.UBAfiuba
FACULTAD DE INGENIERÍA

TA047 - Ciencia de Datos Dr. Ing. Rodríguez - 2°C 2024

Trabajo Práctico 1 - Grupo 05

EJ1: Análisis Exploratorio

El objetivo del ejercicio fue realizar el análisis exploratorio de los datos, aplicar técnicas de exploración y de preprocesamiento (de valores nulos y outliers) para poder responder algunas preguntas planteadas sobre dichos datos, utilizando 3 dataset (que unificamos en 1 para el análisis) de viajes de taxi amarillos en EEUU del año 2023 en los meses de Enero, Febrero y Marzo.

El dataset (unificado) está conformado por 21 columnas (features) con 9.384.487 de registros.
Incluye campos que registran las fechas y horas de salida y llegada, las distancias de los viajes, las tarifas detalladas, los tipos de tarifas, los tipos de pago y los recuentos de pasajeros informados por el conductor. Los datos utilizados en los conjuntos de datos adjuntos fueron recopilados y proporcionados a la Comisión de Taxis y Limusinas de la Ciudad de Nueva York (TLC) por proveedores de tecnología autorizados en virtud de los Programas de Mejora de Pasajeros de Taxis y Limusinas (TPEP/LPEP).

Se analizaron los features de variables cualitativas y cuantitativas.
Para las variables cualitativas las que destacaron fueron las fechas de los viajes (la de salida y llegada coinciden en el mismo día), el tipo de pago y VendorID (Un código que indica el proveedor de TPEP que proporcionó el registro).
Para las variables cuantitativas las que destacaron fueron cantidad de pasajeros informados por el conductor, distancia recorrida por cada viaje y el importe total.

Analizando la correlación entre las variables, lo cuál fue muy poca, las variables que tuvieron más correlación positiva fuerte fueron entre la VendorID con cantidad de pasajeros (0,1) y con el día de la semana (0,051, que se obtuvo de la fecha del viaje). Una correlación negativa fuerte entre importe total con tipo de pago (-0,077) y con día de la semana (-0,012) .
La distancia recorrida apenas tenía un poco de correlación con el importe total del viaje (0,018).

Preprocesamiento de Datos

1. ¿Se eliminaron columnas (Nombre de la columna y motivo de eliminación?

   Las columnas que se eliminaron fueron los features airport_fee y Airport_fee (Tarifa por recogida únicamente en los aeropuertos LaGuardia y John F. Kennedy) ya que tenían el 70% y 30% aprox. respectivamente de datos faltantes, y era un porcentaje muy significativo.

2. ¿Detectaron correlaciones interesantes (entre qué variables y qué coeficiente)?

   Como ya mencionamos, no hubo mucha correlación entre las variables, pero la de mayor coeficiente (0,1) fue entre las variables VendorID y cantidad de pasajeros.

3. ¿Generaron nuevos features?

   Generamos 7 features nuevas (6 cualitativas y 1 cuantitativa):

   Nombre del día de la semana (**weekday**), identificador del día de la semana (**weekofday**) y nombre del día del mes (**month**). Estos 3 features se generaron a partir de la fecha de llegada del viaje.

   El nombre del tipo de pago (**payment_type_name**), que se utilizó la documentación para reemplazar los IDs de payment_type.

   La descripción del VendorID (**vendor_description**), también obtenida de la documentación.

   La descripción del código de la tarifa (**ratecode_description**), utilizando el feature RatecodeID (código de tarifa final vigente al final del viaje) con su documentación.

   Duración del viaje desde la hora de salida y llegada en minutos (**duration_in_minutes**), utilizando dichas fechas del viaje.

4. ¿Encontraron valores atípicos?¿Cuáles?¿Qué técnicas utilizaron y qué decisiones tomaron?

Las variables de tipo monto tenían todas outliers. Arreglamos montos negativos que no tienen sentido en conceptos de impuesto e importe del viaje. Consideramos los negativos como mal cargados y los convertimos en positivos. En cuanto al importe total había casos donde el monto era 0 y claramente era un error de carga, ya que representaba un viaje gratis, así que tomamos la decisión de eliminar los registros ya que sólo representaban el 0.018% de los datos.

Una vez corregidos los datos los outliers ya eran por valores extremos y para mantener los registros (aunque alterando su distribución) aplicamos el método de recorte (capping).

Para el feature de cantidad de pasajeros vimos que había más de 6 pasajeros por viaje, y dedujimos que era un caso extremo que se podía dar en pocos casos donde el taxi era un tipo de vehículo de camioneta y podían entrar más de 6 pasajeros. Eran sólo 55 registros así que tomamos la decisión de aplicar el método de eliminación (trimming), ya que no influye demasiado en el análisis.

Para el feature de distancia recorrida vimos que habían viajes muy largos, más de 100000 km donde en EEUU la distancia más larga entre ciudades es de aprox 5000 km. Por lo tanto, los viajes que superan dicha distancia las consideramos mal cargadas. Las cuales eran el 0.0023% del dataset y no influye demasiado, así que las eliminamos. Luego vimos casos extremos que superan los 100 km, por lo que aplicamos el método de recorte para no perderlos.

Por último, para el feature que creamos para la duración de los viajes, nos encontramos que estaban quedando algunas en negativo, así que buscamos los casos mal cargados donde la fecha de llegada era menor o igual a la de salida. Encontramos 3630 registros y al ser pocos, fueron eliminados. En cuanto a los casos extremos, encontramos varios casos que superan los 1500 min, lo que representa aprox. 1 día de viaje sin descanso como máximo, a los cuales les aplicamos el método de recorte (aunque igualmente eran pocos).

5. ¿Qué columnas tenían datos faltantes?
   ¿En qué proporción? ¿Qué se hizo con estos registros?

Los primeros que encontramos y eliminamos fueron los ya mencionados **impuestos del aeropuerto**.

