Principal Component Analysis (PCA) and K-Means Clustering of Greenhouse Gas Emissions by Country

Introduction

Greenhouse gas (GHG) emissions are a critical driver of climate change, and understanding their distribution across countries is essential for developing effective mitigation strategies. This analysis employs Principal Component Analysis (PCA) and K-Means clustering to categorize countries based on their GHG emissions and key economic indicators, including GDP, population, industrial value, and regulatory quality for 2023. The study leverages data from the Emissions Database for Global Atmospheric Research (EDGAR) and the World Bank Development Indicators, providing a comprehensive dataset for analysis.

Methodology

The study follows a structured approach to analyze and cluster countries based on their GHG emissions and economic indicators:

1. **Data Collection and Preprocessing:**
   - The dataset was sourced from EDGAR and the World Bank Development Indicators.
   - Data cleaning and merging were performed using Excel, resulting in a consolidated dataset named `WORLDDATA`.
   - The dataset was imported into a Python environment using libraries such as ‘pandas’, ‘numpy’, and ‘scikit-learn’.
2. **Data Transformation:**
   - The data exhibited high skewness, necessitating a logarithmic transformation to stabilize variance.
   - Standardization was applied to ensure all features contributed equally to the analysis.
3. **Principal Component Analysis (PCA):**
   - PCA was employed to reduce dimensionality while retaining the maximum variance in the data.
   - The optimal number of principal components was determined using a scree plot, which indicated that three principal components captured 100% of the variance.
   - The loadings of the principal components were visualized using a heatmap, revealing the contribution of each original feature to the principal components.
4. **K-Means Clustering:**
   - K-Means clustering was applied to group countries based on the principal components.
   - The optimal number of clusters was determined using both the elbow method and silhouette score analysis, with the silhouette score suggesting three clusters as the optimal choice.
5. **Visualization:**
   - The clusters were visualized using 2D and 3D scatter plots, with country codes annotated for clarity.

- The 3D plot provided additional insights by incorporating the third principal component, highlighting the relationships between countries across all three dimensions.

Key Findings

- **Cluster 0 (Low Emissions, Low Activity Economies):** This cluster includes countries with low GHG emissions, low industrial activity, low GDP, and low regulatory quality. These countries are likely economically underdeveloped, with minimal industrial output and governance. Examples include Uganda and Luxembourg.
- **Cluster 1 (High Emission, Industrial Economies):** This cluster comprises countries with high GHG emissions, high GDP, and significant industrial activity. These countries are economically developed but contribute heavily to global emissions. Examples include Russia and China.
- **Cluster 2 (Low Emission, High Regulation Economies):** This cluster includes countries with low emissions and industrial activity but high regulatory quality. These nations prioritize environmental governance despite having smaller economies. Examples include the Maldives and Palau.

Discussion

The study successfully categorized countries into three distinct clusters based on their GHG emissions and economic indicators. The use of PCA allowed for effective dimensionality reduction, while K-Means clustering provided a clear grouping of countries. The findings highlight the diversity in emissions profiles and economic structures across countries, offering valuable insights for policymakers.

Future Work

Future Work: Future studies could explore the use of alternative clustering methods, such as DBSCAN or hierarchical clustering, to validate the findings. Additionally, incorporating sector-specific emissions data could reveal more nuanced patterns in emissions profiles.

Conclusion

This study demonstrates the effectiveness of PCA and K-Means clustering in categorizing countries based on their GHG emissions and economic indicators. The findings provide a foundation for targeted climate mitigation strategies, emphasizing the need for tailored approaches based on the unique characteristics of each cluster.

References

- Crippa, M., Guizzardi, D., Pagani, F., Banja, M., Muntean, M., Schaaf, E., Becker, W., Monforti-Ferrario, F., Quadrelli, R., Risquez Martin, A., Taghavi-Moharamli, P., Grassi, G., Rossi, S., Melo, J., Oom, D., Branco, A., San-Miguel, J., Manca, G., Pisoni, E., Vignati, E., Pekar, F. (2024). *GHG emissions of all world countries – JRC/IEA 2024 Report. Luxembourg. https://data.europa.eu/doi/10.2760/4002897