Maria Pinzon

Final Report:
Energy Distribution in Latin America

Problem

Energy distribution in Latin America is something that appeals to many leaders pushing for infrastructure investment as it can be beneficial for the public to have equal access to energy throughout the regions. In Latin America, there is a large in inequality in access to basic resources, such as energy, referring to coverage and efficiency. The data found was collected from 28 countries in Latin America and it done to analyze the coverage through many different factors from 1973 to 2008. The measure of coverage was

Data Wrangling

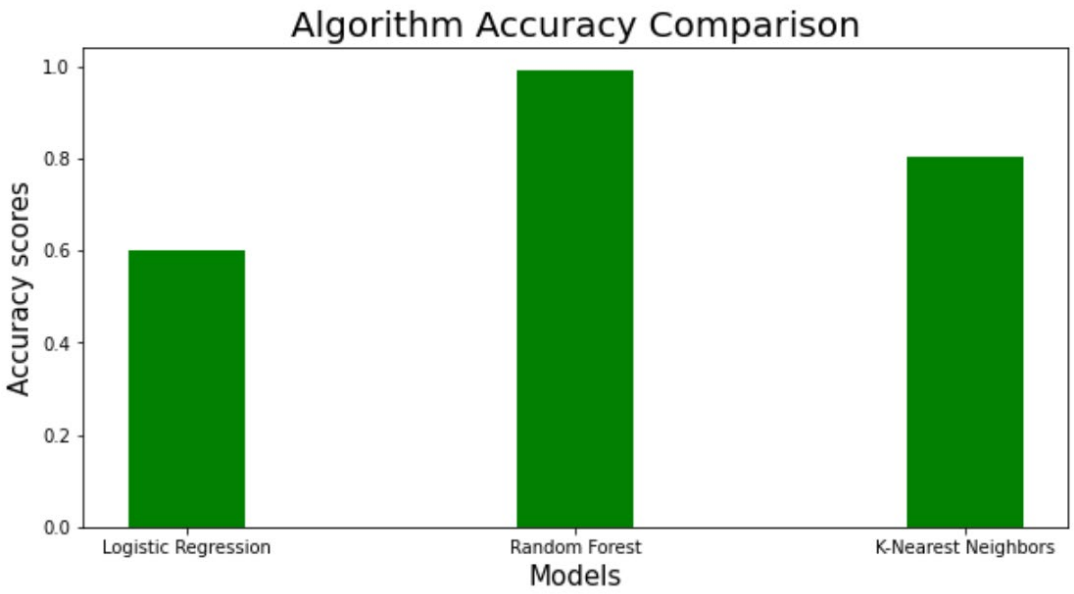
This data was collected by the World Bank Group and was named "Latin America and Caribbean – Utility Benchmarked Database". The raw dataset contained 31 columns and 4074 rows. When exploring the data, I first started by cleaning the column names as most of them were extensive and had to be condensed. Then I analyzed the data by identifier, in this case being each country that was included. Then I moved on to the data cleaning, which in this case showed that there was a large amount of missing data throughout the file. I first dropped the columns that had over 80% missing values. Then the rest of the data was grouped by the utility column and filled in with the mean of the group for the corresponding columns. The rest of the null values were filled in with 0. Then the numeric columns were made numeric to keep consistent.

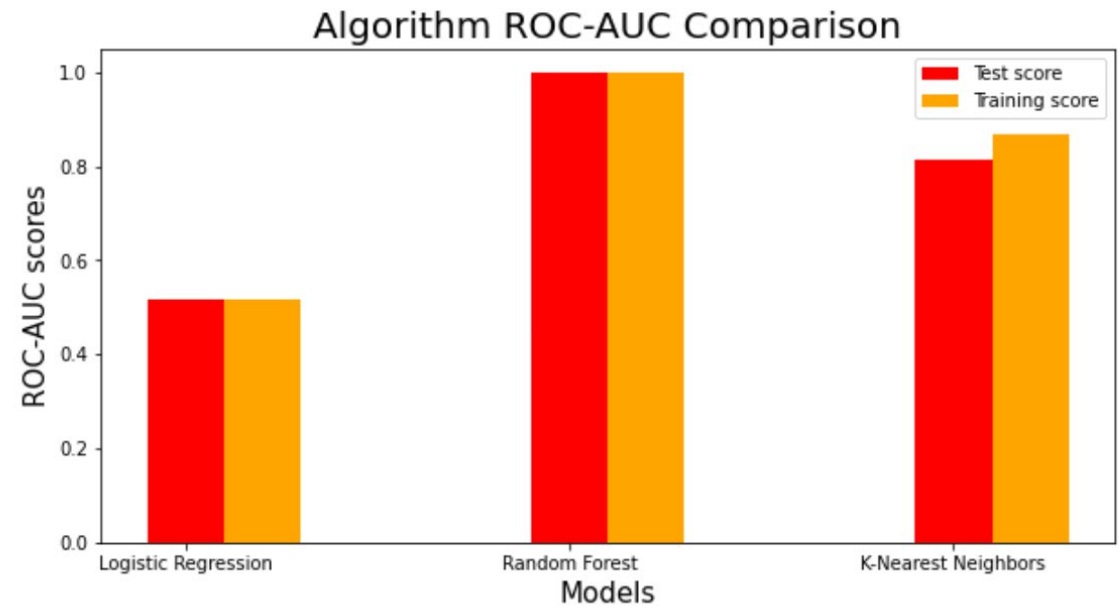
Exploratory Data Analysis

The data, after cleaning, yielded 4073 rows and 24 columns. Most of the columns where numerical and the rest were indicator such as the countries indicated and the energy company it belongs to. The data was then grouped by the utility code to be able to analyze by country.

Modeling

Three models were chosen to analyze and they were Logistic Regression, Random Forest Classifier and K-Nearest Neighbors. A column was created that showed if the residential coverage was above or below 80% (above = 1, below = 0). This column was the target and the rest of the columns were the features analyzed. After running the models and analyzing the accuracy scores and confusion matrices, the Random Forest Classifier model served as the better model to use for this data. The figures below show the comparisons.





Conclusion and Further Analysis

After completing the data analysis and model analysis, the model Random Forest Classifier proved to be the best model for analyzing this data. After this analysis, the project could be further extended by adding to the dataset with more current data and projecting towards the future.