Luego vimos que la **cantidad de pasajeros** habían casos con 0 (156806 registros), el cual tenía sentido un viaje sin pasajeros sólo en los casos donde el viaje haya sido ANULADO (payment_type = 5). Si no se daba esa condición lo consideramos como un valor mal ingresado. Analizando dicho casos, ninguno se encontraba anulado, así que tomamos la decisión de rellenarlo con el valor promedio, que era 1 pasajero por viaje. No perdimos los datos aunque si modificamos la distribución de los datos, pero manteniendo el promedio.

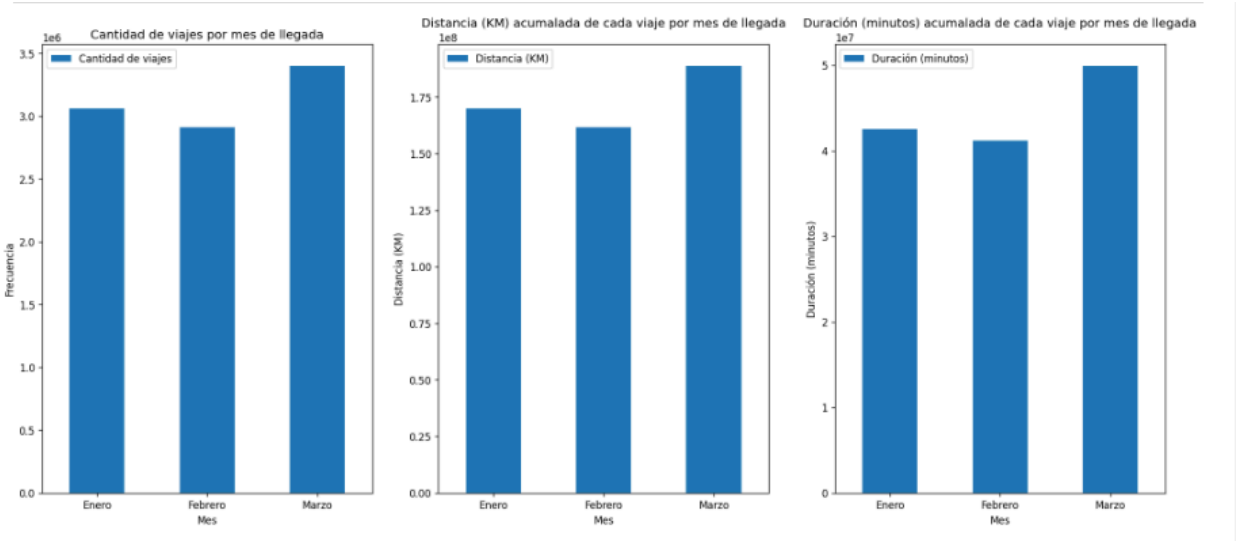
Para el feature **RatecodeID** tenía el 2,5% de los datos faltantes y que también rellenamos con el promedio, tarifa estándar. Además, encontramos un valor mal ingresado (que no se encontraba en la documentación), el ID 99 y viendo que el último ID de la lista el 6, tenía solamente 16 registros, interpretamos que era un valor mal cargado y que correspondía al ID 6 (por ser el último en la lista y además por ser números “parecidos”).

Lo mismo para el feature **store_and_fwd_flag** ("almacenar y reenviar": indica si el registro del viaje se mantuvo en la memoria del vehículo antes de enviarlo al proveedor, porque el vehículo no tenía una conexión con el servidor) tenía el 2,5% de datos faltantes que rellenamos con el valor promedio (“N”), ya que consideramos que no afecta demasiado al análisis.

En el caso del feature **VendorID** no tenía datos faltantes pero si tenía un valor mal cargado, el 6. Ya que según la documentación de los datos los valores posibles son el 1 (Creative Mobile Technologies, LLC.) y el 2 (VeriFone Inc.). Así que lo reemplazamos por el promedio que es el 2.

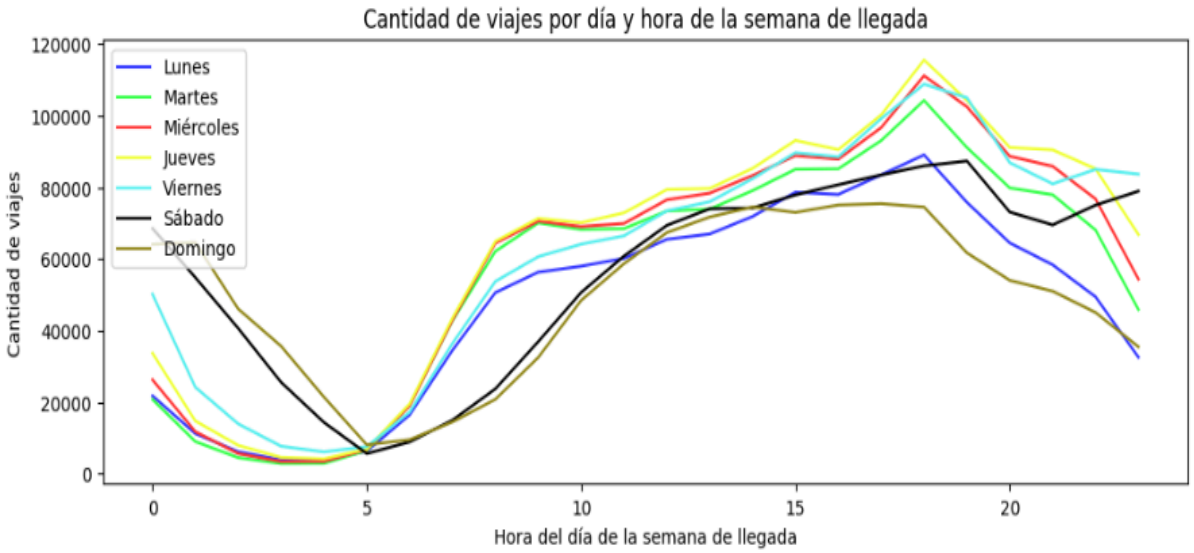
<u>Visualizaciones</u>

● ¿Cantidad de viajes por distancia y duración, respecto al mes de llegada?



