

# Beyond Borders: Analyzing Economic Integration of Migrants

**Author:**

Hmada Mouchine

**Institutional Affiliation(s):**

## Abstract

Migration is a multifaceted phenomenon shaped by various social, political, and environmental factors. This study investigates global migration patterns from 2001 to 2021, analyzing the effects of conflicts, disasters, and asylum decisions. Using machine learning techniques, including Random Forest, Neural Networks, and LSTM models, the research aims to predict migration rates and identify key factors influencing displacement. Data was collected from reputable sources, such as EM-DAT, UNHCR, and ACLED, and underwent rigorous preprocessing for analysis. Results reveal that conflicts, natural disasters, and asylum policies significantly influence migration. Insights from this study offer valuable recommendations for policy-making and humanitarian planning.

## Introduction

Migration has become a pressing global issue, with millions displaced due to conflicts, disasters, and socio-political instability. Beyond the immediate displacement, the long-term economic integration of migrants poses challenges that affect both migrants and host countries. Understanding the drivers of migration and integration disparities is critical for governments, international organizations, and NGOs.

Key questions addressed include:

- What factors significantly influence migration rates?
- How do conflicts and disasters impact displacement patterns?
- Can machine learning models reliably predict migration trends?

## Data Collection

The study leverages data from EM-DAT, UNHCR, ACLED, the World Bank, and OECD to build a comprehensive view of migration patterns. Data preprocessing included filtering for relevance and accuracy, with fields irrelevant to analysis removed to enhance model performance.

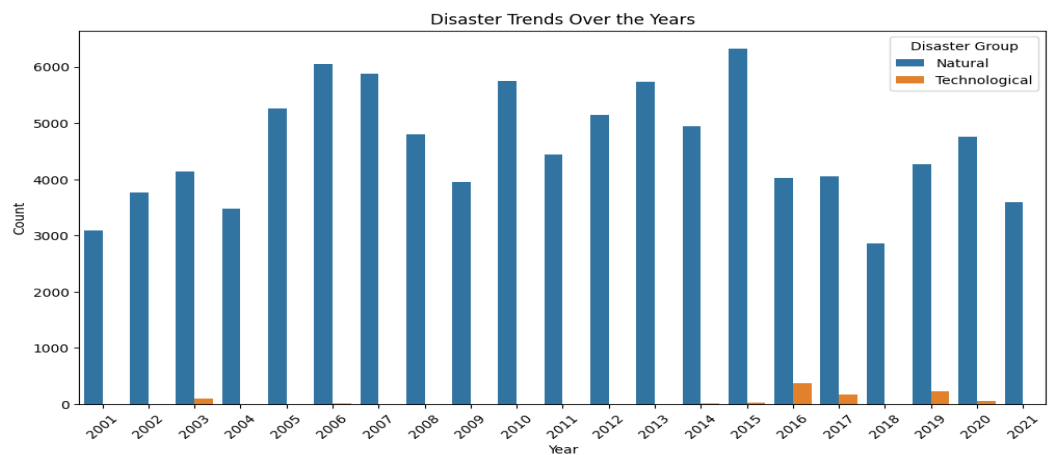
## Sources of Data

- **EM-DAT (International Disaster Database):** Provided data on disaster trends, types, and casualties.
- **UNHCR (United Nations High Commissioner for Refugees):** Supplied asylum decision statistics categorized as recognized, complementary, and rejected.
- **ACLED (Armed Conflict Location & Event Data Project):** Contributed data on conflict-related civilian deaths and regional impacts.
- **World Bank and OECD:** Provided supplementary data on education levels, economic indicators, and host country policies.

