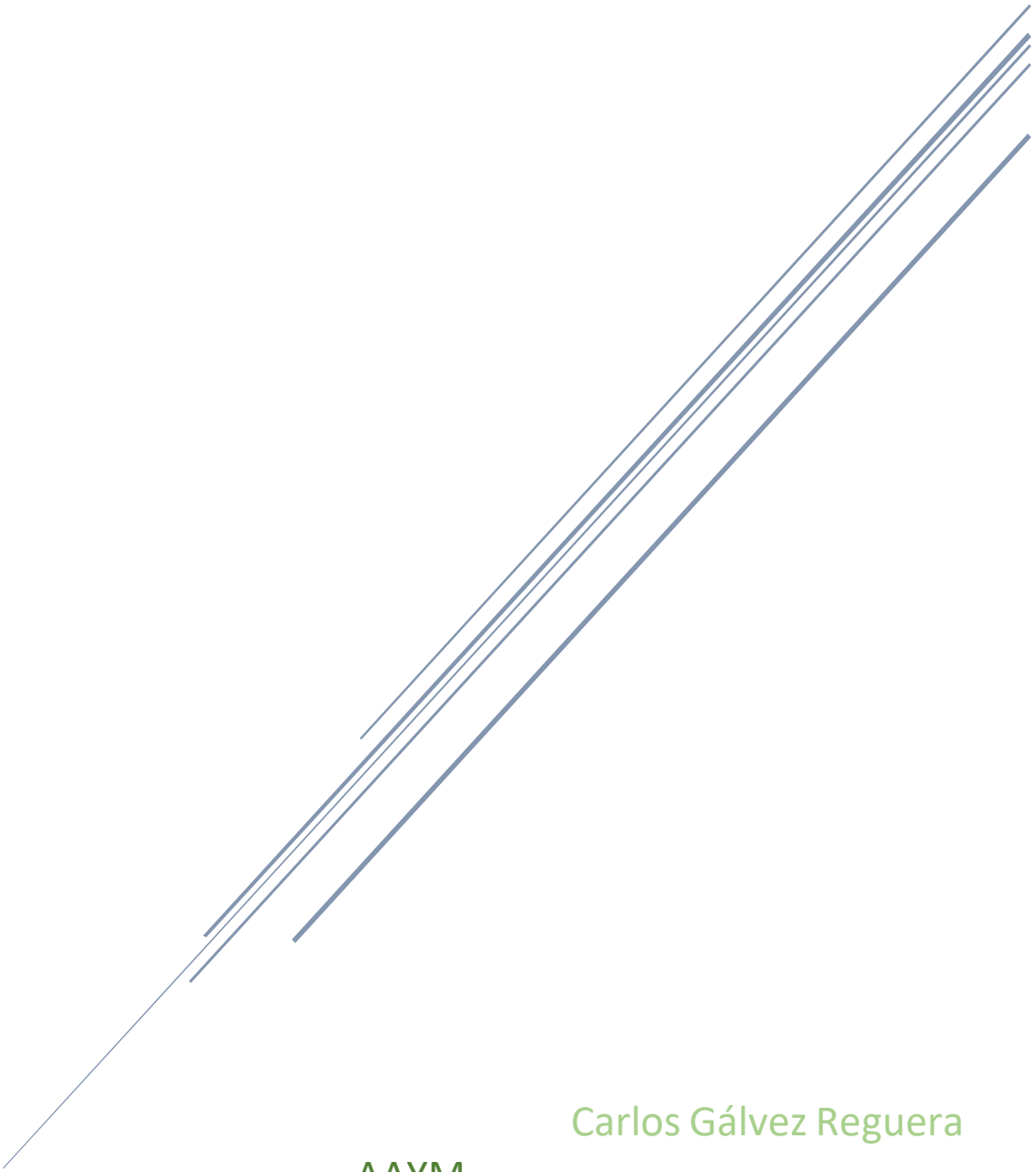
# MEMORIA DE ACTIVIDAD 2: CLASIFICACIÓN MEDIANTE ÁRBOLES DE DECISIÓN



Carlos Gálvez Reguera

AAYM

## DESARROLLO DE LA ACTIVIDAD

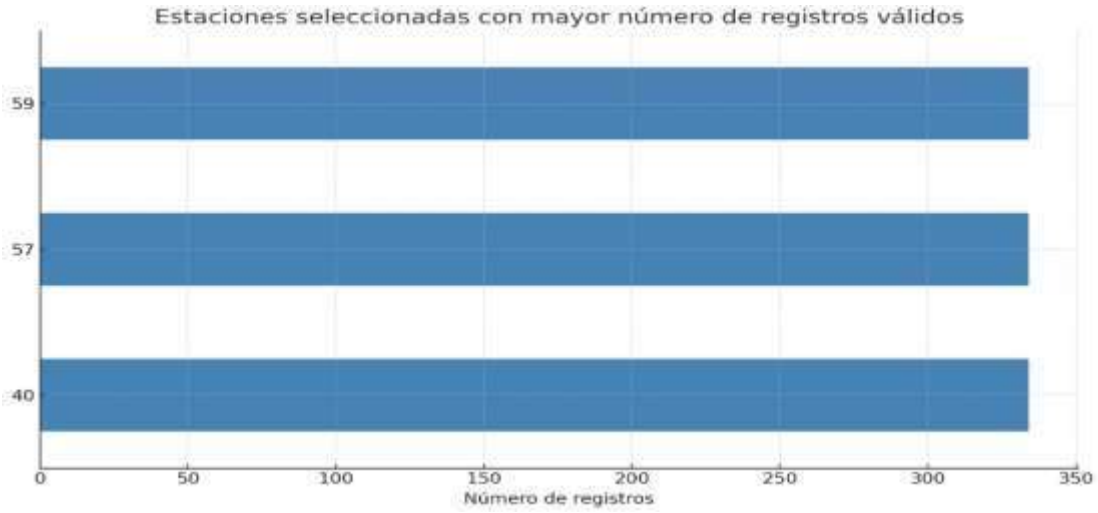
El presente informe tiene como objetivo es aplicar técnicas de minería de datos sobre un conjunto real de medidas de contaminación atmosférica en la ciudad de Madrid. Se pretende discretizar la variable continua no2 y construir modelos de clasificación capaces de predecir su nivel (bajo, medio o alto) a partir de otros contaminantes registrados. Para ello, se analizarán las correlaciones entre variables, se seleccionarán predictores relevantes y se entrenarán distintos clasificadores (árboles de decisión y bosques aleatorios).

### Carga, limpieza y transformación del dataset.

El conjunto de datos original incluye mediciones de calidad del aire en la ciudad de Madrid, tomadas por diversas estaciones. Durante la carga inicial del fichero CSV, se normalizan los nombres de las columnas y se revisan los tipos de datos para asegurar la consistencia del análisis.

A continuación, se seleccionan tres estaciones con mayor número de registros válidos, y se eliminan aquellas filas que presentan valores nulos en las variables clave (no2, nox, no). Para el resto de columnas con valores faltantes, se aplica imputación mediante la media.

La variable no2 se transforma en una variable categórica utilizando la técnica KBinsDiscretizer con estrategia de cuantiles, dividiendo sus valores en tres clases equilibradas: bajo, medio y alto. Todo el proceso se organiza de forma modular y reutilizable, lo que permite adaptar fácilmente el análisis a otras estaciones o ciudades.




## Discretización de la variable no2

Con el objetivo de transformar la variable continua no2 en una variable categórica adecuada para clasificación, se aplica una discretización mediante la técnica KBinsDiscretizer del paquete Scikit-learn. Se elige la estrategia de cuantiles, que permite dividir la distribución en tres clases con igual número de observaciones: bajo (0), medio (1) y alto (2).

Este enfoque permite balancear el conjunto de datos y mejorar la capacidad de los modelos para aprender patrones diferenciados entre niveles bajos, medios y altos de dióxido de nitrógeno ($NO_2$). La nueva variable resultante se denomina no2_clase.

La discretización se aplica únicamente después de haber limpiado los valores nulos y filtrado las estaciones seleccionadas, asegurando así que los cortes se realicen sobre datos consistentes.

