Predicting GDP per capita Using Socioeconomic Indicators: A Machine Learning Approach

Student: Kaylum Smith, 710044098

November 2024

1 Introduction

This project investigates the "countries of the world" dataset from Kaggle [1] and whether GDP per capita can be predicted using regression models. Regression techniques were complemented with clustering model to help group features and enhance model performance. The core research question driving this analysis is: "Can GDP per capita be predicted using socioeconomic and environmental indicators?"

2 Methodology

2.1 Data Preprocessing: Numerical data was standardised to ensure consistency and compatibility with machine learning algorithms, albeit with some computational overhead. Missing values were fixed by using the regional means or averages to preserve geographical trends, while rows with excessive missing data were dropped, balancing data quality and dataset size. Outliers were removed using the Z-score method, reducing skewed predictions but potentially excluding valid extreme cases. Finally, logical inconsistencies, such as feature percentages not summing to 100% (e.g., "Arable," "Crops," "Other"), were corrected to ensure data integrity, though approximations introduced minor deviations from the original data.

2.2 Feature Engineering: Feature engineering began with Variance Inflation Factor (VIF) analysis to remove features with high multicollinearity, improving robustness and interpretability. This process eliminated features with minimal impact on the model (low VIF scores) but risked discarding potentially valuable non-linear relationships. Graphs in Figure 1.e and 1.f, were used to identify similar patterns. Features were grouped to reduce redundancy and multicollinearity while retaining their collective impact on GDP. Hyperparameter tuning was used, especially for the Random Forest,through functions like grid search to optimize parameters. This help make adjustments to parameters like the number of trees and maximum depth. This highlighted a trade-off between computational cost and model performance.

2.3 Critical Reflection: Overall, the preprocessing and feature engineering steps involved careful trial and error of trade-offs between data quality and model interpretability. Such as retaining rows with missing values and implementing an average versus dropping them balanced the preservation of information against potential model bias from incomplete data. The decision to use Z-scores for outlier removal was effective for improving model predictions but potentially removed valid outlier information. Similarly, grouping features reduced complexity and redundancy but required assumptions about feature relationships that might oversimplify underlying patterns. These reflections underline the balance between optimising for performance and acknowledging the limitations of the chosen approaches.

3 Machine Learning Models

3.1 Linear regression: Selected as a baseline for its simplicity and interpretability, it provided an initial assessment of whether the relationship between predictors and GDP is predominantly lin-

ear. The model was trained using 90% of the data for testing (test_size=0.1) with a random seed for reproducibility (random_state=42). Feature scaling was applied using StandardScaler, and 6-fold cross-validation (cv=6) evaluated performance using the R^2 metric ('scoring='r2'). While Linear Regression stabilised at a training score of $R^2 = 0.7$, it initially overfitted small datasets. Its inability to capture non-linear relationships limited predictive power (Figure 1.a, Table 1).

3.2 Ridge regression: Addressed the limitations of Linear Regression by introducing regularisation to mitigate overfitting and multicollinearity. Hyperparameter tuning using GridSearchCV optimised the alpha parameter over the grid [0.01, 0.1, 1, 10, 100]. This model achieved a training score of 0.85 and a test score improving from 0.2 to 0.65 as the training size increased. Although it improved generalisation, Ridge Regression still struggled with non-linear relationships (Figure 1.b, Table 1).

3.3 Random forest regression: The best-performing model, it captured complex non-linear interactions and adapted to feature diversity. The dataset was split into 80% training and 20% testing (test_size=0.2), using the top 10 features identified through feature importance analysis with a preliminary Random Forest model. Hyperparameter tuning adjusted the number of trees (n_estimators=200), maximum depth (max_depth=10), minimum samples per leaf that was used is (min_samples_leaf=1), and minimum samples required to split a node (min_samples_split=5). The model achieved a consistently high training score (0.95) and test score (0.8878), demonstrating minimal overfitting and strong generalisation (Figure 1.c,Table 1).

