Trabajo Práctico 1 - Grupo 2

EJ1: Análisis Exploratorio de Datos

Descripción del dataset:

- Cantidad de registros y columnas:
  - Dataset original: ~14.5 millones de registros, 19 columnas
  - Dataset procesado: ~13.8 millones de registros (tras limpieza), 17 columnas
- Variables más relevantes: tipo de dato, significado y por qué son importantes.

Variable	Tipo de Dato	Significado	Importancia
`tpep_pickup_datetime`	Datetime	Fecha y hora de inicio del viaje	Permite análisis temporales (patrones por hora, día, mes)
`tpep_dropoff_datetime`	Datetime	Fecha y hora de fin del viaje	Cálculo de duración y validación de datos
`trip_distance`	Float (continua)	Distancia recorrida en millas	Principal determinante de la tarifa
`fare_amount`	Float (continua)	Tarifa base calculada	Variable clave para análisis económico
`PULocationID`	Integer (categórica)	ID de zona de origen	Identifica patrones geográficos de demanda
`DOLocationID`	Integer (categórica)	ID de zona de destino	Complementa análisis de rutas frecuentes
`payment_type`	String (categórica)	Método de pago	Influye en el registro de propinas
`passenger_count`	Integer (discreta)	Número de pasajeros	Caracteriza tipo de viaje
`RatecodeID`	String (categórica)	Tipo de tarifa aplicada	Diferencia viajes especiales (aeropuertos, etc.)

`tip_amount`	Float (continua)	Monto de propina	Indicador de satisfacción (con limitaciones)

- Hipótesis iniciales o supuestos de trabajo.
  - La demanda de taxis aumenta en horarios pico (7-9 AM y 18-20 PM)
  - Los días lluviosos generan mayor cantidad de viajes
  - Existe una relación lineal entre distancia recorrida y tarifa
  - Manhattan concentra la mayor actividad de taxis
  - Los jueves son el día de mayor demanda semanal

Preprocesamiento:

- Columnas eliminadas (nombre + motivo).
  - `store_and_fwd_flag`: Información interna del sistema sin relevancia para análisis de viajes
  - `total_amount`: Redundante - es combinación lineal de otras variables (`fare_amount` + `extras` + impuestos)

- Correlaciones destacables (variables + coeficiente).
  - `fare_amount` ↔ `trip_distance` ($r \approx 0.92$)
  - `fare_amount` ↔ `total_amount` ($r = 1.00$)
  - `Airport_fee` ↔ `tolls_amount` ($r \approx 0.46$)

- Nuevos *features* generados.
  - `weekday` (String)
    - Día de la semana del viaje
    - Permite análisis de patrones semanales
    - Valores: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
  - `Rain Today` (Boolean)
    - Indica si llovió el día del viaje
    - Obtenido del dataset de Meteostat para NYC
    - Permite correlacionar demanda con condiciones climáticas
  - `pickup_hour` (Integer)
    - Hora del día (0-23) en que comenzó el viaje
    - Fundamental para identificar horarios pico
    - Facilita análisis de demanda horaria

- Valores atípicos: técnicas usadas, qué se detectó y qué decisión se tomó.
  - Técnicas utilizadas:
    1. Análisis con boxplots para visualización inicial

2. Z-Score para cuantificación estadística (umbral: $|z| > 3$)
3. Filtros basados en conocimiento del dominio

- Detección y decisiones:

Variable	Outliers Detectados	Decisión Tomada
Variables continuas	10% con $\|z\| > 3$	- Eliminados: $\|z\| > 3.5$ (~5%) - Imputados con mediana: $3 < \|z\| \leq 3.5$ (~5%)
`trip_distance`	Negativos y > 100 millas	Eliminados
`fare_amount`	Negativos y $> \$300$	Eliminados
`passenger_count`	0 y > 6 pasajeros	Reemplazados por moda
`tip_amount`	$> \$100$	Eliminados
Fechas	Fuera del período abril-junio 2024	Eliminados

- Datos faltantes: proporción y tratamiento aplicado.
  - `passenger_count`: 11.35% nulos
    - Imputación con la moda
  - `store_and_fwd_flag`: 11.35% nulos
    - Eliminada
  - `congestion_surcharge`: 11.35% nulos
    - Origen / destino en Manhattan (sur de 96th Street) → 2.50
    - Si tarifa es “Group ride” → 0.75
    - Otros casos → 0.00
  - `Airport_fee`: 11.35% nulos
    - Si `PULocationID` = JFK o LaGuardia → 1.75
    - Caso contrario → 0.00
  - `RatecodeID`: 12.45% nulos (variable con más nulos)
    - Viajes entre JFK y Manhattan → “JFK”
    - Viajes entre JFK y NYC (no Manhattan) → “Standard Rate”
    - Viajes desde o hacia Newark → “Newark”
    - Viajes desde o hacia LaGuardia → “Standard Rate”
    - Viajes fuera de Nueva York → “Negotiated Fare”

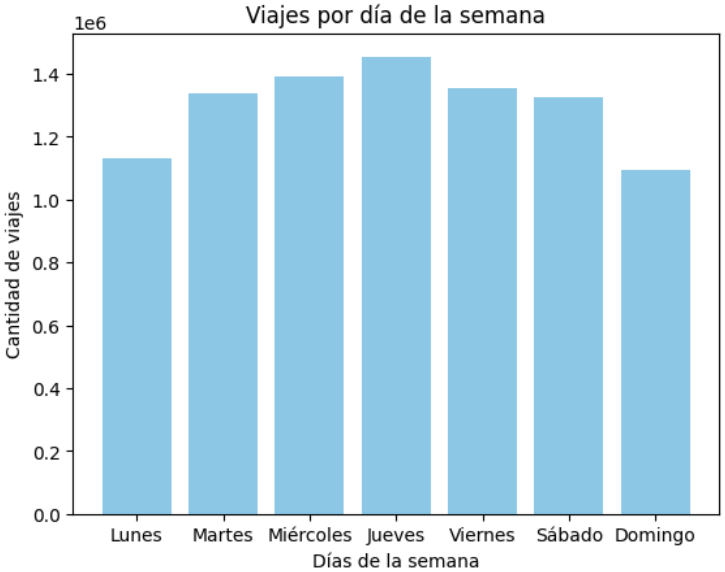
  Las primeras cuatro variables tenían nulos en las mismas filas, sugiriendo un problema sistemático de registro.

