University of Hertfordshire

Predictive modelling, clustering, and exploratory data analysis

Abstract

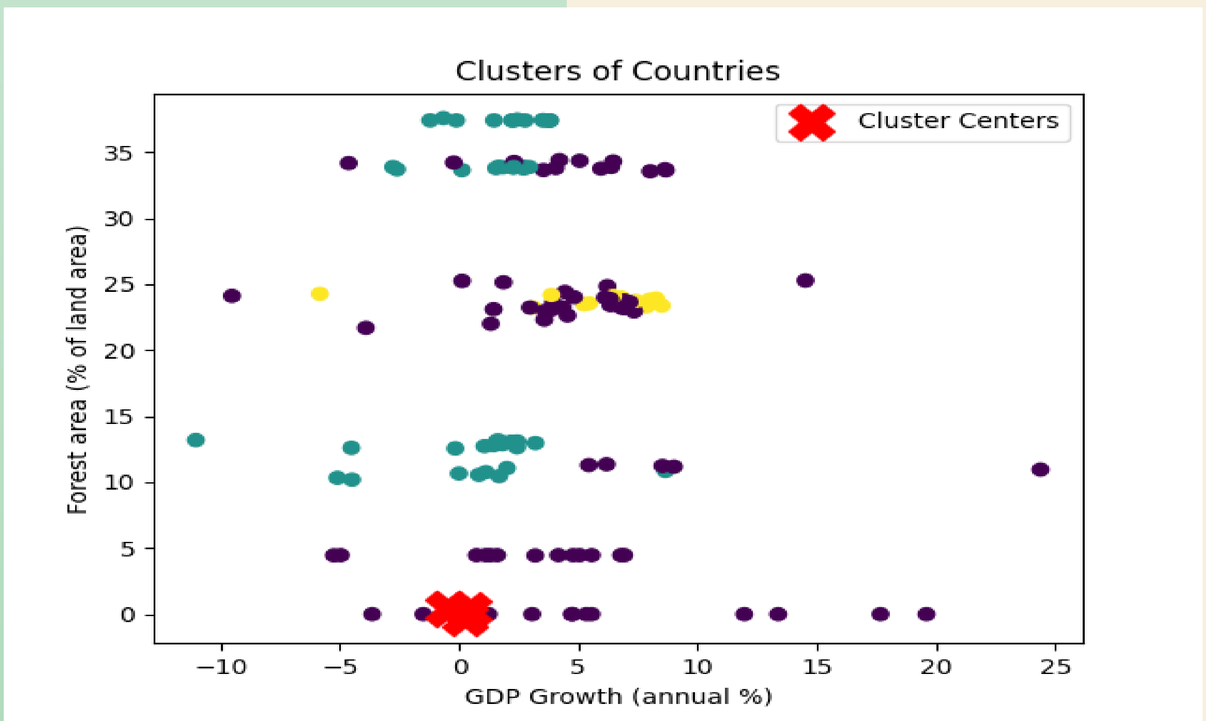
Using a dataset that includes a variety of variables for several nations, this study uses predictive modelling, clustering methods, and exploratory data analysis approaches. Important metrics are examined, including population, net migration, GDP growth, unemployment rate, water productivity, and forest area. Data loading and preparation are the first steps in the analysis, after which KMeans clustering is used to find patterns in the data. The relationship between GDP growth and forest area is then predicted using a predictive model that includes confidence intervals to account for uncertainty.

. Introduction

In a time of increasing global interconnection, it is critical to comprehend the complex dynamics of the socioeconomic indices of many nations. This study goes on an analysis of such indicators, attempting to identify patterns, correlations, and forecast tendencies. By utilizing data analytic technologies, our goal is to simultaneously reveal the relationship between GDP development and forest acreage and cluster countries according to important parameters. Due to its complexity, this research offers important new perspectives on international patterns as well as the possible influence of environmental and economic elements on national development paths.

Cluster Analysis

Using StandardScaler to normalise the data, we next apply KMeans clustering with a predetermined number of clusters (num_clusters = 3 in this example). To evaluate how well the clustering is done, the Silhouette Score is computed. A scatter plot is used to visualise the clusters, with the red centres of each cluster identified.

