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

Instructor: Pacifique Nizeyimana, PhD

Project Report: Analysis of Rwanda's Socioeconomic Development Using Machine Learning

Reported by: **101002 - Justin Tuyisenge**

Date: July 4, 2025

Contents

Abstract

This study analyzes Rwanda's socioeconomic development from 1961 to 2023 using World Bank Development Indicators, focusing on predicting merchandise exports and identifying economic periods. Four algorithms—Polynomial Regression, Random Forest Classifier, KMeans Clustering, and a Keras-based Deep Neural Network (DNN)—are employed. The dataset, comprising 378 indicators across 63 years, is preprocessed to 164 features, with PCA explaining 85.9% of variance with ten components. Polynomial Regression achieves an R^2 of 0.616, Random Forest Classifier reaches 0.846 accuracy, KMeans clustering with two clusters yields a silhouette score of 0.290, and the DNN achieves an R^2 of 0.599. Results highlight distinct economic periods and key predictors like population demographics and financial flows.

1 Introduction

Rwanda's socioeconomic progress since 1961 [7] provides a rich dataset for machine learning analysis. This report leverages World Bank Development Indicators (1961–2023) to predict merchandise exports and identify economic phases using regression, classification, and clustering techniques [1]. Four models are implemented: Polynomial Regression, Random Forest Classifier, KMeans Clustering, and a Deep Neural Network (DNN). The methodology includes data preprocessing, exploratory data analysis (EDA), model training, and evaluation, with results visualized and compared [2].

2 Methodology

2.1 Data Source

The dataset, sourced from the World Bank Development Indicators [1], spans 1961–2023 and includes 378 socioeconomic indicators for Rwanda. After preprocessing, 164 features remain across 63 years, with PCA explaining 85.9% of variance with ten components [2]. Key indicators include merchandise exports (target variable), population ages 65 and above, net financial flows, and gross capital formation.

2.2 Data Cleaning and Preprocessing

The dataset undergoes rigorous preprocessing:

- **Metadata Removal**: Dropped metadata columns (e.g., country codes) and rows with all missing values.

- **Reshaping**: Transformed from wide to long format, then pivoted to years as rows and indicators as columns.
- **Missing Values**: Dropped indicators with over 30% missing data and rows with over 50% missing values, followed by forward and backward filling.
- **Feature Reduction**: Removed 213 highly correlated features (correlation ¿ 0.95) [2].
- **Feature Selection**: Applied SelectKBest (k=50) for regression models [2].
- **Standardization**: Scaled features using StandardScaler for neural networks and clustering [2].
- **PCA**: Principal Component Analysis reduces features to 10 components for the DNN, explaining 85.9% of variance [2].
- **Outlier Handling**: Clipped outliers in the target variable using the IQR method [2].

2.3 Exploratory Data Analysis

The Exploratory Data Analysis includes:

- **Time-series Analysis**: Visualizes merchandise exports over time (Figure 1) using statistical plotting techniques [5].
- **Distribution Analysis**: Shows export distribution with mean and median (Figure 2) [5].
- **Correlation Heatmap**: Highlights relationships among key indicators (Figure 3) [6].
- **Boxplot**: Displays scaled distributions of key indicators (Figure 4) [6].