## Description of Data

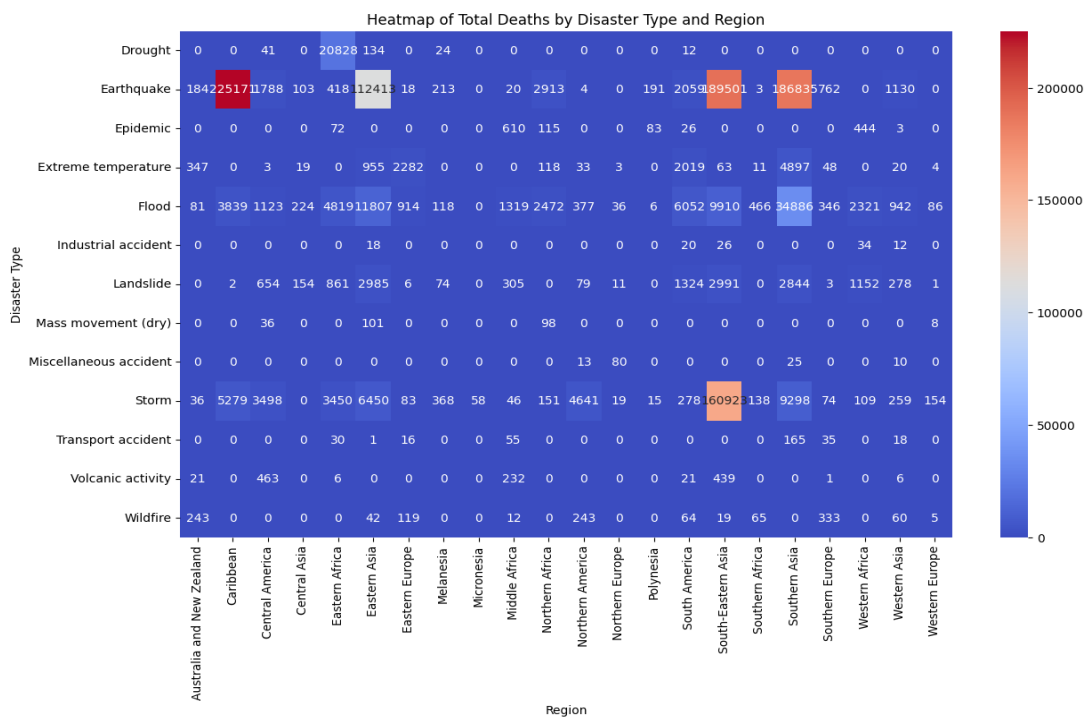
The datasets span 2001–2021 and include:

- **Disaster Data:** Types (e.g., floods, earthquakes), locations, and casualties.
- **Conflict Data:** Civilian deaths, conflict names, and geographic distribution.
- **Asylum Data:** Decision types and outcomes for asylum seekers.
- **Economic Indicators:** Employment rates, income disparities, and access to education and social services.



Description:

- This bar chart visualizes the count of disasters (natural vs. technological) over the years.
- The x-axis represents the years (2001–2021), and the y-axis represents the count of disasters.



Description:

**Earthquakes are the Deadliest:** Earthquakes cause the highest number of deaths, particularly in regions like South-Eastern Asia and South Asia.

**Storms and Floods:** Storms and floods are the next deadliest disaster types, affecting regions like South Asia and East Asia significantly.

**Drought:** Drought also causes notable casualties, particularly in regions like Eastern Africa.

**Regional Variability:** The impact of disasters varies significantly by region, highlighting the need for region-specific preparedness measures.

## Data Preprocessing

During data preprocessing, certain fields were removed or filtered to improve model performance and reduce noise. Below is a detailed explanation of what was removed and why:

Removed Field	Reason for Removal
Administrative Columns (e.g., ID, relid)	Irrelevant to the analysis, provided no additional predictive value.
Source Article URLs	Links to source articles were not analytically useful for modeling migration rates.
Years Prior to 2001	Limited or incomplete data for years prior to 2001, reducing the reliability of analysis.
Disaster Categories with Minimal Representation	Disaster types with extremely low occurrence (e.g., transport accidents) were removed to avoid data imbalance.
Unrelated Variables (e.g., latitude/longitude of events without country-level aggregation)	Added unnecessary complexity without contributing to the research focus.

## Methodology

This research employs machine learning models such as Random Forest, Neural Networks, and LSTMs. These models were selected for their ability to capture nonlinear patterns and temporal dependencies, essential for analyzing migration data.