3.4 Clustering: To better enhance the regression analysis, clustering model was put in place to help group related features. This would reduce multicollinearity and simplify relationships. This approach ended up improving feature interpretability and revealed underlying structures, such as industry or economic trends, that enhanced the analysis. For Random Forest, clustering refined feature selection by highlighting influential groupings, improving R^2 scores and reducing prediction errors (Figure 1.d,Table 1).

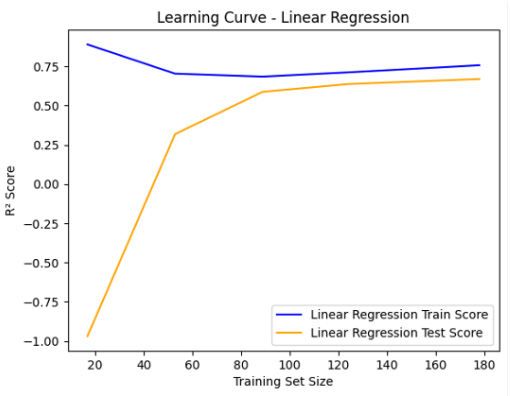
3.5 Evaluation: Model performance was assessed using R^2 and Mean Squared Error (MSE). R^2 quantified the proportion of variance explained, while MSE penalised large errors. The Linear Regression model provided a decent baseline, while Ridge Regression improved generalisation, and Random Forest outperformed both, effectively capturing complex relationships and achieving the most accurate results (Table 1).

4 Conclusion

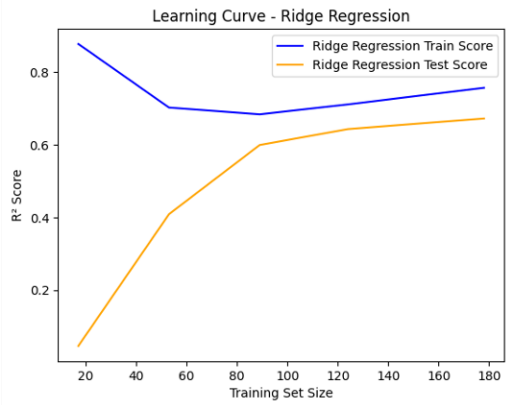
Overall, the best model by far was the Random Forest Regression model. The model outperformed both Linear and Ridge models in accuracy and robustness, making it the most effective method for predicting GDP. For future work, an important issue that could be addressed is the implementation of methods like feature importance analysis to clarify the contributions of individual features. Additionally, increasing the dataset's size and diversity and exploring advanced boosting techniques could further improve performance and mitigate interpretability challenges.
Source of the data: Kaggle: https://www.kaggle.com/datasets/fernandol/countries-of-the-world/data

References

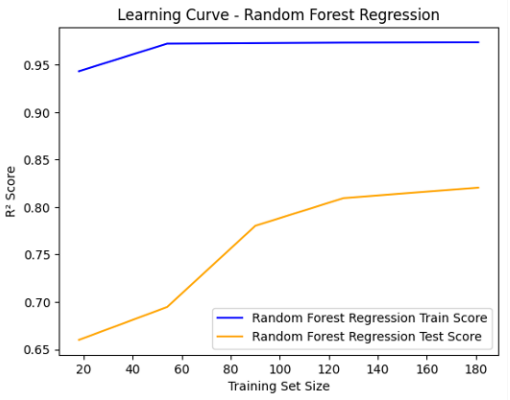
[1] Luis Fernando. Countries of the world dataset, 2023. Accessed: 21/11/2024.



