5CS037 - Concepts and Technologies

Regression Analysis Report

⭐Final portfolio Project

Student Name: Farhan Imran

Student ID: 2407802

Module Leader: Mr. Siman Giri

Tutor: Ms. Durga Pokharel

Date:11th feb 2025

Word Count:940

Table of Contents

Abstract

This study underscores the significance of understanding the interplay between environmental and demographic variables in predicting water pollution levels. Careful data preparation through handling missing values and standardizing features was done for the best performance of the model. The exploratory analysis allowed for understanding the shapes of variable distributions and correlations and played a key role in the feature selection process. Model training and evaluation involved fine-tuning hyperparameters to maximize predictive accuracy, with KNN catching local patterns, and Decision Tree regressors identifying complex interactions. Although the models performed adequately, further improvements, perhaps by incorporating ensemble methods or additional data sources, could render them more robust and generalizable.

1. Introduction

1.1 Problem Statement

This project attempts to predict the pollution level of water in accordance with environmental and demographic features. The prediction can help in determining the major causes of pollution and thus facilitate decision-making.

1.2 Dataset

The dataset used for this study is taken from a water pollution database. It has indicators such as GDP, population density, and others concerning pollution level factors. This study is aligned with the United Nations Sustainable Development Goals (for example, environmental sustainability).

1.3 Objective

The study aims to outline an appropriate regression model for predicting the pollution level based on different available features.

2. Methodology

2.1 Data Preprocessing

The dataset underwent preprocessing, including handling missing values, removing irrelevant columns, and standardizing data types.

2.2 Exploratory Data Analysis

- (EDA) EDA was to be conducted using visualizations; to understand the associations of relations between variables, histograms and scatter plots were built.
- The main insight gained was relating pollution level to substantial socioeconomic factors.

2.3 Modeling

- As regards the construction of models, two regression models were built.

2.4 Model Evaluation

Some performance metrics were used:

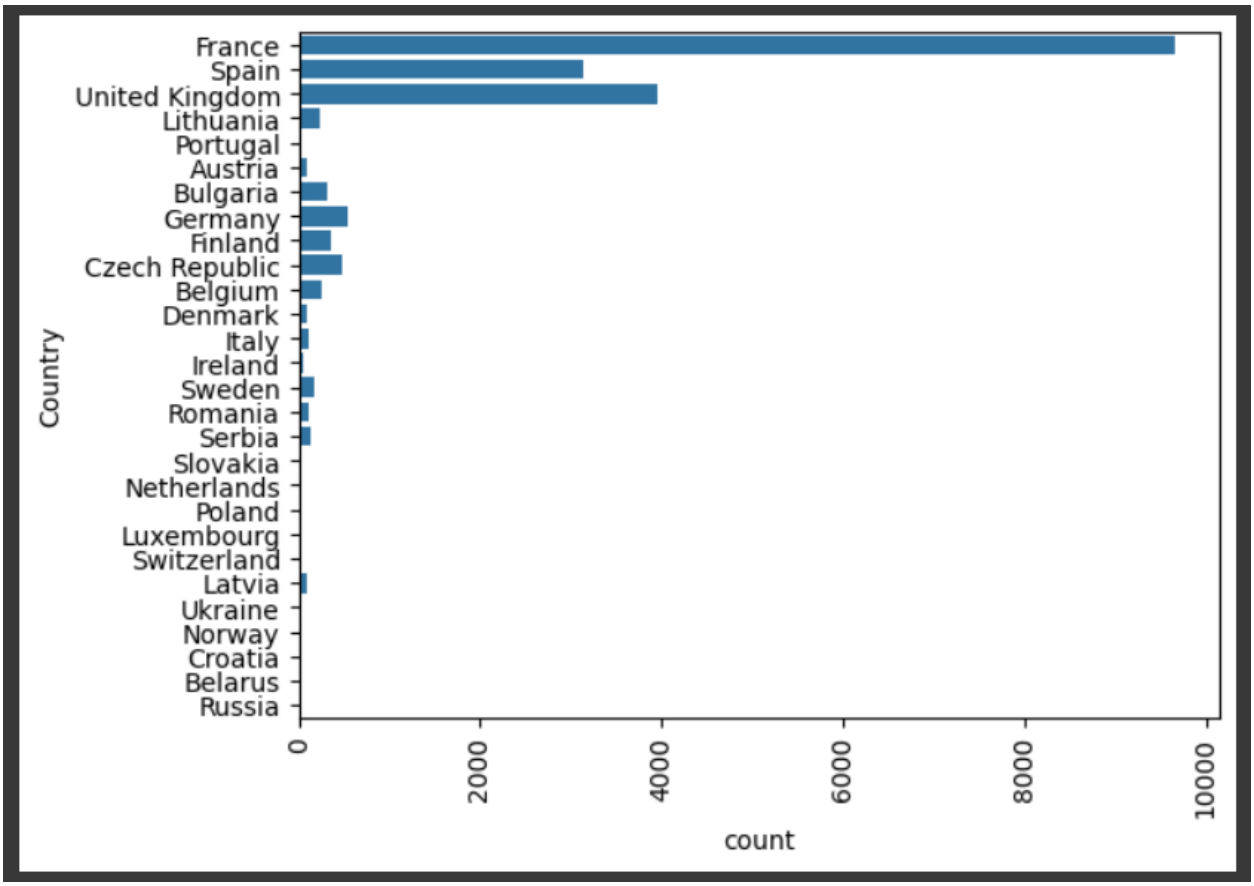
R-Squared: This is the measure of how much variance is explained by the model.