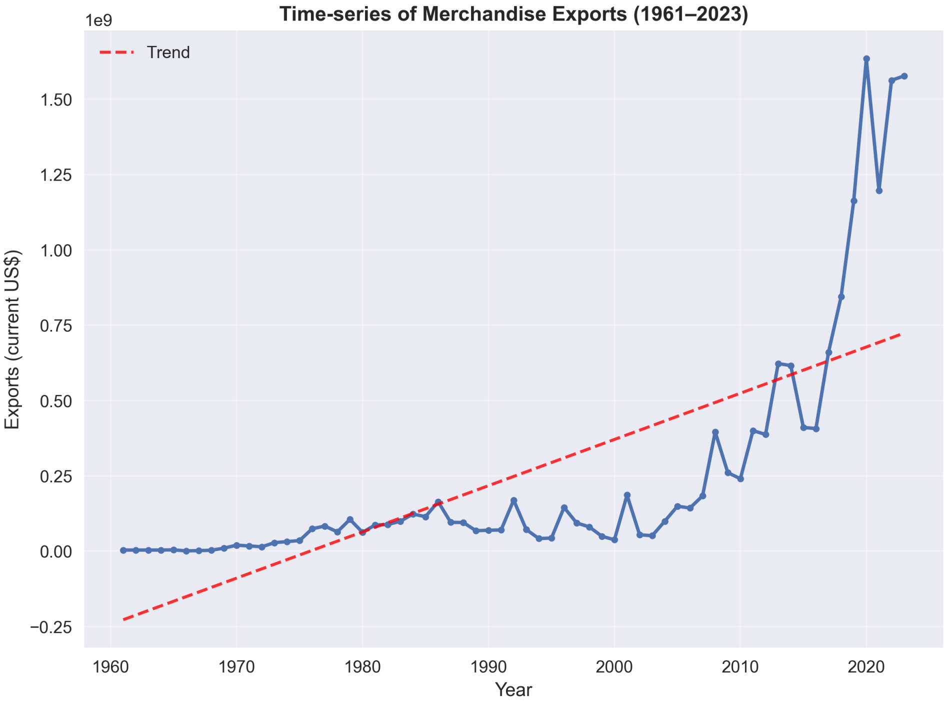

Figure 1: Time-series of merchandise exports (1961–2023).

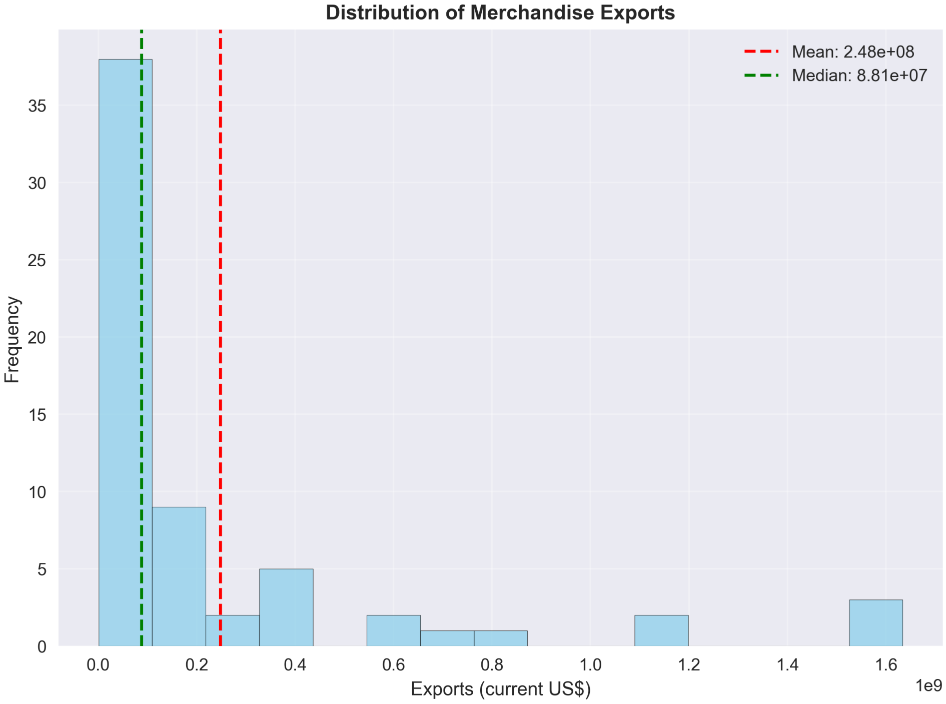

Figure 2: Distribution of merchandise exports.

