Migration Prediction using ML

This report presents a comprehensive analysis of migration trends using a machine learning-based prediction system.

Vadisetti Pranay Satvik Reddy

2201221

System Architecture

The system combines historical data analysis with predictive modeling to forecast migration trends.

The system employs both traditional machine learning techniques and trend-based forecasting for accurate predictions.

Data Processing Module

Handles CSV data import and validation

Implements data preprocessing and feature engineering

Manages categorical encoding and numerical scaling

Machine Learning Engine

Random Forest Regressor for historical pattern learning

Trend-based forecasting for future predictions

Hybrid approach combines both methods for optimal results

Interactive GUI Interface

Real-time data visualization

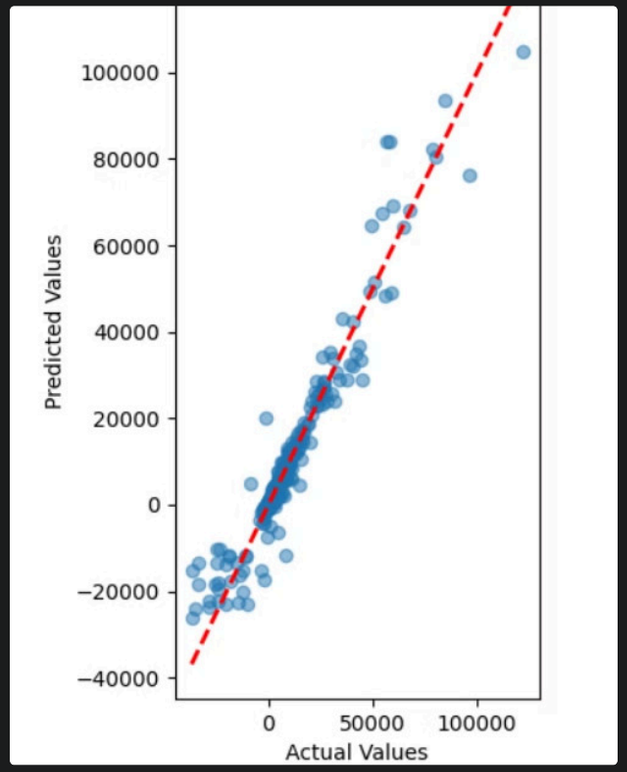
Interactive prediction interface

Model performance metrics display

Data Analysis and Visualization

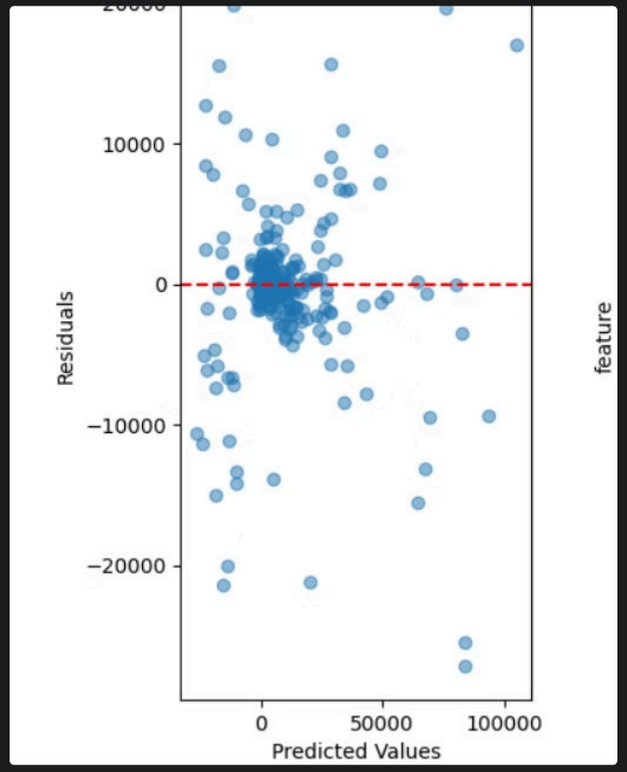
The prediction analysis demonstrates the model's accuracy in forecasting migration patterns.