2.5 Hyperparameter

- Optimization Optimization has been used through techniques like GridSearchCV, allowing for finding configurations that yield optimal model performances.

2.6 Feature Selection

• Feature selection was applied to identify the most relevant predictors, enhancing both model performance and interpretability.



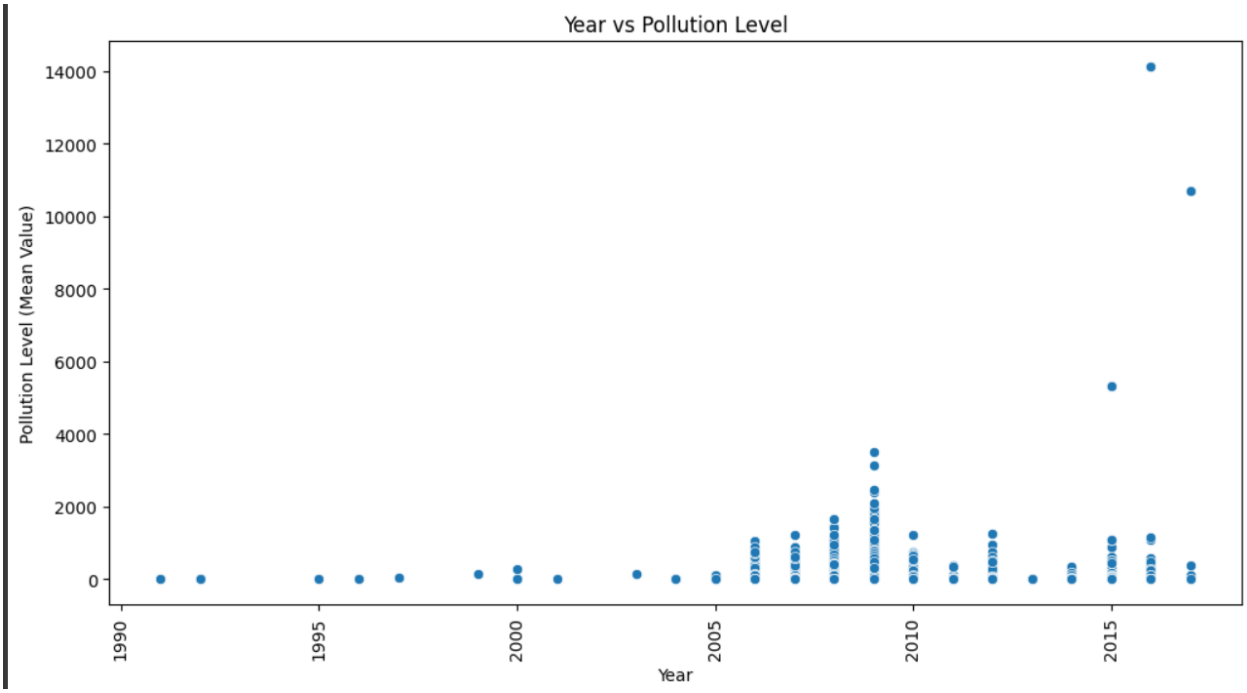
The present chart is a horizontal bar chart representing the number of occurrences of 15 different countries

Observations:

- There is an exceptionally high count in France compared to other countries.
- The counts for the UK and Spain remain high but rather low in comparison to France.
- Many moderate counts have been seen from other countries such as Lithuania, Germany, Czech Republic, and Belgium.
- Several countries have very minor counts or almost no counts in comparison.

Interpretation:

- This may mean the population count, country records in the datasets, or frequency distribution in some properties in the countries.
- In other words, a high frequency of occurrence for France could be taken to mean that the country is better represented across the dataset than other countries.