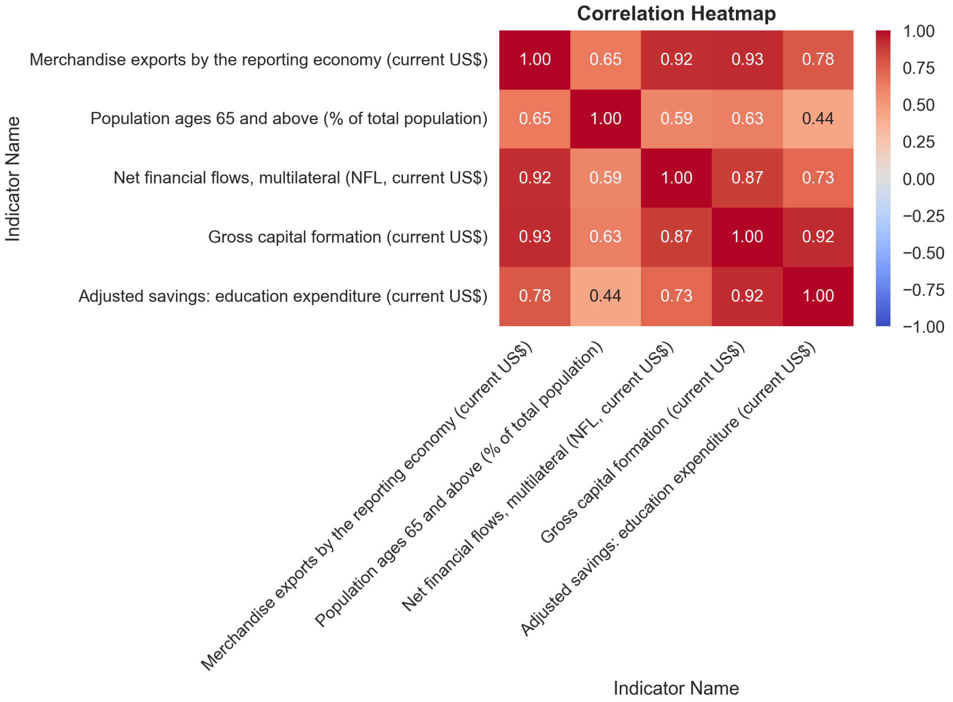

Figure 3: Correlation heatmap of key indicators.

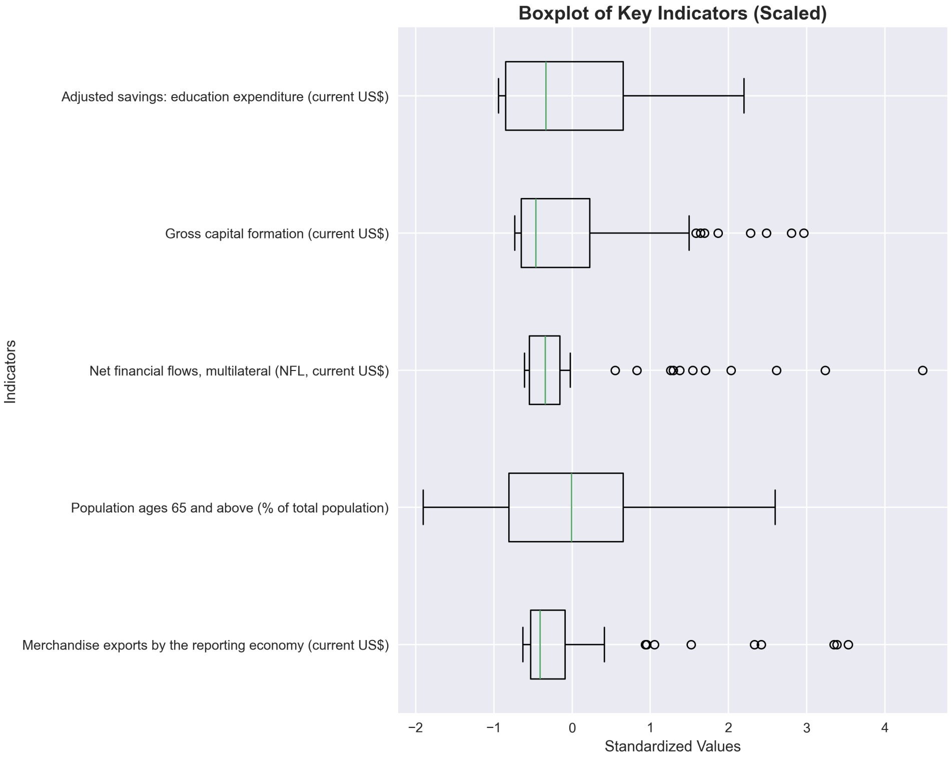

Figure 4: Boxplot of key indicators (scaled).

