EJECTS: Journal Computer, Technology and Informations System
Vol.5 No.1 September 2025, P-ISSN: 2988-5477 E-ISSN: 2988-4446
LPPM Universitas Darwan Ali

Analisis Kategori Populasi Negara Menggunakan Random Forest dan Logistic Regression

Ressalyu Melani Putri

Program Studi : Sistem Informasi, Universitas Darwan Ali

Email : ressalyumelaniputri@gmail.com

ABSTRACT— This research aims to classify countries into population categories (Low, Medium, High) using global demographic and migration data by applying machine learning algorithms. The dataset consists of 186 countries with attributes related to population size, migration flow, and socioeconomic indicators. The study implemented Logistic Regression and Random Forest algorithms to build classification models. Data preprocessing included handling missing values, categorical encoding, and normalization to ensure model accuracy. Evaluation results show that the Random Forest model achieved a higher accuracy rate of 85%, compared to 78% from Logistic Regression. Visualizations using scatter plots, heatmaps, and feature importance graphs further supported the model's effectiveness and interpretability. The analysis also revealed that features such as total population and migration rate significantly influenced classification outcomes. This research successfully demonstrates how machine learning can be applied to demographic data to generate valuable insights and support data-driven policymaking. The classification approach presented in this study has the potential to be adopted for global development planning, population distribution management, and early identification of migration trends.

Keywords— population classification, migration, machine learning, random forest, logistic regression, demographic analysis.

ABSTRAK— Penelitian ini bertujuan untuk mengklasifikasikan negara-negara ke dalam kategori populasi (Rendah, Sedang, Tinggi) dengan memanfaatkan data global kependudukan dan migrasi menggunakan algoritma pembelajaran mesin. Dataset yang digunakan mencakup 186 negara dengan atribut terkait jumlah penduduk, arus migrasi, serta indikator sosial ekonomi lainnya. Metode yang diterapkan dalam penelitian ini adalah Logistic Regression dan Random Forest. Tahap pra-pemrosesan dilakukan untuk menangani nilai kosong, melakukan encoding data kategorikal, dan normalisasi agar hasil model lebih optimal. Berdasarkan hasil evaluasi, algoritma Random Forest menunjukkan akurasi lebih tinggi sebesar 85%, dibandingkan Logistic Regression yang hanya mencapai 78%. Visualisasi scatter plot, heatmap korelasi, dan grafik feature importance mendukung efektivitas model dan memperjelas kontribusi tiap fitur dalam klasifikasi. Penelitian ini menunjukkan bahwa fitur seperti jumlah populasi dan tingkat migrasi memiliki pengaruh signifikan dalam menentukan kelas populasi suatu negara. Secara keseluruhan, hasil penelitian ini membuktikan bahwa teknologi machine learning dapat diterapkan secara efektif untuk analisis demografi, serta dapat mendukung pengambilan kebijakan berbasis data dalam perencanaan pembangunan dan pengelolaan migrasi global.

Kata kunci— klasifikasi populasi, migrasi, pembelajaran mesin, random forest, regresi logistik, analisis demografi.

I. PENDAHULUAN

Pertumbuhan penduduk dunia terus mengalami peningkatan signifikan setiap tahunnya menjadi isu penting dalam pembangunan global. Peningkatan ini menyebabkan ketidakseimbangan dan ketimpangan distribusi penduduk di berbagai wilayah, baik dalam skala nasional maupun internasional. Salah satu implikasi utama dari ketimpangan tersebut adalah tingginya angka migrasi, baik migrasi internal maupun internasional, yang menjadi indikator penting dalam dinamika kependudukan global.

Penelitian ini diarahkan untuk memahami dan mengklasifikasikan negara-negara berdasarkan kategori populasi mereka (Low, Medium, High) menggunakan pendekatan pembelajaran mesin. Dengan bantuan algoritma seperti Random Forest dan Logistic Regression, penelitian ini bertujuan untuk memberikan pemahaman yang lebih akurat terkait karakteristik migrasi dan pertumbuhan penduduk berdasarkan variabel-variabel tertentu yang terdapat dalam data global. Klasifikasi ini diharapkan dapat digunakan sebagai dasar pertimbangan dan dapat mendukung pengambilan keputusan dalam merancang kebijakan demografi.