The actual vs. predicted plot shows the correlation between predicted and actual values.



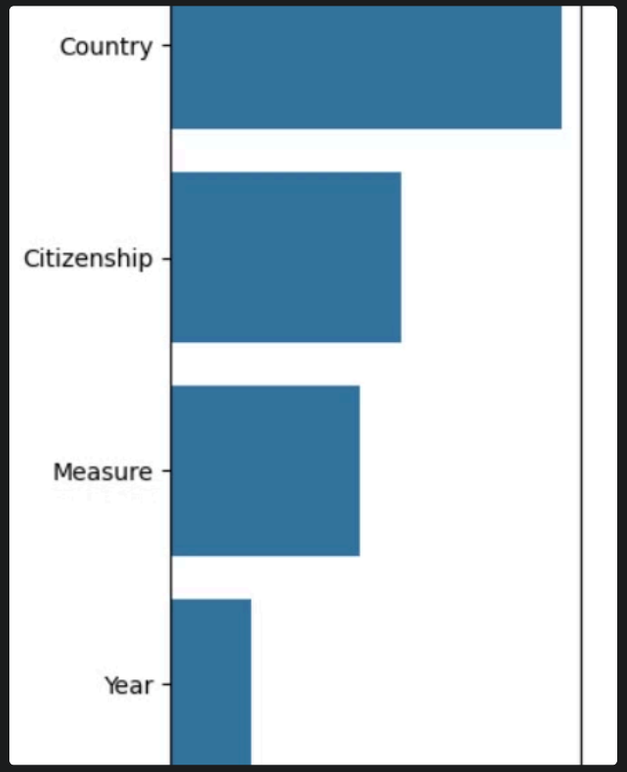
Actual vs. Predicted Values

Comparing predicted and actual migration values.



Residuals Analysis

Helps identify any systematic prediction biases.



Feature Importance Chart

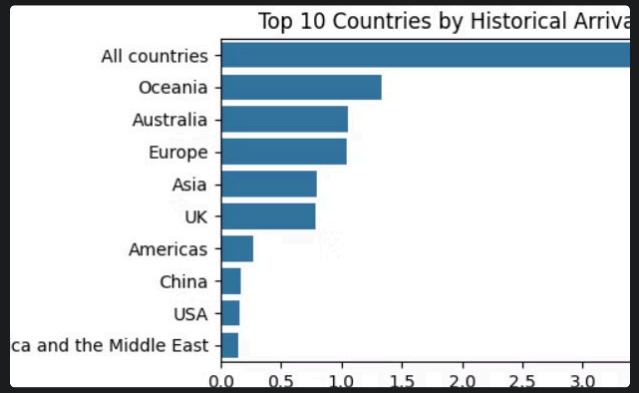
Shows the influence of features on prediction outcomes.

Migration Trends

These visualizations provide insights into historical patterns and projected future trends across different demographic and geographic segments.

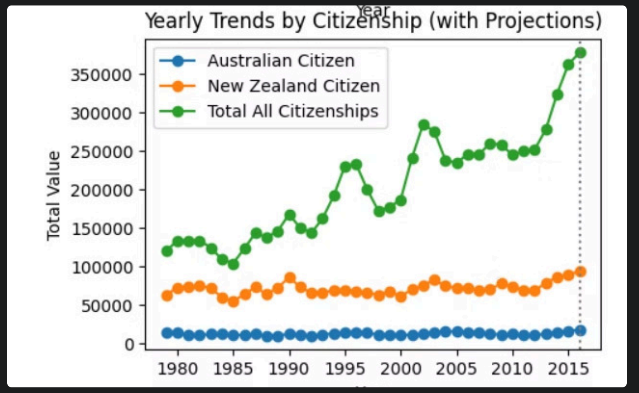
Overall Yearly Trends

Shows historical migration patterns with future projections.



Top 10 Countries by Historical Arrivals

Identifies countries with highest historical migration numbers.



