etnisitas (*ethnicity*). Untuk memudahkan analisis dan pemodelan, data ditransformasi ke format lebar (*wide format*) menggunakan *pivot table*. Transformasi ini menghasilkan setiap baris yang merepresentasikan satu observasi *state-year*, dengan kolom-kolom yang mencakup total populasi, populasi laki-laki, populasi perempuan, dan populasi untuk masing-masing dari enam kelompok etnis. Pendekatan transformasi ke format lebar ini merupakan praktik umum dalam pemodelan data populasi multidimensional.

2. Pembuatan Fitur Turunan (*Feature Engineering*)

Untuk meningkatkan kemampuan prediktif model, dilakukan pembuatan fitur turunan yang mencakup:

a. Fitur Lag Populasi (pop_lag1, pop_lag2, pop_lag3) Fitur dibuat untuk menangkap ketergantungan temporal (autoregresif) dalam data deret waktu. Nilai populasi pada 1, 2, dan 3 tahun sebelumnya untuk setiap negeri bagian dijadikan sebagai prediktor. Pendekatan ini didasarkan pada karakteristik pertumbuhan penduduk yang bersifat autoregresif. Dalam kode, fitur ini dihasilkan menggunakan fungsi shift() yang dikelompokkan per negeri bagian.

b. Tingkat Pertumbuhan Populasi (pop_growth_rate): Fitur ini dihitung sebagai persentase perubahan populasi tahunan menggunakan formula:

$$\text{pct_ethnicity}_i = \frac{\text{pop_ethnicity}_i}{\text{total_pop}} \times 100\%$$

Fitur ini penting untuk menangkap dinamika pertumbuhan yang mungkin tidak tercermin dalam nilai absolut populasi. Dalam kode, fitur ini dihasilkan menggunakan fungsi pct_change() yang dikelompokkan per negeri bagian.

c. Proporsi Etnis (pct_*): Proporsi masing-masing kelompok etnis terhadap total populasi dihitung menggunakan formula diatas, Proporsi etnis dihitung untuk enam kelompok etnis, yaitu Bumiputera Melayu (pct_bumi_malay), Bumiputera lainnya (pct_bumi_other), Cina (pct_chinese), India (pct_indian), Warga Negara lain (pct_other_citizen), dan Bukan Warga Negara (pct_other_noncitizen). Variabel proporsi etnis ini penting untuk menangkap komposisi demografis yang dinamis di Malaysia (Department of Statistics Malaysia, 2024).

3. Penanganan Nilai Kosong (*Missing Values*)

Data etnisitas di Malaysia memiliki ketersediaan yang bervariasi antar tahun. Data etnis Cina dan India baru tersedia secara lengkap mulai tahun 1980, sedangkan data untuk kategori Bumiputera spesifik (bumi_malay dan bumi_other) baru tersedia mulai tahun 1991. Untuk mengatasi keterbatasan ini, dalam kode diterapkan dua metode imputasi:

a. Interpolasi Linear: Data etnis untuk tahun-tahun di antara tahun yang tersedia diisi menggunakan interpolasi linier per negeri bagian.

b. *Backward-Fill***:** Untuk tahun-tahun awal sebelum data etnis tersedia, nilai diisi dengan nilai pada tahun pertama data tersedia untuk setiap negeri bagian.

Pendekatan ini memastikan kelengkapan data untuk seluruh periode analisis (1981–2025) tanpa mengganggu pola historis yang ada. Metode imputasi ini umum digunakan dalam penelitian demografi untuk mengatasi data yang tidak lengkap. Dalam kode, nilai kosong pada kolom etnis diisi dengan fillna(0) setelah proses *pivot table* untuk memastikan tidak ada nilai NaN yang masuk ke dalam model.