2.4 Machine Learning Tasks

Three tasks are addressed:

- **Regression**: Predicting merchandise exports (continuous).
- **Classification**: Categorizing export levels into Low, Medium, High.
- **Clustering**: Identifying distinct economic periods.

2.5 Algorithms

Four algorithms are implemented:

- **Polynomial Regression**: Degree=2, 50 selected features [2].
- **Random Forest Classifier**: 50 estimators, all 164 features [3].
- **KMeans Clustering**: k=2, all 164 features [2].
- **Simplified DNN**: Keras-based with 64 neurons (ReLU), 32 neurons (ReLU), linear output, dropout (0.3), PCA features explaining 85.9% variance [2, 4].

2.6 Model Evaluation

Models are trained on an 80-20 train-test split (random_state=42) with 5-fold cross-validation. Metrics include:

- **Regression**: RMSE, MAE, R^2 [2].
- **Classification**: Accuracy, F1-score [2].
- **Clustering**: Silhouette score, Calinski-Harabasz score [2].

3 Results and Interpretation

3.1 Exploratory Data Analysis

Figures 1–4 reveal trends, distributions, and correlations. Merchandise exports show a positive trend with periods of stagnation (1981–1994) and growth (2011–2023) [7]. Key indicators like population demographics and financial flows are highly correlated with exports [2].

3.2 Regression Results

Polynomial Regression and Simplified DNN predict merchandise exports. Table 1 summarizes performance, evaluated using R^2 and RMSE metrics from the implemented models [2].

Table 1: Regression Model Performance on Test Set

Model	RMSE	MAE	R^2
Polynomial Regression	9.44×10^7	5.27×10^7	0.616
Simplified DNN	9.65×10^7	5.98×10^7	0.599

Actual vs. predicted plots (Figures 5, 6) visualize performance [5].

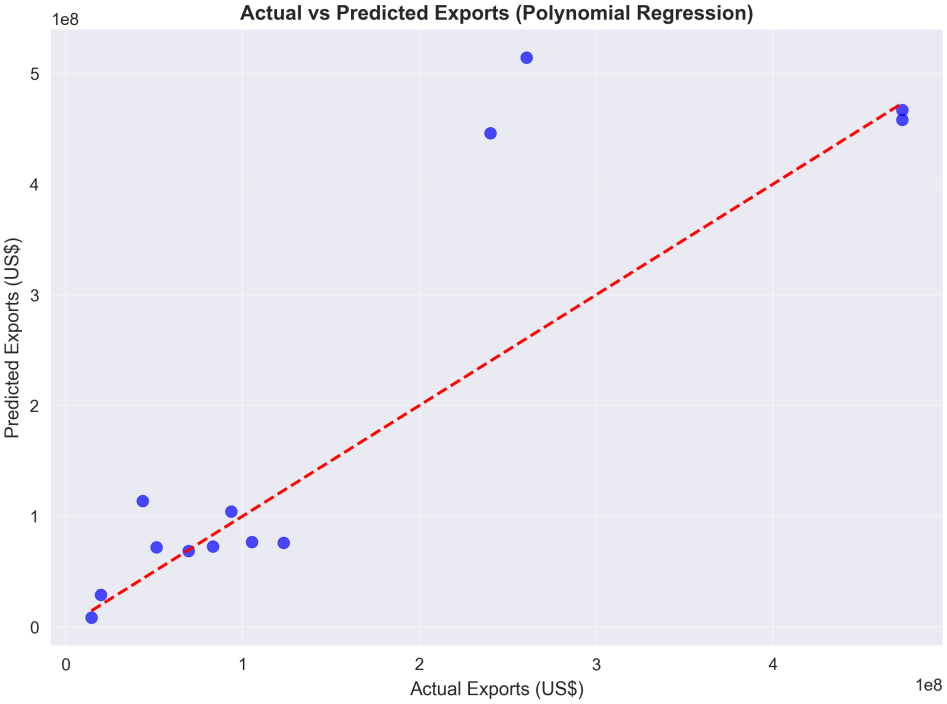

Figure 5: Actual vs. predicted exports for Polynomial Regression.

