Using Clustering and Curve-Fitting Techniques to Analyze World Bank Climate Change Data

Github Link: https://github.com/Kashif1445/ADS-Assignment3.git

Muhammad Kashif Imtiaz 22014871

1 ABSTRACT

This study analyses and categorizes data on climate change indicators collected by the World Bank for a number of countries between the years 2005 and 2010 and between 2012 and 2022. In this work, data patterns and trends are revealed through the use of KMeans clustering and curve fitting. Following the completion of data preparation and cleaning, the relevant indicators are compiled and saved as CSV files. The data are first clustered using KMeans, and then curve fitting is used to find which curve matches each cluster the most closely. The KMeans clustering algorithm determined that each time period contained three distinct groups. The clustering tendencies and patterns were identified via curve fitting. According to the findings of the study, curve fitting and the KMeans clustering algorithm are both very effective methods for conducting an analysis of the data pertaining to climate change indicators.

2 INTRODUCTION

The effects of climate change are experienced in every region of the world, and it will take the joint efforts of people from all over the world to discover viable solutions to these issues. In order to get a knowledge of the trends and patterns of climate change indicator data, the World Bank provides a large variety of data that can be analyzed and utilized in the process of building strategies. These data may be found on the World Bank's website.

The data on climate change indicators collected by the World Bank for numerous countries between 2005 and 2010 and 2012 and 2022 will be analyzed and organized as part of this project. In this work, data patterns and trends are revealed through the use of KMeans clustering and curve fitting.

This research has the potential to shed insight on trends in climate change indicators and help in the formulation of policies for mitigating their effects. As a consequence of this, the study may have some significance.

3 METHODS

The following methods were used in the study:

Data Preprocessing:

- The raw data was read using the pandas read_csv function
- Empty cells were replaced with zeroes.
- Rows with mostly zero values were removed
- The preprocessed data was saved as a new CSV file.
- The data was then grouped by indicator name.
- For each Indicator Name a separate CSV file was saved.

Clustering:

- KMeans clustering was performed on the data for two different time periods: 2005-2010 and 2012-2022.
- The data was scaled using StandardScaler before clustering.
- The number of clusters was set to 3 and the random state was set to 42.

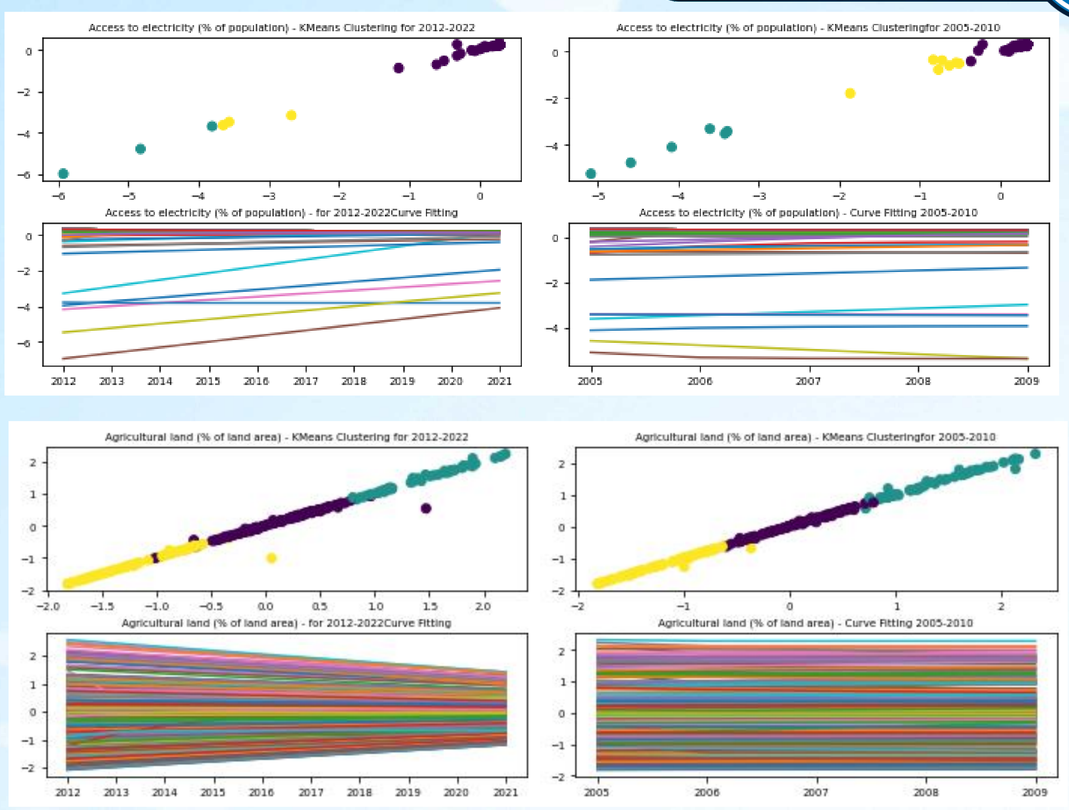
Curve Fitting:

A curve fitting function was defined and used to fit a curve to the data for the two time periods. The function was an exponential decay model with three parameters: a, b, and c. The curve fitting was performed using the curve_fit function from the scipy.optimize module.