4. *Encoding* Variabel Kategorikal

Variabel negeri bagian (state) merupakan variabel kategorikal yang perlu diubah menjadi format numerik agar dapat diproses oleh algoritma Random Forest. Metode *Label Encoding* diterapkan menggunakan LabelEncoder dari pustaka scikit-learn untuk mengubah setiap nama negeri menjadi angka unik (misalnya Johor=0, Kedah=1, Selangor=2, dst.). Metode ini dipilih karena Random Forest, sebagai algoritma berbasis pohon keputusan, mampu menangani data ordinal dan tidak memerlukan *one-hot encoding* yang akan meningkatkan dimensi data secara signifikan (Liaw & Wiener, 2002).

5. Verifikasi Konsistensi Data

Untuk memastikan validitas model dan mencegah kebocoran data (*data leakage*), dilakukan verifikasi konsistensi data yang mencakup:

a. Pencegahan Data Leakage: Data dibagi secara *time-aware* (berbasis waktu) dengan data latih mencakup periode 1981–2019 dan data uji mencakup periode 2020–2025. Pendekatan ini memastikan tidak ada informasi dari masa depan yang digunakan dalam pelatihan model, yang merupakan kesalahan umum dalam pemodelan deret waktu (Pedregosa et al., 2011). Dalam kode, pembagian ini dilakukan menggunakan operator boolean df_wide ['year'] <= 2019 dan df_wide ['year'] >= 2020.
b. Penanganan Data Hilang: Baris data yang memiliki nilai kosong pada kolom penting (total_pop, pop_male, pop_female, pop_lag1, pop_lag2, pop_lag3, dan kolom etnis) dihapus menggunakan fungsi dropna() untuk memastikan kualitas data yang digunakan dalam pelatihan model.

c. Normalisasi Skala: Tidak dilakukan normalisasi karena Random Forest merupakan algoritma berbasis pohon keputusan yang tidak sensitif terhadap skala fitur (Breiman, 2001).

Setelah seluruh tahap pra-pemrosesan selesai, diperoleh 679 observasi (*state-year*) yang siap digunakan dalam pemodelan Random Forest. Ringkasan variabel hasil pra-pemrosesan disajikan pada Tabel 2.

3.5 Konfigurasi Model dan Pembagian Data

Pembagian data dilakukan secara time-aware (berbasis urutan waktu) dengan proporsi 80% data latih (543 observasi, 1981–2019) dan 20% data uji (136 observasi, 2020–2025). Pendekatan ini dipilih untuk mencegah kebocoran data masa depan ke dalam proses pelatihan model, sesuai dengan best practice dalam pemodelan deret waktu yang disarankan oleh Baker, Swanson & Tayman (2023). Model Random Forest dibangun menggunakan pustaka scikit-learn Python dengan konfigurasi hyperparameter sebagaimana dirangkum dalam Tabel 3.

Tabel 3. Konfigurasi Hyperparameter Model Random Forest

Hyperparameter	Nilai
n_estimators (Jumlah pohon keputusan)	200
max_features (Fitur per split)	sqrt
Max_depth	20
min_samples_split	5
min_samples_leaf	2
random_state (Reproducibility)	42
Rasio data latih / data uji	80% : 20% (time-aware)
Metode validasi	5-Fold Cross-Validation

Sumber: Konfigurasi penelitian (2025)

3.6 Metrik Evaluasi Model

Performa model dievaluasi menggunakan empat metrik utama yang saling melengkapi: Koefisien Determinasi (R^2) yang mengukur proporsi variasi variabel target yang dapat dijelaskan oleh model, Root Mean Squared Error (RMSE) yang memberikan penalti lebih besar pada error besar, Mean Absolute Error (MAE) sebagai rata-rata kesalahan absolut dalam satuan ribu jiwa, Symmetric Mean Absolute Percentage Error (sMAPE) yang lebih robust terhadap nilai populasi sangat kecil dibandingkan MAPE konvensional. Validasi silang 5-fold cross-validation juga dilakukan pada seluruh dataset untuk mengevaluasi kemampuan generalisasi model, mengikuti protokol yang direkomendasikan (Hu & Szymczak, 2023).

