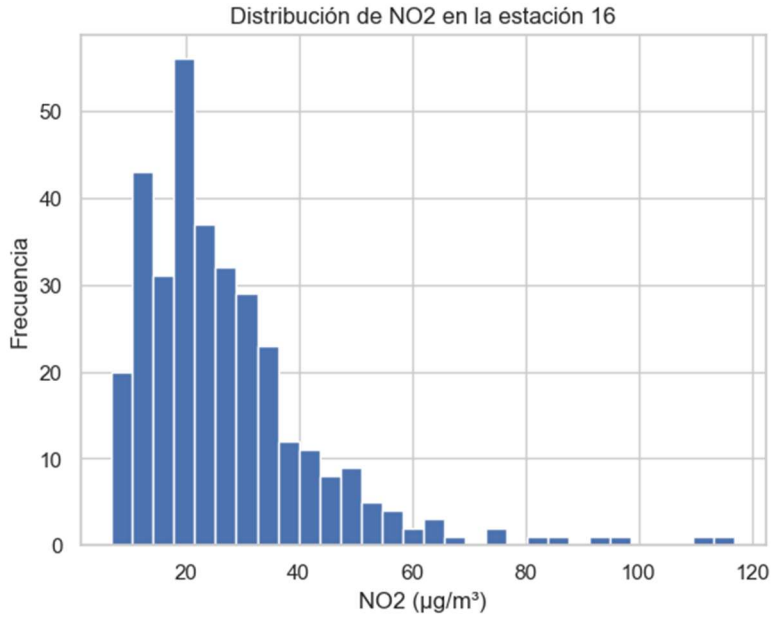
Clasificación mediante árboles de decisión

1. Introducción

Este informe describe el proceso llevado a cabo para aplicar modelos de clasificación basados en árboles de decisión al análisis de datos de calidad del aire en Madrid. Se ha utilizado la biblioteca scikit-learn de Python, y se ha trabajado con la variable NO2, discretizada en tres niveles: bajo, medio y alto. A lo largo del informe se detallan los pasos seguido desde la preparación de los datos hasta la comparación de distintos modelos de clasificación.

