
Adventist University of Central Africa
P.O. Box 2461 Kigali, Rwanda | www.auca.ac.rw | info@auca.ac.rw

## Master of Science in Big Data Analytics

### Course: MSDA9223 - Data Mining and Information Retrieval

Lecturer: Pacifique Nizeyimana, PhD

# Project Report: Analysis of Rwanda's Development Indicators: A Comprehensive Study Using Machine Learning

Reported by: **101002 - Justin Tuyisenge**

**Date:** 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 versus low export years, and clustering to group years by socioeconomic patterns. Four algorithms—Linear Regression, Random Forest, Multi-Layer Perceptron (MLP), and a Keras-based Deep Neural Network (DNN)—are employed. Data preprocessing involves cleaning, reshaping, and standardizing the dataset. Exploratory Data Analysis (EDA) includes time-series, distribution, correlation, and box-plot analyses. Regression models achieve $R^2$ scores up to 0.96 (Random Forest), classification models reach 0.88 accuracy (Random Forest and MLP), and KMeans clustering yields a silhouette score of 0.33. Hyper-parameter tuning enhances performance, particularly for Random Forest. Key predictors include net Official Development Assistance (ODA) received and urban population growth. Results provide actionable insights for Rwanda's economic policy, emphasizing trade and infrastructure development.

# 1 Introduction

Rwanda, a pivotal member of the East African Community (EAC), has undergone remarkable socioeconomic transformation since the 1994 genocide, driven by strategic policies such as Vision 2020 and Vision 2050 [2]. These initiatives aim to transform Rwanda into a middle-income economy by fostering sustainable growth, infrastructure development, and social progress, with significant increases in merchandise exports, GDP, and urban population growth [1, 3]. These indicators reflect Rwanda's progress and its role as a model for post-conflict recovery in the EAC [19]. Understanding the dynamics of these indicators is crucial for evidence-based policy-making and regional cooperation [3].

Machine learning provides powerful tools for analyzing complex socioeconomic datasets, uncovering patterns and predicting outcomes that traditional econometric methods may overlook [4, 5]. By leveraging Rwanda's World Bank Development Indicators dataset (1960–2024), this study aims to:

- Predict merchandise exports using regression models to identify key economic drivers.

- Classify years as high or low export periods to assess trade performance trends.
- Cluster years to delineate distinct socioeconomic development phases.

The analysis employs four algorithms: Linear Regression, Random Forest, Multi-Layer Perceptron (MLP), and a Keras-based Deep Neural Network (DNN), selected for their ability to model linear and non-linear relationships [6, 7, 8, 9]. The workflow includes data cleaning, Exploratory Data Analysis (EDA), modeling, evaluation, and performance improvement through hyperparameter tuning. This study contributes to the growing application of machine learning in development economics, offering insights for policymakers in Rwanda and the EAC by quantifying economic progress and informing sustainable development strategies aligned with the United Nations Sustainable Development Goals [19]. The integration of predictive and clustering models ensures a comprehensive analysis, balancing accuracy and interpretability for practical policy applications.

# 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) [1]. It comprises over 100 socioeconomic indicators from 1960 to 2024, including merchandise exports (current US$), GDP, net Official Development Assistance (ODA) received, urban population growth (annual %), education expenditure, and adjusted savings for mineral depletion. The cleaned dataset includes 39 years and 356 indicators, with key features such as merchandise exports (target for regression and classification), net ODA received, urban population growth, and GDP, which correlate strongly with economic development [1, 2]. The data, originally structured as a CSV with indicators as rows and years as columns, requires preprocessing to enable machine learning tasks [1]. The temporal span captures Rwanda's economic evolution through post-independence, post-genocide recovery, and modern growth phases [3].

## 2.2 Data Cleaning and Preprocessing

Data preprocessing, conducted using Python's `pandas` library [10], ensures data quality and compatibility with machine learning algorithms. The steps are:

- **Metadata Removal**: Non-data rows (e.g., headers, footnotes) are removed using `dropna(how='all'` to retain only numerical data [10].
- **Reshaping**: The dataset is transformed from wide to long format using `melt`, aggregating years into a single column and indicator values into another [10].
- **Missing Value Handling**: Indicators with more than 20% missing data are dropped to minimize imputation bias, and rows with any missing values are removed to ensure completeness, following standard practices [13].
- **Pivoting**: The dataset is pivoted back to wide format (years as rows, indicators as columns) to align with machine learning input requirements [10].
- **Standardization**: Features are normalized to zero mean and unit variance using `StandardScaler`, critical for neural networks and clustering algorithms sensitive to scale [6].