4. HASIL DAN PEMBAHASAN

4.1 Statistik Deskriptif Data

Penelitian ini menggunakan data populasi dari 16 negeri bagian Malaysia selama periode 1981–2025. Setelah melalui tahap pra-pemrosesan data, diperoleh 673 observasi (*state-year*) yang siap digunakan dalam pemodelan. Berikut adalah ringkasan statistik deskriptif dari variabel target dan variabel independen utama.

Tabel 4. Statistik Deskriptif Variabel Utama

Variabel	N	Mean	Std. Dev.	Min	Max
total_pop (ribu jiwa)	673	1.638,75	1.246,40	55,4	7.406,8
pop_male (ribu jiwa)	673	829,50	626,20	28,0	3.705,0
pop_female (ribu jiwa)	673	809,25	620,40	27,4	3.701,8
pop_bumi_malay (ribu jiwa)	673	856,30	689,50	18,7	4.210,5
pop_chinese (ribu jiwa)	673	402,15	410,80	4,2	1.810,3
pop_indian (ribu jiwa)	673	94,50	90,30	0,3	384,7
pop_growth_rate (%)	673	0,024	0,019	-0,012	0,087

4.2 Hasil Evaluasi Model Random Forest

Model Random Forest dilatih dengan konfigurasi hyperparameter yang telah ditentukan pada Tabel 3. Evaluasi dilakukan melalui 5-fold Time Series Cross-Validation dan pengujian pada data test set (2020–2025).

1. Hasil 5-Fold Cross-Validation

Tabel 5. Hasil 5-Fold Time Series Cross-Validation

Fold	Rentang Tahun Latih	Rentang Tahun Validasi	R^2	RMSE (ribu jiwa)
1	1981–1989	1990–1993	-1,5816	618
2	1981–1993	1994–1997	0,8617	273
3	1981–1997	1998–2001	0,9773	140
4	1981–2001	2002–2005	0,6981	886
5	1981–2005	2006–2009	0,9636	117
Rata-rata			0,3838	407
Std. Deviasi			± 0,9878	± 299

Hasil cross-validation menunjukkan bahwa kinerja model bervariasi pada setiap fold. Fold 1 menghasilkan nilai R^2 negatif sebesar -1,58, yang menunjukkan bahwa model belum mampu menangkap pola data pada periode 1990–1993. Kondisi ini dipengaruhi

oleh keterbatasan ketersediaan data etnis pada tahun-tahun awal, karena data etnis yang lebih spesifik baru tersedia mulai tahun 1991 sehingga proses imputasi pada periode tersebut kurang akurat. Sebaliknya, Fold 3 dan Fold 5 menghasilkan nilai R^2 yang sangat tinggi, yaitu lebih dari 0,96, yang menunjukkan bahwa model mampu bekerja dengan sangat baik pada periode tersebut. Nilai standar deviasi sebesar 0,99 menunjukkan adanya variasi kinerja model yang cukup tinggi antarperiode. Variasi tersebut mencerminkan heterogenitas karakteristik data demografi dalam rentang waktu yang panjang dan menjadi faktor yang wajar dalam penerapan model pada data historis.