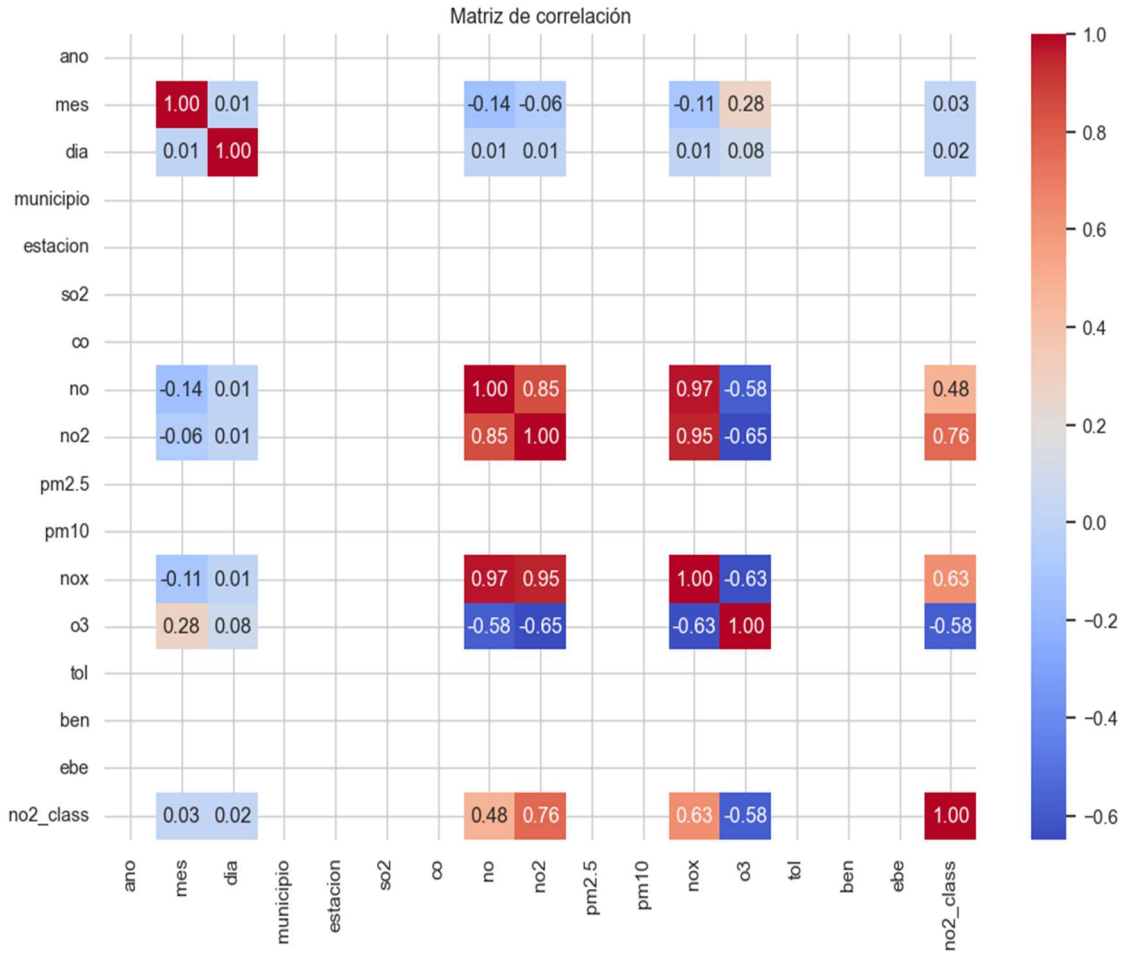
2. Preparación de los datos

Se utilizó un conjunto de datos con medidas diarias de contaminantes. Se seleccionó la estación número 16, que presentaba el mayor número de registros (334). La variable objetivo, NO2, se discretizó usando la técnica de cuantiles para obtener tres clases. Se descartaron los registros con valores nulos y se mantuvieron las columnas numéricas más relevantes.



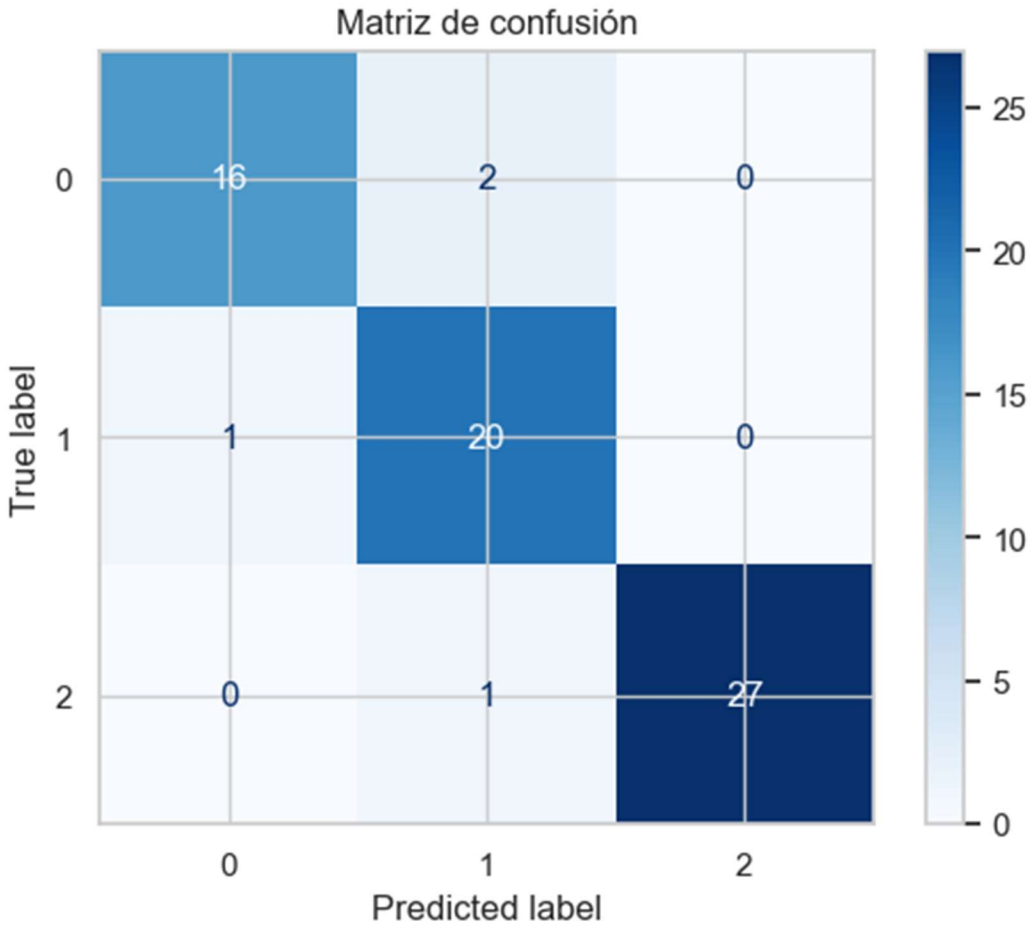
3. Análisis de correlación y selección de atributos