Visualizaciones y preguntas de investigación:

- ¿Los días de lluvia, aumenta la cantidad de viajes en taxi?

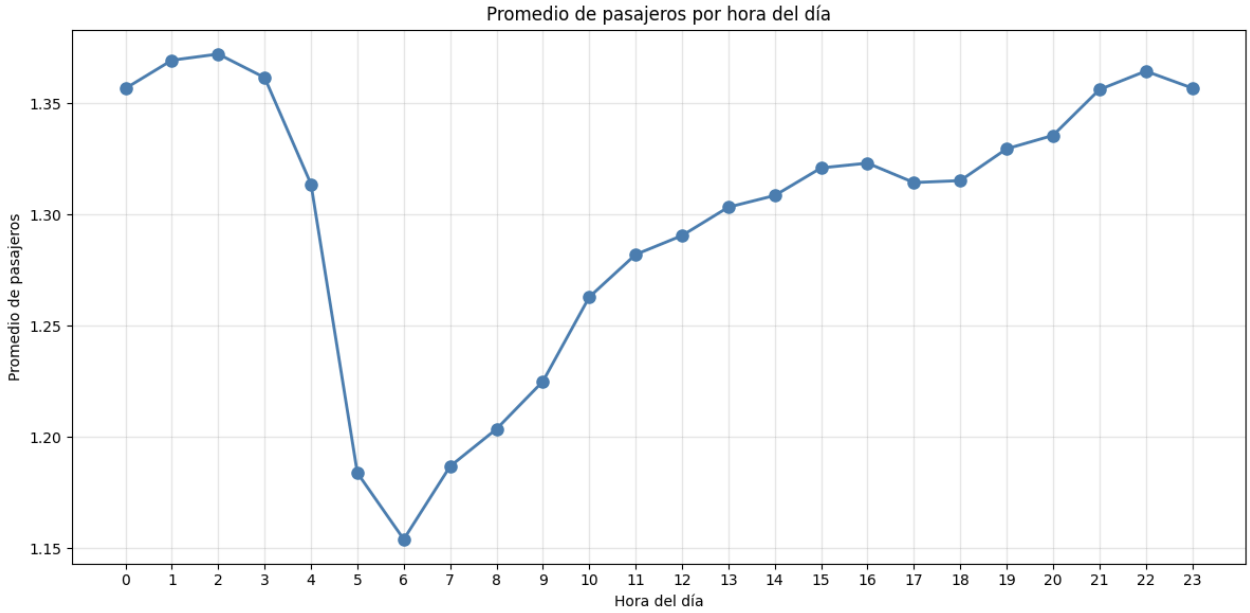
  Como muestra el gráfico podemos ver que los días donde llovió, en promedio, hubo mayor cantidad de viajes que los días que no.

- ¿Qué día de la semana hay más viajes? ¿Varía entre semana y fin de semana?



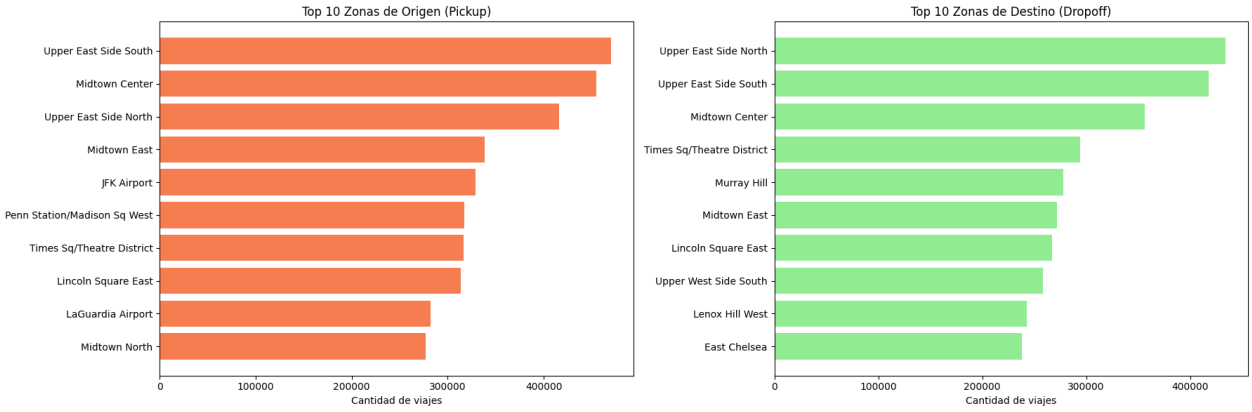
  Vemos que los jueves es el día de mayor cantidad de viajes, mientras que el domingo es el menor. Sorprende los lunes, el resto de los días es parecido.

- ¿Durante qué horario es mayor en promedio la cantidad de pasajeros?



Observamos que el promedio de pasajeros se mantiene relativamente constante a lo largo del día, con un valor cercano a 1.3-1.4 pasajeros por viaje. No se observan diferencias significativas entre horarios, lo que sugiere que la cantidad de pasajeros no varía sustancialmente según la hora del día. La mayoría de los viajes son individuales o con un acompañante.