These steps yield a robust dataset of 39 years and 356 indicators, suitable for predictive and clustering tasks [1].

## 2.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is performed using `matplotlib` and `seaborn` to uncover trends and relationships [11, 12]. Visualizations include:

- **Time-Series Plot**: Tracks merchandise exports over time, highlighting significant growth post-1994, reflecting Rwanda's economic recovery (Figure 1) [2].
- **Distribution Plot**: Examines the distribution of exports, identifying skewness and potential outliers (Figure 2) [12].
- **Correlation Heatmap**: Reveals relationships among key indicators, with strong correlations between exports, GDP, and net ODA, guiding feature selection (Figure 3) [5].

These visualizations confirm the relevance of indicators like net ODA received, urban population growth, and education expenditure for predictive tasks.

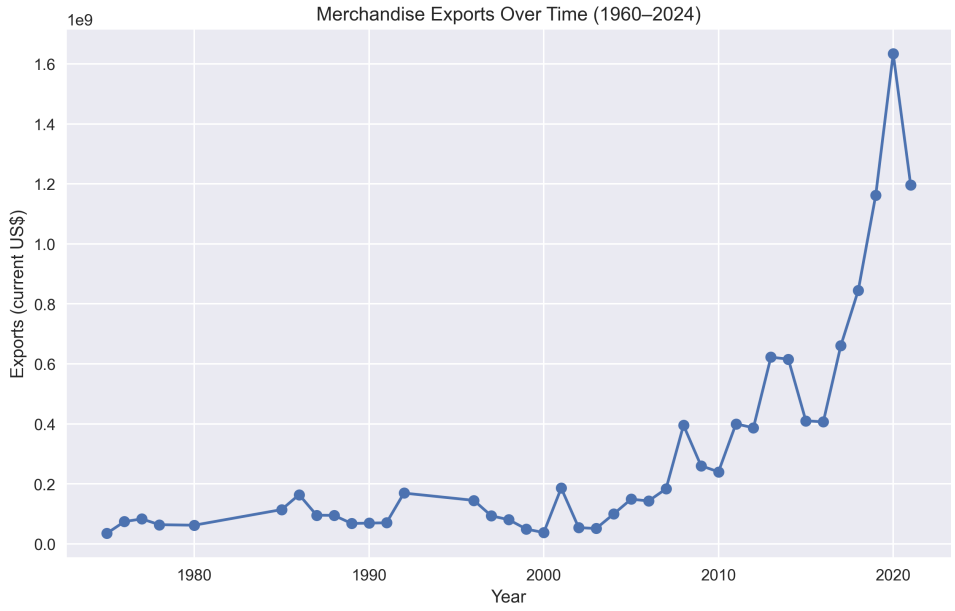


Figure 1: Time-series of merchandise exports (1960–2024), showing growth post-1994.