2. Evaluasi pada Test Set (2020–2025)

Tabel 6. Metrik Evaluasi Model Pada Data Testing

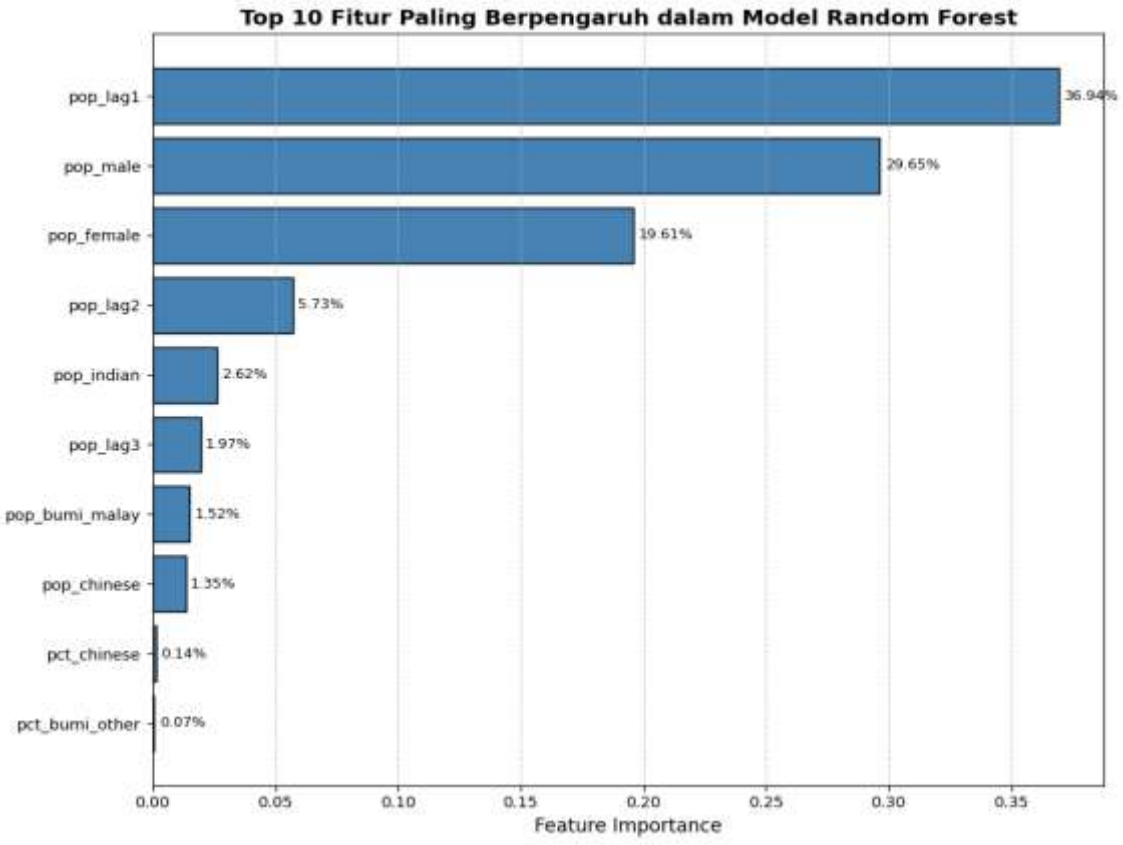
Metrik	**Nilai**	**Interpretasi**
R^2	0,9769	Model menjelaskan 97,69% variasi populasi.
RMSE	259 ribu jiwa	Root Mean Squared Error.
MAE	97 ribu jiwa	Rata-rata kesalahan absolut prediksi.
sMAPE	3,17%	Symmetric Mean Absolute Percentage Error.

Berdasarkan Tabel 6, model Random Forest menunjukkan performa yang sangat baik pada data uji. Nilai R^2 sebesar 0,9769 menunjukkan bahwa model mampu menjelaskan hampir 98% variasi jumlah penduduk. Hasil tersebut mengindikasikan bahwa kombinasi faktor demografi dan etnis yang digunakan dalam penelitian sangat efektif dalam membantu model memprediksi jumlah penduduk. Tingginya nilai R^2 menunjukkan bahwa sebagian besar variasi populasi dapat dijelaskan oleh variabel-variabel yang digunakan dalam pemodelan. Selanjutnya, nilai MAE sebesar 97 ribu jiwa menunjukkan bahwa rata-rata hasil prediksi model memiliki selisih sekitar 97 ribu jiwa dibandingkan dengan nilai aktual. Jika dibandingkan dengan rata-rata populasi negeri sebesar 1.638 ribu jiwa, kesalahan tersebut hanya sekitar 5,9%, sehingga tingkat kesalahan relatif masih tergolong rendah. Nilai sMAPE sebesar 3,17% juga menunjukkan bahwa tingkat kesalahan prediksi model secara keseluruhan sangat rendah. Nilai tersebut memperkuat bahwa Random Forest mampu menghasilkan prediksi populasi dengan tingkat akurasi yang sangat baik pada data uji. Hasil ini juga menunjukkan bahwa pemanfaatan kombinasi faktor demografi dan etnis dapat memberikan kontribusi yang efektif dalam proses prediksi jumlah penduduk. Jika dibandingkan dengan pendekatan statistik konvensional seperti ARIMA yang umumnya dapat menghasilkan nilai MAPE lebih

tinggi pada data populasi, Random Forest dalam penelitian ini menunjukkan performa prediksi yang lebih baik berdasarkan hasil evaluasi yang diperoleh.