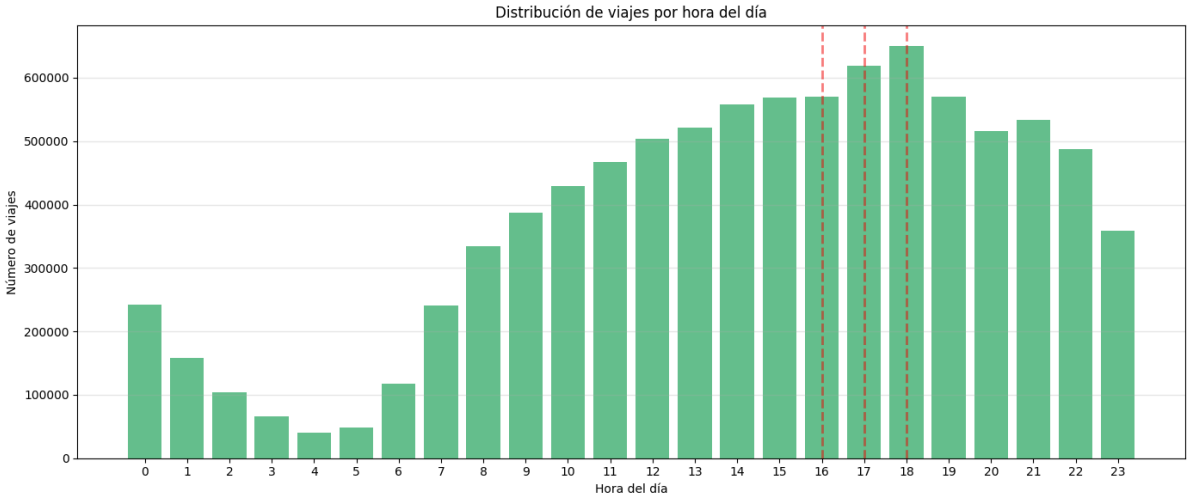
- ¿Cuáles son las zonas más frecuentes de origen y destino? ¿Hay patrones de movilidad?



Como podemos observar, las zonas más frecuentes tanto de origen como de destino se concentran en Manhattan, especialmente en áreas turísticas y comerciales. Las zonas más populares incluyen Upper East Side, Times Square/Theatre District, y Penn Station/Madison Square West. También notamos una alta frecuencia en los aeropuertos (LaGuardia y JFK), lo que

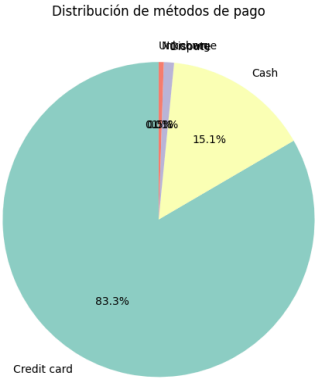
refleja el uso intensivo de taxis para traslados aeroportuarios. Este patrón es consistente entre origen y destino, sugiriendo que estas zonas actúan como hubs de movilidad en la ciudad.

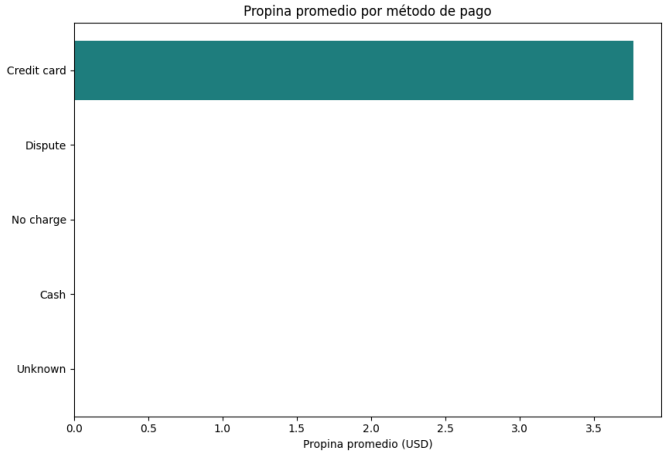
- ¿Cómo varía la demanda de taxis según la hora del día? ¿Hay horas pico?



La demanda de taxis muestra un patrón claro a lo largo del día. Las horas pico se concentran en la tarde-noche (entre las 18:00 y 20:00), con el máximo alrededor de las 19:00 horas, coincidiendo con el horario de salida laboral y actividades nocturnas. También se observa un pico significativo entre las 14:00 y 16:00 horas. La demanda más baja ocurre durante la madrugada (4:00-5:00 AM), cuando la actividad en la ciudad es mínima. Este patrón es consistente con los comportamientos típicos de una ciudad metropolitana donde los taxis son utilizados principalmente para desplazamientos laborales y de entretenimiento.

- ¿Qué método de pago es más utilizado y cómo se relaciona con las propinas?





Los resultados muestran que la **tarjeta de crédito** es el método de pago más utilizado, representando la gran mayoría de los viajes. El efectivo ocupa el segundo lugar, seguido por otros métodos menos frecuentes.

En cuanto a las propinas, existe una diferencia significativa según el método de pago: los pagos con tarjeta de crédito registran propinas con un promedio de más de $3.50. Por otro lado, los pagos en efectivo muestran propinas promedio de $0. Las categorías "No charge" y "Dispute" naturalmente tienen propinas de $0.

Este análisis revela un sesgo importante en los datos: las propinas registradas solo reflejan aquellas pagadas electrónicamente, subestimando las propinas totales realmente otorgadas a los choferes.