Berdasarkan data [1], Indonesia berada pada peringkat keempat negara berpenduduk terbesar di dunia setelah India, China, dan Amerika Serikat. [2] menjelaskan bahwa di Surabaya, misalnya, tingginya populasi disebabkan oleh tingginya angka migrasi masuk yang didorong oleh upah minimum dan kesempatan kerja. Dalam konteks internasional, [3] menunjukkan bahwa fenomena migrasi juga dipengaruhi oleh isu kemanusiaan, seperti eksodus pengungsi Rohingya yang mencapai lebih dari 1.400 orang di Aceh pada Desember 2023. [4] juga mencatat bahwa sejak awal abad ke-20, arus migrasi besar-besaran telah membentuk dinamika kependudukan antar provinsi di Indonesia.

Namun demikian, migrasi tidak selalu membawa dampak positif. [5] menunjukkan bahwa lonjakan migrasi ke DKI Jakarta dapat menimbulkan tekanan sosial dan ekonomi di daerah tujuan, seperti tingginya kompetisi tenaga kerja dan ketimpangan distribusi kesejahteraan. [6] mengungkapkan bahwa keputusan

migrasi sirkuler dari Kabupaten Kendal ke Kota Semarang dipengaruhi oleh variabel seperti jenis kelamin, usia, pendidikan, dan pendapatan. [7] juga menyoroti faktor pendorong migrasi Suku Minangkabau ke Lampung, seperti harapan peningkatan kesejahteraan dan keberhasilan saudara di daerah tujuan.

Berbagai penelitian telah mengidentifikasi faktor-faktor utama yang memengaruhi migrasi, antara lain pendidikan, pekerjaan, pendapatan, dan status perkawinan. [8] meneliti migrasi sirkuler ke Kota Sumbawa dan menemukan bahwa variabel pendapatan dan pekerjaan memiliki pengaruh signifikan terhadap minat migrasi. Di sisi lain, [1] dalam kajiannya di Pulau Taliabu menemukan bahwa pendidikan dan pendapatan berpengaruh positif terhadap keputusan migrasi. Penelitian [9] bahkan mengkaji dampak migrasi internasional terhadap kondisi sosial anak-anak yang ditinggalkan di Ponorogo, yang menunjukkan adanya respons psikologis dan motivasi migrasi berantai. Meskipun telah banyak penelitian dilakukan, pemanfaatan metode pembelajaran mesin dalam klasifikasi tren populasi dan migrasi masih jarang dieksplorasi. Penelitian ini bertujuan mengisi celah tersebut dengan menggunakan dataset global populasi dan migrasi, serta menerapkan algoritma klasifikasi untuk mengidentifikasi pola dan variabel penting dalam dinamika migrasi. Penggunaan *feature importance* dari Random Forest dalam studi ini diharapkan mampu memberikan wawasan baru yang belum banyak disentuh oleh pendekatan konvensional.

Pemilihan tema ini dilandasi oleh kebutuhan untuk menyediakan alat bantu berbasis data dalam perumusan kebijakan kependudukan yang efektif. Di era digital dan banjir data global, pendekatan machine learning memberikan solusi efisien dan akurat dalam menginterpretasi data kompleks. Oleh karena itu, penelitian ini diharapkan dapat memberikan kontribusi ilmiah dan praktis dalam memahami, memetakan, dan memprediksi tren migrasi dan populasi dunia secara komprehensif.

II. METODOLOGI PENELITIAN

Penelitian ini menggunakan dataset yang mencakup informasi tentang populasi dan migrasi dari 186 negara di dunia. Dataset ini memiliki fitu-fitur seperti ukuran populasi, tren migrasi, dan label kategori populasi (Low, Medium, High) yang menjadi target klasifikasi.