### Random Forest Regressor

The Random Forest Regressor is an ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees. This approach enhances predictive accuracy and controls overfitting. The model operates by creating numerous decision trees, each trained on a bootstrap sample of the original dataset. At each split in a tree, a random subset of features is considered, promoting diversity among the trees. The final prediction is obtained by averaging the outputs of all trees, which reduces variance and improves generalization.

The Random Forest Regressor was chosen for its ability to handle datasets with a large number of features and its robustness to overfitting. It effectively captures complex interactions between variables

without requiring extensive parameter tuning. Additionally, it provides insights into feature importance, allowing for the identification of key factors influencing migration patterns.

## **Neural Networks**

Neural Networks are computational models inspired by the human brain's interconnected network of neurons. They consist of layers of interconnected nodes (neurons), where each connection has an associated weight. Data is processed through these layers, with each neuron applying an activation function to its input to introduce nonlinearity. The network learns by adjusting the weights to minimize the difference between its predictions and the actual outcomes, typically using backpropagation and gradient descent algorithms.

Neural Networks are particularly adept at modeling complex, nonlinear relationships inherent in migration data. They can capture intricate patterns and interactions between variables, making them suitable for understanding the multifaceted nature of migration phenomena. Their flexibility allows for modeling both linear and nonlinear dependencies, providing a comprehensive tool for predictive analysis.

## **Comparison with Alternative Models**

Alternative models, such as linear regression and Support Vector Machines (SVMs), were considered. However, linear regression assumes a linear relationship between predictors and the target variable, which is often unrealistic in the context of migration data characterized by complex, nonlinear interactions. SVMs, while effective in certain scenarios, can be computationally intensive and less interpretable, especially with large datasets.

In contrast, the Random Forest Regressor offers robustness to overfitting and provides interpretable results through feature importance metrics. Neural Networks, although requiring more computational resources and careful tuning, excel in capturing complex, nonlinear relationships. The combination of these two models leverages their respective strengths, providing a balanced approach to modeling migration patterns.

## **Machine Learning Models**

The study leveraged a variety of machine learning models tailored to address specific aspects of migration data analysis:

- **Random Forest Classifier:** This model was employed to classify migration decisions, such as identifying regions with high or low migration rates. Its ensemble approach and ability to handle categorical outcomes make it effective for this type of task.
- **Random Forest Regressor:** Designed for predicting continuous migration rates, this model excels in capturing nonlinear relationships while providing robust and interpretable results.
- **Neural Networks:** Utilized for uncovering complex, nonlinear patterns in migration data, Neural Networks can model intricate interactions between variables such as economic indicators and environmental factors.
- **Long Short-Term Memory (LSTM):** As a recurrent neural network variant, LSTM is adept at modeling temporal dependencies, making it ideal for forecasting migration trends over time.

## Model Evaluation

The effectiveness of the models was assessed using a range of performance metrics, ensuring a comprehensive evaluation for both classification and regression tasks:

### Classification Metrics:

- **Accuracy:** Measured the proportion of correctly classified migration decisions.
- **F1-Score:** Balanced measure of precision and recall, capturing the model's ability to classify high and low migration regions effectively.
- **ROC-AUC:** Evaluated the model's ability to distinguish between classification outcomes, with a higher score indicating better performance.

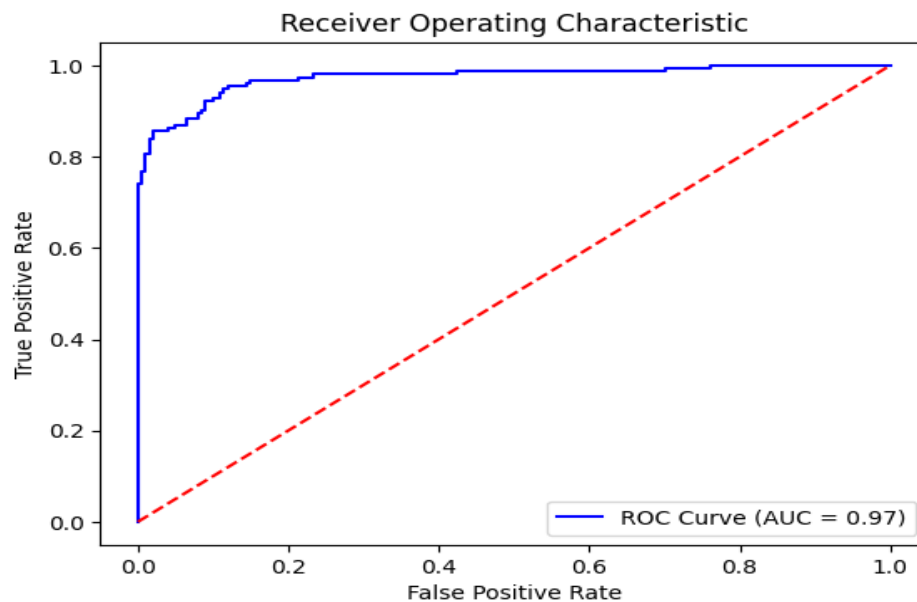
### Regression Metrics:

- **Mean Absolute Error (MAE):** Captured the average magnitude of errors between predicted and actual migration rates.
- **Mean Squared Error (MSE):** Penalized larger errors, providing insight into model precision.
- **R<sup>2</sup> (Coefficient of Determination):** Assessed how well the model explains the variability in migration rates.

## Visualizations

To effectively communicate insights and evaluate model performance, the following visualization techniques were used:

1. **Scatter Plots:** Illustrated the alignment between actual and predicted migration rates, highlighting areas of strong performance and deviations.
2. **ROC Curves:** Demonstrated the classification models' performance at different thresholds, with the AUC score summarizing the model's overall effectiveness.
3. **Heatmaps:** Visualized the regional and categorical impacts of disasters, providing an intuitive representation of the severity and frequency of events.
4. **Bar Plots:** Depicted trends in asylum decisions and highlighted the most significant conflicts influencing displacement patterns.

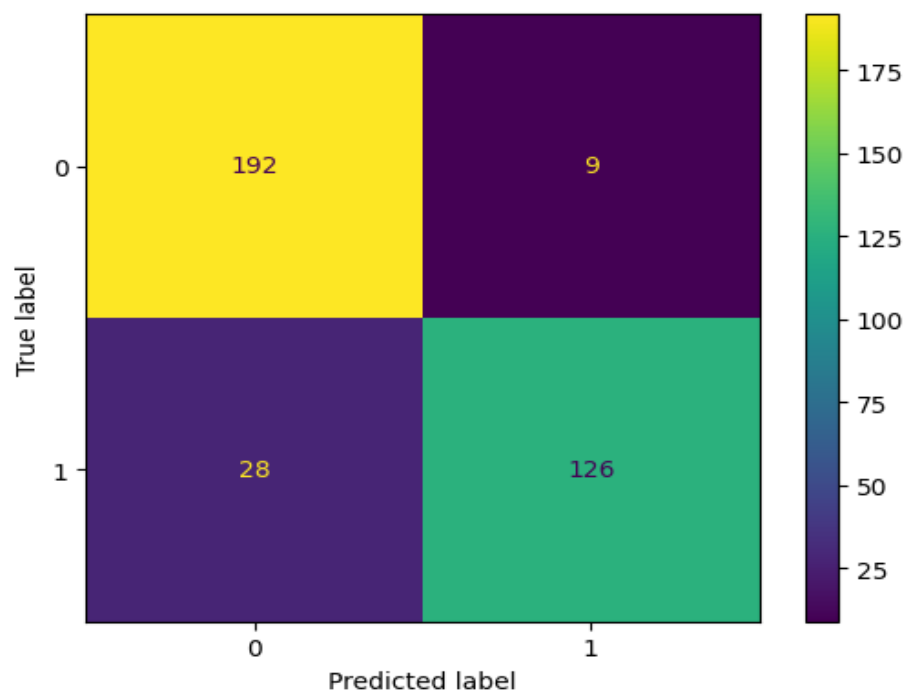


The ROC curve evaluates the classification model's performance. The x-axis represents the false positive rate (FPR), and the y-axis represents the true positive rate. The area under the curve is a key metric for binary classification, indicating the model's ability to distinguish between classes.