EJ2: Modelos de Clasificación Binaria

Descripción del dataset:

- Cantidad de registros: 71.595 (final)
- Cantidad de columnas: 23 (final)

Columnas:

Variable	Descripción	Tipo
`Date`	Fecha del registro	Cualitativa Nominal
`Location`	Localidad en la cual se midieron los datos	Cualitativa Nominal
`MinTemp`	Temperatura mínima en grados celsius	Cuantitativa Continua
`MaxTemp`	Temperatura máxima en grados celsius	Cuantitativa Continua
`Rainfall`	Cantidad en mm de precipitación que cayó	Cuantitativa Continua
`Evaporation`	Cantidad en mm de la evaporación de agua	Cuantitativa Continua
`Sunshine`	Cantidad de horas de luz solar	Cuantitativa Continua
`WindGustDir`	La dirección de la ráfaga de viento más fuerte.	Cualitativa Nominal
`WindGustSpeed`	La velocidad en km/h de la ráfaga de viento más fuerte	Cuantitativa Continua
`WindDir9am`	La dirección del viento a las 9 am	Cualitativa Nominal
`WindDir3pm`	La dirección del viento a las 3 pm	Cualitativa Nominal
`WindSpeed9am`	La velocidad en km/h del viento a las 9 am	Cuantitativa Continua

`WindSpeed3pm`	La velocidad en km/h del viento a las 3 pm	Cuantitativa Continua
`Humidity9am`	Porcentaje de humedad a las 9 am	Cuantitativa Discreta
`Humidity3pm`	Porcentaje de humedad a las 3 pm	Cuantitativa Discreta
`Pressure9am`	Presión atmosférica en hpa a las 9 am	Cuantitativa Continua
`Pressure3pm`	Presión atmosférica en hpa a las 3 pm	Cuantitativa Continua
`Cloud9am`	Octas del cielo obstruido por nubes a las 9 am	Cuantitativa Discreta
`Cloud3pm`	Octas del cielo obstruido por nubes a las 3 pm	Cuantitativa Discreta
`Temp9am`	Temperatura en grados celsius a las 9 am	Cuantitativa Continua
`Temp3pm`	Temperatura en grados celsius a las 3 pm	Cuantitativa Continua
`RainToday`	Indica si llovió el día de la fecha (Sí/No)	Cualitativa Nominal
`RainTomorrow`	Indica si llovió el día siguiente (Sí/No)	Cualitativa Nominal

Pre procesamiento de datos:

- Eliminamos los registros con la columna RainTomorrow o RainToday nulas para no afectar al modelo.
- Se imputaron los valores faltantes por la moda para variables categóricas y por la mediana para variables cuantitativas.
- Para detección de outliers se utilizó Local Outlier Factor, al ser pocos, estos fueron descartados.
- One Hot Encoding de variables categóricas para el entrenamiento de los modelos de clasificación.
- Generación de la feature **Season** en función del mes en el hemisferio sur.

Modelos entrenados:

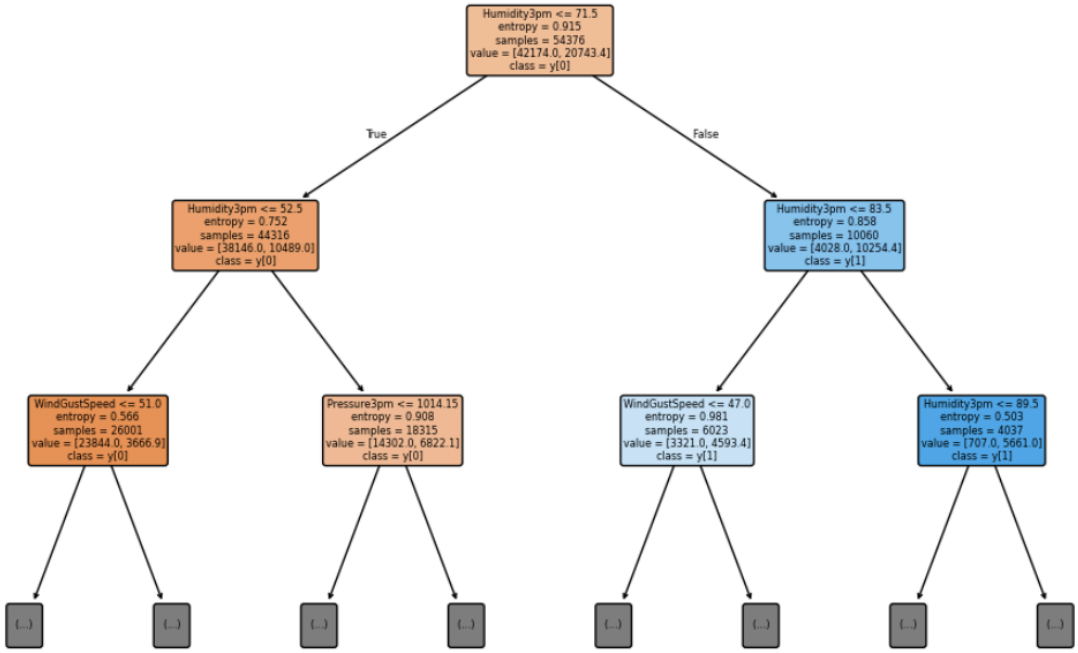
Debido a la naturaleza del dataset, la métrica elegida para elegir el mejor modelo en cada caso fue el F1-Score. Esta decisión fue tomada debido al desbalance entre clases presente en el dataset, donde aproximadamente un 78% pertenece a la clase “No llueve mañana”, mientras que un 22% pertenece a “Llueve mañana”. El F1-Score es sensible a estos desbalances ya que representa una medida armónica entre precisión y recall.

1. **Árbol de decisión:**

Hiperparametros optimizados:

{
'max_depth': 8,
'criterion': 'entropy',
’min_samples_leaf’: 29,
‘min_samples_split’ = 18
}

La técnica de validación utilizada fue K-Fold CV, con K = 5.