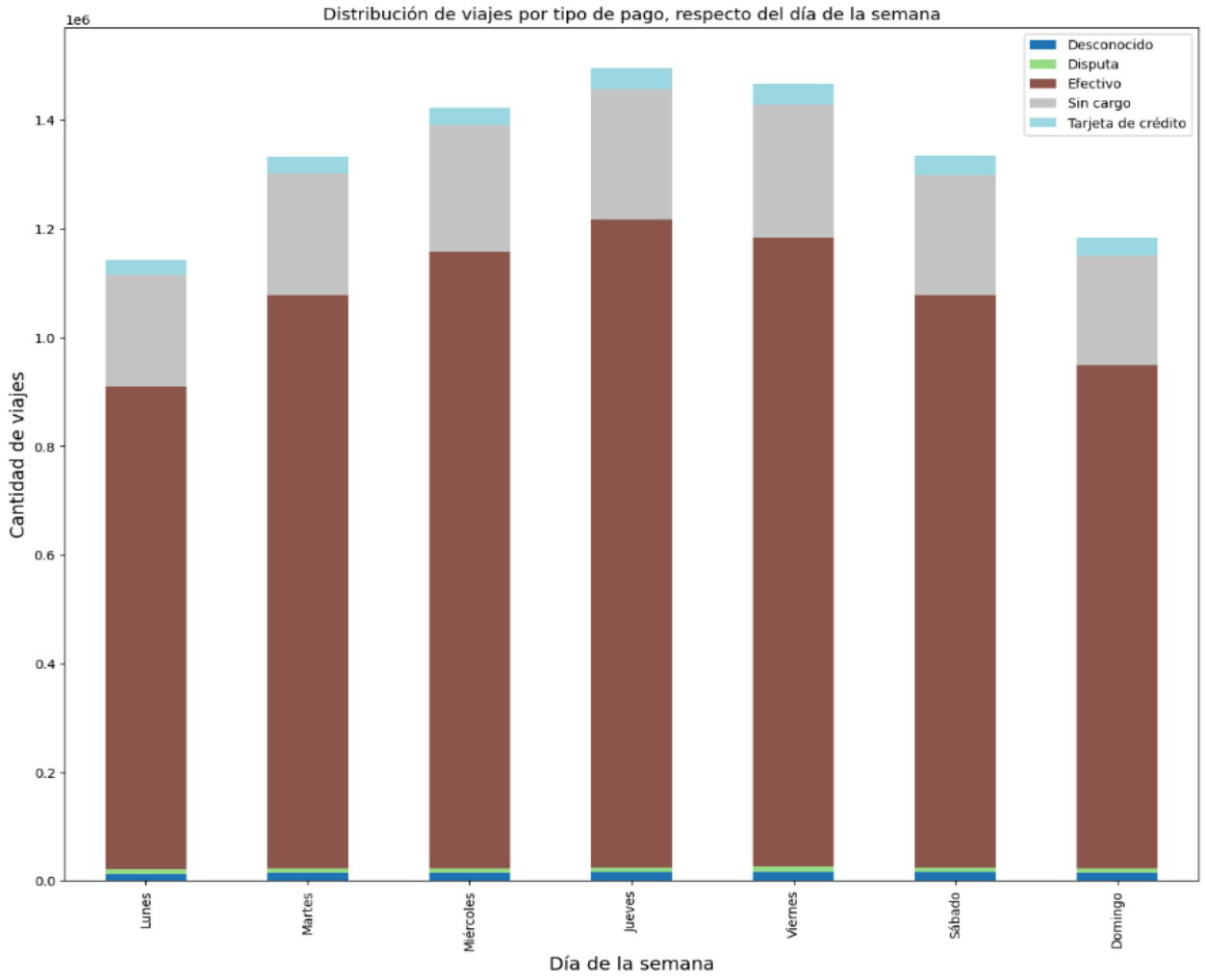
Se puede apreciar que el mes con la mayor cantidad de viajes y con más distancia y duración acumulada es Marzo. Con una cantidad de 3,500,000 viajes acumuló una distancia de 175,000,000 KM y una duración de 50,000,000 minutos (como habíamos visto, un promedio de 14 minutos por viaje).
Luego lo sigue Enero y por úlitmo Febrero con la menor cantidad y acumulación.

● ¿Cantidad de viajes por hora y día de la semana de llegada?



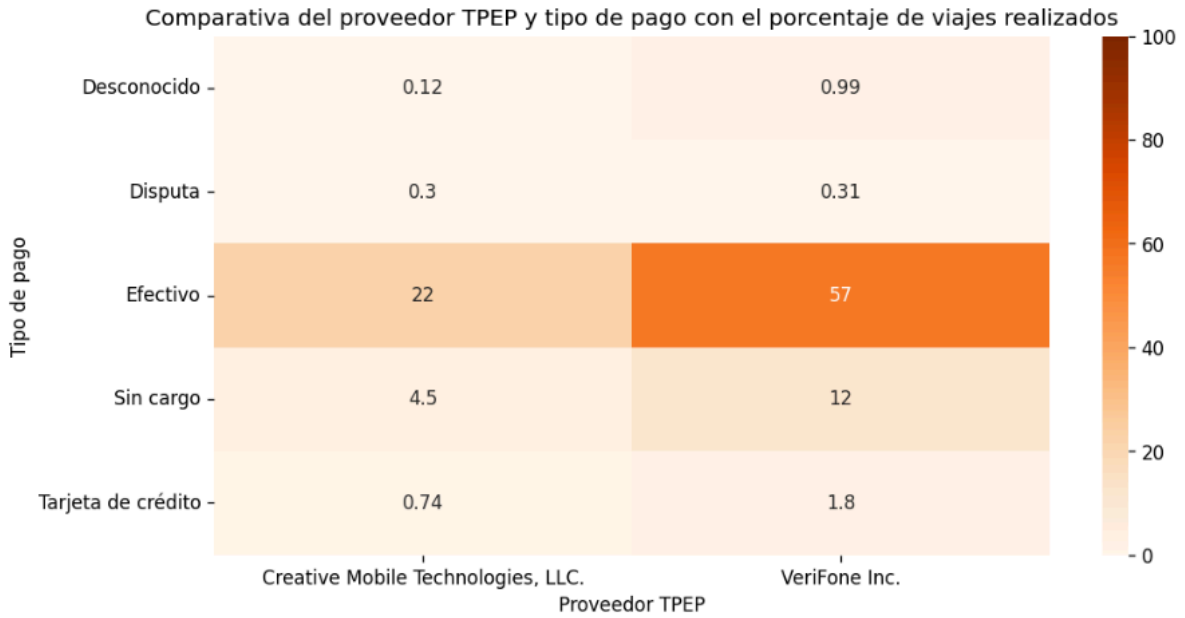
Se puede apreciar que para todos los días, la hora donde hay más viajes es entre las 18 y 19 hs, y las 5hs hay la menor cantidad de viajes. Respecto a los días,Jueves es el que mayor pico tiene, seguido de Miércoles y Viernes, respectivamente. A la madrugada, entre las 00 y 5hs, los fines de semanas son los que tienen mayor cantidad de viajes.

