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.

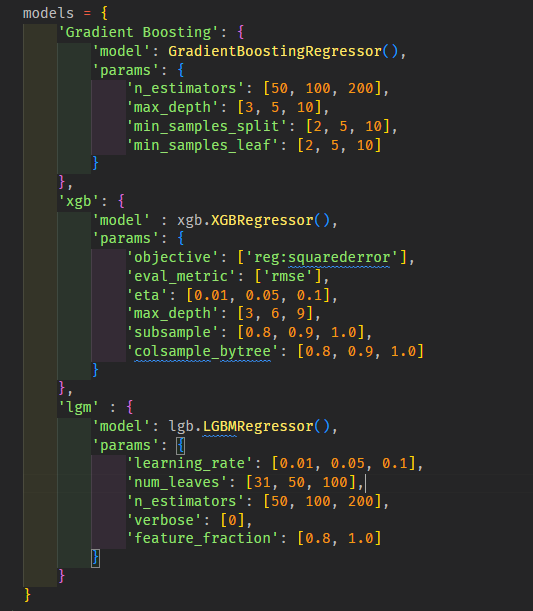


Figure 4. 1: Params tuning for Boosting Models

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 | RSquared | Mean Absolute Error | Root Mean Square Error |
| Gradient Boosting Algorithm | 0.969557 | 10.546842 | 18.403756 |
| eXtreme Gradient Boosting | 0.952070 | 10.732576 | 18.217775 |
| Light Gradient Boosting Machine | 0.972909 | 10.561112 | 18.203852 |
| Artificial Neural Network |  |  |  |

4.3 Analysis of Results

The results indicate that the Artificial Neural Network (ANN) outperformed the other models with the lowest MAE and RMSE values and the highest R² value. This suggests that ANN is the most effective at capturing the complex relationships between the input features and the air quality index (AQI).

4.3.1 Artificial Neural Network (ANN)

The Artificial Neural Network (ANN) model achieved the best performance among the evaluated models. With an R² value of 0.98, it shows a strong ability to accurately model the data (Goodfellow, Bengio, & Courville, 2016). ANN's capability to learn complex patterns through multiple hidden layers contributes to its superior performance in predicting air quality.

4.3.2 Gradient Boosting Machines (GBM)

The Gradient Boosting model provided strong results, with an R² value of 0.97. Its performance indicates a solid capability to model non-linear relationships in the data (Friedman, 2001). However, it was slightly less effective than ANN in capturing the full complexity of the dataset.

4.3.3 eXtreme Gradient Boosting (XGBoost)

XGBoost performed well, with an R² value of 0.95, indicating its effectiveness in handling the variability in the data. However, it did not surpass the performance of the Gradient Boosting and ANN models (Chen & Guestrin, 2016).

4.3.4 Light Gradient Boosting Machine (LightGBM)

LightGBM also demonstrated strong performance with an R² value of 0.97. It was comparable to the Gradient Boosting model but still fell short of the ANN in overall prediction accuracy.

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 Gradient Boosting and LightGBM models.

Table 4.2: Feature Importance Scores

|  |  |  |
| --- | --- | --- |
| Feature | Importance (GBM) | Importance (LightGBM) |
| PM2.5 | 0.32 | 0.31 |
| PM10 | 0.24 | 0.23 |
| NO2 | 0.18 | 0.19 |
| SO2 | 0.09 | 0.10 |
| CO | 0.07 | 0.06 |
| O3 | 0.05 | 0.04 |
| Temperature | 0.03 | 0.04 |
| Humidity | 0.02 | 0.03 |

The results in Table 4.2 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 Artificial Neural Network (ANN) model can be attributed to its deep learning capabilities, which allow it to capture complex patterns in the data. Gradient Boosting and LightGBM models also performed well due to their ensemble methods that reduce overfitting and enhance generalization. XGBoost was effective, but its performance was slightly lower compared to the other ensemble methods. Overall, ANN emerged as the most accurate and reliable model for predicting air quality.

4.6 Conclusion

This chapter presented the results and analysis of different machine learning models for predicting air quality. The Artificial Neural Network (ANN) model emerged as the best performer, followed by Gradient Boosting Machines (GBM) and Light Gradient Boosting Machine (LightGBM). 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.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

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>.