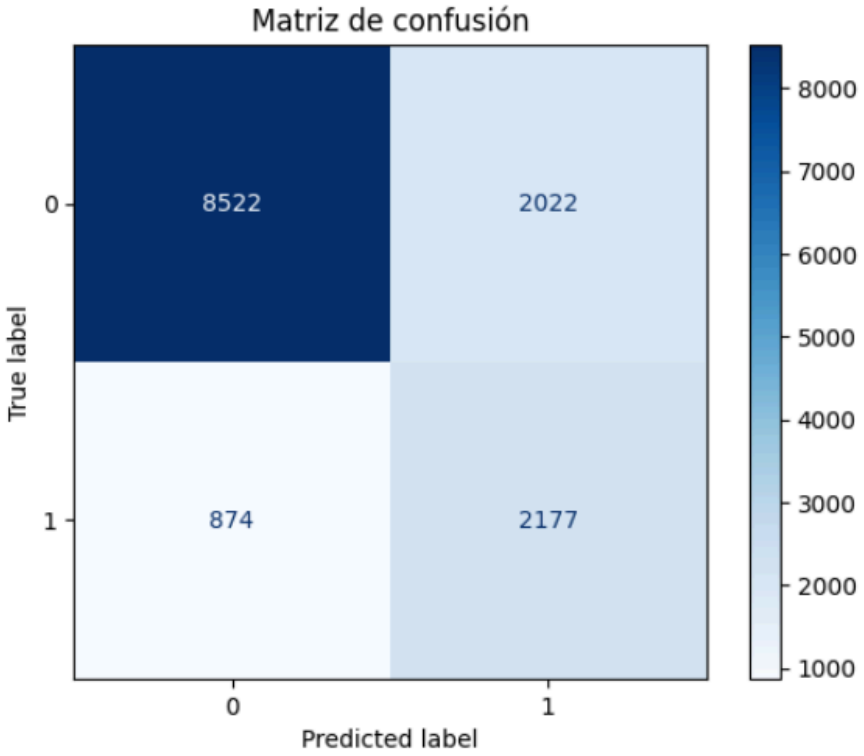
Esta imagen representa las primeras divisiones del árbol de decisión.

Podemos ver que si la humedad a las 3 pm es alta (mayor a 71.5), entonces es muy probable que al día siguiente llueva, esto resulta lógico ya que la humedad en el momento más tardío del cual se tiene registro es un gran predictor.

Cuadro comparativo, Train vs Test:

	Accuracy	Precision	Recall	F1-Score
Train	0.802	0.542	0.749	0.629
Test	0.787	0.518	0.714	0.601

Vemos que ambos valores de F1-Score son relativamente bajos, esto indica un posible underfitting del modelo, es decir, el modelo no es capaz de predecir correctamente, muy posiblemente debido a que es demasiado simple para el problema.

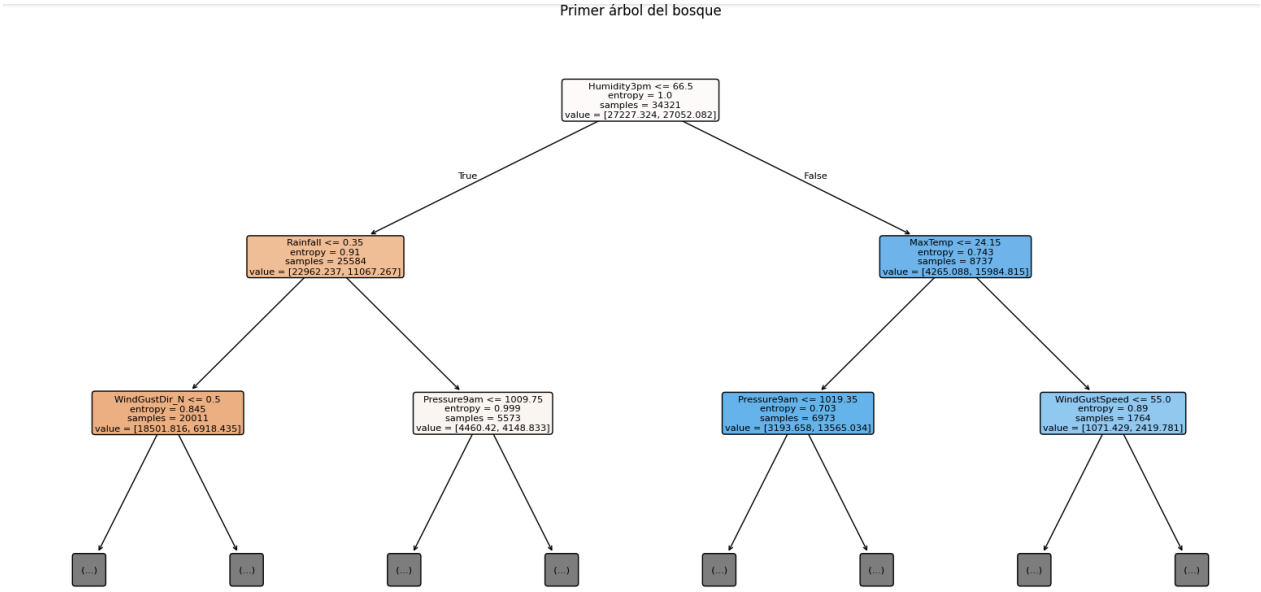


2. Random Forest:

Hiperparametros optimizados:

{
'criterion': 'entropy',
'max_depth': 18,
'min_samples_leaf': 5,
'min_samples_split': 5,
'n_estimators': 120
}

La técnica de validación utilizada fue K-Fold CV, con K = 5.



