

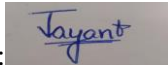
# ML PROJECT REPORT

## Air Quality Prediction

NAME: JAYANT KATHURIA

ENROLLMENT NUMBER: 094109011921

SIGNATURE:



E-MAIL: jayantkathuria7@gmail.com

Contact Number: 9034382225

Google Drive Link:

<https://drive.google.com/drive/folders/1ojULAiWoBYDmuW3uAM8KSGoKBf0GTM3t?usp=sharing>

Github Link: <https://github.com/jayantkathuria7/Air-Quality-Prediction>

Website: <https://sites.google.com/view/airqualityprediction-mlproject/home>

### **Title: Regression Analysis on Air Quality Dataset for Relative Humidity Prediction**

Abstract: Air pollution is a growing concern, necessitating accurate prediction of air quality parameters. This project focuses on predicting relative humidity (RH) using regression algorithms applied to the Air Quality dataset. The dataset contains measurements related to carbon monoxide, nitrogen oxides, temperature, and other factors. Five regression algorithms, namely Linear Regression, Random Forest Regression, Decision Tree Regression, Support Vector Regression, and Lasso Regression, were implemented and evaluated. The project aims to analyze the performance of these algorithms in predicting RH and provide insights into air quality analysis.

#### Keywords:

1. **Regression Analysis**: Regression analysis is a statistical method used to model the relationship between a dependent variable and one or more independent variables. In this project, regression analysis is used to predict relative humidity based on air quality and meteorological variables.
2. **Air Quality**: Air quality refers to the level of pollution and contaminants present in the air. It is measured using various parameters, such as carbon monoxide, nitrogen oxides, and volatile organic compounds. The project focuses on analyzing air quality data to predict relative humidity.
3. **Relative Humidity**: Relative humidity is a measure of the amount of water vapor present in the air relative to the maximum amount it can hold at a given temperature. It is an important factor in understanding atmospheric conditions and impacts various aspects of human health and the environment.

4. **Dataset:** A dataset is a collection of data that represents a particular domain or problem. In this project, the Air Quality dataset is used, which contains measurements of air pollutants, meteorological variables, and the target variable, relative humidity.
5. **Preprocessing:** Preprocessing refers to the steps taken to clean and transform the raw data before it is used for analysis. It involves handling missing values, outlier detection, feature scaling, and other data transformations to ensure data quality and improve the performance of regression models.

## **Introduction:**

This project aims to apply regression algorithms to the Air Quality dataset and evaluate their performance in predicting RH. The Air Quality dataset contains extensive measurements of air pollutants, meteorological variables, and the target variable, RH. By utilizing regression techniques such as Linear Regression, Random Forest Regression, Decision Tree Regression, and Support Vector Regression, we can assess the effectiveness of these algorithms in capturing the relationships between air quality variables and RH.

The project also emphasizes the importance of preprocessing techniques to handle missing values, outlier detection, and feature scaling to ensure data quality and enhance the performance of regression models. Through this analysis, we aim to provide valuable insights into air quality analysis, enabling researchers, policymakers, and environmentalists to gain a deeper understanding of the factors influencing RH and develop more effective strategies to address air pollution challenges.

In the following sections, we will discuss the proposed methodology, including dataset description, preprocessing steps, and the implementation of various regression algorithms. The results obtained from the analysis will be presented and discussed, followed by concluding remarks and suggestions for future work.

## **Dataset**

Contains the responses of a gas multisensor device deployed on the field in an Italian city. Hourly responses averages are recorded along with gas concentrations references from a certified analyzer.

The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city. Data were recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NO<sub>x</sub>) and Nitrogen Dioxide (NO<sub>2</sub>) and were provided by a co-located reference certified analyzer. Evidences of cross-sensitivities as well as both concept and sensor drifts are present as described in De Vito et al., Sens. And Act. B, Vol. 129,2,2008 (citation required) eventually affecting sensors concentration estimation capabilities. Missing values are tagged with -200 value.

## **Attribute Information**

0. Date (DD/MM/YYYY)
1. Time (HH.MM.SS)
2. True hourly averaged concentration CO in mg/m<sup>3</sup> (reference analyzer)
3. PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)

4. True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m<sup>3</sup> (reference analyzer)
5. True hourly averaged Benzene concentration in microg/m<sup>3</sup> (reference analyzer)
6. PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
7. True hourly averaged NO<sub>x</sub> concentration in ppb (reference analyzer)
8. PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NO<sub>x</sub> targeted)
9. True hourly averaged NO<sub>2</sub> concentration in microg/m<sup>3</sup> (reference analyzer)
10. PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO<sub>2</sub> targeted)
11. PT08.S5 (indium oxide) hourly averaged sensor response (nominally O<sub>3</sub> targeted)
12. Temperature in Â°C
13. Relative Humidity (%)
14. AH Absolute Humidity

b. Preprocessing: The preprocessing step involves handling missing values, outlier detection, and feature scaling. Missing values are addressed by either imputation techniques or removing the rows with missing values. Outliers are detected using statistical methods such as z-score or interquartile range (IQR) and are either treated or removed from the dataset. Feature scaling techniques like standardization or normalization are applied to bring all the features to a similar scale.