● ¿Cómo es la distribución de los viajes por tipo de pago, respecto del día de la semana de llegada?



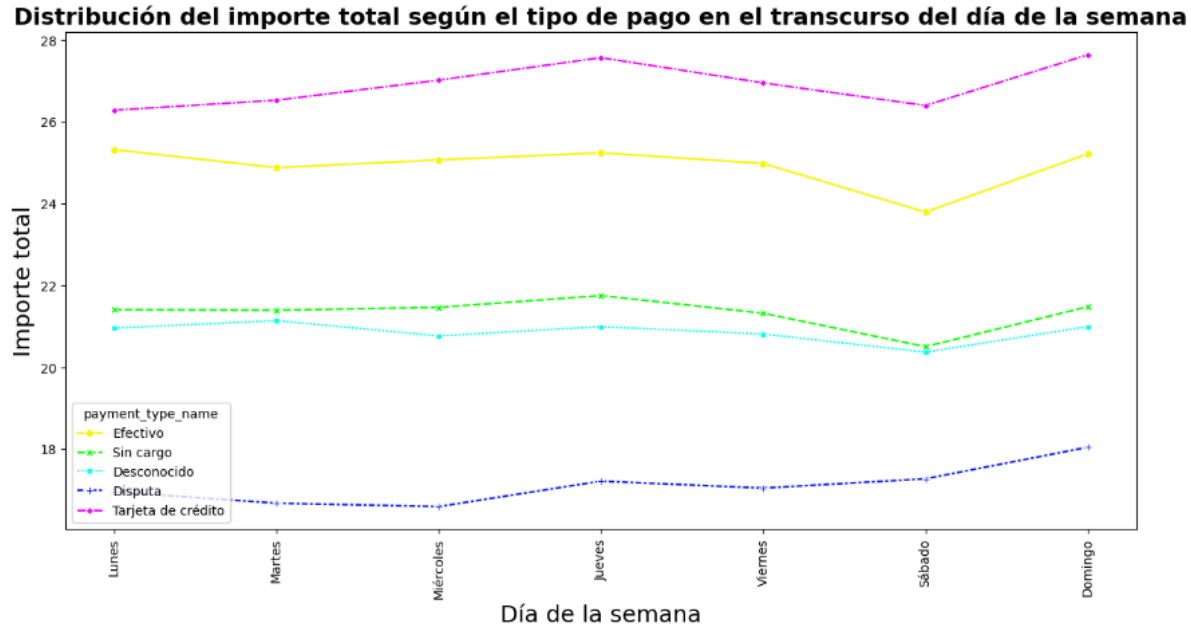
Se puede apreciar que el tipo de pago "Efectivo" es el más utilizado en los 7 días de la semana. Como ya vimos, el Jueves es el que tiene mayor cantidad de viajes, igualmente la distribución de los tipos de pagos son muy similares entre los días.

● ¿Cómo son los viajes por el proveedor de TPEP que proporcionó el registro y tipo de pago?



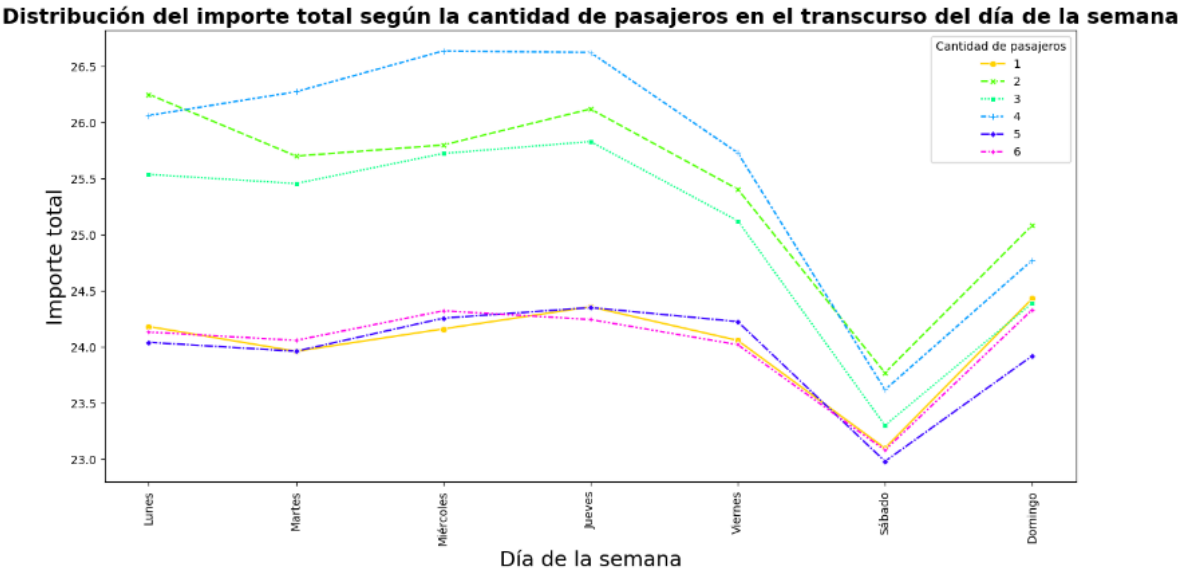
Se puede apreciar, como veníamos viendo que “Efectivo” es el tipo de pago más utilizado para pagar los viajes con el 79%, 57% fue a través del proveedor TPEP VeriFone Inc.