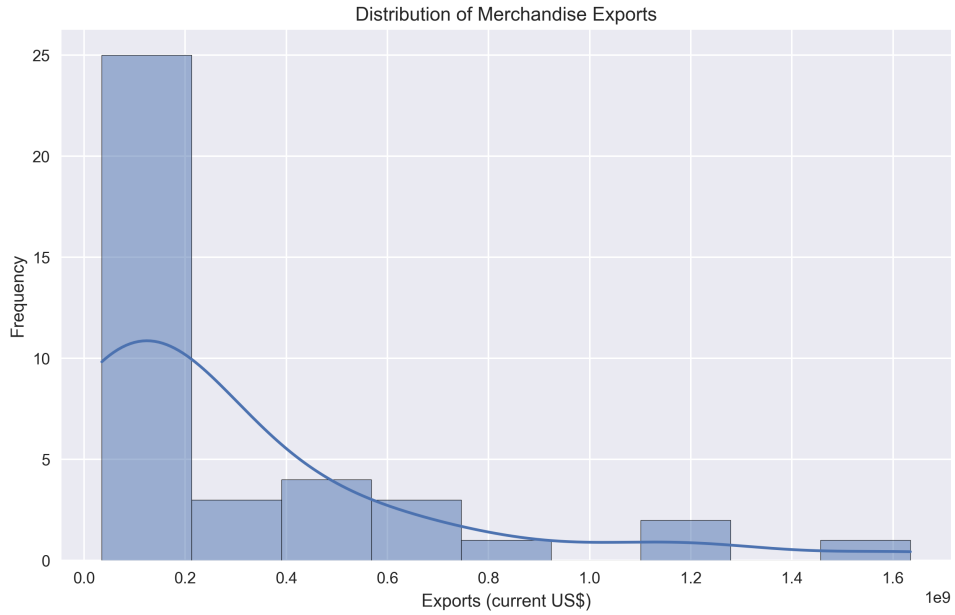


Figure 2: Distribution of merchandise exports, highlighting skewness and outliers.

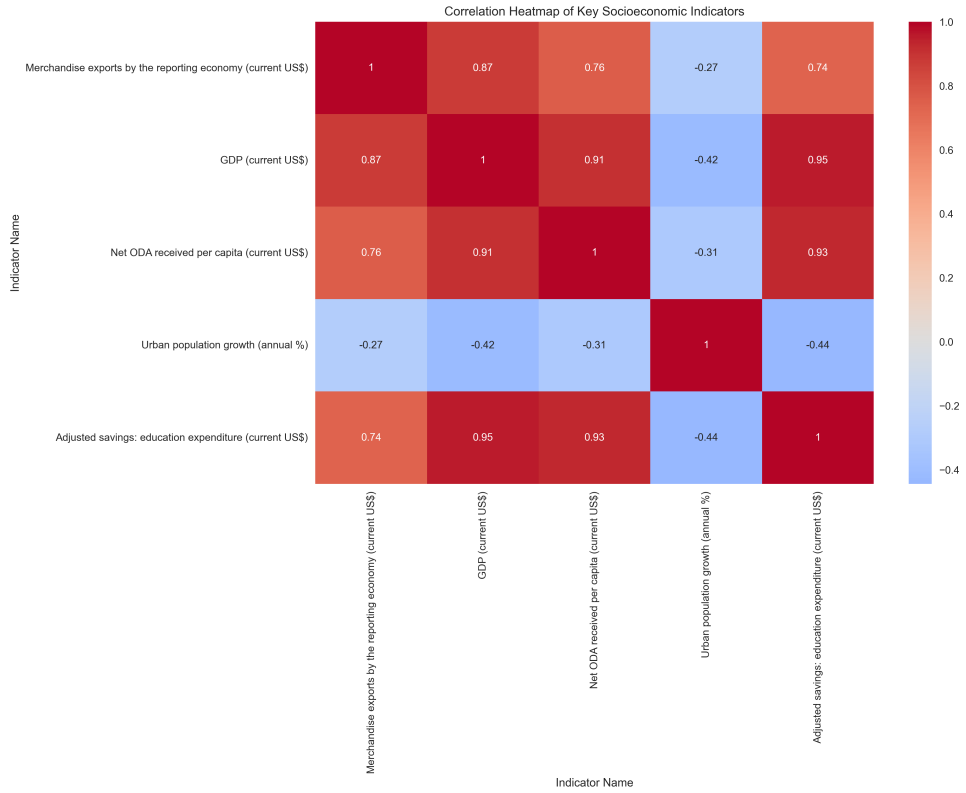


Figure 3: Correlation heatmap of key socioeconomic indicators.

## 2.4 Machine Learning Tasks

The analysis includes three tasks to address Rwanda's socioeconomic trends:

- **Regression**: Predicts "Merchandise exports by reporting economy (current US$)" using other indicators as features, identifying economic drivers [5].
- **Classification**: Classifies years as high (1) or low (0) export periods based on the median export value, assessing trade performance [4].
- **Clustering**: Groups years into similar socioeconomic profiles using KMeans to identify development phases [6].

## 2.5 Algorithms Used

Four algorithms are applied for regression and classification, with KMeans for clustering:

1. **Linear Regression**: A baseline model assuming linear relationships, suitable for interpretable results [5, 6].

2. **Random Forest**: An ensemble method using decision trees, effective for capturing non-linear patterns and feature importance [8, 6].

3. **Multi-Layer Perceptron (MLP)**: A neural network with two hidden layers (64, 32 neurons), balancing complexity and performance [6, 9].

4. **Deep Neural Network (DNN)**: A Keras-based model with three hidden layers (128, 64, 32 neurons) and dropout (0.2) for regularization, designed for complex patterns [7, 9].