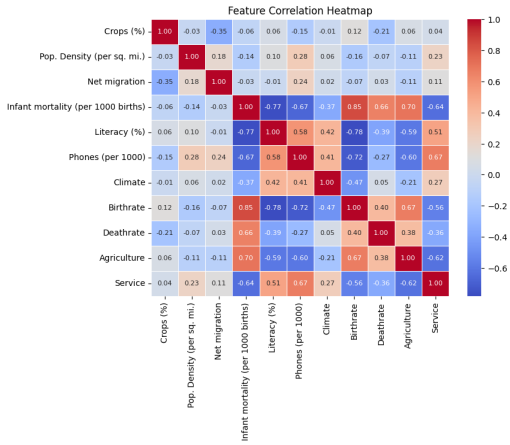
(a) Linear Regression



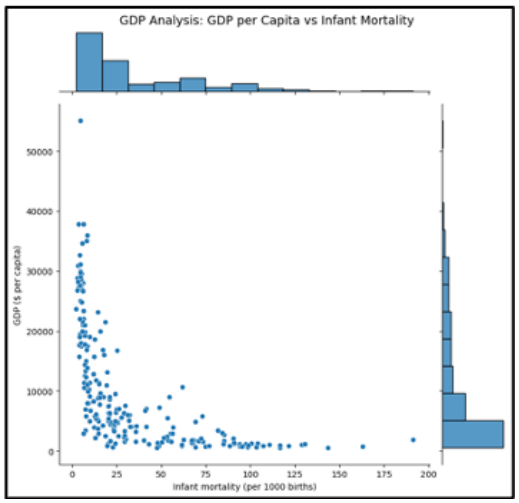
(b) Ridge Regression



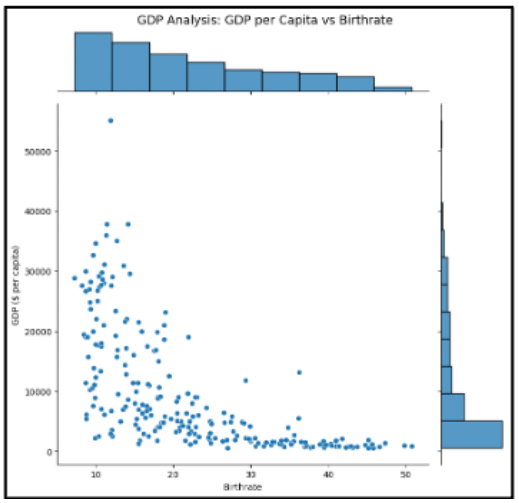
(c) Random Forest



(d) Clustering



(e) Random Forest



(f) Clustering

Figure 1: Learning Curves and Clustering Analysis

A Appendix

Metric	Linear Regression	Ridge Regression	Random Forest
R^2 Score	0.8488	0.8613	0.8878
Cross-Validated R^2	0.6941	0.7369	0.8174
Mean Squared Error (MSE)	9,891,394.68	9,074,738.24	2,022.52

Table 1: Regression Results Summary for Linear, Ridge, and Random Forest Models

Feature	VIF Value
Economic Activity	65.807275
Literacy (%)	57.628361
Climate	17.906210
Deathrate	9.495749
Birthrate Mortality	7.633062
Phones (per 1000)	4.865493
Climate Agriculture	4.136529
Coastline (coast/area ratio)	1.197011
Net Migration	1.177047

Table 2: Ranking of Feature Importance Based on Variance Inflation Factors (VIF)

Generative AI Statement

AI-supported/AI-integrated use is permitted in this assessment. I acknowledge the following uses of GenAI tools in this assessment:

- (YES / NO) I have used GenAI tools for developing ideas.
- (YES / NO) I have used GenAI tools to assist with research or gathering information.
- (YES / NO) I have used GenAI tools to help me understand key theories and concepts.
- (YES / NO) I have used GenAI tools to identify trends and themes as part of my data analysis.
- (YES / NO) I have used GenAI tools to suggest a plan or structure for my assessment.
- (YES / NO) I have used GenAI tools to give me feedback on a draft.
- (YES / NO) I have used GenAI tool to generate image, figures or diagrams.
- (YES / NO) I have used GenAI tools to proofread and correct grammar or spelling errors.
- (YES / NO) I have used GenAI tools to generate citations or references.
- (YES / NO) Other: [please specify]
- I have not used any GenAI tools in preparing this assessment.

I declare that I have referenced use of GenAI outputs within my assessment in line with the University referencing guidelines.