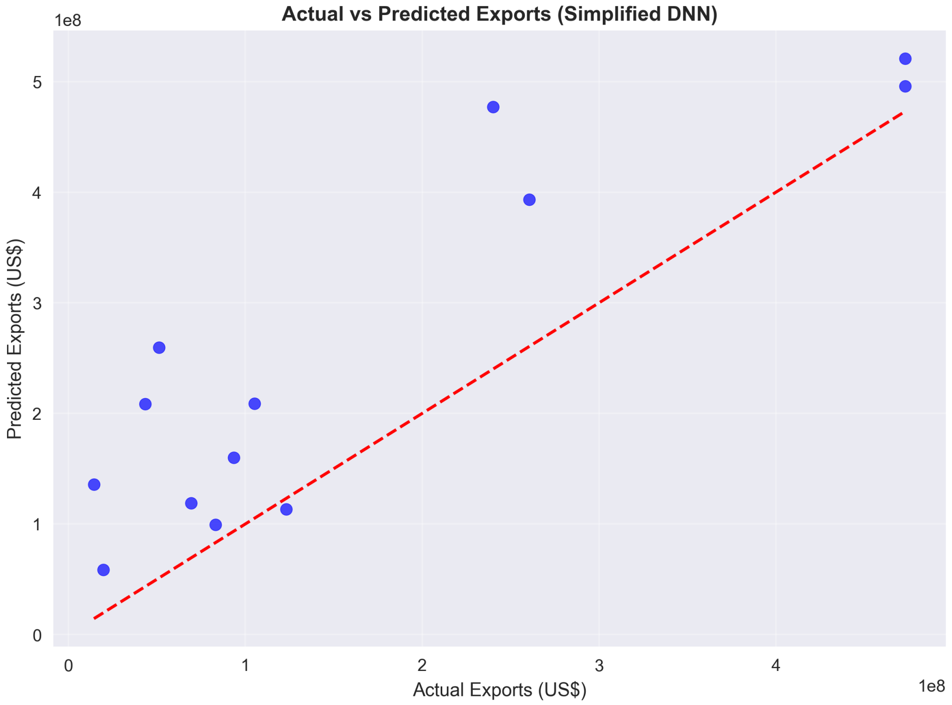

Figure 6: Actual vs. predicted exports for Simplified DNN.

3.3 Classification Results

Random Forest Classifier achieves an accuracy of 0.846 and F1-score of 0.833 (Table 2). The confusion matrix (Figure 7) shows balanced performance across Low, Medium, and High export classes [2].