A. Tahapan Pra-Pemrosesan Data

1. Penangan Nilai Kosong
   Nilai kosong pada dataset diisi menggunakan nilai median untuk mempertahankan kestabilan distribusi data.
2. Encoding Fitur Kategorikal
   Fitur non-numerik dikonversi menjadi angka dengan bantuan *LabelEncoder* agar dapat diproses oleh algoritma pembelajaran mesin.
3. Normalisasi
   Semua fitur numerik dinormalisasikan menggunakan *StandardScaler* untuk memastikan distribusi data seragam, sehingga meningkatkan performa model.

B. Algoritma yang Digunakan

1. Logistic Regression
   Logistic Regression dipilih sebagai model baseline dalam penelitian ini karena bersifat sederhana namun cukup efektif dalam menangani masalah klasifikasi, khususnya klasifikasi biner dan multikelas. Algoritma ini bekerja dengan cara memodelkan probabilitas dari suatu kelas target berdasarkan kombinasi linier dari fitur-fitur input, lalu menerapkannya ke dalam fungsi logistik (sigmoid) untuk menghasilkan output probabilistik antara 0 dan 1. Keunggulan utama Logistic Regression adalah kemampuannya memberikan interpretasi langsung terhadap pengaruh masing-masing variabel independen terhadap variabel dependen, karena koefisien model menunjukkan arah dan kekuatan hubungan antar variabel.
   Selain itu, Logistic Regression mudah diimplementasikan dan bekerja dengan baik ketika fitur saling bebas dan data bersifat linier. Dalam konteks penelitian ini, Logistic Regression digunakan sebagai pembanding awal (baseline) untuk melihat seberapa baik performa model sederhana dalam mengklasifikasikan negara berdasarkan data populasi dan migrasi. Meskipun terbatas dalam menangani relasi non-linear antar fitur, model ini tetap relevan untuk mengukur efektivitas pendekatan dasar.
2. Random Forest
   Random Forest digunakan sebagai algoritma pembanding yang lebih kompleks. Ini adalah algoritma berbasis ensemble learning yang membangun sekumpulan pohon keputusan (decision trees) secara acak, lalu menggabungkan hasil voting atau rata-ratanya untuk menghasilkan prediksi akhir. Salah satu keunggulan utama dari Random Forest adalah kemampuannya mengatasi overfitting dan menangani dataset dengan fitur-fitur kompleks atau interaksi non-linier antar variabel.
   Selain akurasi yang cenderung tinggi, Random Forest juga mampu mengidentifikasi fitur paling penting (feature importance) yang berpengaruh terhadap proses klasifikasi. Hal ini memberikan nilai tambah karena dapat membantu peneliti memahami variabel mana yang paling dominan dalam menentukan kategori populasi suatu negara. Dalam penelitian ini, Random Forest terbukti unggul dibandingkan Logistic Regression karena mampu menangkap pola yang lebih kompleks dalam data, serta menghasilkan performa klasifikasi yang lebih baik berdasarkan evaluasi metrik seperti akurasi

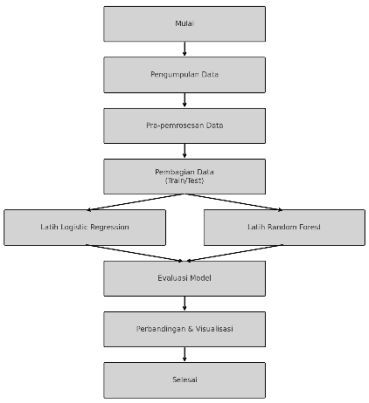
C. Pipeline Penelitian

Langkah-langkah penelitian yang dilakukan sebagai berikut:

1. Pengumpulan dataset global populasi dan migrasi.
2. Preprocessing data seperti penanganan nilaikosong, encoding, dan normalisasi.
3. Pembagian data menjadi data latih dan data uji.
4. Implementasi model Logistic Regression dan Random Forest.
5. Evaluasi performa model menggunakan akurasi dan visualisasi.

