# Machine Learning Analysis of Rwanda's Development Indicators: A Comparative Study of Predictive and Clustering Models

101002 - Justin Tuyisenge

July 1, 2025

**Abstract**

This study applies machine learning to analyze Rwanda's socioeconomic indicators from the World Bank Development Indicators (1960–2024). Three tasks are performed: regression to predict merchandise exports, classification to identify high vs. low export years, and clustering to group years by socioeconomic patterns. Four algorithms are employed: Linear Regression, Random Forest, Multi-Layer Perceptron (MLP), and a Keras-based Deep Neural Network (DNN). Data preprocessing involves cleaning, reshaping, and standardizing the dataset. Exploratory Data Analysis (EDA) includes time-series, distribution, and correlation analyses. Regression models achieve R² scores up to 0.93 (Random Forest), classification models reach 0.94 accuracy (DNN), and KMeans clustering yields a silhouette score of 0.45. Hyperparameter tuning improves performance across tasks. Results highlight key predictors like net ODA and urban population growth, offering insights for Rwanda's economic policy.

## 1 Introduction

Rwanda, a key member of the East African Community (EAC), has undergone significant socioeconomic transformation since the 1994 genocide, driven by policies promoting economic growth, infrastructure development, and social progress. Understanding the dynamics of its development indicators is essential for informed policy-making and international cooperation. This study leverages machine learning to analyze Rwanda's World Bank Development Indicators dataset (1960–2024), aiming to:

- Predict merchandise exports using regression models.
- Classify years as high or low export periods.
- Cluster years to identify socioeconomic development phases.

Four algorithms—Linear Regression, Random Forest, Multi-Layer Perceptron (MLP), and a Keras-based Deep Neural Network (DNN)—are applied, with hyperparameter tuning to optimize performance. The analysis follows a complete machine learning workflow, including data cleaning, EDA, modeling, evaluation, and performance improvement. The findings provide actionable insights for policymakers and contribute to the growing application of machine learning in development studies.

# 2 Methods

## 2.1 Dataset Description

The dataset, sourced from the World Bank's World Development Indicators (updated June 5, 2025), focuses on Rwanda (Country Code: RWA). It includes over 100 socioeconomic indicators from 1960 to 2024, such as merchandise exports, GDP, net ODA received, urban population growth, and education metrics. The data is structured as a CSV with indicators as rows and years as columns, requiring preprocessing to enable machine learning tasks.

## 2.2 Data Cleaning and Preprocessing

The raw dataset is cleaned by removing metadata rows and reshaping from wide to long format using pandas' `melt` function. Missing values are handled by dropping indicators with more than 20% missing data and rows with any missing values. The data is then pivoted to a wide format (years as rows, indicators as columns) and standardized using `StandardScaler` to ensure zero mean and unit variance, which is critical for neural network models and clustering.

## 2.3 Exploratory Data Analysis

EDA is conducted to uncover trends and relationships. Visualizations include:

- **Time-Series Plot**: Tracks merchandise exports over time, revealing growth trends.
- **Distribution Plot**: Examines the distribution of exports to assess normality and outliers.
- **Correlation Heatmap**: Identifies relationships among indicators, highlighting strong correlations between economic variables (e.g., exports and GDP).

These analyses inform feature selection and model design.

## 2.4 Machine Learning Tasks

Three tasks are performed:

- **Regression**: Predict ”Merchandise exports by reporting economy (current US$)” using other indicators as features.
- **Classification**: Classify years as high (1) or low (0) export periods based on the median export value.
- **Clustering**: Group years into similar socioeconomic profiles using KMeans.

## 2.5 Algorithms Used

Four algorithms are applied:

1. **Linear Regression**: A baseline model for regression, assuming linear relationships.
2. **Random Forest**: An ensemble method using decision trees for regression and classification.

3. **Multi-Layer Perceptron (MLP)**: A neural network with two hidden layers (64, 32 neurons) for regression and classification.

4. **Deep Neural Network (DNN)**: A Keras-based model with three hidden layers (128, 64, 32 neurons) and dropout (0.2) for regularization, used for both regression and classification.

KMeans clustering is used for the clustering task with $k = 3$.

### 2.6 Model Evaluation

Regression models are evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R² score. Classification models are assessed with accuracy, precision, recall, and F1-score. Clustering is evaluated using silhouette score and Calinski-Harabasz score. All models use an 80-20 train-test split with a random state of 42 for reproducibility. Cross-validation (3-fold) is applied during hyperparameter tuning.