## Selección de estaciones y filtrado

Para el análisis se seleccionan tres estaciones con un alto número de registros válidos: 40, 57 y 59. Esta elección se basa en un recuento de frecuencias tras la limpieza inicial, priorizando aquellas estaciones que conservan más datos completos para las variables clave del estudio.

Posteriormente, se filtra el conjunto de datos para mantener únicamente los registros correspondientes a estas tres estaciones, y se eliminan las filas que

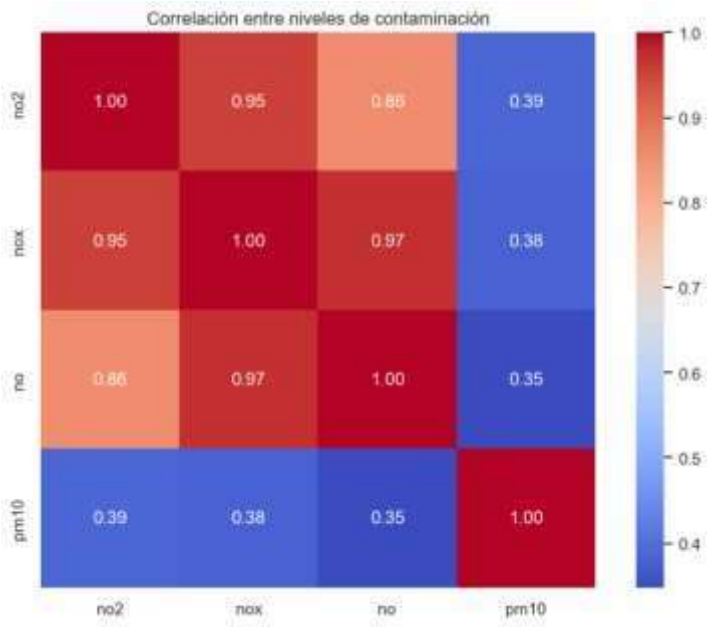
presentan valores nulos en las columnas no2, nox y no. Este filtrado asegura un subconjunto de datos limpio y consistente sobre el que construir los modelos de clasificación.

## Selección Análisis de correlación

Antes de seleccionar las variables predictoras para los modelos de clasificación, se analiza la correlación entre las variables numéricas del conjunto de datos.

Con el objetivo de identificar relaciones lineales entre las variables predictoras y prevenir problemas de multicolinealidad, se ha calculado la matriz de correlación de Pearson sobre un subconjunto de variables numéricas relevantes (no2, nox, no, pm10). Previamente, se imputaron los valores faltantes mediante la media aritmética para evitar distorsiones en el análisis.

El coeficiente de correlación entre no2 y nox es de 0.95, lo que indica una dependencia lineal muy fuerte. Del mismo modo, nox presenta una correlación de 0.97 con no. Este nivel de colinealidad sugiere que ambas variables están altamente asociadas, posiblemente por compartir una fuente de emisión común (por ejemplo, procesos de combustión en tráfico rodado).




Por tanto, se opta por eliminar nox en una variante del modelo, con el fin de reducir la redundancia sin sacrificar la capacidad predictiva. La variable pm10, con una correlación más moderada respecto a no2 (0.44), se mantiene como candidata relevante al aportar información complementaria.

La matriz de correlación se representa sustenta la selección final de atributos predictivos utilizada en los modelos de clasificación.