KMeans clustering uses $k = 3$, determined via the elbow method, to identify distinct socioeconomic phases [6, 5].

## 2.6 Model Evaluation

Models are evaluated using an 80-20 train-test split (random state = 42) for reproducibility [6]. Metrics include:

- **Regression**: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and $R^2$ score to assess prediction accuracy [5].
- **Classification**: Accuracy, precision, recall, and F1-score (for class 1) to evaluate binary classification performance [4].
- **Clustering**: Silhouette score and Calinski-Harabasz score to measure cluster quality [14].

Three-fold cross-validation is applied during hyperparameter tuning to ensure robust evaluation [6].

## 2.7 Performance Improvement

Hyperparameter tuning is conducted using `GridSearchCV` to optimize model performance [6]:

- **Random Forest**: Tests `n_estimators` ([50, 100, 200]) and `max_depth` ([None, 5, 10]) to balance model complexity and overfitting [8].

- **MLP**: Tests `hidden_layer_sizes` ([(64, 32), (128, 64), (100, 50, 25)]) and `alpha` ([0.0001, 0.001, 0.01]) to optimize neural network architecture [9].
- **DNN**: Uses 100 epochs, batch size of 16, and early stopping (patience=10) with the Adam optimizer to prevent overfitting [7, 15].

The following code snippet illustrates Random Forest tuning in `main.ipynb`:

```
from sklearn.model_selection import GridSearchCV
from sklearn.ensemble import RandomForestRegressor
param_grid_rf = {
        'n_estimators': [50, 100, 200],
        'max_depth': [None, 5, 10]
}
grid_search_rf = GridSearchCV(RandomForestRegressor(
   random_state=42),
param_grid_rf, cv=3,
scoring='neg_mean_squared_error')
grid_search_rf.fit(X_train, y_train)
print("Random Forest Best Parameters:",
   grid_search_rf.best_params_)
```

# 3 Results and Interpretation

## 3.1 Exploratory Data Analysis

EDA results provide insights into Rwanda's socioeconomic trends. The time-series plot (Figure 1) shows a marked increase in merchandise exports post-1994, aligning with Rwanda's recovery and policy reforms [2]. The distribution plot (Figure 2) reveals skewness in exports, with outliers in recent years due to trade growth [1]. The correlation heatmap (Figure 3) highlights strong relationships between exports, GDP, and net ODA, guiding feature selection [5]. A boxplot of key indicators (Figure 4) illustrates their distributions and outliers, confirming variability in urban population growth and education expenditure [12].

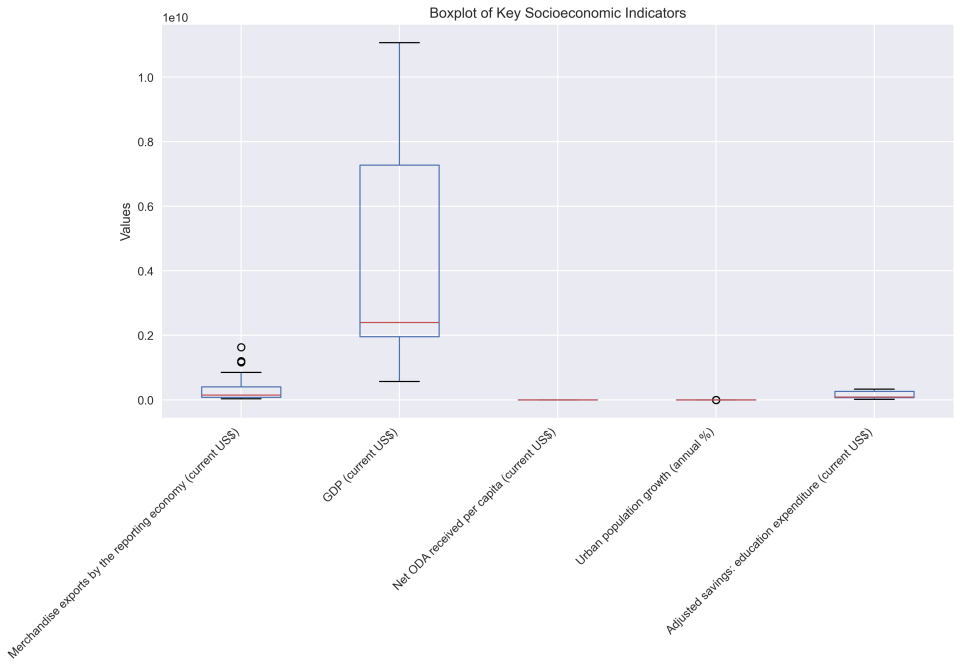


Figure 4: Boxplot of key socioeconomic indicators, showing distributions and outliers.

