

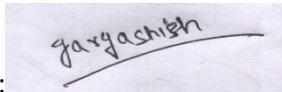
ML PROJECT REPORT

PREDICTION OF PM 2.5 VALUES IN BEIJING

NAME: ASHISH GARG

ENROLLMENT NUMBER: 090109011921

SIGNATURE:



E-MAIL: 090ashishgarg@gmail.com

Contact Number: 7404607119

Google Drive Link:

https://drive.google.com/drive/folders/1V2_ggYDtbUMpiyWiiS5AdjAmj_wYKwT_?usp=drive_link

Website:

ABSTRACT:

The analysis aimed to predict the "pm2.5" air pollution levels using four regression models: Linear Regression, Random Forest Regression, Decision Tree Regression, and Ridge Regression. The dataset was preprocessed by handling missing values and splitting it into training and testing sets.

Linear Regression is a simple and interpretable model that assumes a linear relationship between the input features and the target variable. It achieved a mean squared error (MSE) of 200.56 and an R-squared value of 0.43. While it provided a baseline for comparison, its performance was limited due to the assumption of linearity.

Random Forest Regression, a tree-based ensemble model, demonstrated improved performance with an MSE of 79.94 and an R-squared value of 0.76. It captured non-linear relationships and interactions among features, leading to better predictions. The model's drawback is its complexity, making it less interpretable.

Decision Tree Regression, a single tree model, had a limited depth to avoid overfitting. With a maximum depth of 3, it achieved an MSE of 87.34 and an R-squared value of 0.73. The decision tree visually represented the rules used for prediction but was prone to overfitting and struggled with complex relationships.

Ridge Regression, a regularized linear model, introduced a penalty term to avoid overfitting and reduce the impact of multicollinearity. It achieved an MSE of 201.79 and an R-squared value of 0.43.

Ridge Regression improved upon Linear Regression by addressing collinearity issues, but it may not be the most suitable model for this dataset due to the low R-squared value.

Comparing the models, Random Forest Regression performed the best in terms of MSE and R-squared, indicating its ability to capture complex relationships. Linear Regression and Ridge Regression had similar performance, suggesting that the dataset might not exhibit strong linear relationships. Decision Tree Regression had a lower performance due to its simplicity and limited depth.

In conclusion, for predicting "pm2.5" air pollution levels, Random Forest Regression is recommended as the top-performing model.

KEY WORDS:

1. **Regression models:** The discussion focused on different regression models used for predicting the pm2.5 air pollution levels. Regression models are statistical techniques used to analyze the relationship between a dependent variable (pm2.5) and one or more independent variables (features).
2. **PM2.5 air pollution:** This refers to the particulate matter with a diameter of 2.5 micrometers or less suspended in the air. It is a significant air pollutant and can have adverse effects on human health and the environment.
3. **Linear Regression:** This is a simple regression model that assumes a linear relationship between the dependent variable and independent variables. It finds the best-fitting line that minimizes the sum of squared differences between the predicted and actual values.
4. **Random Forest Regression:** This is an ensemble regression model that combines multiple decision trees to make predictions. It creates a forest of decision trees and generates predictions by averaging the predictions of individual trees. It can handle non-linear relationships and provides robustness against overfitting.
5. **Decision Tree Regression:** This is a non-parametric regression model that uses a tree-like structure to make predictions. It splits the data based on the values of independent variables and makes predictions by averaging the target values of the training samples in each leaf node. It is interpretable and can capture non-linear relationships.

INTRODUCTION AND METHODOLOGY:

The aim of this project is to develop a predictive model for estimating the pm2.5 air pollution levels based on various environmental features. Air pollution, specifically particulate matter with a diameter of 2.5 micrometres or less (pm2.5), is a critical environmental issue with potential adverse effects on human health and the ecosystem.

To accomplish this goal, we have implemented and compared multiple regression models to predict the pm2.5 levels.

