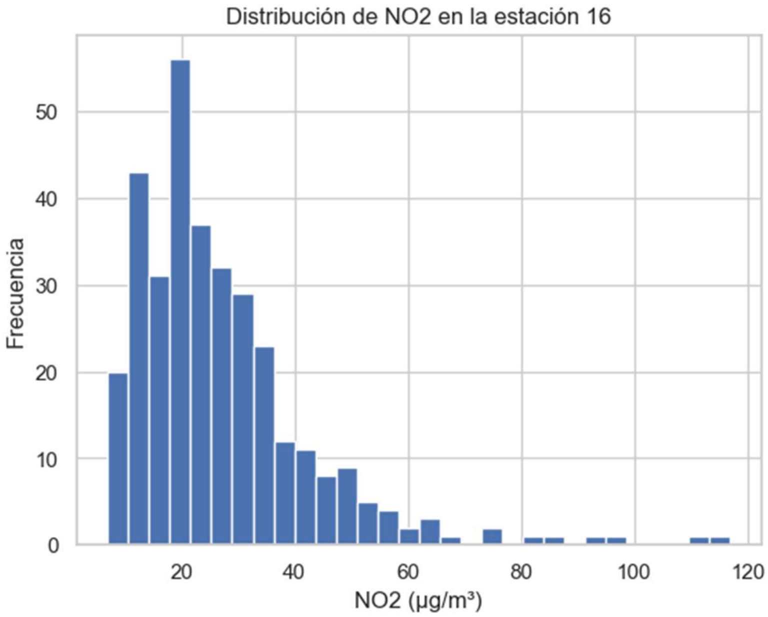
Classification Using Decision Trees

1. Introduction

This report describes the process carried out to apply classification models based on decision trees to the analysis of air quality data in Madrid. The Python library **scikit-learn** was used, and the focus was placed on the variable **NO2**, which was discretized into three levels: low, medium, and high. Throughout the report, the steps are detailed from data preparation to the comparison of different classification models.