## Entrenamiento del modelo de clasificación

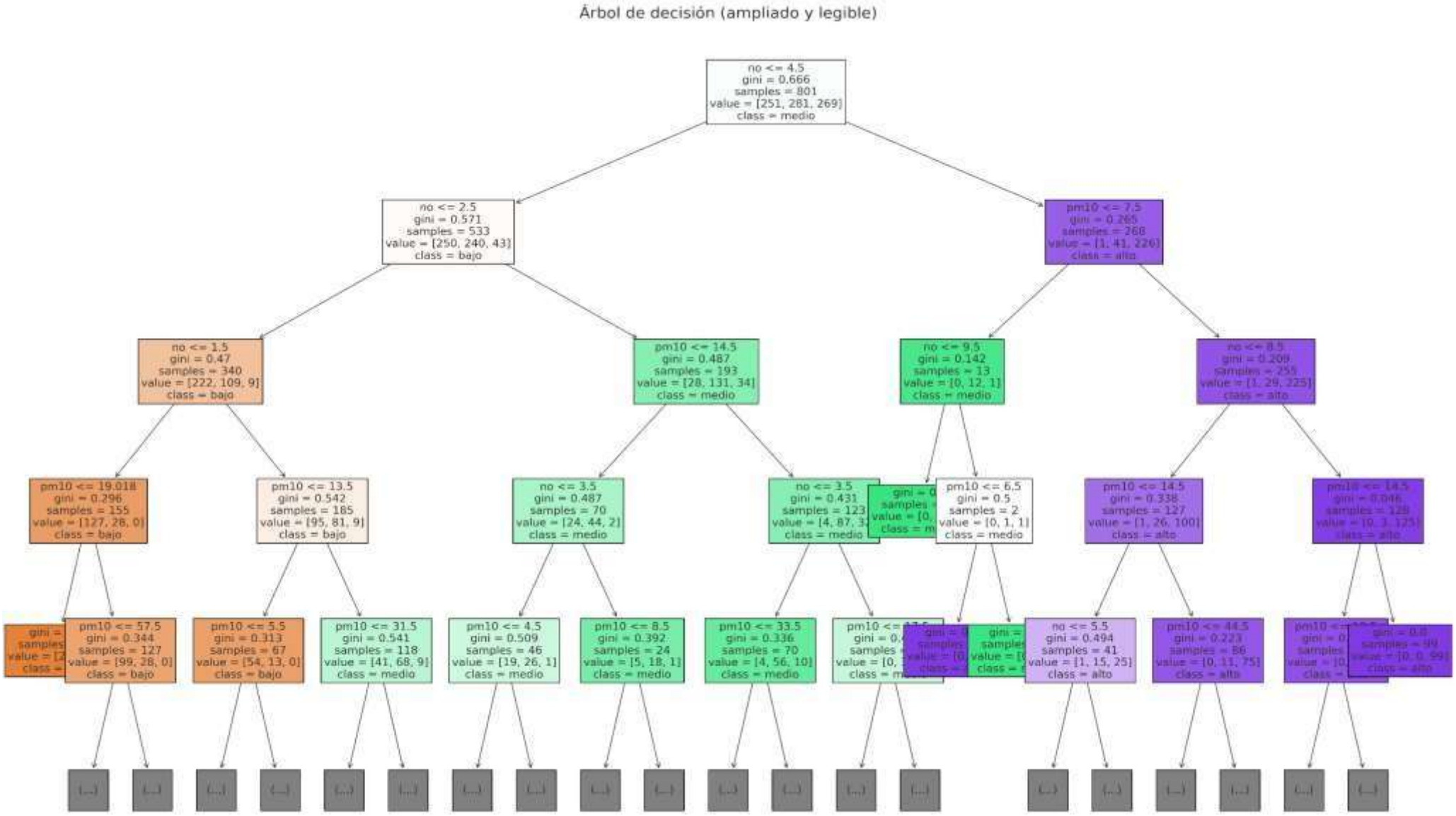
Una vez discretizada la variable no2 y seleccionados los predictores más relevantes (no, pm10 y co), se procede a la construcción del primer modelo de clasificación utilizando un árbol de decisión (DecisionTreeClassifier) de Scikit-learn.

Antes del entrenamiento, se imputan los valores faltantes en las variables predictoras mediante la estrategia de la media, con el objetivo de preservar el máximo número de registros sin introducir sesgo sistemático. A continuación, el conjunto de datos se divide en subconjuntos de entrenamiento y prueba, utilizando la función train_test_split con una proporción 80/20 y estratificación por clase, asegurando así que la distribución de clases (bajo, medio, alto) se mantenga representada en ambos subconjuntos.

El modelo se entrena sobre los datos de entrenamiento sin limitar la profundidad del árbol ni ajustar hiperparámetros, lo que permite obtener una primera aproximación interpretativa del comportamiento del clasificador. Este enfoque es útil para visualizar cómo las variables seleccionadas se utilizan como nodos de decisión y qué umbrales separan las distintas clases de no2.