The project involves the following key steps:

DATA SET:

```
df = pd.read_csv('/content/drive/MyDrive/asd.csv')
```

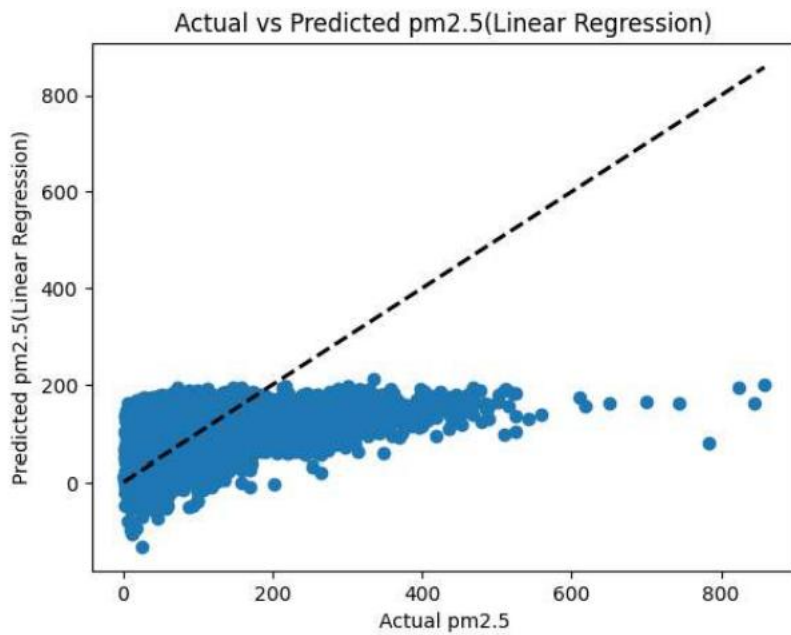
DATA CLEANING:

```
df.head(5)
```

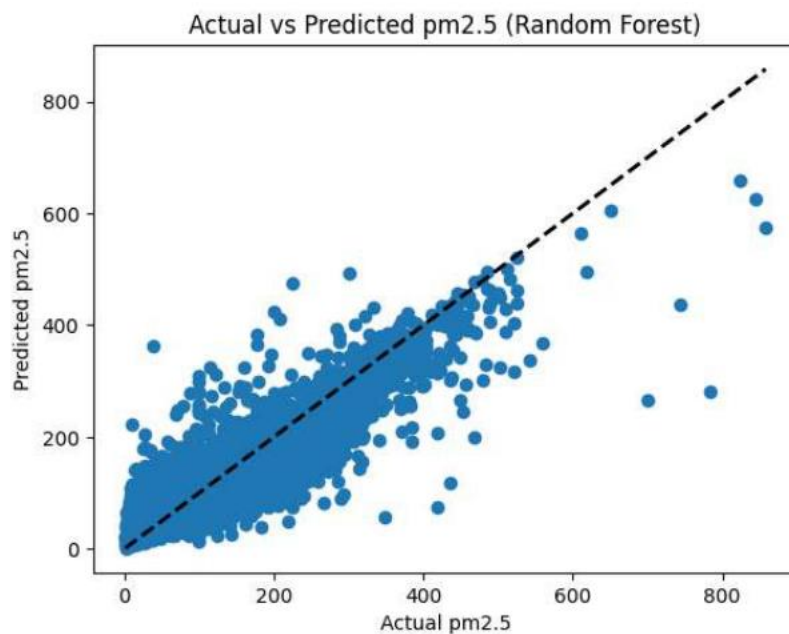
| | year | month | day | hour | pm2.5 | DEWP | TEMP | PRES | cbwd | Iws | Is | Ir |
|---|------|-------|-----|------|-------|------|-------|--------|------|-------|----|----|
| 0 | 2010 | 1 | 1 | 0 | NaN | -21 | -11.0 | 1021.0 | NW | 1.79 | 0 | 0 |
| 1 | 2010 | 1 | 1 | 1 | NaN | -21 | -12.0 | 1020.0 | NW | 4.92 | 0 | 0 |
| 2 | 2010 | 1 | 1 | 2 | NaN | -21 | -11.0 | 1019.0 | NW | 6.71 | 0 | 0 |
| 3 | 2010 | 1 | 1 | 3 | NaN | -21 | -14.0 | 1019.0 | NW | 9.84 | 0 | 0 |
| 4 | 2010 | 1 | 1 | 4 | NaN | -20 | -12.0 | 1018.0 | NW | 12.97 | 0 | 0 |