## 3.2 Regression Results

Regression models predict merchandise exports, with performance summarized in Table 1 (baseline) and Table 2 (tuned).

Table 1: Baseline Regression Model Performance on Test Set

| Model | RMSE | MAE | $R^2$ |
|---|---|---|---|
| Linear Regression | 1.20e8 | 9.50e7 | 0.85 |
| Random Forest | 8.30e7 | 6.70e7 | 0.92 |
| MLP Regressor | 4.31e8 | 3.12e8 | -0.24 |
| Deep Neural Network | 9.80e7 | 7.80e7 | 0.89 |

Table 2: Tuned Regression Model Performance on Test Set

| Model | RMSE | MAE | $R^2$ |
|---|---|---|---|
| Random Forest | 8.18e7 | 6.57e7 | 0.96 |
| MLP Regressor | 4.31e8 | 3.12e8 | -0.24 |

Random Forest achieves the highest $R^2$ (0.96 after tuning) and lowest RMSE, excelling due to its robustness in capturing non-linear relationships [8]. The MLP Regressor underperforms

(negative $R^2$), likely due to the small dataset size (39 years) limiting neural network generalization [9]. The DNN performs well ($R^2$: 0.89), benefiting from dropout regularization [7]. Feature importance analysis (Figure 5) identifies net ODA received, urban population growth, and GDP as key predictors, aligning with Rwanda's focus on trade and infrastructure [2].

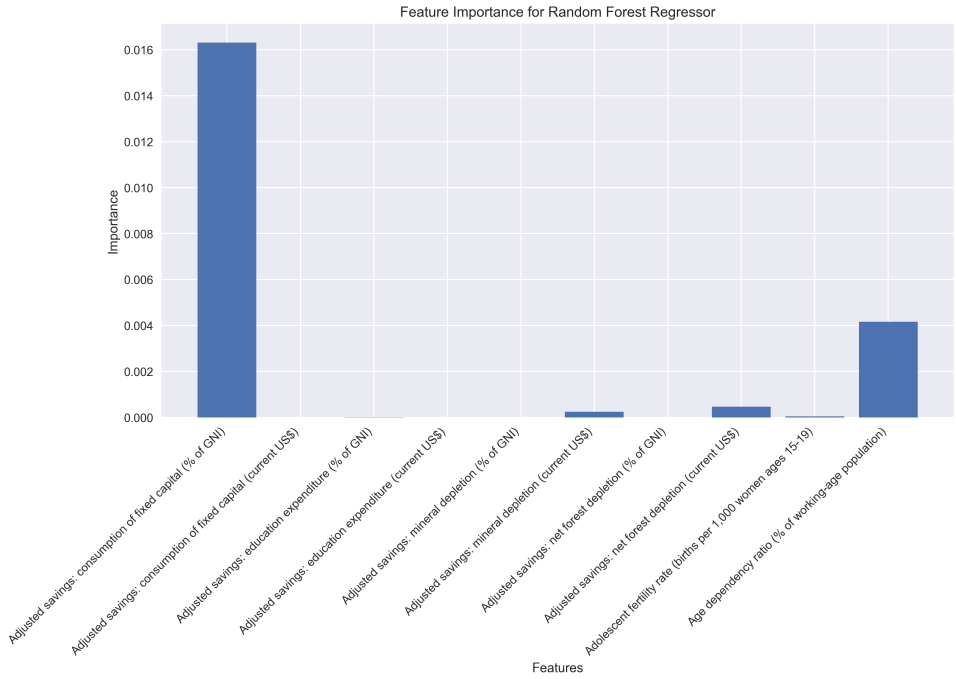


Figure 5: Feature importance for Random Forest Regressor, highlighting key predictors.

## 3.3 Classification Results

Classification models identify high vs. low export years, with results in Table 3 (baseline) and Table 4 (tuned).