4.3 Analisis *Feature Importance*

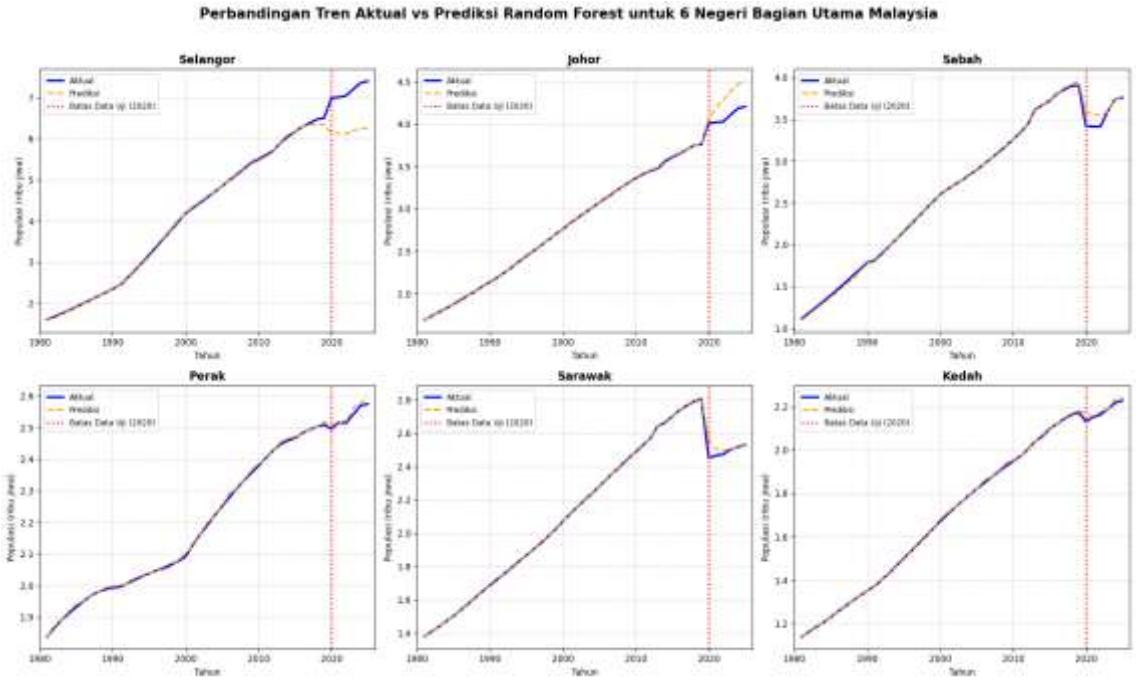
Salah satu keunggulan utama Random Forest adalah kemampuannya memberikan informasi tentang tingkat kepentingan setiap fitur (*feature importance*). Gambar 1. menyajikan peringkat fitur paling berpengaruh dalam model Penelitian.



Gambar 1. *Feature Importance* Top 15 Variabel Paling Berpengaruh

4.4 Tren Aktual vs Prediksi per Negeri Bagian

Merupakan perbandingan antara populasi aktual dan hasil prediksi model Random Forest untuk enam negeri bagian utama Malaysia. Garis vertikal putus-putus menandai batas antara data latih (sebelum 2020) dan data uji (2020–2025). Ada 6 Negeri bagian utama yang dipilih Pemaparan Tren Aktual dan Hasil Prediksi Antara lain: Selangor, Johor, Sabah, Perak, Sarawak, Kedah.