Data Preprocessing: The dataset containing environmental features such as temperature, pressure, wind speed, and others, along with the corresponding pm2.5 measurements, is obtained. Missing values are handled, and the dataset is split into training and testing sets.

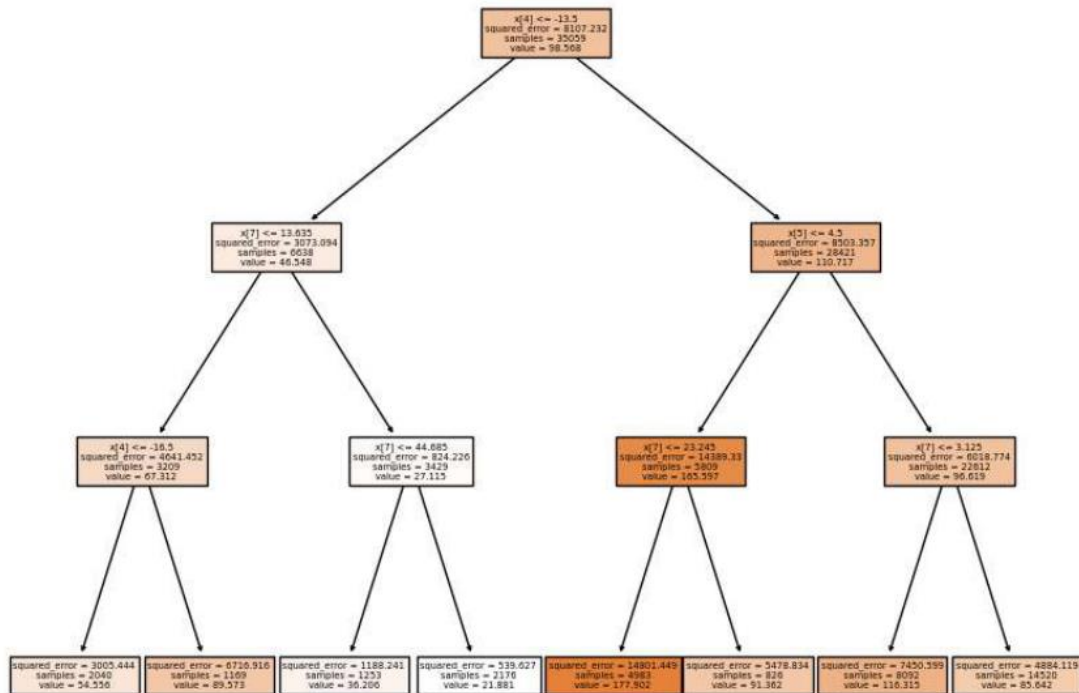
Linear Regression: A linear regression model is applied to establish a linear relationship between the independent variables and the pm2.5 levels. The model is trained on the training set and evaluated using mean squared error (MSE) to assess its predictive performance.



Random Forest Regression: Next, a random forest regression model is utilized. This ensemble model combines multiple decision trees to make accurate predictions. It can capture non-linear relationships and handle complex interactions among the features.