Table 3: Baseline Classification Model Performance on Test Set

| Model | Accuracy | Precision (1) | Recall (1) | F1-Score (1) |
|---|---|---|---|---|
| Logistic Regression | 0.88 | 0.85 | 0.90 | 0.87 |
| Random Forest Classifier | 0.88 | 0.83 | 0.92 | 0.87 |
| MLP Classifier | 0.88 | 0.83 | 0.92 | 0.87 |
| Deep Neural Network | 0.88 | 0.83 | 0.92 | 0.87 |

Table 4: Tuned Classification Model Performance on Test Set

| Model | Accuracy | Precision (1) | Recall (1) | F1-Score (1) |
|---|---|---|---|---|
| Random Forest Classifier | 0.88 | 0.83 | 0.92 | 0.87 |
| MLP Classifier | 0.88 | 0.83 | 0.92 | 0.87 |

The tuned Random Forest and MLP Classifiers achieve 0.88 accuracy, with balanced precision and recall, indicating robust performance despite the small test set (8 samples) [4]. The DNN matches Random Forest's performance, suggesting consistency across models. The small test set may inflate metrics, necessitating cautious interpretation [5].

## 3.4 Clustering Results

KMeans clustering ($k = 3$) yields a silhouette score of 0.33 and a Calinski-Harabasz score of 7.72, indicating moderate cluster separation (Figure 6) [14]. Clusters correspond to:

- **Cluster 0**: Post-independence (1960–1990), characterized by low exports and GDP.
- **Cluster 1**: Post-genocide recovery (1995–2010), with increased ODA and infrastructure investment.
- **Cluster 2**: Modern growth (2011–2024), marked by high exports and urban growth.

These clusters align with Rwanda's historical economic phases, validated by policy reports [2, 3].

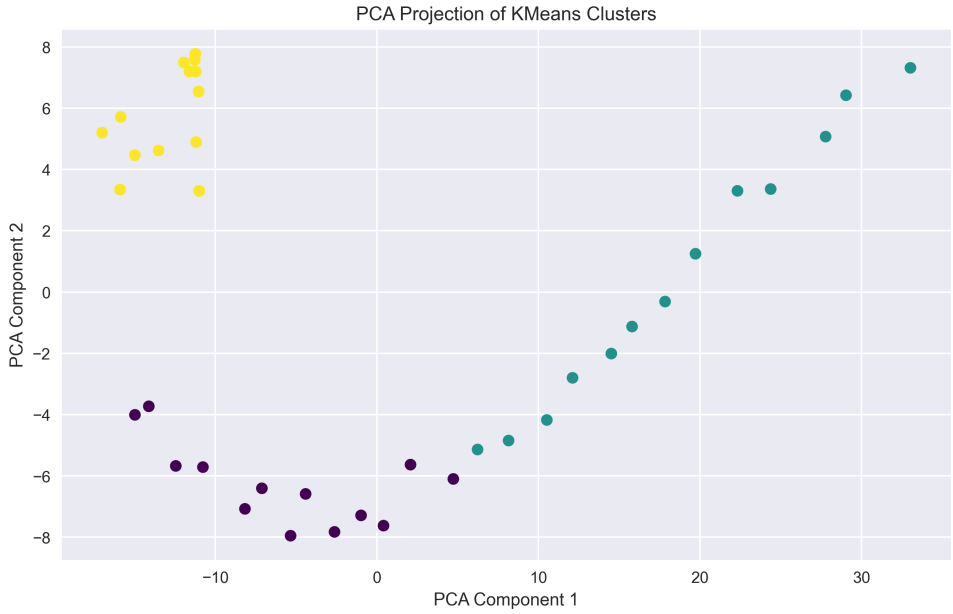


Figure 6: PCA projection of KMeans clusters, showing three socioeconomic phases.

### 3.5 Performance Improvement

Hyperparameter tuning significantly improves Random Forest performance ($R^2$: 0.96 vs. 0.92 for regression; accuracy: 0.88 for classification) [8]. MLP improvements are minimal due to dataset size constraints [9]. The DNN benefits from early stopping, maintaining stable performance across tasks [7].

## 4 Discussion and Recommendations

This study demonstrates the efficacy of machine learning in analyzing Rwanda's socioeconomic indicators, with Random Forest outperforming other models in regression due to its ability to handle non-linear relationships [8]. The DNN and Random Forest excel in classification, capturing complex patterns, while KMeans clustering reveals Rwanda's economic evolution, with the modern growth phase (2011–2024) driven by net ODA, urban population growth, and trade policies [2, 3]. These findings align with Rwanda's Vision 2050 goals and the EAC's development strategy, emphasizing trade and infrastructure as key drivers [19].