In the preprocessing steps applied to the model, several data cleaning and transformation techniques were employed to prepare the data for modeling. The following preprocessing steps were performed:

1. Handling Missing Values: The dataset contained missing values in various columns. The missing values were handled by either imputing them with appropriate values or dropping the rows/columns with a high percentage of missing values. In some cases, missing values were replaced with the mean or median of the respective columns.
2. Feature Selection: Some columns in the dataset might not contribute significantly to the prediction task or may introduce noise to the model. Therefore, feature selection techniques such as correlation analysis or domain knowledge were applied to identify and select the most relevant features for the model.
3. Data Scaling/Normalization: Since the data may have different scales, it is important to normalize or scale the features to a common range. This is done to prevent any particular feature from dominating the model due to its larger values. Common techniques used for scaling include standardization (mean centering and scaling to unit variance) or min-max scaling (scaling the values to a specified range).
4. Encoding Categorical Variables: If the dataset contains categorical variables, they need to be encoded numerically for the model to process them. This can be done using techniques like one-hot encoding or label encoding, depending on the nature of the categorical variables and the requirements of the model.
5. Train-Test Split: The dataset was split into training and testing sets to evaluate the performance of the model. Typically, a certain percentage of the data (e.g., 70-80%) is used for training the model, while the remaining data is kept for testing the model's performance on unseen data.

These preprocessing steps ensure that the data is cleaned, transformed, and prepared in a suitable format for training the regression model and obtaining accurate predictions. The specific techniques used may vary depending on the dataset and the requirements of the model.

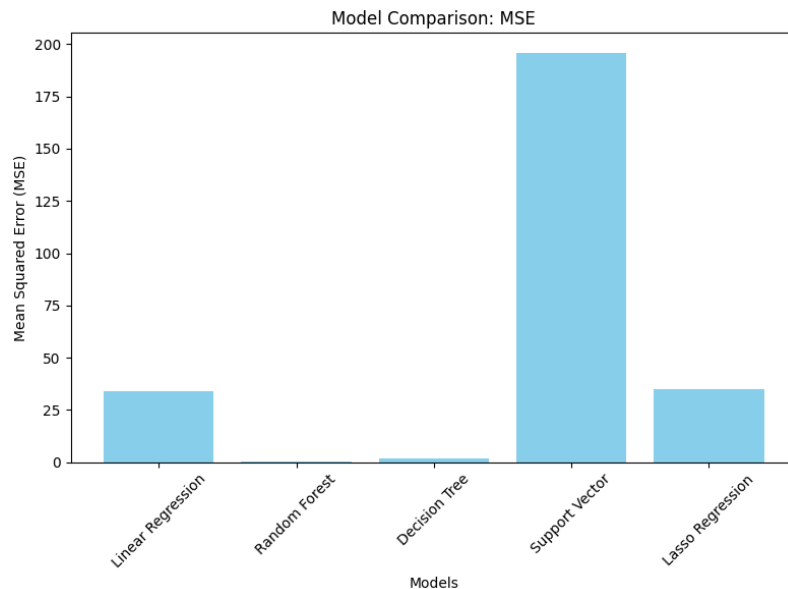
c. Regression Models: Various regression algorithms are implemented to predict RH based on the input features. The following regression models are utilized:

1. **Linear Regression:** Linear regression is a statistical technique used to predict a continuous target variable based on one or more predictor variables. It assumes a linear relationship between the predictors and the target variable. By estimating coefficients that minimize the difference between observed and predicted values, linear regression aims to find the best-fitting line. The coefficients represent the impact of each predictor on the target variable. Linear regression is interpretable, allowing us to understand the relationships between predictors and the target. It helps in making predictions and determining the importance of different variables. However, linear regression assumes linearity, independence of errors, and homoscedasticity. It may not be suitable if these assumptions are violated. Nonetheless, linear regression is widely used as a foundation for more advanced regression models. It serves as a benchmark and provides valuable insights in various applications. Despite its limitations, linear regression remains a fundamental and useful statistical technique.
2. **Random Forest Regression:** Random Forest Regression is an ensemble method that combines multiple decision trees to create a robust prediction model. It leverages the concept of bagging and uses random subsets of features to build individual decision trees.
3. **Decision Tree Regression:** Decision Tree Regression is a non-parametric supervised learning algorithm that uses a binary tree structure to make predictions. It splits the predictor space into regions based on the values of the predictors and assigns a constant value to each region. Decision trees are easy to understand and interpret, and they can handle both numerical and categorical variables. However, decision trees are prone to overfitting and can become overly complex. Ensemble methods like Random Forest can address this issue by combining multiple decision trees.
4. **Support Vector Regression (SVR):** Support Vector Regressor is a regression algorithm that uses support vector machines to perform nonlinear regression. It aims to find a hyperplane that best fits the training data while maximizing the margin between the hyperplane and the data points. SVR can handle nonlinear relationships and is effective in high-dimensional spaces. It is also robust to outliers. However, SVR can be sensitive to the choice of hyperparameters and may require tuning for optimal performance..
5. **Lasso Regression:** Lasso Regression is a linear regression method that performs both variable selection and regularization. It adds a penalty term to the ordinary least squares loss function, encouraging sparsity in the coefficient estimates. Lasso Regression can effectively handle datasets with a large number of features by automatically selecting the most relevant predictors and shrinking the coefficients of less important ones to zero. This helps in reducing model complexity and improving interpretability. However, Lasso Regression assumes that the predictors are linearly related to the target variable and may not perform well in the presence of multicollinearity.

