Student Name: Muhammad Furqan ul Hassan
Student ID :22075982

University of Hertfordshire UH

Title: Exploring Arable Land Dynamics: Clustering Nations and Predicting Growth with an Exponential Model

Introduction

Arable land is important, influencing food production, economic sustainability, and environmental health. In this study, an exponential growth model is employed to simulate the dynamic changes in arable land over time. The implemented code utilizes curve fitting techniques to analyze and predict the evolution of arable land, contributing to a comprehensive understanding of its trends and implications on various aspects.

Abstract

This study delves into the modeling of arable land expansion over time through the utilization of an exponential growth model. A comprehensive dataset was curated for in-depth analysis following meticulous data loading, cleaning, and imputation processes. Key features, including the percentage of arable land, annual freshwater withdrawals, adjusted net national income growth, the contribution of agriculture and fisheries to GDP, and forest area, were selected for further exploration. The derived Silhouette Score of 0.4270 signifies a robust grouping structure. The predictions for arable land percentage in 2025, 2030, and 2035 are 21.90, 23.22, and 24.62, respectively. Employing K-Means clustering on these parameters revealed distinct patterns across nations. The primary focus was to elucidate the relationship between the growth in adjusted net national income and the percentage of arable land.

Data Loading and Cleaning:

- The dataset was loaded into a Pandas Data Frame for further analysis.
- Non-numeric values represented by '..' were replaced with NaN to ensure data consistency and integrity.

Missing Data Handling:

- Simple imputation was employed to fill missing values in specific columns with the mean value, ensuring a complete dataset for analysis.

Feature Selection:

Several crucial columns were selected for in-depth investigation, including:

- Percentage of arable land.
- Yearly freshwater withdrawals.
- Adjusted net national income growth.
- GDP contribution from agriculture and fishery.
- Forest area.

Transposing Data for Analysis:

- The data was transposed for a more organized and insightful exploration of selected features.
- This comprehensive data pre-processing and cleaning procedure laid the foundation for subsequent analysis and modeling in the research.

K-Means Clustering

Initialization:

- Randomly select K initial cluster centroids: \(C = \{c_1, c_2, ..., c_K\} \).

Assignment:

- Assign each data point to the nearest cluster centroid, minimizing Euclidean distance.

Update:

- Recalculate centroids based on the average of data points in each cluster.

Repeat:

- Iteratively repeat assignment and update until convergence.

Analysis:

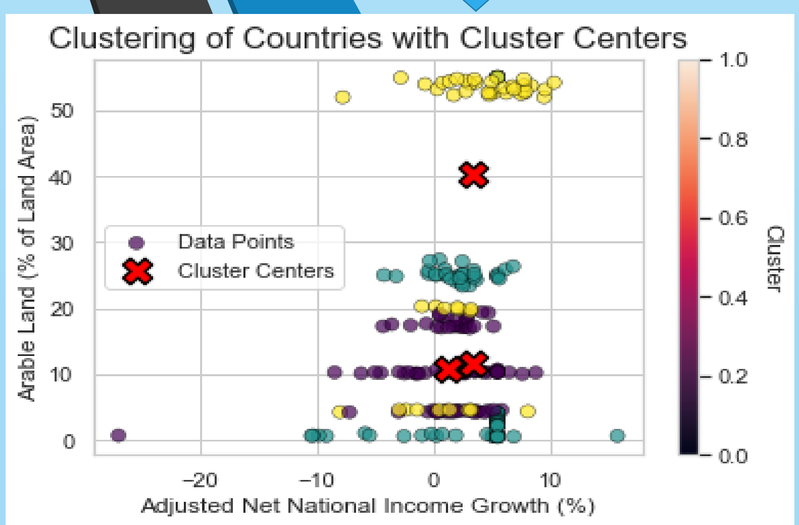
- K-Means clusters nations based on economic and environmental features in the code.

Mathematical Notation:

- Let X be the dataset, C represent centroids, and minimize Σ ||x_i - c_j||^2.

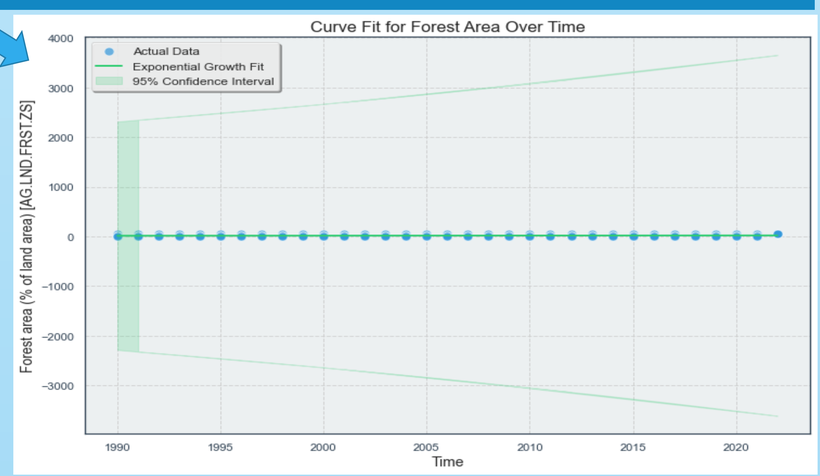
Cluster and Visualize:

- Utilizes K-Means clustering on key features: forest area, adjusted net national income growth, agriculture contribution to GDP, and arable land percentage.
- Standardizes data using StandardScaler and applies K-Means with 3 clusters, enhancing interpretability.
- Silhouette Score measures clustering effectiveness, ensuring well-defined clusters.
- Visualization plots nations based on adjusted net national income growth and arable land, highlighting cluster centers (X markers) and indicating cluster membership with colors.