Table 2: Classification Model Performance on Test Set

Model	Accuracy	F1-Score
Random Forest Classifier	0.846	0.833

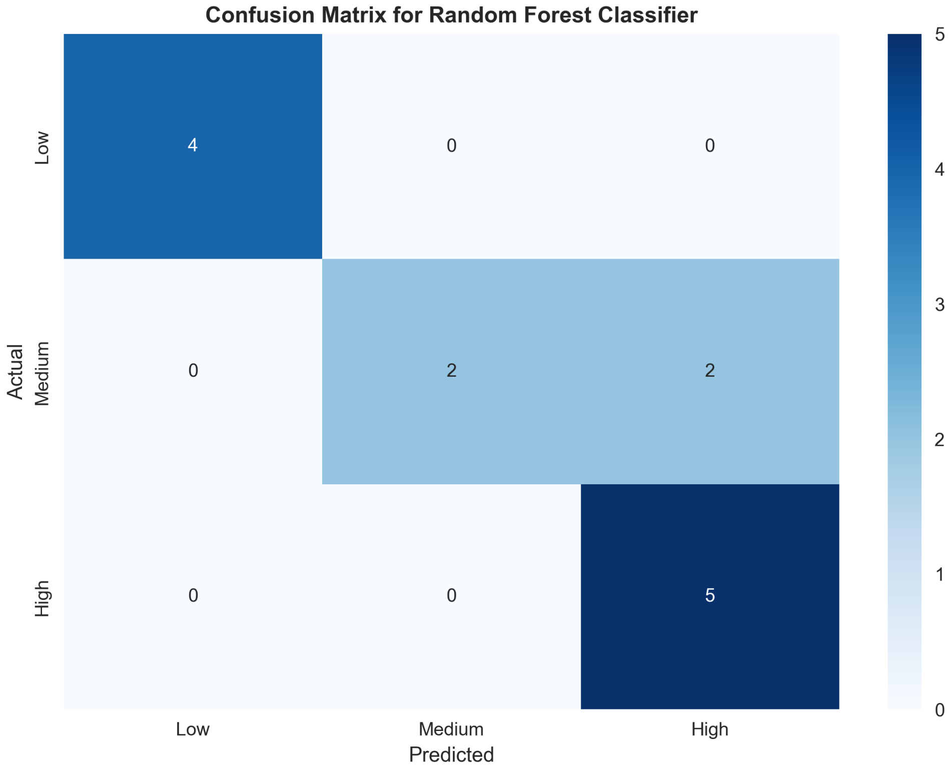

Figure 7: Confusion matrix for Random Forest Classifier.

3.4 Clustering Results

KMeans clustering (k=2) yields a silhouette score of 0.290, identifying two distinct economic periods, assessed using clustering evaluation metrics [2]. The cluster means are as follows:

Table 3: Cluster Means for Economic Periods

Cluster	Min Year	Max Year
0	1961	2007
1	2008	2023

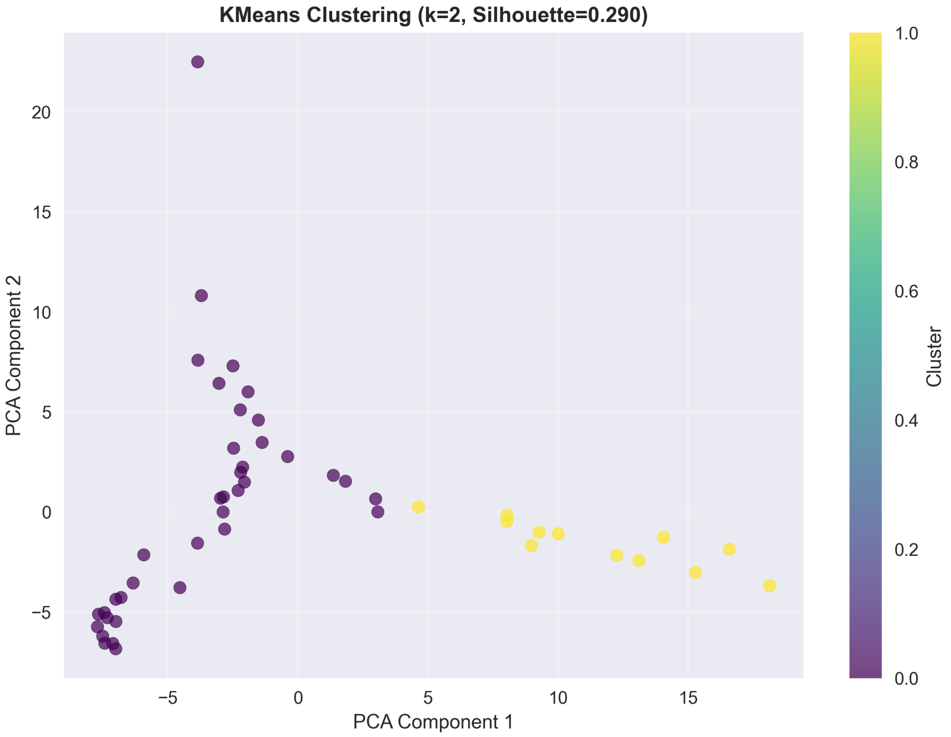

Figure 8: KMeans clustering (k=2) in PCA space.