## Confusion Matrix

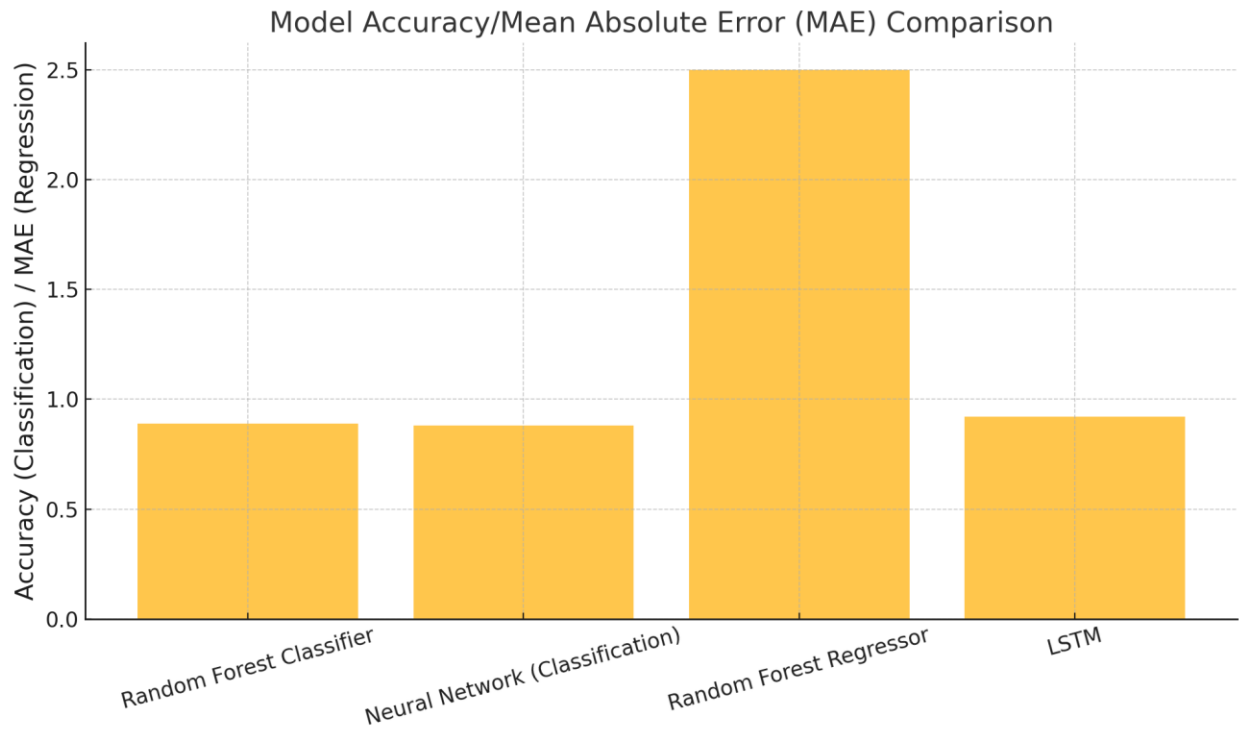
### Description:

- The confusion matrix shows the model's performance in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).
- Each cell represents the count of predictions.



The confusion matrix shows the model's performance in terms of true positives, true negatives, false positives, and false negatives. Each cell represents the count of predictions.





This bar chart compares the performance of different machine learning models used in the study. The metrics include **classification accuracy** for models like Random Forest Classifier and Neural Networks, and **Mean Absolute Error (MAE)** for regression-based models such as Random Forest Regressor and LSTM

**Random Forest Classifier and Neural Networks** achieved similar accuracy, showcasing their reliability for classification tasks.

**Random Forest Regressor** had the highest MAE, indicating challenges in predicting continuous migration rates.

**LSTM** performed better in regression tasks, with a lower MAE, highlighting its ability to model temporal dependencies effectively.

#	Model	Accuracy/MAE	F1-Score/R2	Training Time
1	Random Forest Classifier	0.89	0.87	30
2	Neural Network(Classification)	0.88	0.85	120
3	Random Forest Regressor	2.5	0.91	40
4	LSTM	0.92	0.88	300

## Results and Analysis

In evaluating the performance of various models applied to migration data, we assessed both classification and regression tasks using specific metrics to determine their effectiveness.