Citizenship Trends Analysis

Tracks changes in citizenship categories over time.

Model Performance

The model demonstrates a high level of accuracy in predicting migration patterns, as evidenced by the performance metrics.

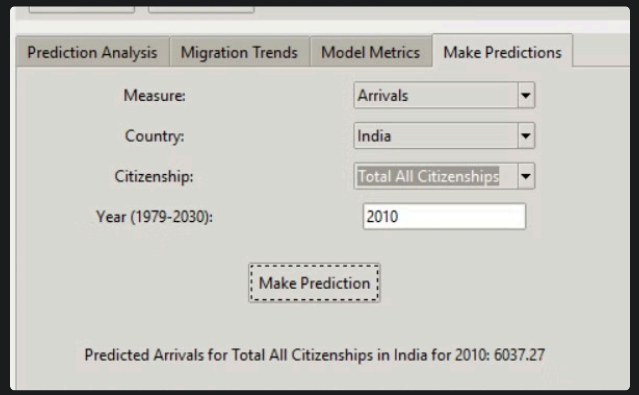
The performance metrics were calculated on training, validation, and test sets.

Set	R-squared (R^2)	RMSE	MAE
Training	0.9723	515.6471	48.9259
Validation	0.9311	817.5573	79.2210
Test	0.9375	812.3082	79.7909

Prediction Testing

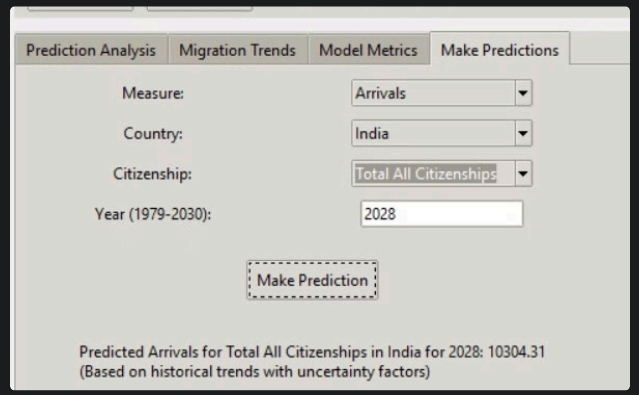
Sample prediction results for the number of migrants from India to New Zealand illustrate the system's capabilities.

The test cases represent different time periods: historical, near-future, and long-term.



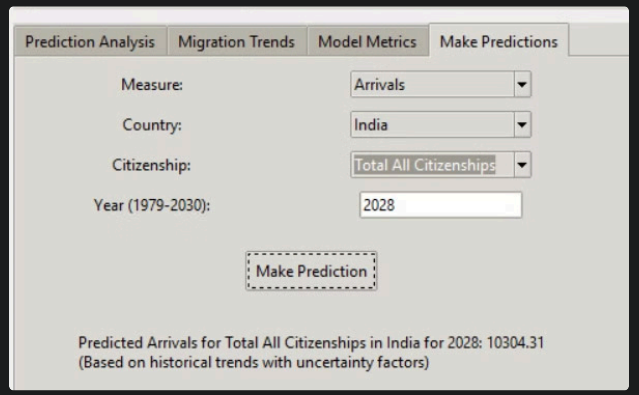
Test Case l: 2010

Historical period (pre-2016).



Test Case 2: 2017

Near-future prediction (2016-2023).



Test Case 3: 2028

Long-term forecast (2024-2030).

Technical Features

The system incorporates advanced capabilities to enhance prediction accuracy and user experience.

The system leverages dynamic trend modeling, data validation, and visualization components.