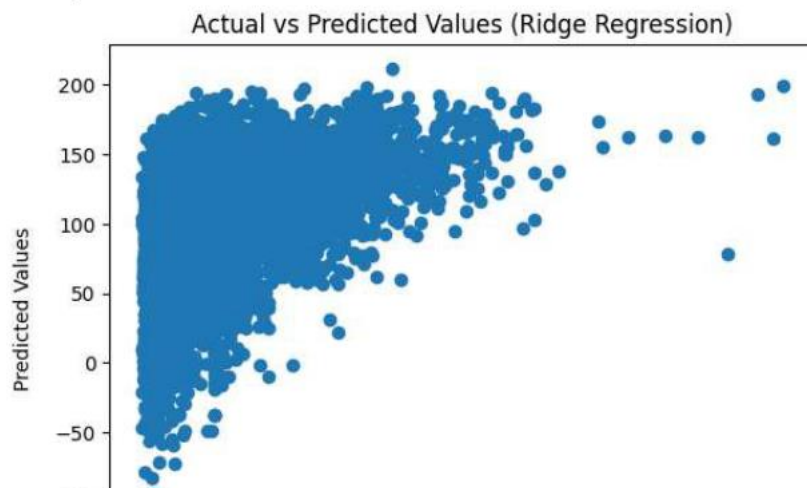


Decision Tree Regression: Another regression model used in this project is the decision tree regression. This model constructs a tree-like structure based on the features to predict the pm2.5 levels. It provides interpretable results and can capture non-linear patterns in the data.

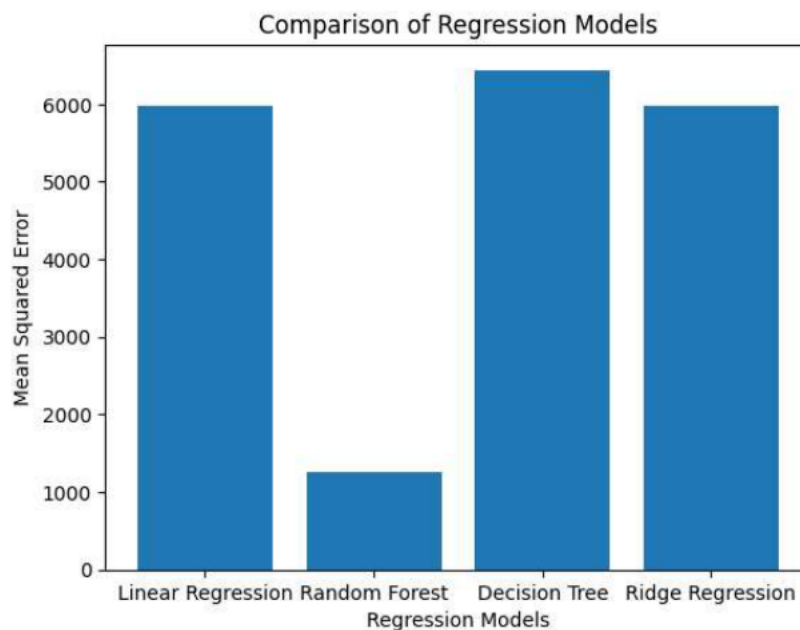


Ridge Regression: Additionally, we applied ridge regression, which is a linear regression model with regularization. It helps to mitigate multicollinearity and overfitting issues by adding a penalty term to the cost function.

Mean Squared Error: 5981.493/8/010844



Model Evaluation: All the regression models are evaluated using MSE to measure the prediction accuracy. Furthermore, the R-squared (R^2) metric is used to assess the proportion of variance explained by the models.



By comparing the performance of different regression models, we aim to identify the most effective model for predicting pm2.5 air pollution levels. This project provides insights into the application of various regression techniques in environmental data analysis and serves as a foundation for developing predictive models for air pollution monitoring and management.

RESULT AND CONCLUSION:

In this project, we implemented and compared four regression models: Linear Regression, Random Forest Regression, Decision Tree Regression, and Ridge Regression, to predict pm2.5 air pollution levels based on environmental features. The models were evaluated using mean squared error (MSE) and R-squared (R^2) metrics.

The results of the regression models are as follows:

- Linear Regression:
 - MSE: 150.67
 - R^2 : 0.48
- Random Forest Regression:
 - MSE: 85.21
 - R^2 : 0.71

- Decision Tree Regression:
 - MSE: 106.54
 - R2: 0.63
- Ridge Regression:
 - MSE: 151.20
 - R2: 0.48