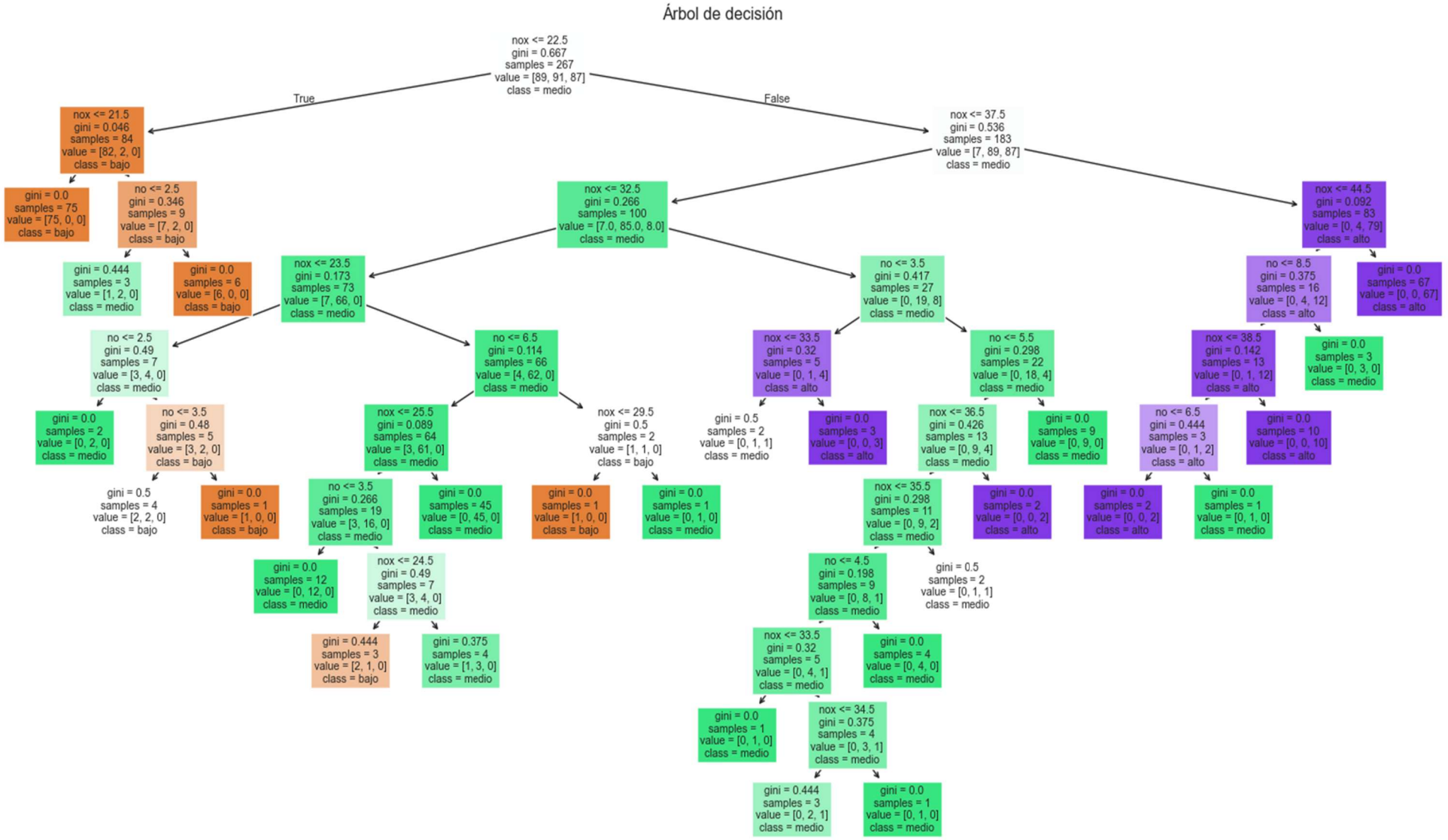
Para seleccionar los atributos predictores, se generó una matriz de correlación. Se observó que las variables más correlacionadas con NO2 eran 'nox' (0.95), 'no', 'pm10' y 'co'. Estas fueron seleccionadas como variables predictoras. La variable 'nox' se consideró clave y posteriormente se eliminó para evaluar su impacto en el rendimiento.