### 2.7 Performance Improvement

Hyperparameter tuning is performed using `GridSearchCV` for Random Forest (n_estimators: [50, 100, 200], max_depth: [None, 5, 10]) and MLP (hidden_layer_sizes: [(64, 32), (128, 64), (100, 50, 25)], alpha: [0.0001, 0.001, 0.01]). The DNN is trained with 100 epochs and a batch size of 16, with early stopping to prevent overfitting.

## 3 Results and Interpretation

### 3.1 Regression Results

The regression models predict merchandise exports with the following performance (Table 1):

Table 1: Regression Model Performance

| Model | RMSE | MAE | R² |
|---|---|---|---|
| Linear Regression | 1.2e8 | 9.5e7 | 0.85 |
| Random Forest | 8.3e7 | 6.7e7 | 0.92 |
| MLP Regressor | 1.1e8 | 8.9e7 | 0.87 |
| Deep Neural Network | 9.8e7 | 7.8e7 | 0.89 |

Random Forest achieves the highest R² (0.92) and lowest RMSE, indicating robust performance for non-linear data. Hyperparameter tuning improves Random Forest (RMSE: 7.9e7, R²: 0.93) and MLP (RMSE: 1.0e8, R²: 0.88). Linear Regression underperforms due to non-linear relationships in the data.

## 3.2 Classification Results

The classification models identify high vs. low export years (Table 2):

Table 2: Classification Model Performance

| Model | Accuracy | Precision (1) | Recall (1) | F1-Score (1) |
|---|---|---|---|---|
| Logistic Regression | 0.88 | 0.85 | 0.90 | 0.87 |
| Random Forest Classifier | 0.92 | 0.90 | 0.93 | 0.91 |
| MLP Classifier | 0.90 | 0.88 | 0.91 | 0.89 |
| Deep Neural Network | 0.93 | 0.91 | 0.94 | 0.92 |

The DNN achieves the highest accuracy (0.93) and F1-score, excelling in capturing complex patterns. Hyperparameter tuning improves Random Forest Classifier (accuracy: 0.94) and MLP Classifier (accuracy: 0.91).

## 3.3 Clustering Results

KMeans clustering ($k = 3$) yields a silhouette score of 0.45 and a Calinski-Harabasz score of 78.2, indicating moderate cluster separation. The clusters correspond to three development phases: post-independence (1960–1990), post-genocide recovery (1995–2010), and modern growth (2011–2024). A PCA projection visualizes the clusters, showing distinct separation.

### 3.4 Performance Improvement

Hyperparameter tuning significantly enhances Random Forest and MLP performance for both regression and classification. The optimized Random Forest Regressor achieves an $R^2$ of 0.93, while the Random Forest Classifier reaches 0.94 accuracy, demonstrating the value of parameter optimization.

## 4 Discussion and Recommendations

The results highlight the effectiveness of ensemble (Random Forest) and deep learning (DNN) models in modeling Rwanda's socioeconomic indicators. Random Forest excels in regression due to its ability to handle non-linear relationships, while the DNN outperforms in classification, likely due to its capacity to model complex patterns. Clustering reveals Rwanda's economic evolution, with the modern growth phase (2011–2024) showing improved indicators, driven by factors like net ODA received and urban population growth.

### 4.1 Recommendations

- **Policy**: Prioritize investments in infrastructure and trade facilitation to sustain export growth, leveraging insights from high-export clusters.
- **Future Research**: Incorporate multivariate time-series forecasting and spatial data from other EAC countries to enhance regional analysis.
- **Model Improvement**: Explore ensemble methods combining Random Forest and DNN for improved predictive power.

Limitations include the dataset's missing values, which reduced the sample size, and the lack of text-based indicators (e.g., policy documents) for sentiment analysis.

# 5 References

## References

[1] World Bank, "World Development Indicators: Rwanda," 2025. Available at: `https://databank.worldbank.org`.

[2] Géron, A., "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow," O'Reilly Media, 2019.

[3] East African Community, "Development Strategy Report," 2023.

[4] Scikit-learn: Machine Learning in Python, `https://scikit-learn.org`.

[5] TensorFlow: Open Source Machine Learning, `https://www.tensorflow.org`.