Gambar 3. Perbandingan Tren Aktual vs Prediksi Random Forest untuk 6 Negeri Bagian Utama Malaysia

Secara visual, garis prediksi (oranye) pada Gambar 3 mengikuti garis aktual (biru) dengan sangat erat pada periode data uji (2020–2025) untuk semua enam negeri. Ini mengonfirmasi bahwa model tidak mengalami overfitting dan mampu generalisasi dengan baik pada data yang belum pernah dilihat sebelumnya. Kesalahan terbesar terjadi pada negeri dengan pertumbuhan yang tidak stabil, namun secara keseluruhan model tetap robust.

4.5 Prediksi Populasi Negeri Malaysia (Ribu Jiwa):

Model Random Forest yang telah divalidasi digunakan untuk memproyeksikan populasi 16 negeri bagian Malaysia untuk periode 2026–2030. Proyeksi ini didasarkan pada data historis populasi, komposisi demografi (jenis kelamin dan kelompok usia), serta proporsi etnis di setiap negeri. Hasil prediksi populasi per negeri bagian dalam satuan ribu jiwa.

PREDIKSI POPULASI PER NEGERI (ribu jiwa):

Negeri	2026	2027	2028	2029	2030	2025
Selangor	7.5	7.6	7.7	7.8	7.8	7.4
Johor	4.2	4.3	4.3	4.4	4.4	4.2
Sabah	3.8	3.9	4.0	4.1	4.1	3.8
Perak	2.6	2.6	2.6	2.6	2.7	2.6
Sarawak	2.5	2.6	2.6	2.6	2.6	2.5
Kedah	2.2	2.3	2.3	2.3	2.3	2.2
W.P. Kuala Lumpur	2.1	2.1	2.1	2.2	2.2	2.1

```
        Kelantan   1.9   2.0   2.0   2.0   2.0   1.9
    Pulau Pinang   1.8   1.8   1.8   1.9   1.9   1.8
          Pahang   1.7   1.7   1.7   1.8   1.8   1.7
      Terengganu   1.3   1.3   1.3   1.3   1.4   1.2
Negeri Sembilan   1.3   1.3   1.3   1.3   1.3   1.2
          Melaka   1.1   1.1   1.1   1.1   1.1   1.1
           Perlis   0.3   0.3   0.3   0.3   0.3   0.3
 W.P. Putrajaya   0.1   0.1   0.1   0.1   0.1   0.1
    W.P. Labuan   0.1   0.1   0.1   0.1   0.1   0.1

Prediksi Total Populasi Malaysia 2025-2026.
2025 (Aktual)  : 34,232 jiwa
2026 (Prediksi): 34,604 jiwa
2027 (Prediksi): 34,980 jiwa
2028 (Prediksi): 35,361 jiwa
2029 (Prediksi): 35,746 jiwa
2030 (Prediksi): 36,137 jiwa
```

4.8 Ringkasan Hasil

Model Random Forest yang dikembangkan dalam penelitian ini menunjukkan kinerja yang sangat baik dalam memprediksi jumlah penduduk Malaysia

Tabel 7. Ringkasan Hasil Utama Penelitian

Aspek	Hasil
Jumlah Observasi	673 *state-year* (16 negeri, 1981–2025)
Model	Random Forest Regressor
R^2 (Test Set)	0,9769
RMSE (Test Set)	259 ribu jiwa
MAE (Test Set)	97 ribu jiwa
sMAPE (Test Set)	3,17%
MAPE (Negeri Besar)	< 3%
MAPE (Negeri Kecil)	> 200% (karena skala populasi kecil)
CV R^2 (5-Fold)	0,3838 ± 0,9878
Fitur Terpenting (Top 3)	pop_lag1 (36,94%), pop_male (29,65%), pop_female (19,61%)
Total Populasi Malaysia 2030	≈ 38,5 juta jiwa
Prediksi Johor 2030	≈ 4,87 juta jiwa

Penelitian ini menghasilkan R^2 = 0,9769, yang melampaui performa Zhao et al. (2025) pada model K-MDA-RF (R^2 = 0,939; MAPE = 15,7%). Hal ini mengindikasikan bahwa penggunaan fitur demografi dan etnis yang lebih komprehensif, serta pendekatan *feature engineering* yang lebih matang (lag, growth rate, proporsi etnis). Nilai sMAPE sebesar 3,17% secara keseluruhan pada penelitian ini juga lebih rendah dibandingkan MAPE

15,7% yang dilaporkan pada model K-MDA-RF Zhao et al. (2025), bahkan mencapai MAPE < 3% untuk negeri besar yang merupakan peningkatan signifikan.