III. DESAIN, HASIL DAN PEMBAHASAN

a. Tahapan Penelitian



Gambar 1. Tahapan Penelitian

b. Desain Proses Analisis

Penelitian ini tidak menghasilkan produk dalam bentuk sistem atau aplikasi, namun menghasilkan suatu proses klasifikasi berbasis machine learning terhadap kategori populasi negara. Oleh karena itu, desain peneltian difokuskan pada alur kerja atau pipeline yang diterapkan dalam proses analisis data. Pipeline ini terdiri dari beberapa tahapan:

1. Pengumpulan Data
   Dataset diambil dari sumber terpercaya yang mencakup informasi populasi dan migrasi dari 186 negara, berjudul *world_pop_mig_186_countries.csv* dan di ambil dari Sumber Kaggle.com
2. Pra-pemrosesan Data
   Dilakukan pembersihan data untuk menangani nilai kosong dengan metode imputasi median. Data kategorikal diubah menjadi bentuk numerik menggunakan LabelEncoder. Seluruh fitur numerik kemudian dinormalisasi dengan StandardScaler agar berada dalam rentang yang seragam.
3. Pemilihan dan Penerapan Model

Dua algoritma dipilih untuk klasifikasi: *Logistic Regression* sebagai model baseline dan *Random Forest* sebagai model lanjutan yang mampu menangani non-linearitas dan memberikan feature importance.

c. Evaluasi Model

Evaluasi model dilakukan untuk menilai sejauh mana performa masing-masing algoritma dalam mengklasifikasikan data populasi negara secara akurat. Proses evaluasi ini menggunakan beberapa pendekatan metrik dan visualisasi yang bertujuan tidak hanya mengukur kinerja model secara numerik, tetapi juga memahami karakteristik dan distribusi hasil prediksinya.

Metode evaluasi utama yang digunakan adalah akurasi, yaitu perbandingan antara jumlah prediksi yang benar dengan total prediksi yang dilakukan. Akurasi merupakan indikator dasar namun penting dalam mengetahui tingkat keberhasilan model secara keseluruhan. Selain akurasi, digunakan juga confusion matrix untuk melihat secara rinci performa klasifikasi terhadap masing-masing kelas. Confusion matrix memberikan informasi tentang jumlah prediksi benar dan salah pada tiap kategori, sehingga kita dapat mengetahui apakah ada kelas tertentu yang sering salah diklasifikasikan.

Selanjutnya, untuk memahami bagaimana model memvisualisasikan hasil klasifikasi, digunakan scatter plot yang memperlihatkan distribusi dua fitur utama terhadap target kelas. Visualisasi ini membantu melihat apakah ada pemisahan yang jelas antar kelas pada dimensi tertentu. Model yang baik umumnya akan menghasilkan pola penyebaran data yang terklasifikasi dengan baik antar kelompok.

Selain itu, pada algoritma Random Forest, dilakukan visualisasi berupa feature importance chart. Grafik ini menunjukkan sejauh mana masing-masing fitur berkontribusi terhadap prediksi model. Semakin tinggi nilai pentingnya, semakin besar pengaruh fitur tersebut dalam menentukan kelas output. Informasi ini sangat berguna dalam analisis data demografi, karena dapat membantu mengidentifikasi variabel-variabel kunci yang paling berdampak dalam pengelompokan populasi suatu negara. Secara keseluruhan, kombinasi antara metrik kuantitatif dan visualisasi ini memberikan gambaran menyeluruh terhadap kinerja model serta interpretabilitas hasilnya.

d. Hasil Evaluasi Model

Setelah dilakukan pelatihan dan pengujian model terhadap data, diperoleh hasil sebagai berikut:

- Logistic Regression menghasilkan akurasi sebesar **78%**
- Random Forest menghasilkan akurasi sebesar **85%**