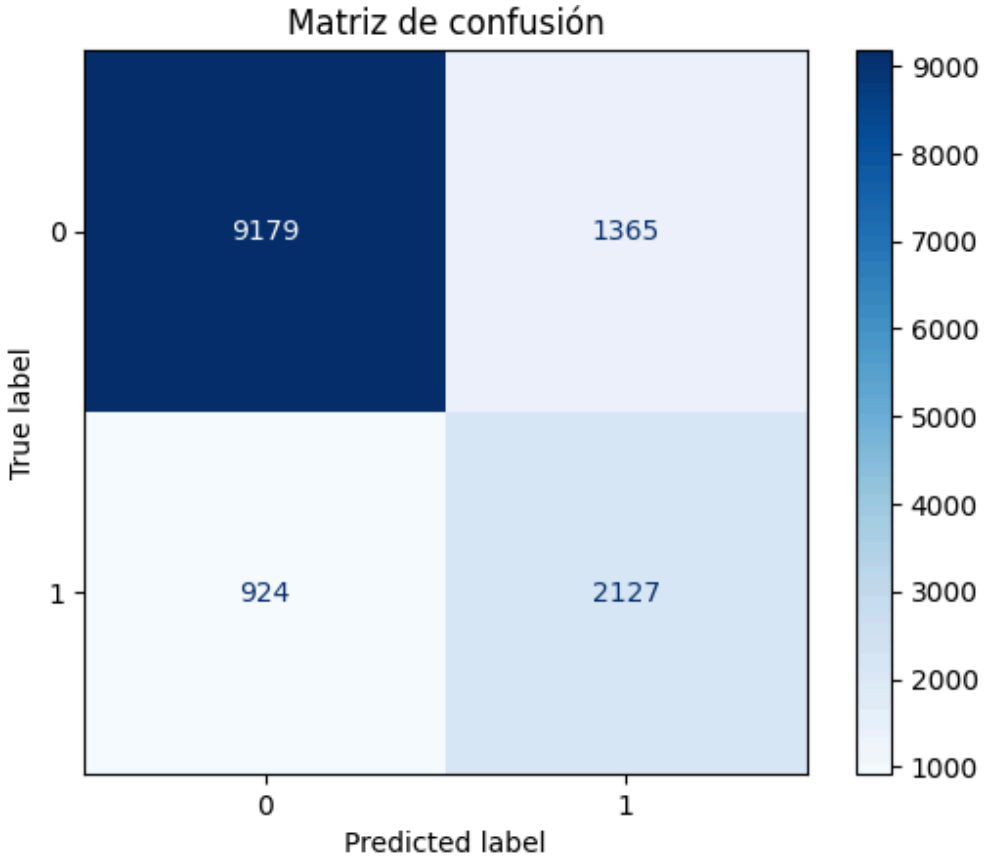
Esta imagen representa las primeras divisiones de uno de los árboles de Random Forest. Vemos que la primer variable predictora vuelve a ser la humedad a las 3 pm, sin embargo, en los niveles inferiores aparecen otras nuevas, tales como rainfall y

pressure.

Cuadro comparativo, Train vs Test:

	Accuracy	Precision	Recall	F1-Score
Train	0.9	0.74	0.853	0.792
Test	0.832	0.609	0.697	0.65

En este cuadro vemos que el F1-Score en train es bastante más elevado que el de test, lo que indica un posible overfitting del modelo.
Sin embargo, se observa una mejora en el F1-Score de test en comparación con el Árbol de Decisión previamente entrenado, lo cual se debe a que Random Forest combina múltiples árboles entrenados reduciendo la varianza y haciendo el modelo más robusto y generalizable.



3. XGBoost:

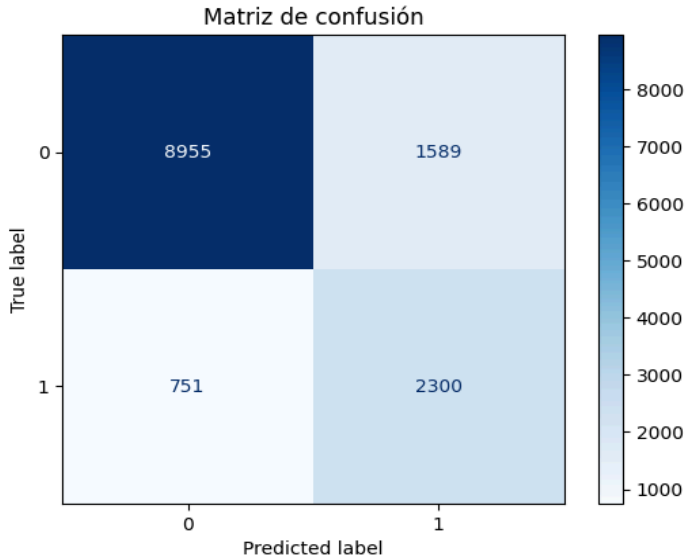
Hiperparametros optimizados:
{
'n_estimators': 200,
'learning_rate': 0.1,
'max_depth': 7,
'subsample': 0.8
}

La técnica de validación utilizada fue K-Fold CV, con K = 5.

Cuadro comparativo, Train vs Test:

	Accuracy	Precision	Recall	F1-Score
Train	0.894	0.705	0.91	0.794
Test	0.828	0.591	0.754	0.663

XGBoost posee el F1-Score más alto entre los modelos, aunque debido a la diferencia entre Train y Test, indica un overfitting.



Cuadro comparativo de resultados (TEST):

Modelo	F1	Precision	Recall	Accuracy
Arbol de Decision	0.601	0.518	0.714	0.787
Random Forest	0.65	0.609	0.697	0.832
XGBoost	0.663	0.591	0.754	0.828

Elección del modelo

En base a los resultados obtenidos, el modelo que elegimos para predecir la lluvia es XGBoost. Esto se debe a que este modelo posee el F1-Score más alto entre los modelos entrenados, además de tener el recall más elevado. Esto causa que el número de falsos negativos sea bajo, lo cual es acertado en este caso, ya que es peor no detectar una lluvia y que en realidad llueva.