1 Dynamic Trend Modeling

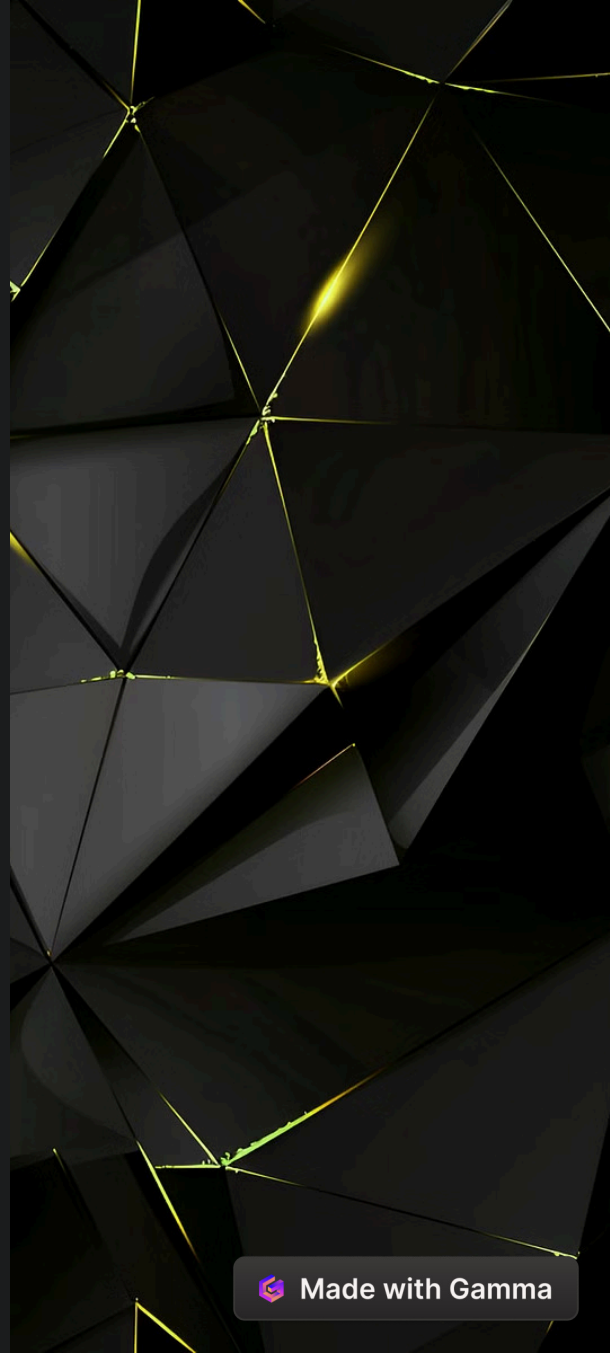
Automatic trend detection for each category combination.

2 Data Validation

Comprehensive error handling.

3 Visualization Components

Interactive plotting.



Implementation Details

The system employs a combination of algorithms and a data processing pipeline for effective prediction.

The system relies on Random Forest Regressor and Trend-based Forecasting algorithms for prediction.

Random Forest Regressor

100 estimators

Maximum depth of 20

Minimum samples split of 5

Minimum sample leaf of 2

Trend-based Forecasting

Linear regression for trend calculation

Dampening factor for long-term predictions

Random variation incorporation for realistic projections

Data Processing Pipeline

Initial data loading and validation

Feature preprocessing and encoding

Model training and validation

Hybrid prediction system implementation

System Limitations and Considerations

The system has limitations and considerations, including historical data cutoff, prediction constraints, and model assumptions.

These limitations and considerations need to be taken into account when interpreting predictions.

1 Historical Data Cutoff

Primary model training uses data up to 2016.

2 Prediction Constraints

Valid prediction range: 1979-2030.

3 Model Assumptions

Linear trend continuity.

Conclusions and Recommendations

The system demonstrates robust prediction capability for historical data, and future predictions incorporate uncertainty factors.

Regular model retraining, continuous monitoring of prediction accuracy, and periodic adjustment of trend parameters are recommended.

1 Key Findings

The system demonstrates robust prediction capability for historical data.

2 Recommendations

Regular model retraining with new data.

3 Future Enhancements

Integration of external factors (economic indicators, policy changes).

The system requires Python 3.12.1 and specific libraries, including pandas, numpy, scikit-learn, matplotlib, seaborn, and tkinter.