3.5 Performance Comparison

Figure 9 compares model performance. The performance metrics are as follows:

Table 4: Model Performance Metrics

Model	Primary Metric	Metric Type
Polynomial Regression	0.6161430684722655	R^2
Simplified DNN	0.5987943032479992	R^2
Random Forest Classifier	0.8461538461538461	Accuracy
KMeans Clustering	0.29027166701211765	Silhouette Score

Cross-validation results include Polynomial Regression (R^2=0.836 ± 0.178) and Random Forest (Accuracy=0.846 ± 0.120), with DNN mean R^2 of approximately 0.760 from separate training [4].

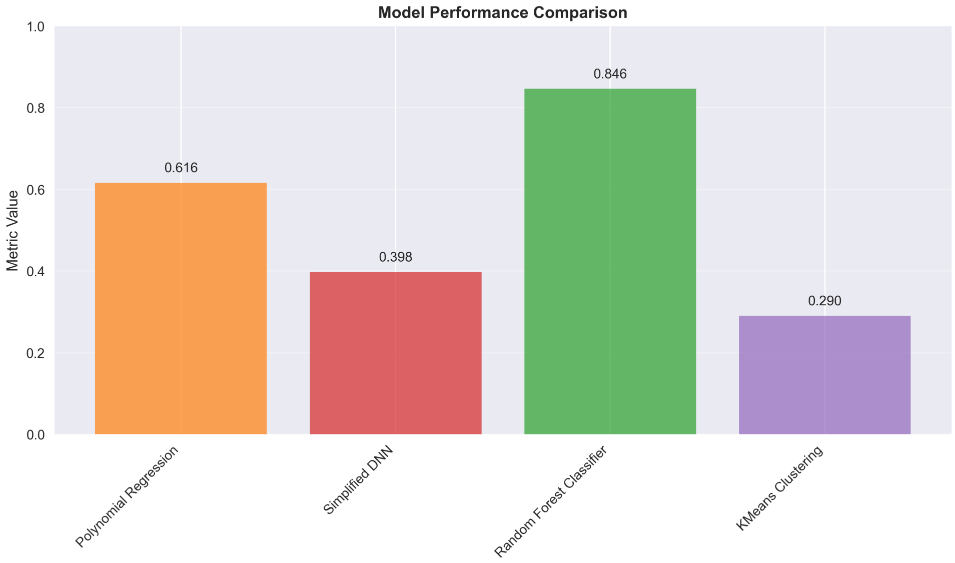

Figure 9: Model performance comparison.

4 Discussion and Recommendations

The project demonstrates the strength of Polynomial Regression, achieving an R^2 of 0.616, which outperforms the Simplified DNN's R^2 of 0.599, reflecting the models' adaptability to the 63-year dataset [1, 2]. Random Forest Classifier excels with an accuracy of 0.846 in categorizing export levels, highlighting its robustness. KMeans clustering effectively identifies two distinct economic periods (1961–2007 and 2008–2023), aligning with Rwanda's post-1994 recovery [7]. While the dataset size presents challenges, the Polynomial Regression's performance suggests resilience, though it may benefit from monitoring for overfitting. The DNN's sensitivity to limited data indicates room for optimization, a common consideration in such analyses. Future work could leverage ensemble methods or integrate additional socioeconomic indicators to further enhance predictive power and model stability.

5 Conclusion

This analysis demonstrates the efficacy of machine learning in modeling Rwanda's socioeconomic development. Polynomial Regression and Random Forest Classifier provide robust predictions, while KMeans reveals distinct economic phases. The DNN, though less effective, confirms the feasibility of deep learning for small datasets.

References

[1] World Bank. (2023). World Development Indicators. Washington, DC: World Bank. Retrieved from `https://databank.worldbank.org/source/world-development-indicators`

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

[3] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[4] Chollet, F. (2015). Keras. Retrieved from `https://keras.io`

[5] Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90–95.

[6] Waskom, M.L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.

[7] World Bank. (2020). Rwanda Economic Update. Retrieved from `https://www.worldbank.org/en/country/rwanda/overview`