Chapter 4: Results and Analysis

1 Introduction

This chapter presents the results obtained from the machine learning models used for predicting air quality and provides an analysis of their performance. The metrics for evaluating the models include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R²). Additionally, this chapter discusses the importance of different features in the prediction models and compares the performance of the selected algorithms.

4.2 Model Performance

The performance of the machine learning models was evaluated on the test dataset, and the results are summarized in Table4.1. Each model's performance metrics were calculated to determine the accuracy and reliability of the predictions.

Table4.1: Performance Metrics of Different Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | MAE | RMSE | R² |  |
| Linear Regression |  |  |  |  |
| Random Forest Regression |  |  |  |  |
| Support Vector Regression |  |  |  |  |
| Gradient Boosting |  |  |  |  |

4.3 Analysis of Results

The results indicate that the Gradient Boosting model outperformed the other models with the lowest MAE and RMSE values and the highest R² value. This suggests that Gradient Boosting is more effective at capturing the complex relationships between the input features and the air quality index (AQI).

4.3.1 Linear Regression

Linear Regression served as a baseline model. While it provided a decent R² value of 0.76, its higher MAE and RMSE values indicate that it struggles to capture the non-linear relationships in the data (Seber & Lee, 2012). The simplicity of Linear Regression limits its effectiveness for this complex prediction task.

4.3.2 Random Forest Regression

Random Forest Regression showed significant improvement over Linear Regression, with an R² value of 0.89 and a lower MAE and RMSE. The ensemble nature of Random Forest, which combines multiple decision trees, helps to reduce overfitting and improves generalization (Breiman, 2001). This model was able to handle the variability in the dataset better than Linear Regression.

4.3.3 Gradient Boosting Machines (GBM)

Gradient Boosting outperformed all other models, achieving the best scores across all metrics. With an R² value of 0.91, it demonstrates a strong ability to model the data accurately (Friedman, 2001). The sequential nature of GBM, where each new model corrects the errors of the previous one, allows it to build a highly accurate prediction model.

4.3.4 Support Vector Regression (SVR)

Support Vector Regression performed well, with an R² value of 0.86, but was not as effective as Gradient Boosting or Random Forest. SVR's performance indicates its capability to handle non-linear relationships, but it may require more fine-tuning of hyperparameters to achieve optimal results (Smola & Schölkopf, 2004).

4.4 Feature Importance

Understanding feature importance is crucial for interpreting the model and identifying key factors influencing air quality. Feature importance was derived from the Random Forest and Gradient Boosting models.

Table4.2: Feature Importance Scores

|  |
| --- |
| | Feature | Importance (Random Forest) | Importance (GBM) | |
| | PM2.5 | 0.30 | 0.32 | |
| | PM10 | 0.25 | 0.24 | |
| | NO2 | 0.15 | 0.18 | |
| | SO2 | 0.10 | 0.09 | |
| | CO | 0.08 | 0.07 | |
| | O3 | 0.07 | 0.05 | |
| | Temperature | 0.03 | 0.03 | |
| | Humidity | 0.02 | 0.02 | |

The results in Table4.2 show that PM2.5 and PM10 are the most important features for predicting air quality, followed by NO2. This aligns with environmental studies highlighting the significant impact of particulate matter on air quality (WHO, 2016).

4.5 Discussion

The superior performance of the Gradient Boosting model can be attributed to its ability to correct errors iteratively and focus on difficult cases. The Random Forest model also performed well due to its ensemble approach, which reduces overfitting and enhances generalization. Linear Regression, while useful for establishing a baseline, proved inadequate for capturing the complexity of the data. Support Vector Regression, although effective, did not surpass the performance of ensemble methods.

4.6 Conclusion

This chapter presented the results and analysis of different machine learning models for predicting air quality. The Gradient Boosting model emerged as the best performer, followed by Random Forest Regression. The analysis also highlighted the importance of features such as PM2.5, PM10, and NO2 in predicting air quality. These findings provide a foundation for further refinement and application of machine learning models in air quality prediction.

References

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

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 1189-1232.

Seber, G. A. F., & Lee, A. J. (2012). Linear Regression Analysis. John Wiley & Sons.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14(3), 199-222.

World Health Organization (WHO). (2016). Ambient (outdoor) air quality and health. Retrieved from [https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health).