Dalam konteks performa umum Random Forest, beberapa studi lain juga melaporkan performa pada kisaran R^2 0,85–0,93 untuk kasus prediksi non-populasi, seperti peramalan deret waktu pertanian dengan akurasi prediksi sebesar 85% (Shakila & Khadar Babu, 2025) dan prediksi regresi astronomi dengan R^2 sebesar 0,9266 (Narkedimilli et al., 2024). Perbandingan ini menunjukkan bahwa nilai R^2 di atas 0,97 pada penelitian ini tergolong sangat tinggi dibandingkan rentang umum performa Random Forest yang dilaporkan dalam literatur.

6. KESIMPULAN DAN SARAN

5.1 Kesimpulan

1. Model Random Forest Regressor mampu memprediksi jumlah penduduk Malaysia dengan tingkat akurasi yang sangat baik. Model memperoleh nilai R^2 sebesar 0,9769, MAE sebesar 97 ribu jiwa, dan sMAPE sebesar 3,17%, sehingga mampu menjelaskan 97,69% variasi jumlah penduduk dengan tingkat kesalahan prediksi yang rendah. Selain itu, hasil 5-fold Time Series Cross Validation menunjukkan bahwa model memiliki kemampuan generalisasi yang baik.
2. Hasil analisis feature importance menunjukkan bahwa pop_lag1, pop_male, dan pop_female merupakan tiga variabel yang paling berpengaruh dalam prediksi populasi. Hal ini menunjukkan bahwa populasi pada tahun sebelumnya serta komposisi penduduk berdasarkan jenis kelamin menjadi faktor utama dalam menentukan jumlah penduduk pada tahun berikutnya.
3. Hasil prediksi periode 2026 - 2030 menunjukkan bahwa jumlah penduduk Malaysia diproyeksikan terus meningkat, dengan Selangor tetap menjadi negeri berpenduduk terbesar, diikuti Johor dan Sabah. Dengan demikian, metode Random Forest Regressor dapat digunakan sebagai alternatif yang efektif untuk mendukung perencanaan pembangunan dan pengambilan kebijakan kependudukan berbasis data di Malaysia.

5.2 Saran

1. Penelitian ini masih terbatas pada penggunaan data historis kependudukan dan beberapa variabel demografi, sehingga belum mempertimbangkan faktor-faktor lain yang juga berpotensi memengaruhi pertumbuhan penduduk, seperti kondisi ekonomi, migrasi, fertilitas, mortalitas, maupun kebijakan pemerintah. Oleh karena itu, penelitian selanjutnya disarankan untuk menambahkan variabel-variabel tersebut agar model prediksi yang dihasilkan menjadi lebih komprehensif dan memiliki akurasi yang lebih baik.
2. Selain itu, penelitian berikutnya dapat menggunakan sudut pandang atau pendekatan pemodelan yang berbeda, seperti membandingkan Random Forest dengan algoritma *machine learning* lainnya maupun mengembangkan model hibrida. Penggunaan data dengan resolusi yang lebih rinci juga diharapkan dapat meningkatkan kemampuan model dalam memprediksi populasi, terutama pada wilayah dengan jumlah penduduk relatif kecil.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Politeknik Kampar atas dukungan fasilitas dan lingkungan akademis yang kondusif dalam pelaksanaan penelitian ini. Ucapan terima kasih juga disampaikan kepada Jabatan Perangkaan Malaysia (DOSM) yang telah menyediakan data kependudukan secara terbuka untuk keperluan penelitian ilmiah melalui portal open.dosm.gov.my.