EJ3: Modelos de Regresión

Descripción del dataset: registros, columnas, variables destacables y transformaciones realizadas.

Modelos entrenados:

1. **Regresión lineal múltiple**
   - Variables utilizadas.
   - Evaluación en train/test (MSE, RMSE, otras métricas).
2. **XGBoost**
   - Cross-validation, hiperparámetros y métrica optimizada.
   - Evaluación en train/test.

3. **Modelo a elección**
   - Justificación y métricas de evaluación.

Ingeniería de Features y Preprocesamiento de Datos

En primer lugar, se realizó una selección de variables eliminando aquellas columnas que no aportaban información relevante para la predicción del precio de alquiler, tales como identificadores, URLs y atributos que contenían información no disponible en una instancia futura de predicción (por ejemplo, datos derivados de evaluaciones posteriores, información sobre reviews si no existía antes).

Posteriormente, se abordó el tratamiento de las variables cualitativas nominales. Para este tipo de variables se aplicó One Hot Encoding, tanto en aquellas con pocas categorías como en las de alta cardinalidad, limitando en este último caso la codificación a las N categorías más frecuentes. En el caso de las variables relacionadas con la ubicación geográfica, se optó por Label Encoding, con el fin de conservar una representación más compacta y evitar una expansión excesiva del espacio de características.

Los valores faltantes se completaron según la naturaleza de la variable: en algunos casos con ceros (por ejemplo, conteos o indicadores binarios), y en otros con la media del atributo correspondiente.

Para garantizar una escala uniforme entre variables y mejorar la estabilidad de los modelos, se aplicó normalización mediante Standard Scaler (basado en el z-score) para aquellas variables con distribución aproximadamente normal. En los casos con alta presencia de outliers, se utilizó Robust Scaler, que escala los datos en función del rango intercuartílico.

Finalmente, se realizó una detección y tratamiento de valores atípicos, aplicando una transformación logarítmica para reducir la asimetría de las distribuciones y posteriormente recortando los valores extremos a los percentiles 1 y 99, con el objetivo de limitar el impacto de observaciones anómalas sobre los modelos de regresión.

Entrenamiento y Evaluación de Modelos

Para evaluar el desempeño de los modelos de regresión se utilizaron tres métricas principales: Error Absoluto Medio (MAE), Raíz del Error Cuadrático Medio (RMSE) y Coeficiente de Determinación (R^2), calculadas tanto sobre el conjunto de entrenamiento como sobre el de prueba. Además, se analizaron las variables más influyentes en cada modelo con el objetivo de interpretar la relevancia de los distintos predictores en la estimación del precio.

Modelo 1 – Regresión Lineal

El modelo de Regresión Lineal sirvió como punto de partida, proporcionando una primera aproximación simple y fácilmente interpretable.

Los resultados obtenidos fueron los siguientes:

Conjunto	MAE	RMSE	R^2

Entrenamiento	1.037	1.653	0.295
Prueba	1.040	1.645	0.281

El bajo valor de R^2 sugiere que el modelo sólo explica una parte limitada de la variabilidad del precio, lo que indica que las relaciones entre las variables son probablemente no lineales.

Entre las 20 variables más influyentes destacan características como has_availability, room_type_Hotel room, host_response_time, bathrooms, y amenities _count, lo cual refuerza la importancia de la disponibilidad, el tipo de alojamiento y el equipamiento en la determinación del precio.

Modelo 2 – XGBoost

Para el modelo XGBoost, se realizó una optimización de hiper parámetros mediante validación cruzada (k-fold), ajustando parámetros como learning_rate, max_depth, alpha y n_estimators, entre otros, con el objetivo de encontrar el mejor balance entre sesgo y varianza.

Los resultados fueron:

Conjunto	MAE	RMSE	R^2
Entrenamiento	0.333	0.472	0.942
Prueba	0.344	0.490	0.936

Este modelo logró un ajuste significativamente superior al de la regresión lineal, con valores de R^2 cercanos a 0.94 tanto en entrenamiento como en prueba, mostrando una excelente capacidad de generalización y baja diferencia entre ambos conjuntos.

Entre las variables más importantes para este modelo podemos nombrar, bathrooms, beds, longitude, accommodates, latitude, neighbourhood_cleansed, property_type, entre otras, las cuales tienen un enfoque mucho más orientado a la calidad de la propiedad y los amenities que tiene para ofrecer.

Modelo 3 – Random Forest

Como tercer enfoque se implementó un Random Forest Regressor, también ajustado mediante validación cruzada para optimizar parámetros como la cantidad de árboles y la profundidad máxima.

Los resultados obtenidos fueron:

Conjunto	MAE	RMSE	R^2
Entrenamiento	0.449	0.637	0.895
Prueba	0.459	0.650	0.888

El rendimiento fue ligeramente inferior al de XGBoost, aunque mantuvo una muy buena capacidad predictiva y estabilidad entre entrenamiento y prueba.

En cuanto a las variables más influyentes, se destacan bathrooms, availability_365, neighbourhood_cleansed y beds, lo que coincide con la intuición de que la cantidad de baños, la disponibilidad y la zona son factores determinantes en el precio de un alojamiento.

Comparación de Resultados

A continuación se resumen las métricas de desempeño obtenidas para los tres modelos evaluados:

Modelo	Train MAE	Train RMSE	Train R^2	Test MAE	Test RMSE	Test R^2
LR	1.037	1.653	0.295	1.040	1.645	0.281

XGBoost	0.333	0.472	0.942	0.344	0.490	0.936
R Forest	0.449	0.638	0.895	0.459	0.650	0.888

Observaciones

- Regresión Lineal:

  Presentó un desempeño considerablemente inferior, con un R^2 en torno al 0.28 -- 0.29, lo que indica que el modelo logra explicar sólo una fracción limitada de la variabilidad del precio. Su simplicidad lo hace útil como punto de referencia o para interpretar relaciones lineales, pero no resulta adecuado para capturar la complejidad del problema.