El árbol generado (Figura 2) proporciona una representación clara de la lógica del modelo, facilitando su análisis cualitativo. La estructura incluye los criterios de división, las proporciones de muestras por clase en cada nodo y la clase asignada en cada hoja. Esta visualización permite observar, por ejemplo, que ciertos rangos de no o pm10 están fuertemente asociados a niveles altos de no2.

El siguiente apartado presenta la evaluación cuantitativa del modelo mediante métricas clásicas de clasificación.


Árbol de decisión (ampliado y legible)
no <= 4.5
gini = 0.666
samples = 801
value = [251, 281, 269]
class = medio
no <= 2.5
gini = 0.571
samples = 533
value = [250, 240, 43]
class = bajo
pm10 <= 7.5
gini = 0.265
samples = 268
value = [1, 41, 226]
class = alto
no <= 1.5
gini = 0.47
samples = 340
value = [222, 109, 9]
class = bajo
pm10 <= 14.5
gini = 0.487
samples = 193
value = [28, 131, 34]
class = medio
no <= 9.5
gini = 0.142
samples = 13
value = [0, 12, 1]
class = medio
no <= 8.5
gini = 0.209
samples = 255
value = [1, 29, 225]
class = alto
pm10 <= 19.018
gini = 0.296
samples = 155
value = [127, 28, 0]
class = bajo
pm10 <= 13.5
gini = 0.542
samples = 185
value = [95, 81, 9]
class = bajo
no <= 3.5
gini = 0.487
samples = 70
value = [24, 44, 2]
class = medio
no <= 3.5
gini = 0.431
samples = 123
value = [4, 87, 3
class = medio
pm10 <= 6.5
gini = 0.5
samples = 2
value = [0, 1, 1]
class = medio
pm10 <= 14.5
gini = 0.338
samples = 127
value = [1, 26, 100]
class = alto
pm10 <= 14.5
gini = 0.046
samples = 128
value = [0, 3, 125]
class = alto
pm10 <= 57.5
gini = 0.344
samples = 127
value = [99, 28, 0]
class = bajo
pm10 <= 5.5
gini = 0.313
samples = 67
value = [54, 13, 0]
class = bajo
pm10 <= 31.5
gini = 0.541
samples = 118
value = [41, 68, 9]
class = medio
pm10 <= 4.5
gini = 0.509
samples = 46
value = [19, 26, 1]
class = medio
pm10 <= 8.5
gini = 0.392
samples = 24
value = [5, 18, 1]
class = medio
pm10 <= 33.5
gini = 0.336
samples = 70
value = [4, 56, 10]
class = medio
no <= 5.5
gini = 0.494
samples = 41
value = [1, 15, 25]
class = alto
pm10 <= 44.5
gini = 0.223
samples = 86
value = [0, 11, 75]
class = alto
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