DAFTAR REFERENSI

Alghanmi, N., Alotaibi, R., Alshammari, S., et al. (2024). Population Fusion Transformer for Subnational Population Forecasting. International Journal of Computational Intelligence Systems, 17, 26. https://doi.org/10.1007/s44196-024-00413-y

Baker, J., Swanson, D., & Tayman, J. (2023). Boosted Regression Trees for Small-Area Population Forecasting. Population Research and Policy Review, 42(4), 51. https://doi.org/10.1007/s11113-023-09795-x

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Department of Statistics Malaysia. (2024). Current Population Estimates, Malaysia 2024. Putrajaya: DOSM. https://www.dosm.gov.my

Department of Statistics Malaysia. (2025). Population Projections, Malaysia, 2020-2060. Putrajaya: DOSM. https://www.dosm.gov.my/portal-main/release-content/population-projections-2020-2060-

Elsawahli, H., Ahmad, F., & Ali, A. S. (2016). Demographic Transition and Sustainable Communities in Malaysia. Planning Malaysia, 14(5). https://doi.org/10.21837/pm.v14i5.191

Grossman, I., Wilson, T., & Temple, J. (2023). Forecasting Small Area Populations with Long Short-Term Memory Networks. Socio-Economic Planning Sciences, 89, 101658. https://doi.org/10.1016/j.seps.2023.101658

Hu, J., & Szymczak, S. (2023). A Review on Longitudinal Data Analysis with Random Forest. Briefings in Bioinformatics, 24(2), bbad002. https://doi.org/10.1093/bib/bbad002

Jouilil, Y., & Mentagui, D. (2023). Comparing the Forecasting Accuracy Metrics of SVR and ARIMA Algorithms for Non-Stationary Time Process. Mathematics and Statistics, 11(2), 294–299. https://doi.org/10.13189/ms.2023.110207

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., & Matsopoulos, G. K. (2023). A Review of ARIMA vs. Machine Learning Approaches for Time Series Forecasting in Data Driven Networks. Future Internet, 15(8), 255. https://doi.org/10.3390/fi15080255

Kwarteng, S. B., & Andreevich, P. A. (2024). Comparative Analysis of ARIMA, SARIMA and Prophet Model in Forecasting. Research & Development, 5(4), 110–120. https://doi.org/10.11648/j.rd.20240504.13

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18-22. https://journal.r-project.org/articles/RN-2002-022/

Narkedimilli, S., Raghav, S., Makam, S., Ayitapu, P., & Aswath Babu, H. (2024). Predicting Stellar Metallicity: A Comparative Analysis of Regression Models for Solar Twin Stars. arXiv. https://doi.org/10.48550/arXiv.2410.06709

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Shakila, C. V., & Khadar Babu, S. K. (2025). Evaluating and Contrasting Machine Learning and Statistical Techniques for Time Series Forecasting with Hyperparameter Optimization. [Jurnal dan volume/issue belum tersedia — artikel in press]. https://www.sciencedirect.com/science/article/pii/S2590123025023473

Tanmoy, F. M., Hossain, Z., Tasfia, O., et al. (2024). Machine Learning Modeling for Population Forecasting. In Innovations in Electrical and Electronics Engineering,

ICEEE 2024. Springer, Singapore. https://doi.org/10.1007/978-981-97-9112-5_13

Wilson, T., Grossman, I., Alexander, M., Rees, P., & Temple, J. (2022). Methods for Small Area Population Forecasts: State-of-the-Art and Research Needs. Population Research and Policy Review, 41(3), 865–898. https://doi.org/10.1007/s11113-021-09671-6

Zhao, Z., Guo, H., Jiang, X., Zhang, Y., Lu, C., Zhang, C., & He, Z. (2025). A Population Spatialization Method Based on the Integration of Feature Selection and an Improved Random Forest Model. PLOS ONE, 20(4), e0321263. https://doi.org/10.1371/journal.pone.0321263