- XGBoost:

  Alcanzó el mejor desempeño global, con valores muy bajos de MAE y RMSE, y un R^2 cercano a 0.94 tanto en entrenamiento como en prueba. Esto demuestra una excelente capacidad de generalización y una alta precisión predictiva. No obstante, requirió un proceso de ajuste de hiperparámetros más detallado, dada su sensibilidad al overfitting.

- Random Forest:

  También mostró un buen rendimiento (R^2 de 0.89 aproximadamente), ligeramente inferior al de XGBoost. Sin embargo, resultó más estable y menos propenso al sobreajuste, manteniendo un equilibrio adecuado entre precisión y robustez. Es una alternativa sólida cuando se busca un modelo más simple de ajustar que XGBoost.

Conclusiones

En síntesis, la Regresión Lineal fue útil como modelo base, pero insuficiente para representar las relaciones no lineales y las interacciones entre variables presentes en los datos.

Los modelos de ensamble, en cambio, demostraron una capacidad mucho mayor para capturar dicha complejidad. Entre ellos, XGBoost se destacó como el mejor predictor del precio de alquiler, logrando el mayor poder explicativo y los errores más bajos. Random Forest, aunque ligeramente menos preciso, ofreció un rendimiento muy competitivo y una mayor resistencia al sobreajuste, por lo que también podría considerarse una buena opción en contextos donde se priorice la estabilidad frente a la optimización extrema.

Si tuviera que elegir un modelo para predecir precios de alquileres de la plataforma AirBnb en distintas ciudades XGBoost es la mejor opción.

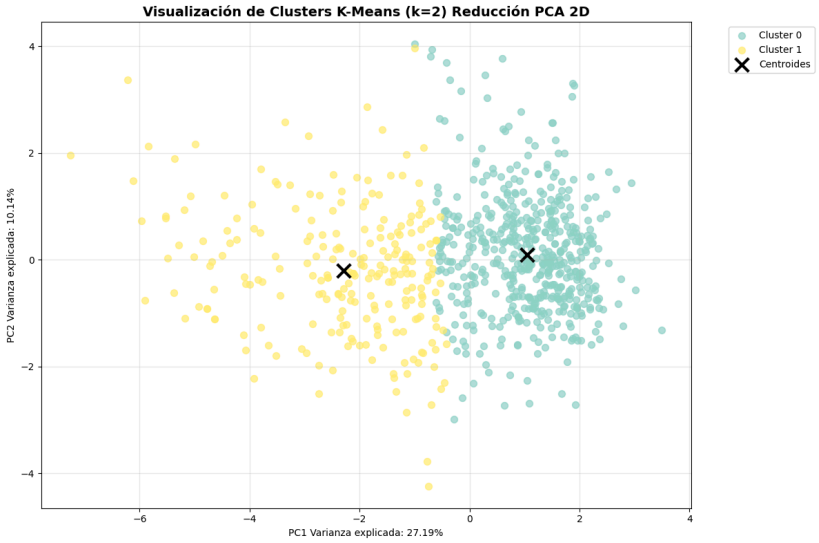
EJ4: Clustering

Tendencia al clustering y número de grupos

Se evaluó la estructura de los datos mediante el estadístico de Hopkins, obteniendo un valor de 0.97, lo que indica una fuerte tendencia a formar grupos naturales. Para determinar el número óptimo de clusters, se aplicaron el método del codo y el índice de Silhouette, ambos sugiriendo que k = 2 es el número más apropiado.

Visualización de los grupos

La siguiente figura muestra la proyección bidimensional (PCA) de los datos y la asignación de clusters obtenida con K-Means (k=2). Se observa una separación clara entre ambos grupos, aunque los dos primeros componentes principales sólo explican el 37.3% de la varianza total, por lo que esta representación es una simplificación visual del espacio original.



Características de los grupos

Cluster 0 (31.1% de los datos): agrupa canciones más acústicas y relajadas, con bajo nivel de energía y danceability, y un tempo medio (~113 BPM).

Cluster 1 (68.9% de los datos): reúne canciones más enérgicas y bailables, con alta energía, danceability y tempo elevado (~124 BPM).

Conclusión

El análisis nos permitió agrupar las observaciones en dos clusters bien diferenciados, demostrando la utilidad de las técnicas de clustering para identificar patrones en datos sin etiquetas. En términos metodológicos, K-Means mostró un buen rendimiento y simplicidad interpretativa y la reducción de dimensionalidad mediante PCA facilitó la visualización y comprensión de los grupos.

Tiempo dedicado

Indicar brevemente en qué tarea trabajó cada integrante del equipo durante estas semanas. Si trabajaron en las mismas tareas lo detallan en cada caso (como en el ejemplo el armado de reporte). Deben indicar el promedio de horas semanales que dedicaron al trabajo práctico.

Integrante	Tarea	Prom. Hs Semana
Calderan, Facundo Andres	Pre procesamiento de Datos Modelos de Clasificación Binaria	4
Merlinsky Camins, Mariano Gabriel	Armado de Reporte Pre Procesamiento de Datos Modelos de Clasificación Binaria	6
Castellano Bogdan, Benjamin	Armado de Reporte, Modelos de Regresión y clasificación	6
Yu, Fernando	Armado de Reporte Formulación de preguntas	3
Pons Echeverria, Tomas	Redacción informe, clustering, revisión modelos regresión y clasificación	4.5