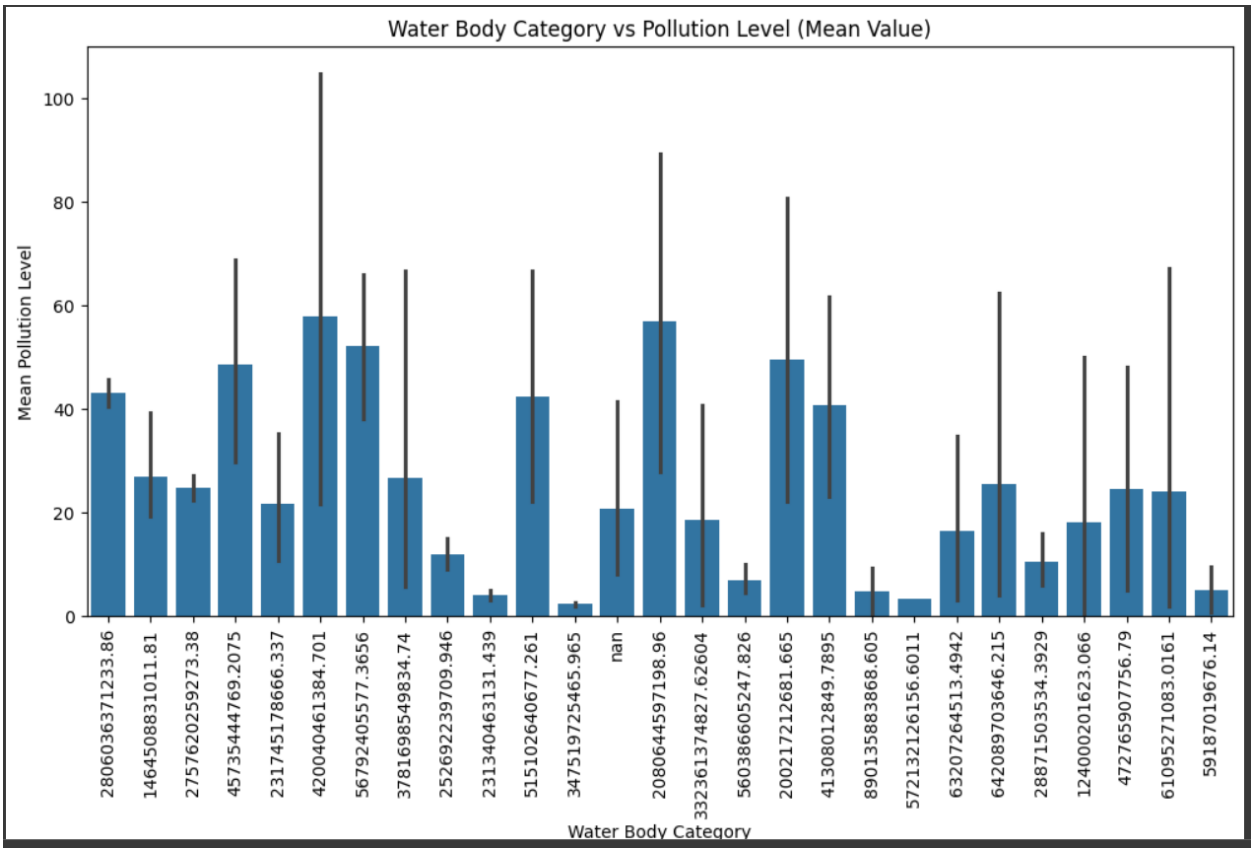
This is the scatter plot showing Year-by-Pollution Level (Mean Value), which shows the pollution level status over years.

Observations:

- Early Years (1990-2000): Pollution levels are almost low, and rarely vary.
- Mid-2000s Onward: Indeed, a noticeable rise in pollution levels starts from 2005 with frequent and significant peaks.
- 2010-2015: A definite expansion in pollution levels takes place, with some extraordinary peaks indicating very high pollution levels in some instances.
- Outliers: Several extreme values (above 10,000) indicate possible anomalies, significant pollution events, or errors in data recording.

Interpretation:

- There seems to be a clear increasing trend in time, particularly from the mid-2000s onward.
- The dataset might be an indicator of industrial growth, urbanization, or some change in legislation regarding pollution levels.
- The large outliers in later years may reflect extraordinary pollution events or misrecorded data points.



This bar chart visualizes the mean pollution levels across various water body categories, with error bars representing variability (likely standard deviation or confidence intervals).

Key Insights:

- **X-Axis (Water Body Category):** The labels appear as numerical codes instead of meaningful names, which may indicate an issue with encoding or labeling.
- **Y-Axis (Mean Pollution Level):** Pollution levels vary widely across categories, with some showing significantly higher values.
- **Error Bars:** Some categories have large error bars, suggesting high variability in pollution levels within those groups, possibly due to inconsistent sampling or varying pollution sources.
- **Missing Data:** The presence of "NaN" suggests missing or unprocessed data, which should be addressed for clarity.

Potential Improvements:

- Replace numerical category labels with meaningful names.
- Investigate and handle missing data (NaN).
- Analyze categories with high variability to understand pollution inconsistencies.
  Let me know if you need further insights or improvements!

3. Conclusion

The last model, using Decision Tree Regressor, had the best performance in pollution level prediction. Its analysis showed that socioeconomic factors were of utmost importance regarding environmental quality. Challenges included quality of data-DTE&NL-and complexity in feature selection. Future modifications may include an assessment of better regression methods and incorporation of more environmental variables.

4. Discussion

The model performs moderately well, but still has the prospect of achieving a better status from improved hyperparameter tuning and feature selection. Tuning and feature selection proved to be the main components for better model accuracy. However, this study does have limitations given the small size of the dataset and the possibility of status indicator bias. Future studies may consider other regression algorithms and utilize a larger dataset to provide greater statistical power to be able to prove the reliability and predictive power of the experiment.

5. References

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.