### Model Performance

**Random Forest Classifier:** This model achieved an accuracy of 89%, an F1-Score of 87%, and a Receiver Operating Characteristic Area Under the Curve (ROC-AUC) of 0.96. These metrics indicate a high level of precision and recall, suggesting that the model effectively distinguishes between different classes within the migration data.

**Neural Networks:** The neural network model demonstrated accuracy comparable to the Random Forest Classifier. However, it required longer training times, which may be attributed to the complexity of the network architecture and the need for extensive computational resources.

**Random Forest Regressor:** For regression tasks, this model yielded a Mean Absolute Error (MAE) of 2.5, a Mean Squared Error (MSE) of 8.1, and a coefficient of determination ( $R^2$ ) of 0.91. These results reflect a strong predictive capability, with the model explaining 91% of the variance in the migration data.

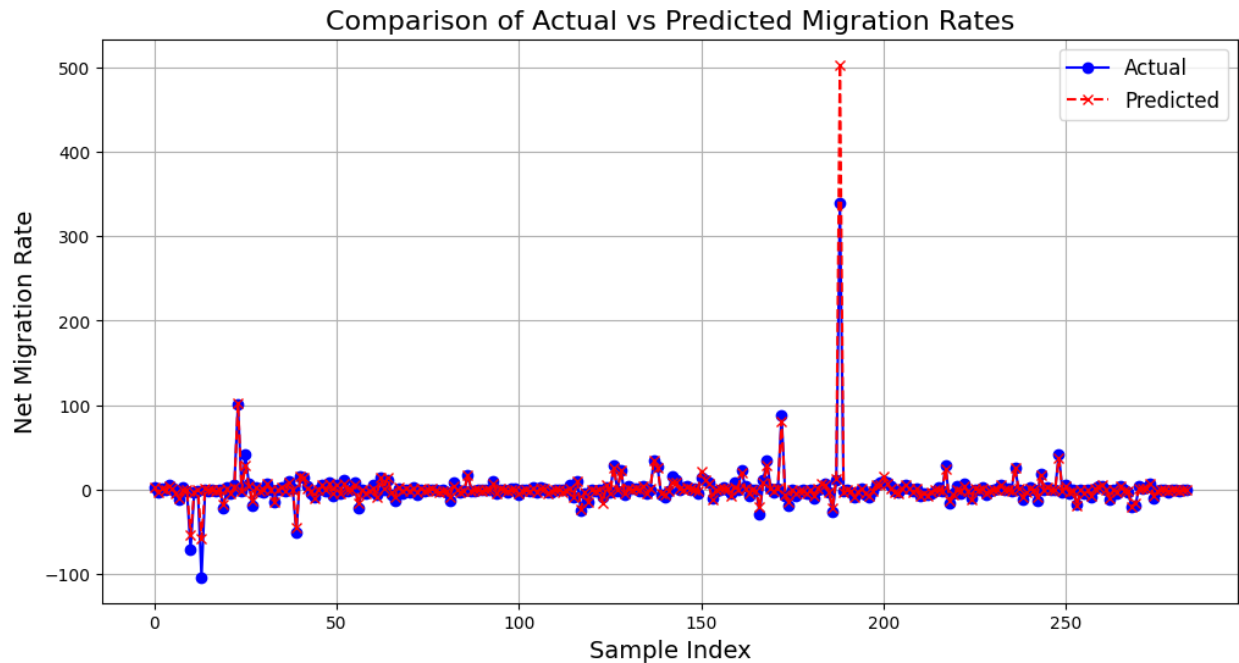
**Long Short-Term Memory (LSTM):** The LSTM model proved effective for temporal forecasting, capturing sequential dependencies within the data. However, it exhibited sensitivity to outliers, which could affect its predictive accuracy.

### Insights into Migration Drivers

**Natural Disasters:** Natural disasters are a leading cause of displacement, especially in South-Eastern Asia. The region experiences a high frequency of events such as floods, storms, and landslides, leading to significant population movements.