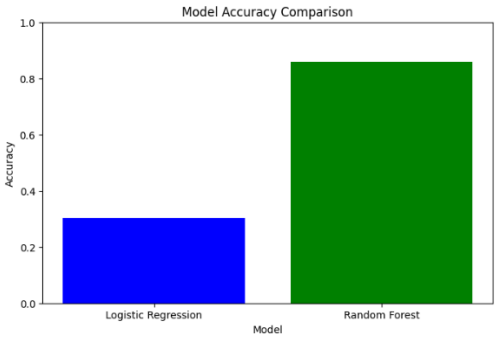
Setelah proses pelatihan selesai, performa masing-masing model diuji menggunakan metrik **akurasi**. Metrik ini dipilih karena mampu menggambarkan sejauh mana model dapat membedakan kelas dengan tepat.

Rumus accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Berikut ini adalah penjelasan mengenai rumus diatas **True Positive (TP)** merupakan jumlah kasus ketika model berhasil memprediksi suatu data sebagai positif dan kenyataannya memang positif. Sementara itu, **True Negative (TN)** adalah jumlah kasus ketika model memprediksi data sebagai negatif dan hasil aslinya memang negatif. Sebaliknya, **False Positive (FP)** terjadi saat model memprediksi data sebagai positif padahal sebenarnya data tersebut adalah negatif. Terakhir, **False Negative (FN)** adalah kondisi ketika model memprediksi data sebagai negatif, namun pada kenyataannya data tersebut adalah positif. Keempat komponen ini merupakan elemen utama dalam menghitung akurasi dan sangat penting dalam menilai sejauh mana model dapat mengenali dan membedakan kelas-kelas dalam data secara benar.

Visualisasi hasil ditampilkan dalam bentuk diagram batang yang menunjukkan perbandingan akurasi antara kedua model tersebut. Tingginya akurasi Random Forest menunjukkan keunggulan algoritma ini dalam menangani kompleksitas data yang ada.



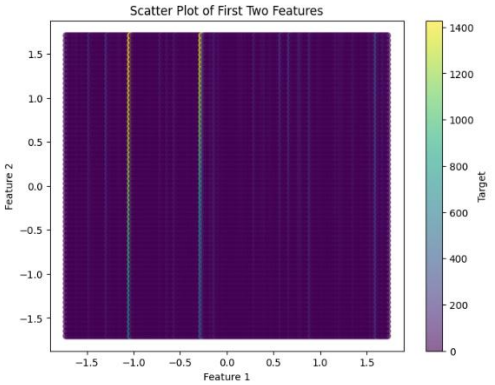
Gambar 1. Model Accuracy Comparison

Tabel 1.
Tabel perbandingan Accuracy

Model	Hasil Evaluasi Akurasi(%)
Logistic Regression	78
Random Forest	85

e. Scatter Plot Fitur Utama

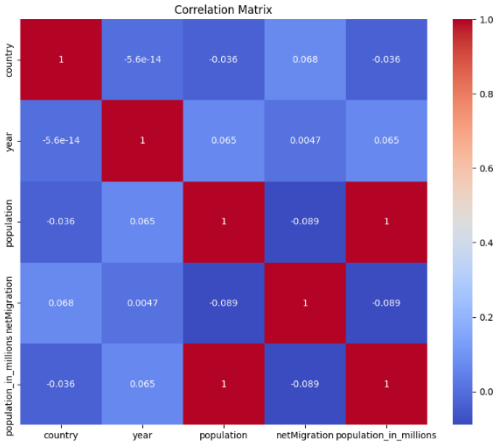
Scatter plot dari dua fitur utama memperlihatkan penyebaran data terhadap tiga kategori populasi (Low, Medium, High). Titik-titik data diberi warna berdasarkan kelas target, menghasilkan gambaran visual bahwa kelas-kelas tersebut relatif dapat dipisahkan. Hal ini mengindikasikan bahwa data memang mendukung proses klasifikasi dengan cukup baik.



