**CRISP-DM Data Preparation, Modelling, and Evaluation Report: Analysis of Air Quality Data in India**

Table of Contents

[1. Introduction: 3](#_Toc142363864)

[2. Data Preparation: 3](#_Toc142363865)

[3. Modelling: 3](#_Toc142363866)

[4. Evaluation: 3](#_Toc142363867)

[5. Insights from Modelling and Evaluation: 4](#_Toc142363868)

[6. Conclusion and Future Steps: 4](#_Toc142363869)

[7. References 4](#_Toc142363870)

# **Introduction:**

This report explores the steps three through five of the CRISP-DM methodology, which follow the Business as well as Data Understanding phases. Preparation, modelling, as well as evaluation of a dataset on India's air quality.

# **Data Preparation:**

* ***Handling Missing Values:*** Due to the nature of the dataset, gaps in the data were unavoidable. We used multiple imputation methods, taking into account the spatial as well as temporal context of the data. For instance, the average of the preceding and succeeding days was used to estimate missing values for a specific day in a city.
* ***Outlier Treatment:*** Box plots as well as z-scores were used to find outliers, which could have been caused by instrument failure or particularly severe pollution events. A cutoff value based on the interquartile range was used to limit the range of extreme values.
* ***Data Transformation:*** Min-Max scaling was used to normalise the variables in the model. Time series data were subjected to seasonal decomposition in order to isolate trends, seasonality, as well as residuals.
* ***Feature Engineering:*** In order to capture the temporal patterns in the data, novel features like rolling averages as well as lag variables were implemented.

# **Modelling:**

* ***Model Selection:*** For this reason, we decided to use a mix of regression models, time series forecasting, as well as machine learning algorithms to analyse and anticipate changes in air quality (Krishnaswamy, et. Al., 2022).
* ***Time Series Forecasting:*** ARIMA (AutoRegressive Integrated Moving Average) as well as Pollutant levels in the future were predicted using prophet models fed with data from the past.
* ***Regression Analysis:*** To learn how various pollutants interact with environmental variables like temperature as well as wind speed, multiple linear regression models were created.
* ***Machine Learning:*** Pollutant levels were predicted using Random Forest as well as Gradient Boosting algorithms due to their flexibility in dealing with non-linear relationships and massive datasets.

# **Evaluation:**

* ***Model Performance Metrics:*** Root Mean Square Error (RMSE) was used to quantify the discord between predicted and observed values in regression models. Accuracy, precision, recall, as well as F1-score were all taken into account when performing classification tasks.
* ***Time Series Forecasting:*** The RMSE was within a reasonable range, indicating that the ARIMA model provided a reasonable fit. When it came to identifying trends and seasons, however, the Prophet model performed better than ARIMA, making it the model of choice for forecasting.
* ***Regression Analysis:*** The factors that contribute to pollution levels were revealed by the regression models. However, variance inflation factors (VIF) were effective in overcoming the problem of multicollinearity between predictors.
* ***Machine Learning:*** Above 85% accuracy was achieved by both the Random Forest as well as Gradient Boosting models. Important predictors of air quality were highlighted in feature importance plots generated by these models.

# **Insights from Modelling and Evaluation:**

* ***Temporal Patterns:*** Seasonal spikes in pollution, especially during the winter months in North India, were captured by both the ARIMA as well as Prophet models (Patil, 2021).
* ***Predictive Factors:*** Significant predictors of air quality were found to be temperature, wind speed, as well as the prior day's pollution levels.
* ***Model Robustness:*** Non-linearities were handled competently by the machine learning models, especially the Gradient Boosting algorithm, which produced accurate forecasts of air quality.

# **Conclusion and Future Steps:**

A thorough understanding of the dynamics of India's air quality has been provided by the Data Preparation, Modelling, as well as Evaluation stages. Future trends can be predicted with confidence using the developed models because they provide a solid grounding in the factors influencing air quality.

# **References**

Krishnaswamy, V., Singh, N., Sharma, M., Verma, N. and Verma, A., 2022. Application of CRISP-DM methodology for managing human-wildlife conflicts: an empirical case study in India. *Journal of Environmental Planning and Management*, pp.1-27.

Patil, R., 2021. *Prediction an air quality index data using machine learning and deep learning* (Doctoral dissertation, Dublin, National College of Ireland).