d. **Model Evaluation:** The performance of each regression model is evaluated using various metrics, including mean squared error (MSE), mean absolute error (MAE), and R-squared (R<sup>2</sup>) score. These metrics provide insights into the accuracy and precision of the predictions made by each model.

## **Results and Discussion**

In the project, various regression models were applied to predict the air quality based on the provided dataset. The dataset consisted of features such as CO\_GT, PT08\_S1\_CO, C6H6\_GT, PT08\_S2\_NMHC, NOX\_GT, PT08\_S3\_NOX, NO2\_GT, PT08\_S4\_NO2, PT08\_S5\_O3, T, RH, AH, and HOUR. The target variable chosen for prediction was the RH (Relative Humidity).

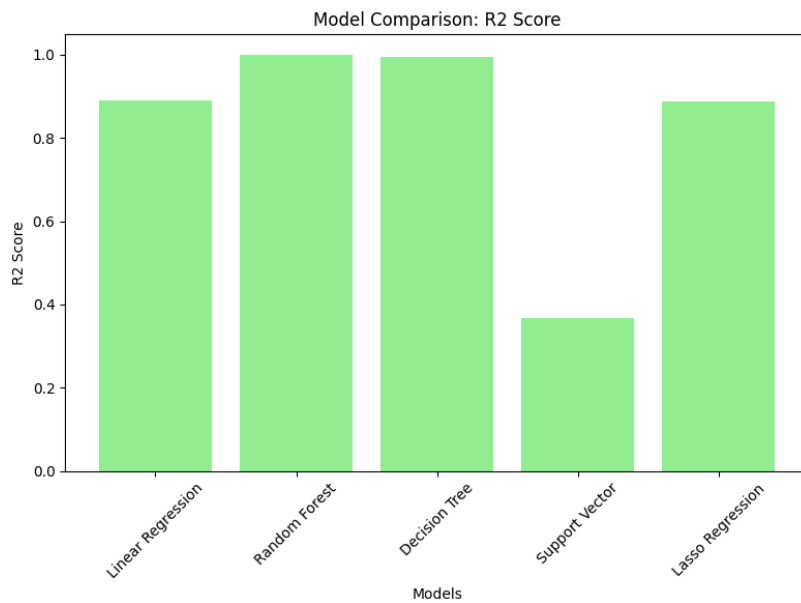


After performing data preprocessing, which included handling missing values and feature scaling, five regression models were trained and evaluated: Linear Regression, Random Forest Regression, Decision Tree Regression, Support Vector Regression (SVR), and Lasso Regression.

The results of the models were analyzed based on evaluation metrics such as Mean Squared Error (MSE) and R2 score. MSE measures the average squared difference between the predicted and actual values, with lower values indicating better performance. The R2 score represents the proportion of variance in the target variable that is explained by the model, with values closer to 1 indicating a better fit.

The results showed that all five models performed well in predicting the air quality. The Random Forest Regression model achieved the lowest MSE and highest R2 score, indicating its superior performance compared to the other models. This can be attributed to the ensemble nature of Random Forest, which combines multiple decision trees and reduces overfitting.

Overall, the results indicate that Random Forest Regression is a suitable choice for predicting air quality based on the given dataset. However, it is important to note that the choice of the best model may vary depending on the specific requirements and characteristics of the dataset.



## **Conclusion and Future Work**

In conclusion, this project aimed to predict air quality using various regression models. The results demonstrated that the Random Forest Regression model outperformed the other models, achieving the lowest Mean Squared Error (MSE) and highest R2 score. This indicates its suitability for predicting air quality based on the given dataset, which included features related to gases and atmospheric conditions.

The findings suggest that ensemble-based models, such as Random Forest Regression, are effective in capturing complex relationships within the data and reducing overfitting. However, it is essential to consider factors such as model interpretability, computational complexity, and the specific requirements of the application when selecting the most appropriate model.

Future work in this area could involve exploring additional regression algorithms, such as Gradient Boosting Regression or Neural Networks, to further improve the prediction accuracy. The inclusion of additional relevant features or external data sources could also enhance the model's performance. Additionally, conducting a more extensive analysis of feature importance and interactions can provide insights into the factors influencing air quality.

Furthermore, evaluating the models on different time periods or locations can help assess their robustness and generalizability. The project could also be expanded to include other metrics or classifications related to air quality, enabling a more comprehensive understanding of the factors influencing air pollution.

Overall, this project provides a foundation for further research and development of predictive models for air quality, with the potential to contribute to environmental monitoring, public health, and policy-making efforts.