4. Entrenamiento y evaluación del árbol de decisión

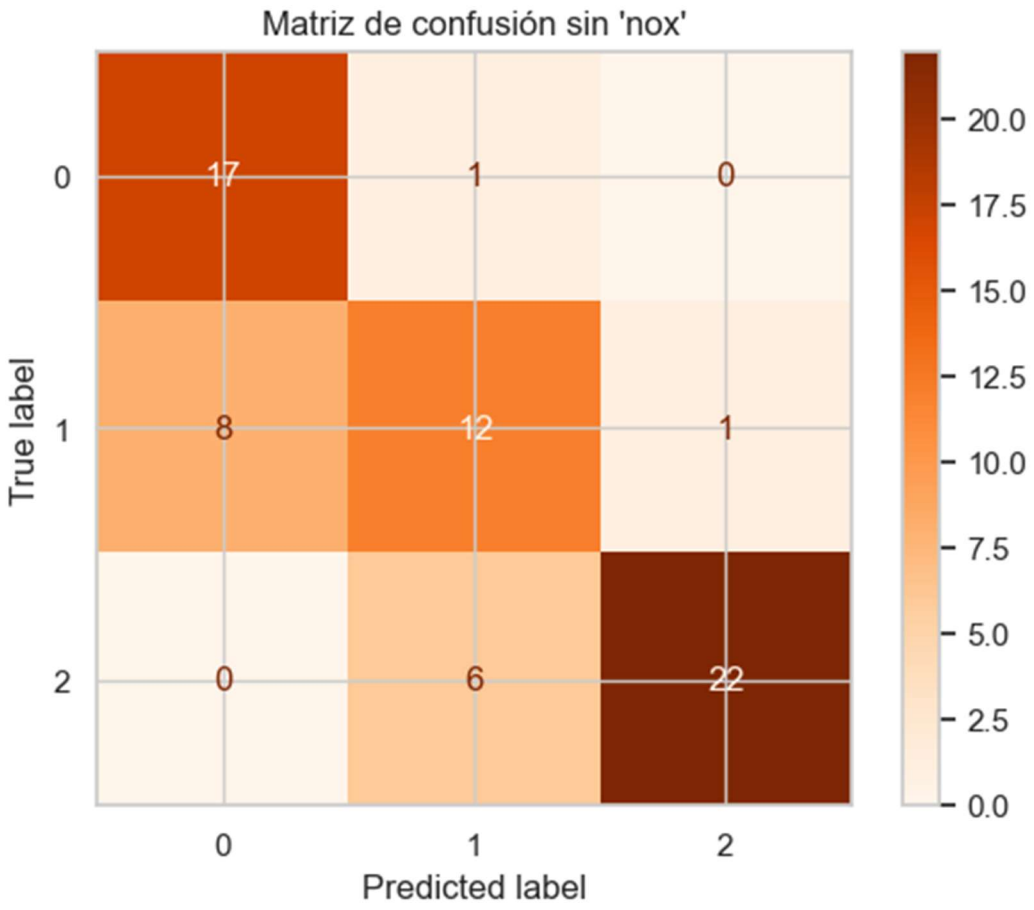
Se entrenó un modelo DecisionTreeClassifier usando un 80% del dataset para entrenamiento y un 20% para test. El modelo obtuvo un accuracy del 94% y mostró un buen equilibrio en las tres clases. A continuación, se visualiza la matriz de confusión y el árbol de decisión generado.