● ¿Cuál es el flujo del importe total según el tipo de pago en el transcurso del día de la semana?



Se puede apreciar que los días donde hay más viajes (Jueves) el importe total con "Efectivo" es un poco más elevado. Se mantiene parejo, excepto los Sábados que son más baratos.
Respecto a los pagos con "Tarjeta de crédito" son un poco más caros, seguramente por el impuesto agregado.

● ¿Cuál es el flujo del importe total según la cantidad de pasajeros en el transcurso del día de la semana?



Se puede apreciar que los días donde hay más viajes (Jueves) el importe total es más elevado con 4 pasajeros por viaje. Se mantiene parejo, excepto los Sábados que son más baratos.
Respecto a la cantidad de pasajeros, los importes más altos son cuando hay 4,2 y 3 pasajeros por viaje.

EJ2: Clasificación

Realizar una breve descripción del dataset: cantidad de registros y columnas, etc. Comentar los features más destacables. Mencionar en cada caso si realizaron transformaciones sobre los datos (encoding, normalización, etc)

Dada un conjunto de datos provenientes de estaciones meteorológicas de Australia, mediante sus datos meteorológicos del día actual, nuestro objetivo es predecir si lloverá o no al día siguiente. Esto se realizará mediante la creación de modelos de clasificación.El conjunto de datos está comprendido por observaciones diarias del clima durante 10 años de distintas localidades de Australia. La variable a predecir es: "RainTomorrow". Es una variable de Yes o No. El criterio de tomar como positivo, o sea Yes, es que en ese día llegó a llover 1mm o más.

En nuestro caso, las Ubicaciones ("Locations") a estudiar son:

- Queensland
- Victoria
- Australia Meridional
- Australia Occidental

El conjunto original del dataset esta compuesto por 23 variables.

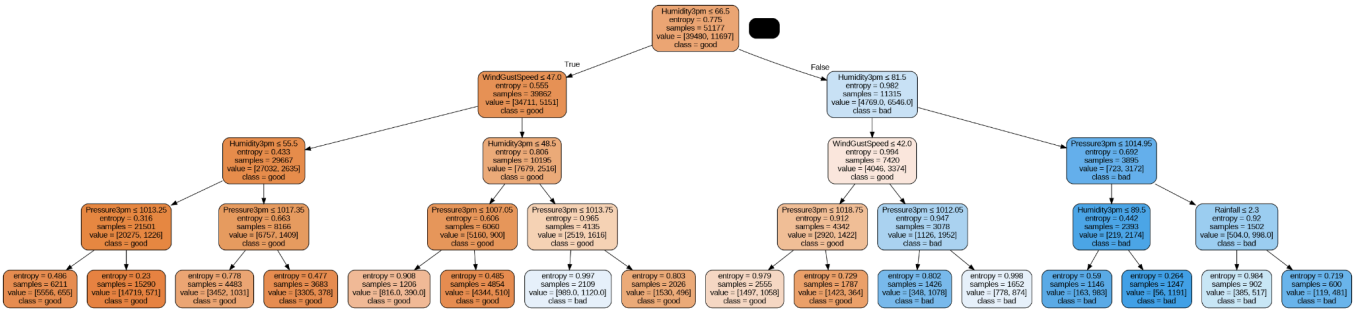
Las variables del dataset son:

- **Date**: Fecha de la observación de los datos meteorológicos.
- **Location**: El nombre común de la ubicación de la estación meteorológica.
- **MinTemp**: La temperatura mínima en grados centígrados.
- **MaxTemp**: Máxima temperatura en grados centígrados.
- **Rainfall**: La cantidad de lluvia registrada durante el día en mm.
- **Evaporation**: La denominada evaporación en tanque de clase A (mm) en las 24 horas hasta las 9 a. m.
- **Sunshine**: El número de horas de sol brillante en el día.
- **WindGustDir**: La dirección de la ráfaga de viento más fuerte en las 24 horas hasta la medianoche.
- **WindGustSpeed**: La velocidad (km/h) de la ráfaga de viento más fuerte en las 24 horas hasta la medianoche.
- **WindDir9am**: Dirección del viento a las 9 am.
- **WindDir3pm**: Dirección del viento a las 3 pm.
- **WindSpeed9am**: Velocidad del viento (km/h) promediada durante 10 minutos antes de las 9 a. m.
- **WindSpeed3pm**: Velocidad del viento (km/h) promediada durante 10 minutos antes de las 3 p. m.
- **Humidity9am**: Humedad (porcentaje) a las 9 am
- **Humidity3pm**: Humedad (porcentaje) a las 3 pm
- **Pressure9am**: La presión atmosférica (hpa) se redujo al nivel medio del mar a las 9 am
- **Pressure3pm**: La presión atmosférica (hpa) se redujo al nivel medio del mar a las 3 pm
- **Cloud9am**: Fracción del cielo oscurecida por nubes a las 9:00 h. Se mide en "oktas", que son una unidad de octavos. Registra cuántos octavos del cielo están oscurecidos por nubes. Un valor de 0 indica un cielo completamente despejado, mientras que un valor de 8 indica que está completamente nublado.
- **Cloud3pm**: Fracción del cielo oscurecida por nubes a las 15:00 h. Se mide en "oktas", que son una unidad de octavos. Registra cuántos octavos del cielo están oscurecidos por nubes. Un valor de 0 indica un cielo completamente despejado, mientras que un valor de 8 indica que está completamente nublado.
- **Temp9am**: Temperatura (grados C) a las 9 am
- **Temp3pm**: Temperatura (grados C) a las 3 pm
- **RainToday**: Booleano: 1 si la precipitación (mm) en las 24 horas hasta las 9 a. m. supera 1 mm, de lo contrario 0
- **RainTomorrow**: Cantidad de lluvia del día siguiente en mm. Se utiliza para crear la variable de respuesta **RainTomorrow**. Una especie de medida del "riesgo".