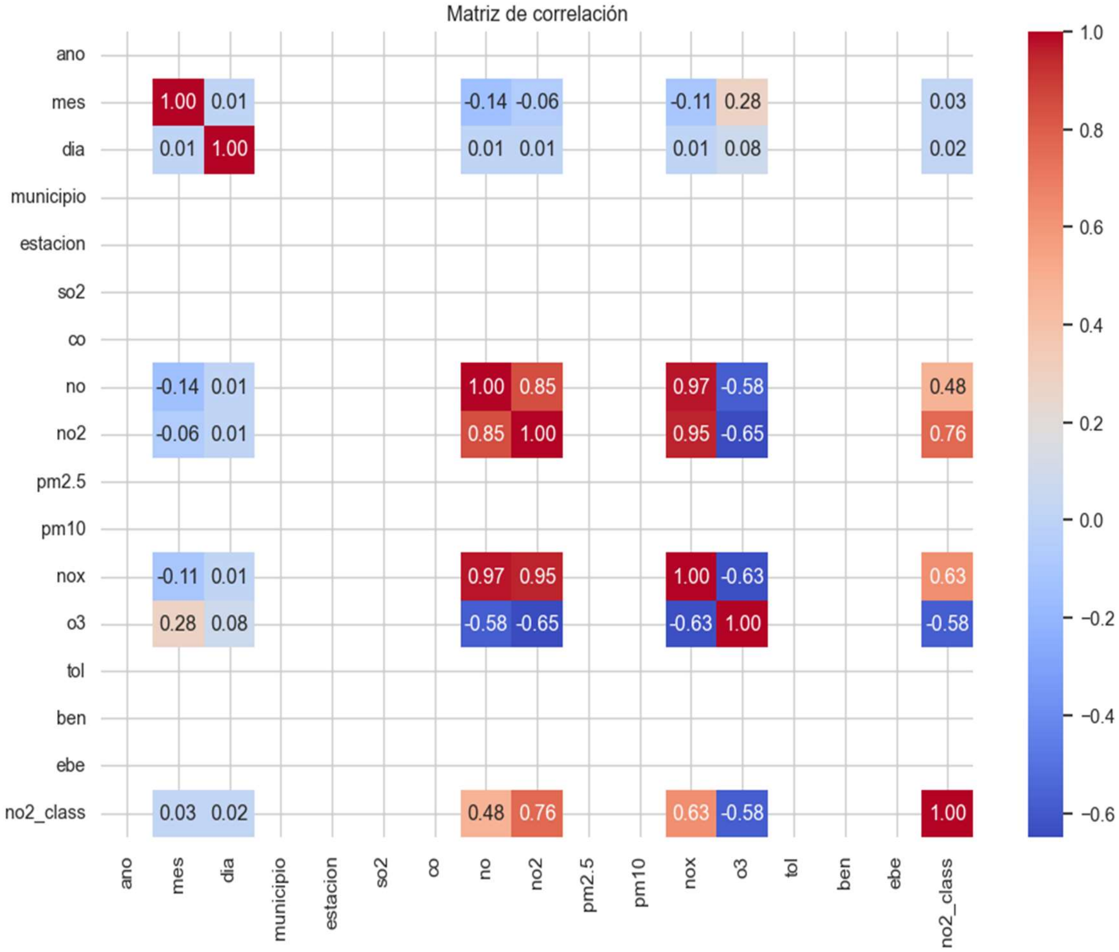
2. Data Preparation

A dataset containing daily measurements of pollutants was used. **Station number 16**, which had the highest number of records (334), was selected. The target variable, **NO2**, was discretized using the **quantile-based technique** to obtain three classes. Records with missing values were removed, and the most relevant numerical columns were retained.



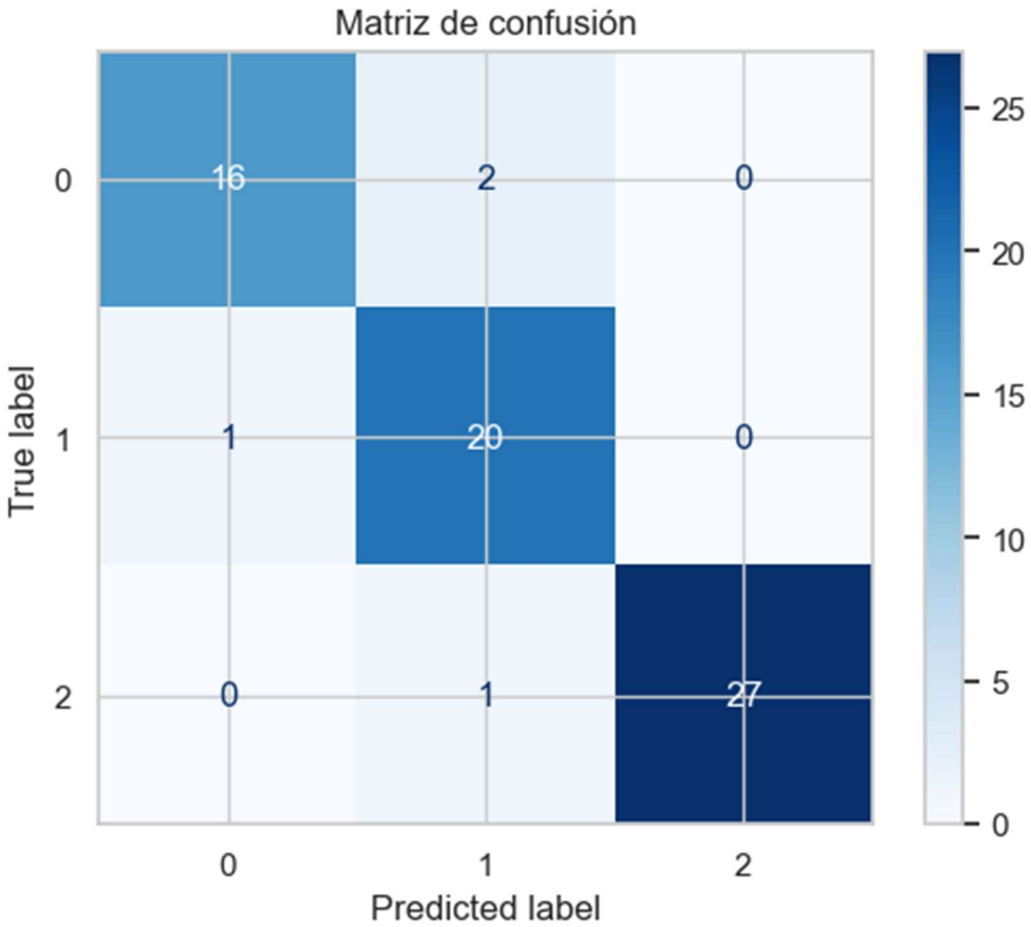
3. Correlation Analysis and Feature Selection