Curve Fitting

- Exponential Growth Model Visualization:
- Utilizes an exponential growth model ($Y=a \cdot e^{bx}$) to capture arable land trends over time.
- Plots actual data points (blue) and the fitted exponential growth curve (green) to visualize the model's fit.
- The shaded area represents the 95% confidence interval around the fitted curve, providing insight into prediction uncertainty.
- Enhances understanding of the forest area's evolution and potential trajectory.



Arable Land Projection and Cluster Analysis:

Curve and Prediction Plot:

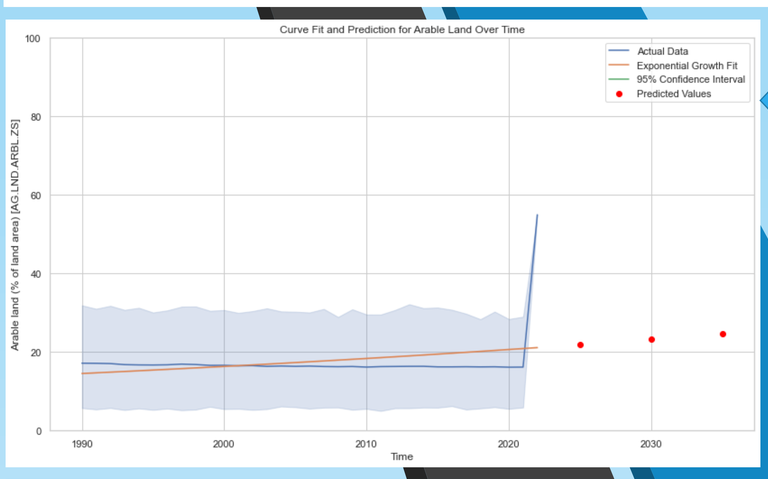
- Displays the fitted curve (green) representing arable land trends.
- The shaded area indicates the 95% confidence interval around the fitted curve, offering insights into prediction uncertainty.
- Red dots represent predicted values for specific future years (2025, 2030, 2035).

Silhouette Score:

- The Silhouette Score of 0.4270 denotes a robust clustering structure, highlighting distinct patterns among nations based on selected features.

Predicted Values:

- Anticipated arable land percentages for 2025, 2030, and 2035 are 21.90%, 23.22%, and 24.62%, respectively.
- These predictions assist in forecasting potential shifts in arable land distribution over the specified time frame.

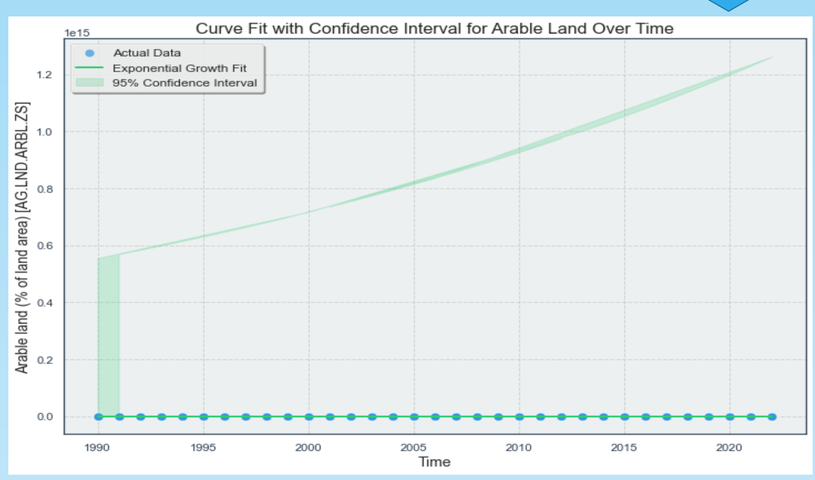


Arable Land Trend with Confidence Interval

- Illustrates the fit of the arable land curve (green line), capturing historical trends.
- The scatter points represent actual data (blue), offering context to the curve's accuracy.
- The shaded region signifies the 95% confidence interval around the fitted curve, providing insights into the model's reliability.

Interpretation:

- The confidence interval aids in assessing the uncertainty associated with the predicted arable land values.
- This visual representation enhances the understanding of potential variations in arable land trends over time.



Data source link:
https://databank.worldbank.org/source/world-development-indicators#
GitHub link :
https://github.com/AndroidFury/Assignment-3-Clustering-and-fitting

Conclusion:

In unraveling the dynamics of arable land, our code's focus on food production, economic vitality, and environmental well-being yields insightful findings. The application of an exponential growth model and K-Means clustering enhances our understanding, offering a glimpse into the intricate relationship between adjusted net national income and arable land percentage. This concise analysis equips us with predictive capabilities, fostering informed decisions for the sustainable management of this indispensable resource.