Haciendo un análisis de correlación, de cantidad de nulos, se eliminaron las variables de alto contenido de datos nulos. Para las variables que tenían poca cantidad de valores nulos se procedió a eliminar dichos registros. También se eliminaron variables que estaban muy correlacionadas, por lo que no aportarían información extra, este es caso de las variables de temperaturas. Luego se transformó Date por estaciones del año. El conjunto de dataset limpio y filtrado quedó de 17 variables. Luego se procedió a hacer One Hot Encoding para las variables categóricas.

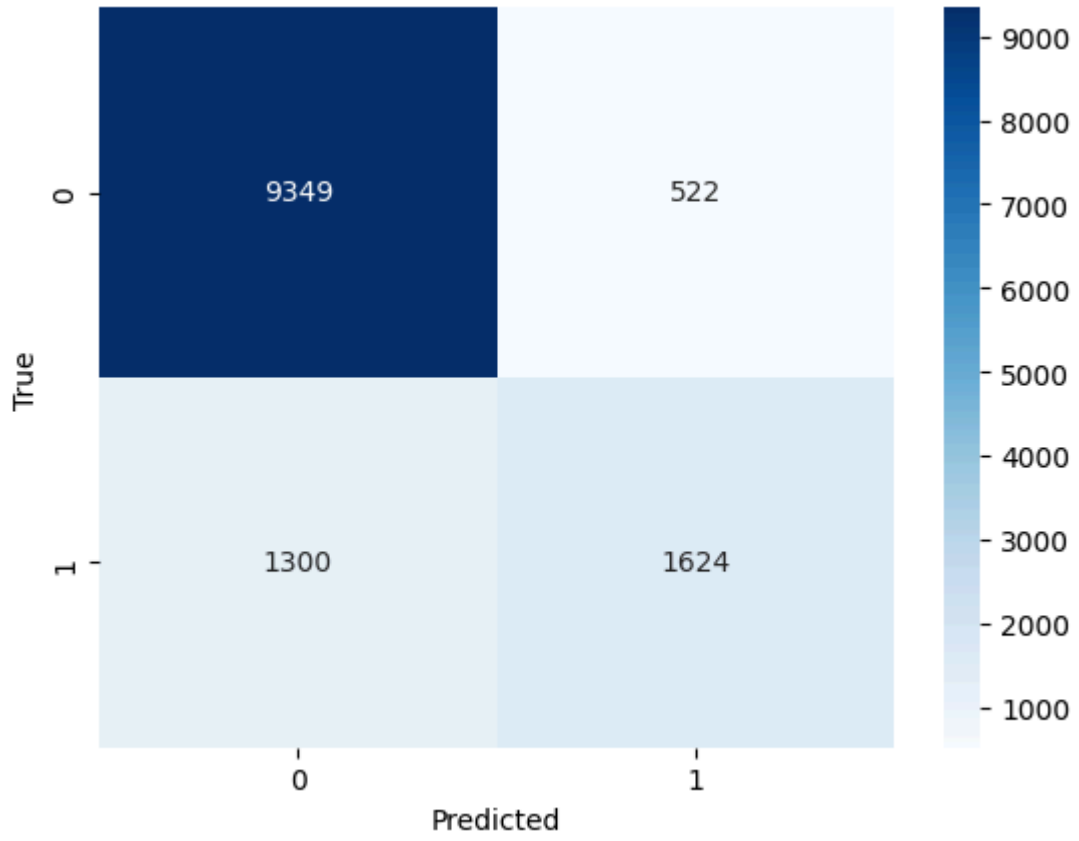
Modelos

1. Arbol de Decision
    - Los hiperparametros a optimizar fueron: criterion, ccp_alpha y max_depth. No se pudieron optimizar más por errores de Timeout.
    - Se utilizó K-fold Cross Validation, de 5 folds.
    - La metrica para optimizar fue f1-score.
    - Mejores hiperparametros encontrados: {'max_depth': 4, 'criterion': 'entropy', 'ccp_alpha': 0.0}
    - Las variables más representativas a la hora de clasificar en el árbol fueron: Humidity3pm, WindGustSpeed y Pressure3pm.



2. Random Forest
    - Por GridSearch se optimizaron los hiperparametros: criterion, max_features, min_samples_leaf, min_samples_split, n_estimators.
    - Los mejores hiperparametros hallados fueron:
        i. 'criterion': 'gini',
        ii. 'max_features': 'sqrt',
        iii. 'min_samples_leaf': 1,
        iv. 'min_samples_split': 10,
        v. 'n_estimators': 20Se utilizó K-fold Cross Validation con 5 folds.
    - La métrica para buscar los hiperparametros fue accuracy, f1 y roc_auc.
    - Las métricas obtenidas fueron:
        i. Accuracy: 0.8569753810082064
        ii. Recall: 0.54890560875513
        iii. f1 score: 0.6369047619047619

- Mostrar la conformación final de uno de los árboles generados. Si es muy extenso mostrar una porción representativa y explicar las primeras reglas.
- Matriz de confusión



3. XGBoost
   - Optimizamos los hiperparametros: learning_rate, max_depth, min_child_weight, subsample, colsample_bytree y n_estimators.
   - Se utilizó K-fold Cross Validation con 5 folds.
   - La métrica que utilizamos para buscar los hiperparametros fue roc_auc.

Cuadro de Resultados

Modelo	F1-Test	Precision Test	Recall Test	Accuracy Test
Arbol de Decision	0.66	0.74	0.60	0.86
Random Forest	0.64	0.76	0.56	0.86
XGBoost	0.64	0.74	0.57	0.86

Árbol de decisión: {'max_depth': 5, 'criterion': 'entropy', 'ccp_alpha': 0.0}

Random Forest: {'criterion': 'gini', 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 20}

XGBoost: {'criterion': 'gini', 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 4, 'n_estimators': 20}

Elección del modelo

Elegimos los modelos de ensamble XGBoost y Random Forest a pesar de que no dieron tan bien como el Árbol de Decision. Creemos que se puede mejorar.