To select the predictor attributes, a **correlation matrix** was generated. It was observed that the variables most correlated with **NO2** were 'nox' (0.95), 'no', 'pm10', and 'co'. These were selected as the predictor features. The variable 'nox' was considered key and was later removed to evaluate its impact on model performance.

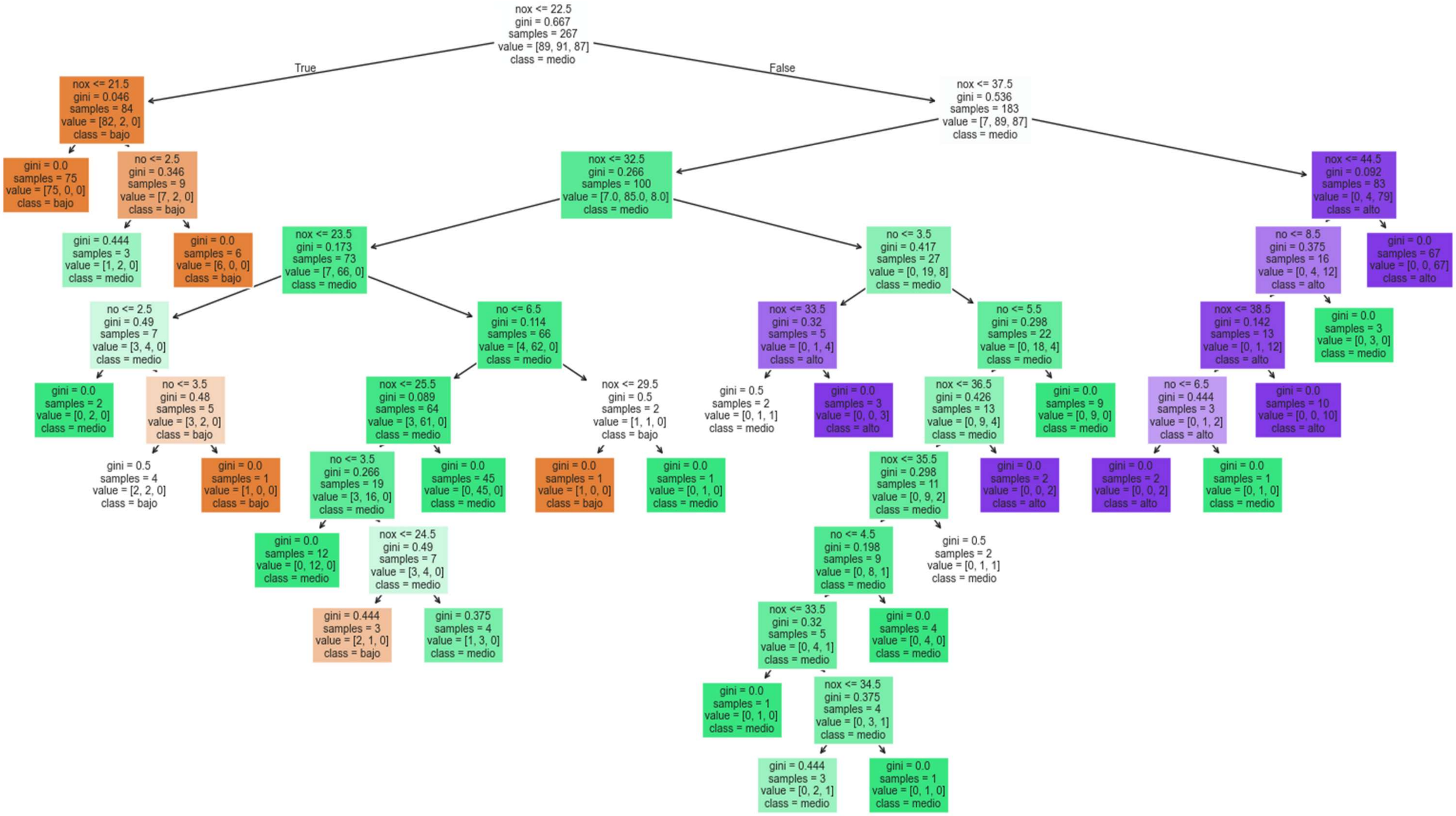


4. Training and Evaluation of the Decision Tree

A **DecisionTreeClassifier** model was trained using 80% of the dataset for training and 20% for testing. The model achieved an **accuracy of 94%** and showed a good balance across the three classes. The following figures display the resulting **confusion matrix** and the **generated decision tree**.