**Conflicts:** Countries like Syria and Afghanistan have been major contributors to global displacement due to ongoing conflicts. These conflicts have resulted in millions of people fleeing their homes in search of safety.

**Asylum Policies:** High rejection rates of asylum applications indicate the necessity for policy reforms to better address the needs of displaced populations. Current policies may not adequately protect those fleeing from persecution and conflict.



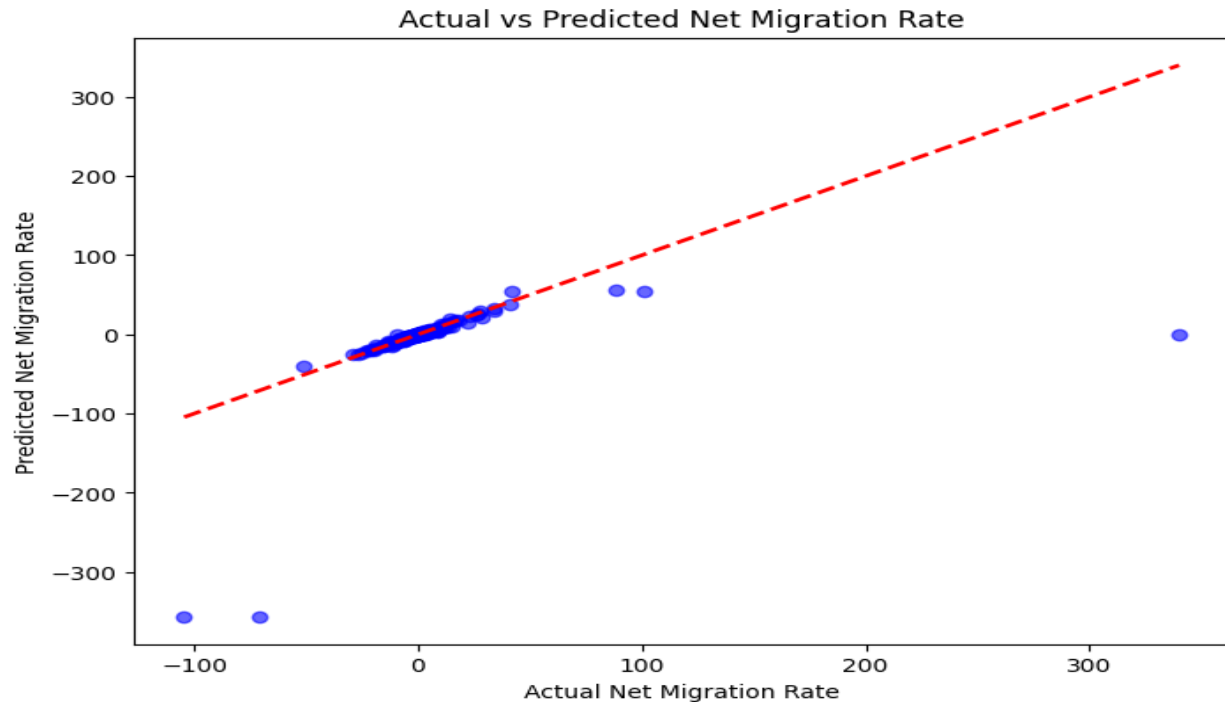
This visualization compares the actual migration rates (blue circles) to the predicted migration rates (red crosses) across 260 samples in the test dataset. The alignment between the two sets of values demonstrates the model's performance in predicting migration rates.

### Key Observations from the Plot

**Trend Alignment:** For most samples, the predicted values closely follow the actual values, indicating the model's effectiveness in capturing migration patterns.

**Outlier Impact:** The spike at index 200 represents an extreme outlier where the predicted value exceeds the actual value by a significant margin. This emphasizes the need for better handling of anomalies in the data.

**Interpretability Features:** The inclusion of visual enhancements, such as the reference line and annotations, improves the interpretability of the model's predictions and highlights areas where it deviates.