EJ3: Regresión

Realizar una breve descripción del dataset: cantidad de registros y columnas, etc. Comentar los features más destacables. Mencionar en cada caso si realizaron transformaciones sobre los datos (encoding, normalización, etc)

Modelos

1. Regresión Lineal
    - ¿Qué features seleccionaron para construir el modelo?
    - Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas. Comparar con la performance de entrenamiento.
2. XGBoost
    - ¿Utilizaron K-fold Cross Validation?¿Cuántos folds utilizaron?
    - ¿Qué métrica utilizaron para buscar los hiperparámetros?
    - Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas. Comparar con la performance de entrenamiento.
3. Modelo a elección
    - ¿Utilizaron K-fold Cross Validation?¿Cuántos folds utilizaron?
    - ¿Qué métrica utilizaron para buscar los hiperparámetros?
    - Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas. Comparar con la performance de entrenamiento.

Cuadro de Resultados
Realizar un cuadro de resultados comparando los modelos que entrenaron (entre ellos debe figurar cuál es el que seleccionaron como mejor predictor).

Medidas de rendimiento en el conjunto de TEST:

- MSE
- RMSE
- XXX: si seleccionaron alguna métrica adicional…

Confeccionar el siguiente cuadro con esta información:

Modelo	MSE	RMSE	XXX
Regresión Lineal			
XGBoost			
…			

En cada caso ¿Cómo resultó la performance respecto al set de entrenamiento?

Nota: indicar brevemente en qué consiste cada modelo de la tabla

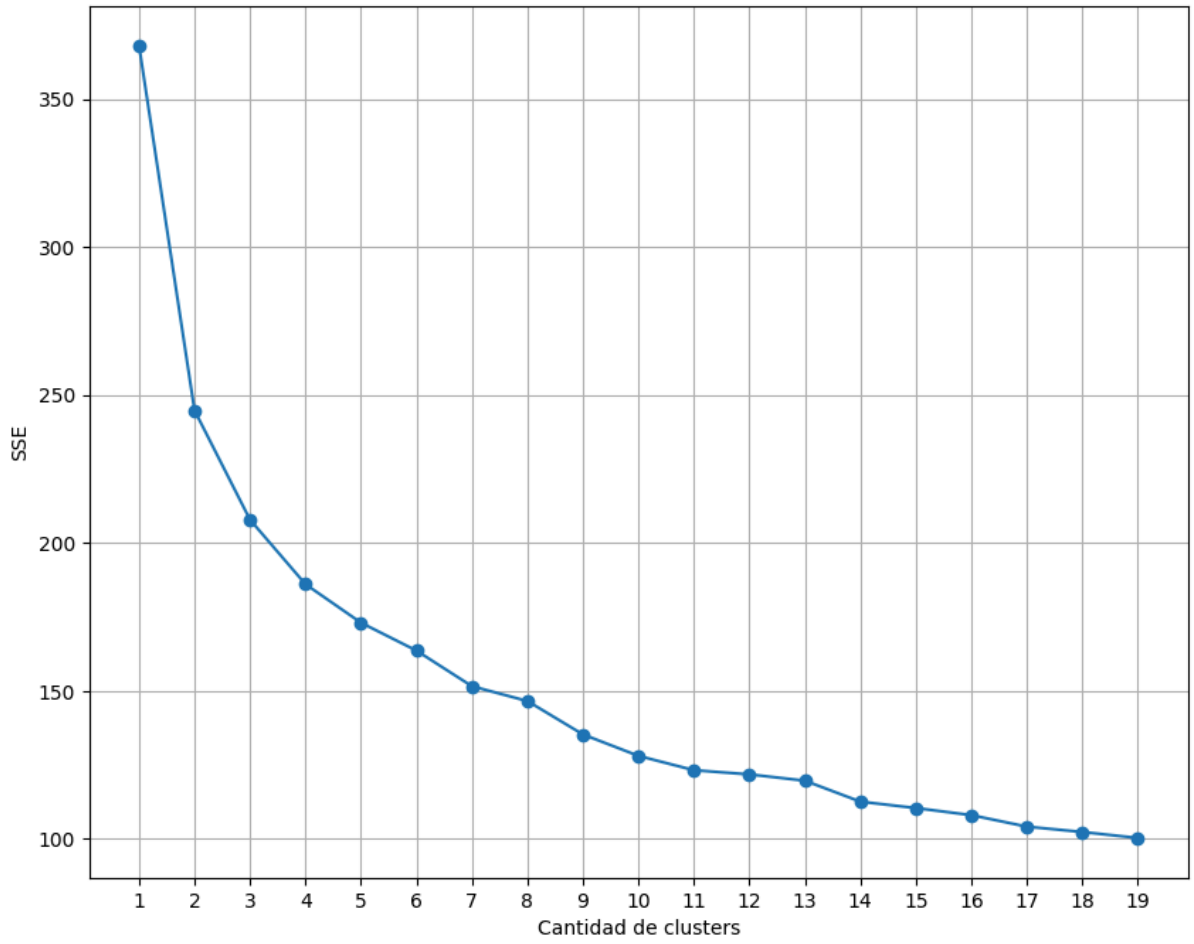
Elección del modelo

En base a los resultados obtenidos ¿Qué modelo elegirían para predecir el precio de una propiedad de AirBnB para la ciudad estudiada?¿Por qué?

