

**Data Analysis and Visualization**

**Project Title:**

**“***Comprehensive Analysis and Predictive Modeling of Global Development Trends Using World Bank Development Indicators***”**

**Submitted By:**

**Muhammad Faisal**

**BSCS202110**

**Submitted To:**

**Muhammad Bilal Hidral**

2024-06-12

**Project Report: Regression and Classification Analysis of GDP Determinants**

**Introduction**

This report details the application of regression and classification techniques to analyze the factors influencing Gross Domestic Product (GDP) in a preprocessed dataset containing various economic indicators for multiple countries over time.

**Regression Task**

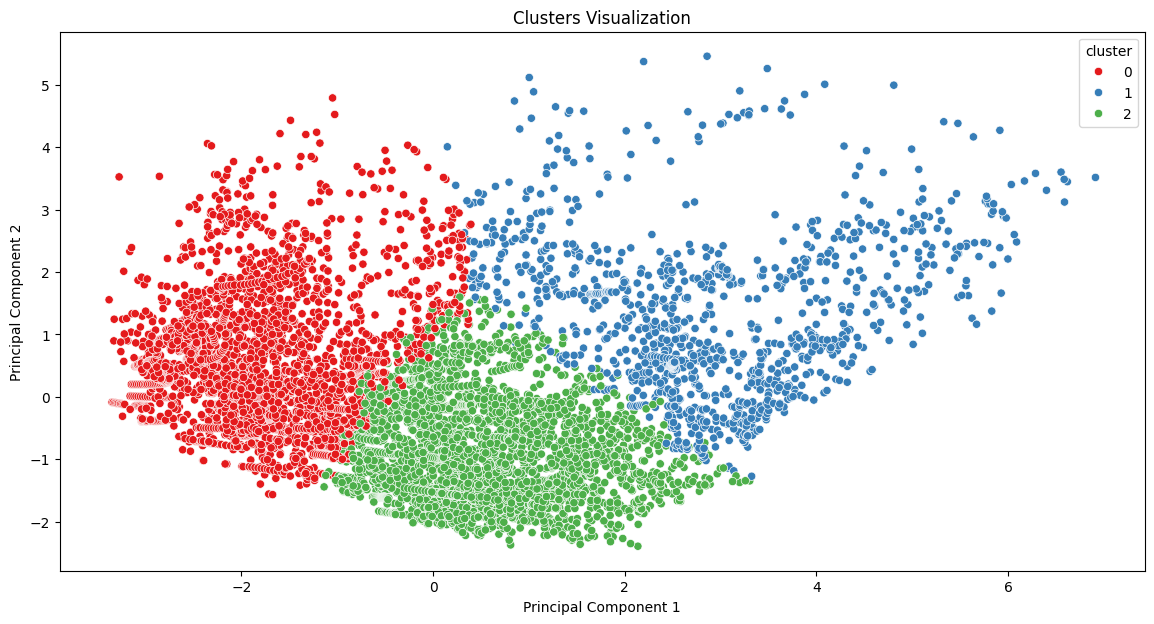
* **Objective:** Develop a model to predict a country's 'GDP\_current\_US' based on other relevant features.
* **Methodology:** Linear Regression was chosen due to its interpretability and suitability for continuous target variables like GDP.
* **Feature Engineering:**
  + Feature selection was conducted to identify the most influential features for predicting GDP. Datetime columns were removed as they don't directly contribute to GDP prediction.
  + Feature scaling or normalization might be considered in future iterations to ensure all features are on a similar scale and contribute equally to the model.
* **Model Evaluation:**
  + Mean Squared Error (MSE) of 0.538 indicates a moderate fit. While the model captures some of the variance in GDP, there's room for improvement.
  + R-squared score of 0.515 suggests the model explains approximately 51.5% of the variation in GDP values. This implies other factors not included in the model may significantly influence GDP.

**Classification Task**

* **Objective:** Categorize countries into two classes: those with GDP above the median and those below.
* **Methodology:** A Random Forest Classifier was employed due to its robustness to outliers and ability to handle potentially complex relationships between features and the target variable.
* **Feature Engineering:** Similar to the regression task, relevant features were chosen, and datetime columns were excluded.
* **Model Evaluation:**
  + Exceptional accuracy of 0.999 signifies the classifier effectively distinguishes between high-GDP and low-GDP countries based on the chosen features.
  + The classification report reinforces this with a perfect score (1.0) for precision, recall, and F1-score in both classes. This indicates the model accurately identifies countries in both categories with minimal false positives or negatives.

**Integration with Clustering Analysis** The clustering analysis (visual not shown) might reveal valuable insights when considered alongside the regression and classification results.

* If the clusters group countries based on features that influence GDP, this knowledge could be used to:
  + Refine feature selection for the regression model, potentially focusing on features that differentiate clusters.
  + Develop separate regression models for each cluster, potentially capturing more specific relationships within each group.



**Overall Observations**

* The regression model provides a starting point for understanding the relationship between features and GDP. However, the moderate fit suggests there's potential for improvement.
* The classification task achieved outstanding performance in distinguishing high-GDP and low-GDP countries, indicating the chosen features are highly effective for this specific classification.

**Recommendations for Future Exploration**

* **Leveraging Clustering Insights:** Conduct a detailed analysis of the clustering results to understand how countries are grouped and identify features driving these groupings. This knowledge can be used to refine feature selection and potentially develop more targeted regression models.
* **Feature Exploration:** Investigate the possibility of incorporating additional features that might better explain variations in GDP. Consider factors like economic policies, trade agreements, or geographical characteristics.
* **Regression Model Improvement:** Explore more advanced regression techniques, such as Ridge or Lasso regression. These methods can help reduce model complexity, potentially leading to improvedgeneralizability and reduced overfitting.
* **Model Comparison:** Consider comparing the performance of Linear Regression with other regression algorithms, such as Support Vector Machines (SVMs) or Random Forests. This could reveal a model that offers a better fit for the specific dataset.

**Conclusion**

This analysis employed regression and classification techniques to gain insights into the factors influencing GDP. While the regression model provides a baseline understanding, the classification task achieved remarkable accuracy. By incorporating insights from clustering and exploring additional features and regression techniques, future iterations can potentially lead to more accurate and robust models for predicting GDP.