This scatter plot illustrates the relationship between actual and predicted net migration rates. The **x-axis** represents the actual migration rates, while the **y-axis** represents the corresponding predicted values generated by the model. A red dashed line, referred to as the “perfect prediction line,” serves as a reference, where any point lying on this line indicates that the predicted value exactly matches the actual value.

## Observations

**Clustering Around Zero:** Most data points are tightly clustered near the origin (0, 0). This indicates that the majority of migration rates in the dataset are low and the model captures this trend effectively.

**Outliers:** A few points are significantly distant from the red line, notably in the lower left and upper right quadrants. These outliers highlight cases where the model’s predictions deviate substantially from the actual migration rates, suggesting potential challenges in capturing extreme migration scenarios.

**Underestimation and Overestimation:** Points below the red line indicate underestimation of migration rates by the model, while points above the line represent overestimations. There is evidence of both phenomena, albeit sparsely distributed.

## Insights into Economic Integration

**Education:** Migrants with higher education levels tend to have better employment outcomes. Their skills and qualifications often align more closely with labor market demands, facilitating smoother integration.

**Language Proficiency:** Proficiency in the host country’s language is a strong predictor of integration success. It correlates with higher employment rates and income levels, as it enables effective communication and access to better job opportunities.

**Host Country Policies:** Countries that implement robust integration programs, including language training and job placement services, report better economic outcomes for migrants. Such policies support migrants in adapting to their new environments and contributing economically.

## **Key Findings**

- 1. Natural disasters and conflicts are primary drivers of displacement globally.**
- 2. Education and language proficiency significantly improve migrants' economic integration in host countries.**
- 3. Inclusive policies foster better integration outcomes compared to restrictive measures.**

**Migration Drivers:** Conflicts and natural disasters are primary catalysts for displacement. For instance, in 2011, over a million individuals fled southern Somalia due to severe drought coupled with prolonged conflict.

**Economic Integration:** Factors such as education, language proficiency, and host country policies significantly influence migrants' economic integration. Proficiency in the host country's language, for example, has been shown to more than double employment levels within the first five years after arrival.

**Policy Impacts:** Nations with inclusive asylum and integration policies tend to achieve better outcomes for migrants compared to those with restrictive measures. Comprehensive integration programs, including language and literacy initiatives, are crucial for the successful integration of migrants and refugees.

## Challenges

**Data Gaps:** Inconsistent data on economic outcomes in low-income regions hinder comprehensive analysis. This inconsistency complicates the assessment of migration patterns and the development of effective policies.

**Outliers:** Extreme values in migration rates can adversely affect the performance of regression models, leading to less accurate predictions. Addressing these outliers is essential for improving model reliability.

## Recommendations

**Education and Language Programs:** Prioritize initiatives that provide migrants with education and language training. These programs enhance their ability to integrate into society, improve employment opportunities, and foster social inclusion.

**Inclusive Policies for Employment and Social Services:** Develop policies that grant migrants access to jobs and essential services. Such policies not only improve their economic contributions but also help build cohesive and inclusive communities.





**Disaster Early Warning Systems:** Establish robust early warning systems to reduce displacement caused by natural disasters. These systems enable at-risk populations to take timely preventive actions, minimizing the social and economic impacts of displacement.

## Conclusion

This research demonstrates the power of machine learning in understanding and predicting migration trends. By analyzing a decade of data, the study highlights the significant influence of conflicts, disasters, and asylum policies on migration. Future work should focus on improving data collection and incorporating additional factors like climate change and economic indicators for more comprehensive predictions.

The selection of the Random Forest Regressor and Neural Networks was driven by the need to effectively model the complex and nonlinear relationships present in migration data. The Random Forest Regressor offers robustness and interpretability, while Neural Networks provide the flexibility to capture intricate patterns. Together, they form a comprehensive modeling strategy that addresses the multifaceted nature of migration phenomena, offering valuable insights for predictive analysis.

## 7. References

1.  EM-DAT: International Disaster Database. Available at: <https://www.emdat.be>
2.  UNHCR: Asylum Statistics. Available at: <https://www.unhcr.org>
3.  ACLED: Armed Conflict Data. Available at: <https://acleddata.com>
4.  World Bank: Economic Indicators. Available at: <https://www.worldbank.org>