Árbol de decisión
nox <= 22.5
gini = 0.667
samples = 267
value = [89, 91, 87]
class = medio
True
False
nox <= 21.5
gini = 0.046
samples = 84
value = [82, 2, 0]
class = bajo
nox <= 37.5
gini = 0.536
samples = 183
value = [7, 89, 87]
class = medio
gini = 0.0
samples = 75
value = [75, 0, 0]
class = bajo
no <= 2.5
gini = 0.346
samples = 9
value = [7, 2, 0]
class = bajo
nox <= 32.5
gini = 0.266
samples = 100
value = [7.0, 85.0, 8.0]
class = medio
nox <= 44.5
gini = 0.092
samples = 83
value = [0, 4, 79]
class = alto
gini = 0.444
samples = 3
value = [1, 2, 0]
class = medio
gini = 0.0
samples = 6
value = [6, 0, 0]
class = bajo
nox <= 23.5
gini = 0.173
samples = 73
value = [7, 66, 0]
class = medio
no <= 3.5
gini = 0.417
samples = 27
value = [0, 19, 8]
class = medio
no <= 8.5
gini = 0.375
samples = 16
value = [0, 4, 12]
class = alto
gini = 0.0
samples = 67
value = [0, 0, 67]
class = alto
no <= 2.5
gini = 0.49
samples = 7
value = [3, 4, 0]
class = medio
no <= 6.5
gini = 0.114
samples = 66
value = [4, 62, 0]
class = medio
nox <= 33.5
gini = 0.32
samples = 5
value = [0, 1, 4]
class = alto
no <= 5.5
gini = 0.298
samples = 22
value = [0, 18, 4]
class = medio
nox <= 38.5
gini = 0.142
samples = 13
value = [0, 1, 12]
class = alto
gini = 0.0
samples = 3
value = [0, 3, 0]
class = medio
gini = 0.0
samples = 2
value = [0, 2, 0]
class = medio
no <= 3.5
gini = 0.48
samples = 5
value = [3, 2, 0]
class = bajo
nox <= 25.5
gini = 0.089
samples = 64
value = [3, 61, 0]
class = medio
nox <= 29.5
gini = 0.5
samples = 2
value = [1, 1, 0]
class = bajo
gini = 0.5
samples = 2
value = [0, 1, 1]
class = medio
gini = 0.0
samples = 3
value = [0, 0, 3]
class = alto
no <= 36.5
gini = 0.426
samples = 13
value = [0, 9, 4]
class = medio
gini = 0.0
samples = 9
value = [0, 9, 0]
class = medio
no <= 6.5
gini = 0.444
samples = 3
value = [0, 1, 2]
class = alto
gini = 0.0
samples = 10
value = [0, 0, 10]
class = alto
gini = 0.5
samples = 4
value = [2, 2, 0]
class = bajo
gini = 0.0
samples = 1
value = [1, 0, 0]
class = bajo
no <= 3.5
gini = 0.266
samples = 19
value = [3, 16, 0]
class = medio
gini = 0.0
samples = 45
value = [0, 45, 0]
class = medio
gini = 0.0
samples = 1
value = [1, 0, 0]
class = bajo
gini = 0.0
samples = 1
value = [0, 1, 0]
class = medio
nox <= 35.5
gini = 0.298
samples = 11
value = [0, 9, 2]
class = medio
gini = 0.0
samples = 2
value = [0, 0, 2]
class = alto
gini = 0.0
samples = 2
value = [0, 0, 2]
class = alto
gini = 0.0
samples = 1
value = [0, 1, 0]
class = medio
gini = 0.0
samples = 12
value = [0, 12, 0]
class = medio
nox <= 24.5
gini = 0.49
samples = 7
value = [3, 4, 0]
class = medio
no <= 4.5
gini = 0.198
samples = 9
value = [0, 8, 1]
class = medio
gini = 0.5
samples = 2
value = [0, 1, 1]
class = medio
gini = 0.444
samples = 3
value = [2, 1, 0]
class = bajo
gini = 0.375
samples = 4
value = [1, 3, 0]
class = medio
nox <= 33.5
gini = 0.32
samples = 5
value = [0, 4, 1]
class = medio
gini = 0.0
samples = 4
value = [0, 4, 0]
class = medio
gini = 0.0
samples = 1
value = [0, 1, 0]
class = medio
nox <= 34.5
gini = 0.375
samples = 4
value = [0, 3, 1]
class = medio
gini = 0.444
samples = 3
value = [0, 2, 1]
class = medio
gini = 0.0
samples = 1
value = [0, 1, 0]
class = medio