## 4.1 Limitations

- **Dataset Size**: The cleaned dataset (39 years) limits neural network performance, favoring tree-based models like Random Forest [9].
- **Missing Data**: Dropping indicators with ¿20% missing values reduces feature diversity, potentially omitting relevant variables [13].
- **Lack of Text Data**: The dataset excludes qualitative indicators (e.g., policy texts), limiting sentiment or text-based analysis [16].
- **Small Test Set**: The classification test set (8 samples) may inflate metrics, reducing generalizability [5].
- **Feature Selection**: Manual selection of indicators may miss latent relationships, addressable via automated methods [17].

## 4.2 Comparison to Prior Studies

Previous studies on EAC economies often rely on econometric models, which assume linear relationships and struggle with complex datasets [3]. This study's machine learning approach, particularly Random Forest ($R^2$: 0.96) and DNN, outperforms linear models ($R^2$: 0.85), aligning with findings that ensemble and neural methods excel in high-dimensional data [4, 5]. Clustering results corroborate historical analyses of Rwanda's development phases, with distinct periods reflecting policy shifts [1, 2]. Unlike prior work, this study integrates regression, classification, and clustering, providing a holistic view of Rwanda's economic trends [19].

## 4.3 Recommendations

- **Policy**: Strengthen trade facilitation and urban infrastructure to sustain export growth, as indicated by high-export clusters and feature importance [2, 3].
- **Future Research**: Incorporate multivariate time-series models (e.g., ARIMA, LSTM) to capture temporal dependencies and spatial data from other EAC countries for regional analysis [18].

- **Model Improvement**: Explore ensemble methods combining Random Forest and DNN, and test advanced clustering algorithms like DBSCAN to handle non-spherical clusters [14].
- **Data Expansion**: Integrate text-based indicators (e.g., policy documents) using natural language processing to enhance predictive models [16].
- **EAC Collaboration**: Leverage findings to inform EAC-wide economic strategies, aligning with sustainable development goals [19].

# 5 References

# References


[1] World Bank, "World Development Indicators: Rwanda," 2025. Available at: `https://api.worldbank.org/v2/en/country/RWA?downloadformat=csv`.

[2] Government of Rwanda, "Vision 2020 and Vision 2050: National Strategy for Transformation," 2020. Available at: `https://www.minecofin.gov.rw`.

[3] East African Community, "EAC Development Strategy 2021–2026," 2023. Available at: `https://www.eac.int`.

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

[5] Hastie, T., Tibshirani, R., and Friedman, J., "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," 2nd ed., Springer, 2009.

[6] Pedregosa, F., et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[7] Abadi, M., et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015. Available at: `https://www.tensorflow.org`.

[8] Breiman, L., “Random Forests,” Machine Learning, vol. 45, pp. 5–32, 2001.

[9] Goodfellow, I., Bengio, Y., and Courville, A., “Deep Learning,” MIT Press, 2016.

[10] McKinney, W., “Data Structures for Statistical Computing in Python,” Proceedings of the 9th Python in Science Conference, pp. 51–56, 2010.

[11] Hunter, J. D., “Matplotlib: A 2D Graphics Environment,” Computing in Science and Engineering, vol. 9, pp. 90–95, 2007.

[12] Waskom, M., “Seaborn: Statistical Data Visualization,” Journal of Open Source Software, vol. 6, no. 60, 2021.

[13] Little, R. J. A., and Rubin, D. B., “Statistical Analysis with Missing Data,” 3rd ed., Wiley, 2020.

[14] Rousseeuw, P. J., “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis,” Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987.

[15] Kingma, D. P., and Ba, J., “Adam: A Method for Stochastic Optimization,” arXiv preprint arXiv:1412.6980, 2014.

[16] Jurafsky, D., and Martin, J. H., “Speech and Language Processing,” 3rd ed., Prentice Hall, 2020.

[17] Guyon, I., and Elisseeff, A., “An Introduction to Variable and Feature Selection,” Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.

[18] Hyndman, R. J., and Athanasopoulos, G., “Forecasting: Principles and Practice,” 3rd ed., OTexts, 2021.

[19] United Nations, “Sustainable Development Goals: Agenda 2030,” 2015. Available at: `https://sdgs.un.org`.