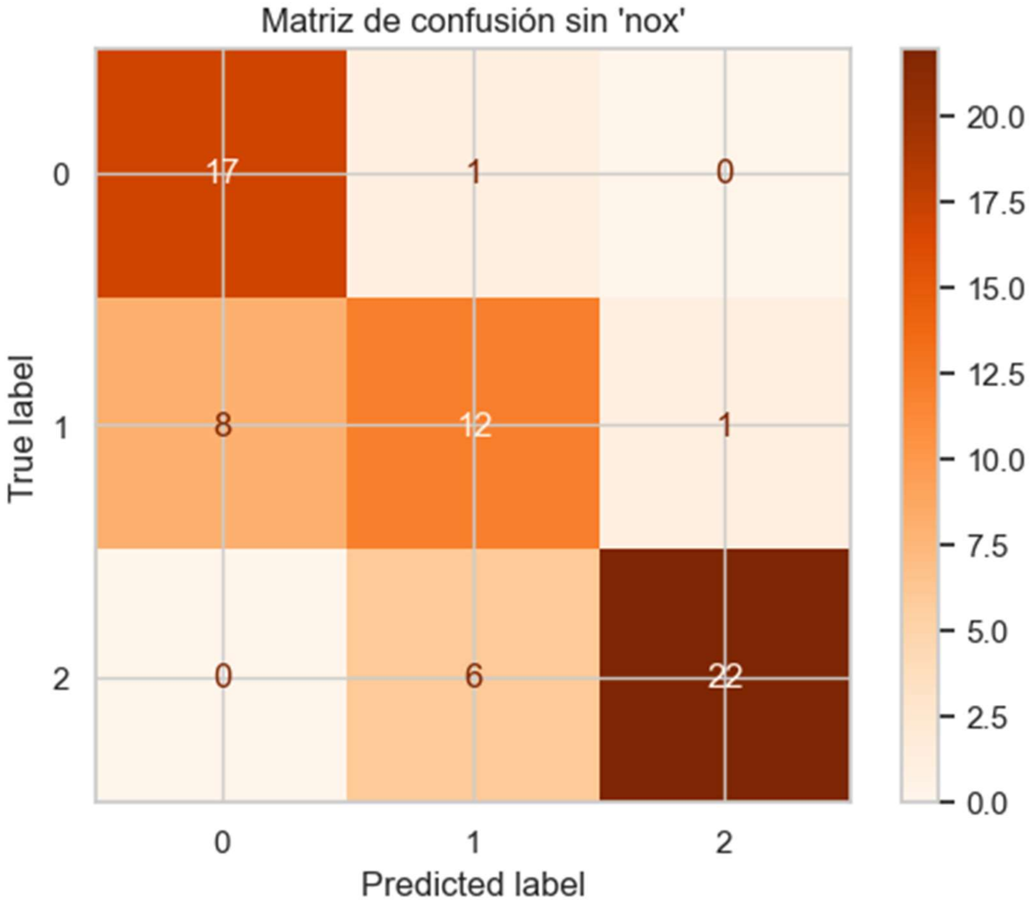


Árbol de decisión



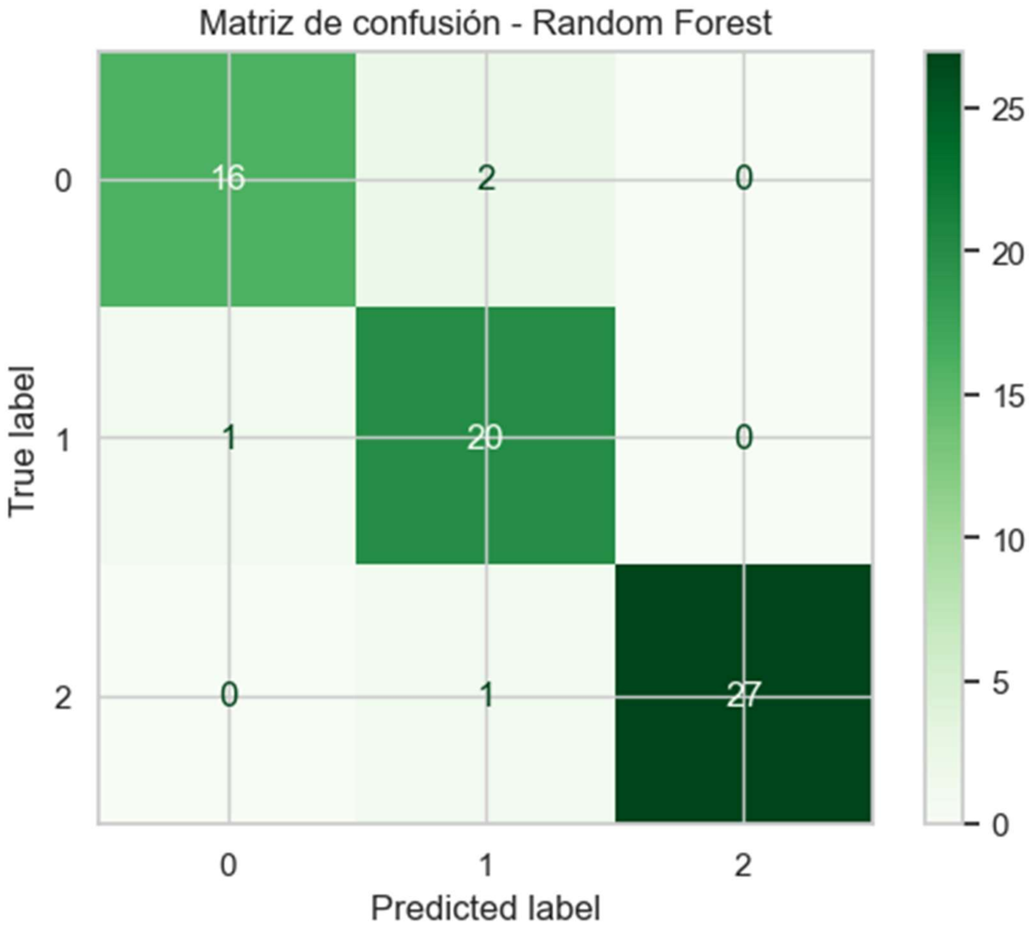
5. Evaluation After Removing the Most Correlated Variable

To analyze the impact of the most correlated variable ('nox'), it was removed from the set of features and the model was retrained. The **accuracy dropped to 76%**, demonstrating the high importance of this variable for the model. The metrics were also affected, especially for the **intermediate class**.



6. Comparison with Random Forest

A **RandomForestClassifier** model was also trained using the same initial set of features. The model achieved an **accuracy of 94%** as well, proving to be a robust alternative to the simple decision tree. **Random Forest** has the advantage of reducing overfitting by combining multiple decision trees.



7. Conclusions

Throughout this analysis, it has been demonstrated that classification models based on decision trees, such as **DecisionTreeClassifier** and **RandomForestClassifier**, are effective tools for predicting air quality based on various pollutant variables. The **correlation analysis** was key in selecting the most relevant features (nox, no, pm10, co), which together enabled highly accurate predictions.

The simple decision tree model achieved **94% accuracy**, showing a good balance across the three defined **NO2 classes**. The subsequent removal of the most correlated variable (nox) resulted in a significant drop in performance (to **76%**), confirming its crucial role in the classification process.

On the other hand, the implementation of a **Random Forest** model using the same initial features also achieved high performance (**94% accuracy**), reinforcing the dataset's stability and showcasing the advantages of ensemble models in terms of **robustness and generalization**.

Overall, this experience provided hands-on application of key concepts in **supervised learning**, **data preparation**, **model evaluation**, and **result visualization**. As future improvements, it would be interesting to expand the analysis by including data from multiple stations, incorporating **meteorological variables**, or applying more advanced techniques such as **XGBoost** or **LightGBM** to compare performance.