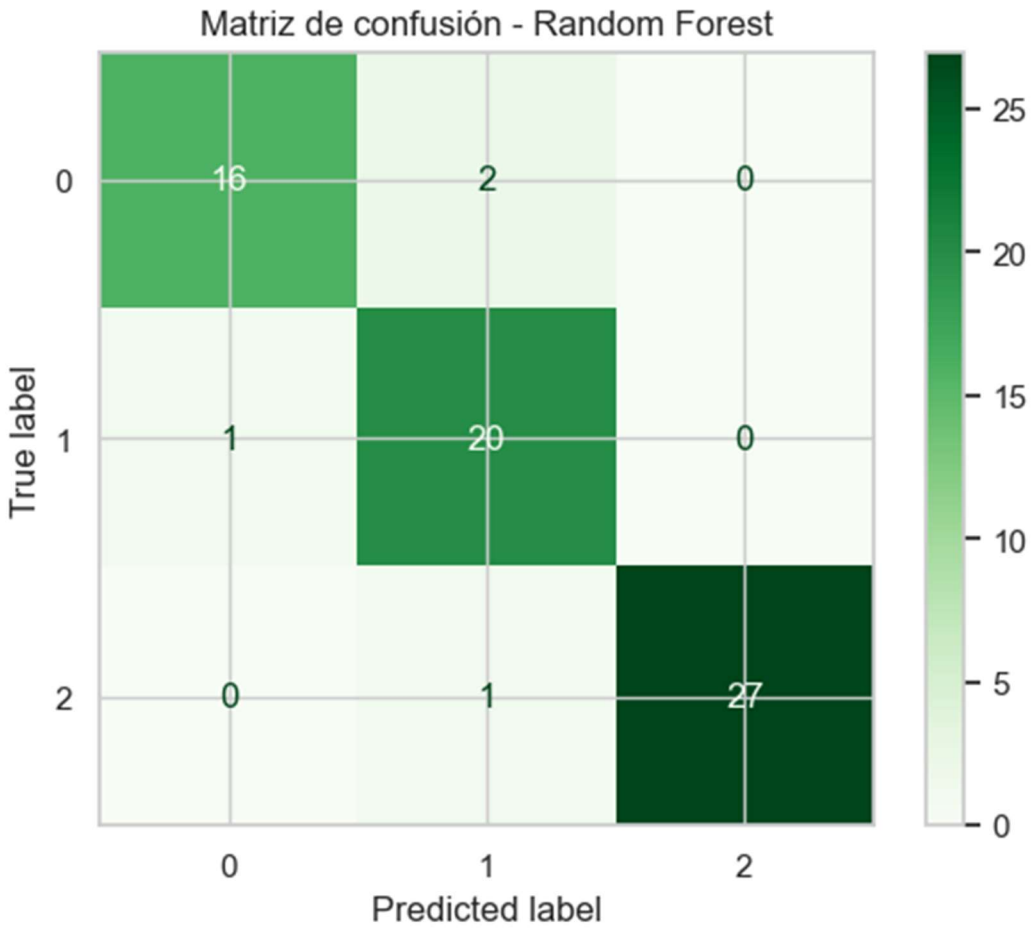
5. Evaluación tras eliminar la variable más correlacionada

Para analizar el impacto de la variable más correlacionada ('nox'), se eliminó del conjunto de atributos y se volvió a entrenar el modelo. El accuracy descendió a 76%, lo cual demuestra la alta relevancia de esta variable para el modelo. Las métricas también se vieron afectadas, especialmente en la clase intermedia.



6. Comparación con Random Forest

Se entrenó también un modelo RandomForestClassifier con los mismos atributos iniciales. El modelo obtuvo también un accuracy del 94%, demostrando ser una alternativa robusta al árbol de decisión simple. Random Forest tiene la ventaja de reducir el sobreajuste gracias a la combinación de múltiples árboles.



7. Conclusiones

A lo largo de este analisis se ha demostrado que los modelos de clasificación basados en árboles de decisión, como DecisionTreeClassifier y RandomForestClassifier, son herramientas eficaces para predecir la calidad del aire en función de diferentes variables contaminantes. El análisis de correlaciones ha sido clave para seleccionar los atributos más relevantes (nox, no, pm10, co) que, en conjunto, permiten realizar predicciones con un alto grado de precisión.

El modelo de árbol de decisión simple logró un **94% de accuracy**, mostrando un buen equilibrio en las tres clases de NO2 definidas. La posterior eliminación de la variable más correlacionada (nox) provocó una caída notable en el rendimiento del modelo (hasta el 76%), lo que confirma su papel fundamental en el proceso de clasificación.

Por otro lado, la implementación de un modelo Random Forest con los mismos atributos iniciales demostró un rendimiento igualmente alto (94%), reforzando la estabilidad del conjunto de datos y mostrando las ventajas de los modelos ensamblados en términos de robustez y generalización.

En conjunto, esta experiencia ha permitido poner en práctica conceptos fundamentales del aprendizaje supervisado, la preparación de datos, la evaluación de modelos y la visualización de resultados. Como futuras líneas de mejora, sería interesante ampliar el análisis incluyendo datos de varias estaciones, incorporar variables meteorológicas, o aplicar técnicas más avanzadas como XGBoost o LightGBM para comparar el rendimiento.