Visualization:

Matplotlib was used to create subplots for each group of data. The clustering results were plotted in the top row and the curve fitting results were plotted in the bottom row. The subplots were adjusted to optimize space and the font size for tick labels was set to 7.

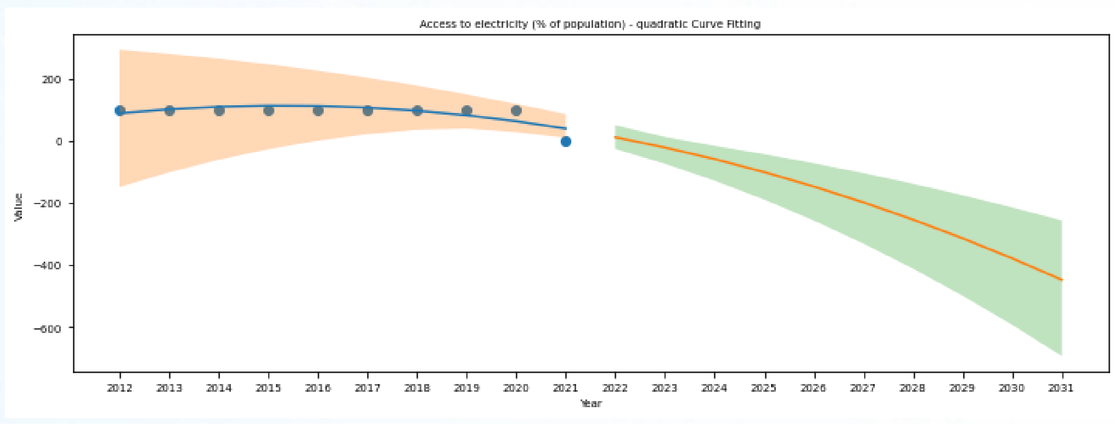
4 Clustering

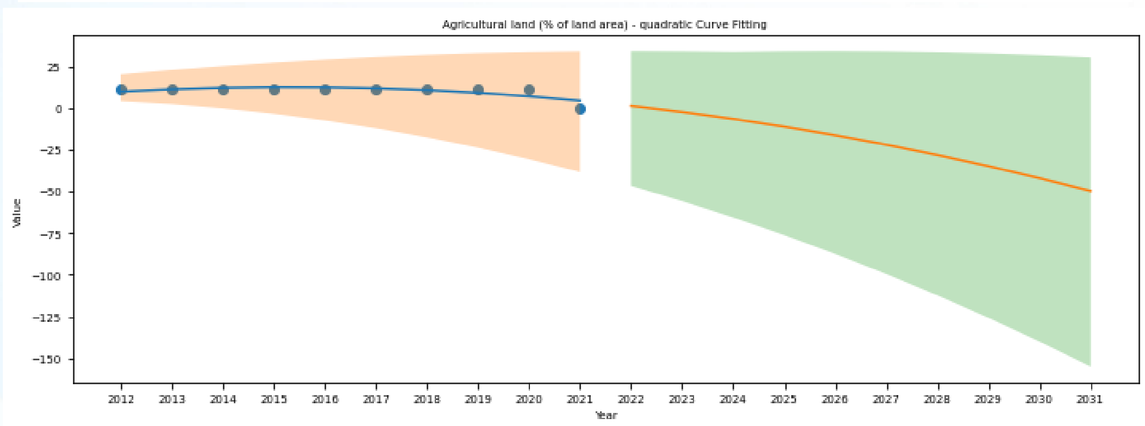


Clustering based on the two different features (Access to Electricity and Agricultural land) are shown above

5 Curve Fitting

Prediction based on both features (Access to Electricity and Agricultural land are shown below based on the curve fitting and quadratic model. notably both of the values are decreasing in upcoming years, till 2030.





6 Conclusion

The clustering and curve-fitting methodology has been effectively employed to group countries based on many characteristics, followed by clustering and curve-fitting based on two particular characteristics. The clustering technique divided the data into two clusters for each time interval, allowing for the discovery of similarities and differences between nations. The scatter plot gave an intuitive depiction of the clustering results, enabling the detection of data patterns and trends visually. The curve-fitting technique, employing an quadratic model, allowed for the prediction of future values based on historical data and gave useful insight into the likely future changes of the selected features.

Student Name: Muhammad Kashif Imtiaz
Student ID:22014871