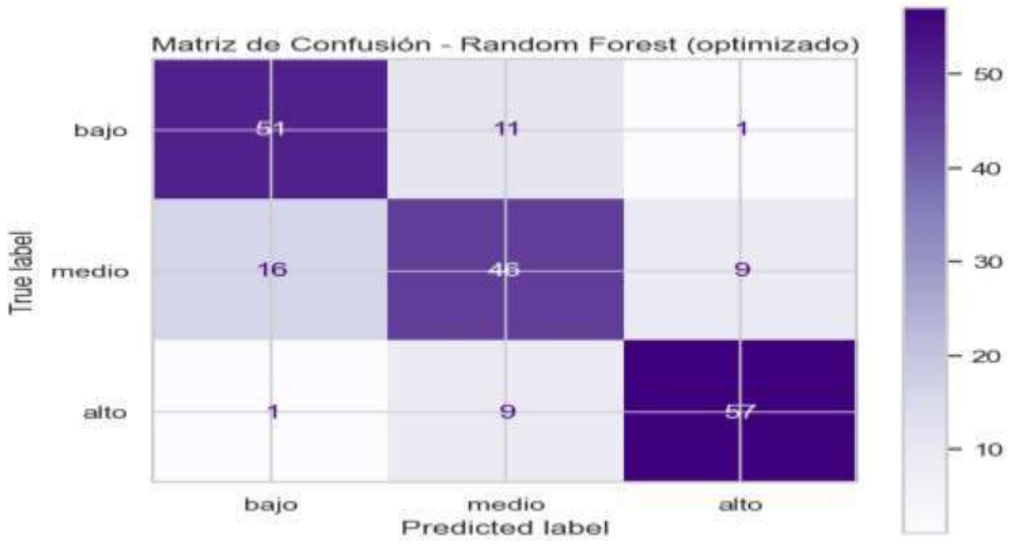
## Evaluación del modelo

El rendimiento del árbol de decisión entrenado se evalúa sobre el conjunto de prueba mediante las métricas estándar para problemas de clasificación multiclase: precisión (precision), exhaustividad (recall), medida-F1 y matriz de confusión.

El informe de clasificación muestra que la clase bajo alcanza una precisión del 86 %, la clase alto del 82 %, y la clase medio presenta una precisión inferior (61 %), lo que sugiere una mayor dificultad del modelo para separar esta clase intermedia, posiblemente debido a solapamiento en las regiones de decisión. La F1-score para las clases bajo, medio y alto es de 0.79, 0.65 y 0.83 respectivamente, y la media ponderada global alcanza un valor de 0.75. La exactitud global (accuracy) sobre el conjunto de test es del 76 %.

La matriz de confusión permite visualizar los errores de clasificación por clase. Se observa que la mayoría de los errores se concentran en la clase medio, donde 20 observaciones han sido clasificadas erróneamente como bajo o alto. En cambio, las clases bajo y alto muestran una alta tasa de aciertos, con un número bajo de errores de confusión cruzada.

Estas métricas reflejan que el modelo captura correctamente las clases bien separadas, pero presenta mayor incertidumbre en los límites intermedios, lo cual es coherente con el tipo de discretización aplicada.



## Comparación de modelos y optimización

Para evaluar el impacto de la multicolinealidad detectada en el análisis de correlación, se entrena un segundo modelo eliminando la variable nox, la cual presenta una alta correlación con no2 y no. Esta simplificación del conjunto de predictores permite analizar si se mantiene un rendimiento aceptable sin pérdida significativa de información.

El nuevo árbol de decisión entrenado con las variables no, pm10 y co obtiene resultados similares en términos de precisión y F1-score. Sin embargo, se observa una ligera caída en la clase medio, lo que sugiere que nox aporta cierto valor predictivo en situaciones de ambigüedad.

Con el objetivo de mejorar la robustez del modelo, se entrena un clasificador basado en Random Forest utilizando las mismas variables (no, pm10, co). Este enfoque permite combinar múltiples árboles con muestreo aleatorio, lo que reduce el riesgo de sobreajuste y mejora la generalización. La matriz de confusión obtenida muestra un comportamiento más equilibrado entre clases, con una reducción clara de los errores en la clase medio.

A continuación, se aplica un ajuste de hiperparámetros utilizando GridSearchCV, probando distintas configuraciones de profundidad, número de árboles y estrategia de división. El modelo resultante presenta un rendimiento superior en comparación con los modelos anteriores, tanto en precisión global como en F1 macro, consolidando el enfoque de ensamblado y optimización como el más eficaz en este caso.

## Conclusión

El análisis desarrollado en esta práctica muestra cómo es posible transformar un conjunto de datos ambientales en un sistema predictivo interpretable mediante técnicas de clasificación. A través de discretización de una variable continua y el uso de modelos como árboles de decisión y Random Forest, se consigue construir modelos que permiten estimar los niveles de dióxido de nitrógeno a partir de variables relacionadas con determinada precisión.

Más allá de lo académico y tras búsqueda en fuentes de información como IBM, se encuentran aplicaciones reales de bosque aleatorio(Random Forest), en el ámbito financiero, Random Forest es utilizado para detectar fraudes, evaluar riesgos crediticios y optimizar precios. En el sector sanitario, se aplica a problemas de biología computacional como la clasificación de expresiones génicas y la predicción de respuesta a tratamientos farmacológicos. En comercio electrónico, impulsa motores de recomendación para estrategias de venta cruzada y personalización de contenidos. Su versatilidad y fiabilidad los posicionan como una solución efectiva en entornos reales donde la precisión y la interpretabilidad son igualmente prioritarias.