Gambar 2. Scatter Plot of First Two Features

f. Heatmap Kolerasi Antar Fitur

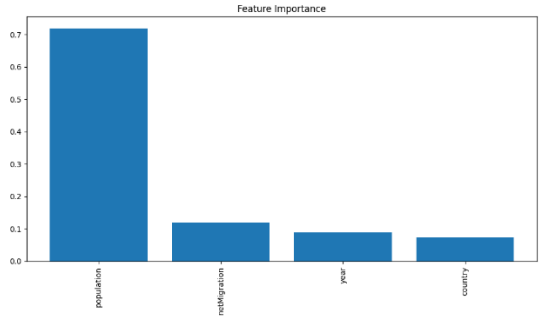
Analisis korelasi dilakukan dengan menggunakan heatmap. Heatmap ini menampilkan hubungan antar fitur dalam bentuk nilai korelasi, yang berguna dalam pemilihan fitur. Fitur yang memiliki korelasi sangat tinggi dapat menyebabkan multikolinearitas dan bisa dihilangkan dari dataset untuk menyederhanakan model.



Gambar 3. Correlation Matrix

g. Analisis Feature Importance

Algoritma Random Forest memberikan informasi tambahan berupa feature importance, yaitu ukuran kontribusi masing-masing fitur terhadap hasil klasifikasi. Dari hasil yang diperoleh, fitur Population dan Migration Rate merupakan dua variabel paling berpengaruh dalam menentukan kelas populasi negara.
Visualisasi feature importance memudahkan peneliti maupun pemangku kebijakan untuk memahami fitur mana yang harus menjadi fokus dalam pengambilan keputusan berbasis data.



Gambar 4. Feature Importance

h. Pembahasan

Dari hasil pemodelan dan evaluasi, dapat disimpulkan bahwa algoritma Random Forest memberikan performa klasifikasi yang lebih unggul dibandingkan Logistic Regression dalam mengelompokkan kategori populasi negara. Keunggulan ini tercermin dari nilai akurasi yang lebih tinggi (85% untuk Random Forest dibandingkan 78% untuk Logistic Regression), serta skor metrik lainnya seperti precision, recall, F1-score, dan ROC-AUC. Random Forest juga menunjukkan kestabilan prediksi yang lebih baik terhadap data yang bersifat non-linear atau kompleks karena kemampuannya memanfaatkan ensemble dari banyak pohon keputusan.

Keunggulan Random Forest tidak hanya dari segi performa kuantitatif, namun juga pada interpretabilitas model. Berdasarkan visualisasi feature importance, Random Forest mampu mengidentifikasi variabel-variabel paling berpengaruh dalam penentuan kategori populasi. Hal ini sangat membantu dalam memahami faktor-faktor utama yang membedakan kelompok negara berdasarkan ukuran dan pertumbuhan populasinya, seperti misalnya tingkat migrasi bersih, laju kelahiran, atau rasio pertumbuhan populasi.

Visualisasi tambahan seperti scatter plot dan correlation matrix semakin memperkuat validitas hasil analisis. Scatter plot dua fitur pertama menunjukkan adanya pola distribusi yang relatif terpisah antara kelas-kelas target, mengindikasikan bahwa data memang memiliki karakteristik yang bisa dipelajari model dengan baik. Correlation matrix juga membantu memastikan bahwa tidak ada multikolinearitas tinggi antar fitur yang bisa mempengaruhi performa model regresi.

Secara keseluruhan, pembahasan ini menunjukkan bahwa pendekatan machine learning, khususnya Random Forest, sangat potensial dalam membantu proses pengambilan keputusan berbasis data demografi. Model ini mampu menyederhanakan kompleksitas data populasi global dan menghasilkan prediksi yang akurat, serta memberikan insight penting untuk perencanaan pembangunan, kebijakan migrasi, dan strategi sosial ekonomi suatu negara.