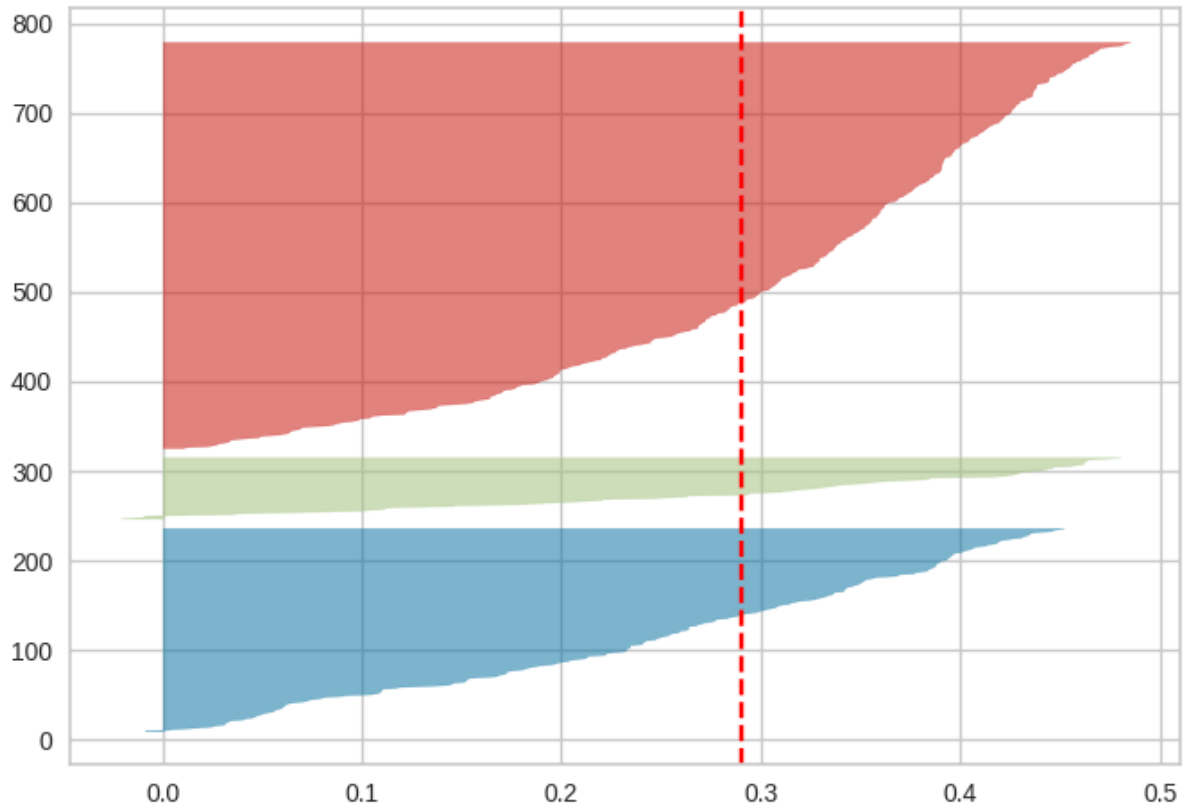
EJ4: Clustering

El objetivo de este ejercicio es encontrar posibles agrupaciones de datos. Para ello primero realizamos un análisis mediante la estadística de Hopkins. El valor hallado dio alrededor de 0,9 , indicando así una fuerte tendencia a clusters de datos.

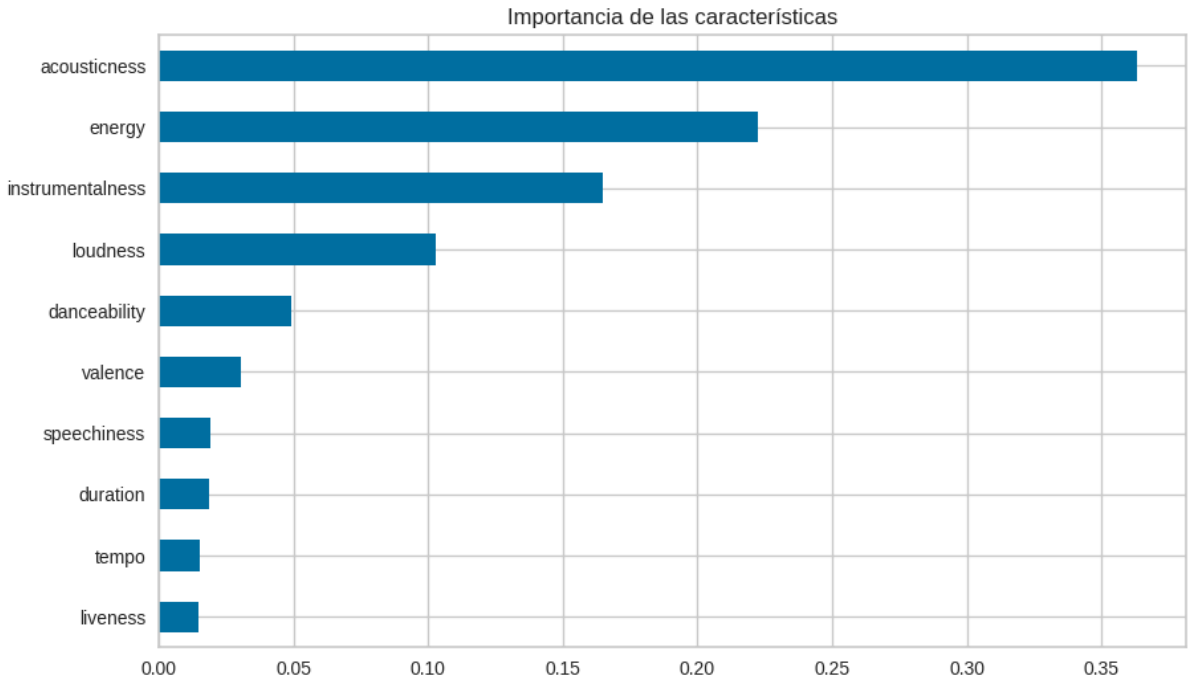
Para hallar la cantidad de grupo de datos óptima, realizamos el análisis del método de Elbow.



Para seguir investigando la cantidad de grupos de datos, usamos el análisis de mediante el índice de Silhouette para cada cantidad. Los de mayor valor dieron para clusters de cantidad n=2 y n=3. Para ambos dieron valores muy similares. Para definir la cantidad realizamos un gráfico de Silhouette y nos quedamos con aquel que tenga mayor cantidad de valores positivos de silueta, en este caso para n=3.



Se analizaron los centroides de dichos clusters y también a través de un análisis de random forest, se busco la importancia de las variables como muestra en el siguiente gráfico:



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
Mejia Alan	Análisis de datos, procesamiento de datos, entrenamiento de modelos, realización de gráficos y análisis de métricas.	8 hs
Prieto Pablo Alejandro		
Sosa Zoraida Flores		