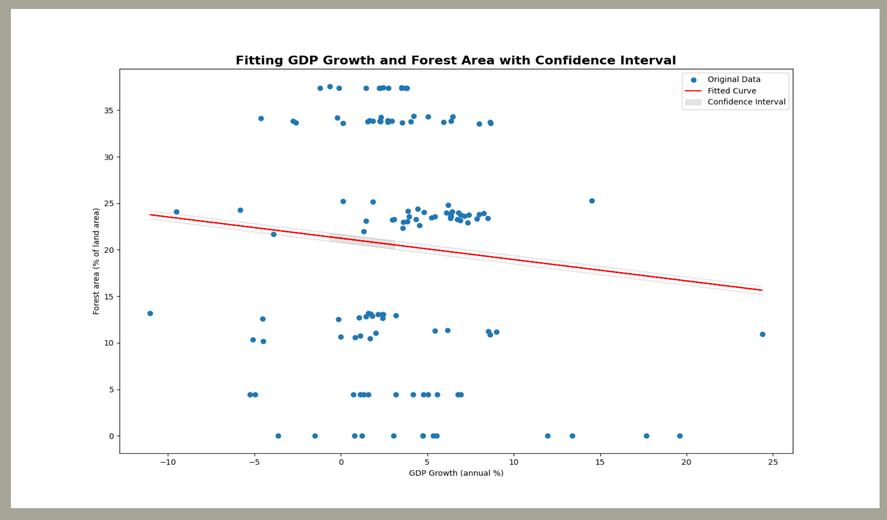


Data Pre-processing

A Pandas Data Frame is filled with the dataset, which is kept in 'DATA.csv'. We concentrate on particular clustering indicators, such as GDP growth, unemployment rate, water productivity, population, net migration, and health care spending, as well as forest area.

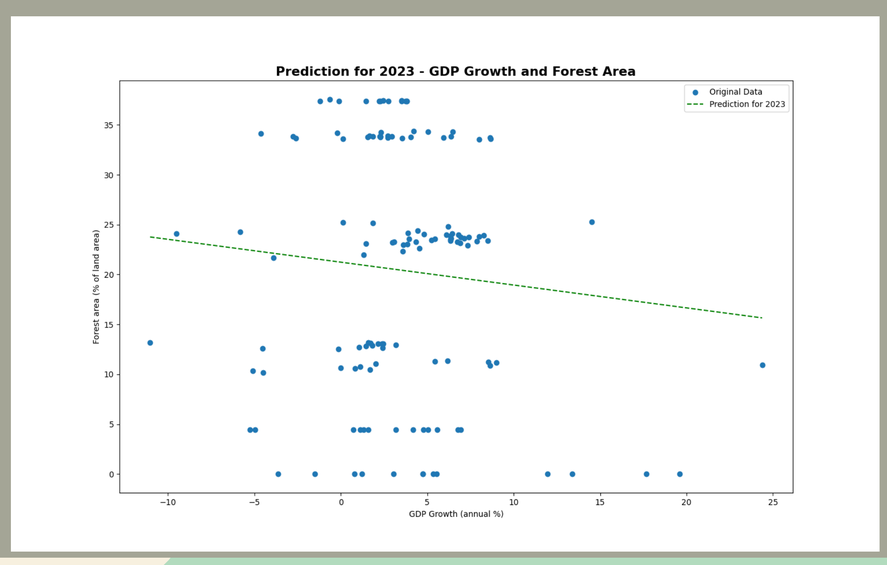
Fitting a Predictive Model:

Using the curve_fit function from SciPy, we fit a straightforward linear model to comprehend the relationship between GDP growth and forest area. The covariance matrix and model parameters are acquired. Furthermore, a function called err_ranges is provided to determine the fitted curve's confidence interval.
A scatter plot displays the original data, the fitted curve, and the confidence interval. This sheds light on the relationship—as well as the inherent uncertainty—between GDP growth and forest area.



Prediction for 2023:

Using the fitted model, we predict the values for the year 2023 for all countries. The prediction is visualized alongside the original data.



Conclusion

This analysis creates a predictive model for the association between GDP growth and forest area and offers insightful information about how nations are clustered based on particular indicators. Understanding the dataset is improved by the fitted model's ability to make predictions and the visualisations' assistance in deciphering patterns.
:

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

GitHub link: https://github.com/pallaramya/ADS-assignment-3.git

Name: Palla Ramya
Student Id: 22063652