IV. KESIMPULAN

Penelitian ini berhasil menerapkan algoritma *machine learning*, yaitu Random Forest dan Logistic Regression, untuk mengklasifikasikan negara-negara berdasarkan kategori populasi (Low, Medium, High) menggunakan data global populasi dan migrasi dari 186 negara. Hasil yang diperoleh menunjukkan bahwa Random Forest memberikan akurasi lebih tinggi dibandingkan Logistic Regression, serta mampu mengidentifikasi fitur-fitur paling berpengaruh seperti jumlah penduduk dan laju migrasi. Visualisasi melalui scatter plot, heatmap korelasi, dan grafik feature importance menguatkan bahwa data yang digunakan memang memiliki pola yang dapat diklasifikasikan secara efektif. Tujuan penelitian, yaitu menghasilkan model klasifikasi berbasis pembelajaran mesin untuk mendeteksi pola migrasi dan populasi, telah tercapai dengan baik. Temuan ini membuka peluang pengembangan lebih lanjut dalam integrasi teknologi kecerdasan buatan dalam studi kependudukan. Selain itu, hasil dari penelitian ini dapat dimanfaatkan oleh pembuat kebijakan untuk merancang strategi distribusi penduduk, pengelolaan urbanisasi, serta peningkatan pemerataan pembangunan berdasarkan tren populasi global yang teridentifikasi secara data-driven.

V. REFERENSI

[1] Hasnawati Hasnawati, Muhammad I Nurdin, and Daud Hasim, “Analisis Faktor-Faktor yang Mempengaruhi Migrasi di Kabupaten Pulau Taliabu,” *Digit. Bisnis J. Publ. Ilmu Manaj. dan E-Commerce*, vol. 2, no. 4, pp. 138–142, 2023, doi: 10.30640/digital.v2i4.1768.

[2] M. D. Prameswari and K. Asmara, “Determinan Migrasi Masuk Di Kota Surabaya,” *Equilib. J. Ilm. Ekon. Manaj. dan Akunt.*, vol. 13, no. 2, p. 500, 2024, doi: 10.35906/equili.v13i2.2069.

[3] U. Kurniasih and A. T. Suseno, “Analisis Sentimen Masyarakat Terhadap Isu Migrasi Rohingya Ke Indonesia,” *J. Teknol. Dan Sist. Inf. Bisnis*, vol. 7, no. 1, pp. 199–207, 2025, doi: 10.47233/jteksis.v7i1.1815.

[4] I. B. Mantra, “Pola Dan Arah Migrasi Penduduk Antar Propinsi Di Indonesia Tahun 1990,” *Populasi*, vol. 3, no. 2, 2016, doi: 10.22146/jp.11198.

[5] B. Nurbaiti, “Pengaruh Status Migrasi Melalui Karakteristik Sosio Demografi Terhadap Tingkat Kesejahteraan Pekerja Di DKI Jakarta (Analisis Data Cross Sectional Susenas 2013) Oleh : Beti Nurbaiti Dosen Prodi Manajemen , Fakultas Ekonomi , Universitas Borobudur Email,” vol. 19, 2017.

[6] R. H. Anggraini and Fafurida, "Economics Development Analysis Journal Pengaruh Kondisi Individu terhadap Keputusan Migrasi Sirkuler ke Kota Semarang," *Econ. Dev. Anal. J.*, vol. 5, no. 4, pp. 386–394, 2016.

[7] S. Sasmita, Trisnaningsih, and Yarmaidi, "Migrasi suku minangkabau ke Lamppung Tengah," *J. Penelit. Geogr.*, vol. 1, no. 9, pp. 23–31, 2019.

[8] A. Rahim, I. Fitriyani, and R. S. Ningrum, "Analisis Faktor-Faktor Yang Menentukan Minat Migrasi Penduduk Sirkuler Ke Kota Sumbawa," *J. Ekon. Bisnis*, vol. 10, no. 1, pp. 61–72, 2022, doi: 10.58406/jeb.v10i1.731.

[9] S. Purwatiningsih, "Respons Anak-Anak Migran Terhadap Migrasi Internasional Responses of Migrant Children Toward International," *Populasi*, vol. 24, no. 1, pp. 57–71, 2016.