University of Hertfordshire UH

Clustering And Fitting

Abstract:
Using this algorithm, country-specific environmental data is modelled and clustered. The principal aims are to ascertain trends among nations by means of particular environmental metrics and to develop a temporal model of the overall emissions of greenhouse gases. After loading, cleaning, and imputed values for missing variables, the dataset is clustered using the k-means algorithm. The entire set of data on greenhouse gas emissions is also fitted with a model, which yields insights into future projections and the confidence intervals that go along with them.

Objective

Analysing environmental data is essential to comprehending global trends and patterns. The main goals of this code are to forecast future total greenhouse gas emissions and cluster countries according to a set of environmental parameters. Preprocessing is done on the data to deal with missing values, and clustering is done to find groups of nations that share common environmental traits. Furthermore, a model is fitted in order to investigate the possible course of overall greenhouse gas emissions.

Data Pre-processing

- **Data Loading:** The Pandas library is used to load the environmental data from a CSV file.
- **Cleaning:** To convert the data to numeric format, non-numeric characters in particular columns are found and eliminated.
- **Imputation:** To preserve data integrity, missing values in particular columns are imputed using the mean.
- **Normalisation:** To guarantee that every feature contributes equally to the clustering process, the imputed data is normalised.

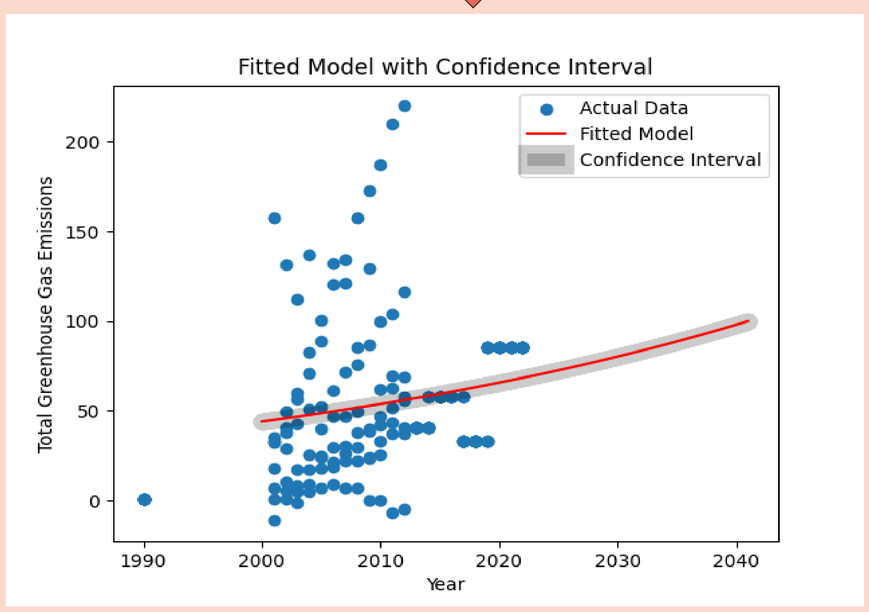
Clustering

- **K-Means Clustering:** Using normalised environmental factors as a basis, the K-Means algorithm is used to group countries together.
- **Silhouette Score:** This metric measures how well-separated the clusters are and is used to evaluate the quality of the clustering results. For the following data, the silhouette score is 0.30.

Visualisation: To show how the countries are clustered, a scatter plot is created,
with the cluster centres indicated in red. The percentage of forest land is shown
on the x-axis, and the total greenhouse gas emissions are shown on the y-axis.

Modeling:

- **Curve Fitting:** The exponential model function is used to fit a curve in order to determine the overall trend in greenhouse gas emissions over time.
- **Confidence Interval:** Confidence intervals give the fitted model an additional degree of uncertainty by providing a range of potential future values.Visualisation: Over a predetermined time range, the fitted model, real data points, and confidence interval are displayed.



Predictions:

year	Green house gas emmission
2036	90.38904169
2037	92.21502161
2038	94.07788878
2039	95.97838837
2040	97.91728059
2041	99.89534105

Conclusion:

With the use of curve fitting to simulate the trajectory of total greenhouse gas emissions and clustering algorithms to identify nation groups, this tool offers a thorough study of environmental data. The findings can guide future policy decisions and advance our understanding of global environmental trends.

Datasource: https://databank.worldbank.org/reports.aspx?source=2&series=AG.LND.FRST.ZS&country=#

GitHub Link: https://github.com/harighanapuram1329/ADS-assignment-3.git

Name : Hari Prasad Ghanapuram
Id : 22051910