Predicción de cancelación de reserva
Entrega 2
Carlos Eduardo Castaño Garzón
Juan Antonio Arango Moreno
Jagler David Velasquez Velasquez
Introducción a la IA
UdeA
2023

Manejo de datos faltantes:

Uno de los requisitos para el dataset del proyecto fue "al menos ha de tener 5% de datos faltantes en al menos 3 columnas" por lo que para la primera entrega decidimos tomar 3 columnas al azar (específicamente "arrival_year", "arrival_month" y "arrival_date") y eliminar el 5,34% de cada una de ellas (equivalente a 1500 datos), lo cual sería suficiente para cumplir el requisito. Ahora, para la segunda entrega nos encontramos con la incógnita de cómo manejar esos datos faltantes conocidos como "missing data" (un problema común en el análisis de datos y en la construcción de modelos de aprendizaje automático), para lo cual tenemos varias alternativas:

1. **Eliminar los datos faltantes:** Si la cantidad de datos faltantes es pequeña en comparación con el tamaño total del conjunto de datos, se puede optar por simplemente eliminar los registros que contengan valores faltantes. Sin embargo, esta estrategia puede no ser adecuada si la cantidad de datos faltantes es significativa, ya que esto podría reducir significativamente el tamaño del conjunto de datos.
2. **Llenar con valores predeterminados:** Si los valores faltantes son de tipo categórico, una estrategia común es llenarlos con la moda de la variable correspondiente. Si los valores faltantes son de tipo numérico, se puede llenarlos con la media o la mediana de la variable correspondiente.
3. **Llenar con valores estimados:** Otra estrategia es utilizar técnicas de imputación, donde los valores faltantes se llenan utilizando métodos estadísticos más avanzados, como regresión, interpolación o k-vecinos más cercanos (KNN). Estos métodos pueden ser más precisos que la simple sustitución por valores predeterminados.
4. **Crear una nueva categoría:** Si los valores faltantes son de tipo categórico, una estrategia es crear una nueva categoría para los valores faltantes, lo que permite que el algoritmo distinga entre los valores faltantes y los valores existentes.

Analizamos cada una de ellas y llegamos a las siguientes conclusiones respectivamente:

1. La idea de tener datos faltantes es aprender a afrontarlos y a encontrar una solución viable, por lo que simplemente eliminar las filas completas no encaja con el objetivo de el trabajo/curso
2. Los datos faltantes no son de tipo categórico ni tampoco un valor numérico estándar, debido a que son fechas se nos imposibilita simplemente sacar un promedio de la columna y llenar dichos datos
3. El método KNN (K-Nearest Neighbor) que se entiende como "un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual." (Ver imagen 1). Nos llamó la atención, pero tampoco nos daba suficiente tranquilidad a la hora de completar la información faltante.

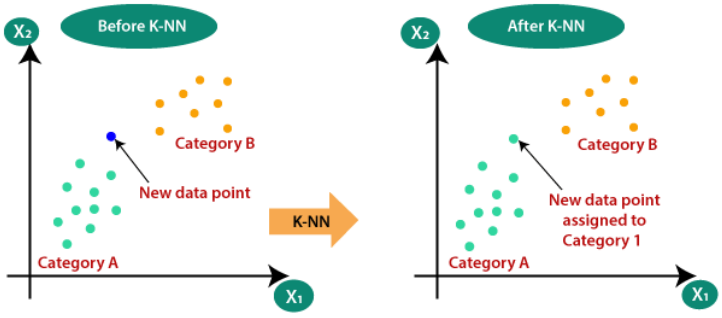

Imagen 1. Método KNN

4. Al no ser una variable categórica éste método tampoco nos brindaba una alternativa viable.

También estuvimos investigando sobre qué librerías tenían alguna función que permitiera solucionar dicha cuestión, encontrando por ejemplo la siguiente:

```
missing_values = data.isnull()
data.fillna(data.mean(), inplace=True)
```

Usando la librería pandas, primero se cargarían los datos y luego se identificarían los valores faltantes (línea 1) para posteriormente llenar los valores faltantes con la media de la columna correspondiente (línea 2); pero claramente no era una solución viable para nosotros.

Así que decidimos organizar el documento directamente en Excel teniendo como referencia diferentes columnas, para luego completar los datos con los valores semejantes acorde a dichas columnas (Semejante al método KNN), para al final comparar los resultados con los valores iniciales y determinar según nuestro criterio si el método empleado nos ofrecía un resultado deseado.

Específicamente lo que hicimos fue tomar el documento CSV y convertirlo a XLS por columnas, posteriormente ejecutamos una tabla dinámica con las 3 columnas en cuestión y unificamos cada fecha de ambos años, es decir, sumamos las reservas de Enero 1 de 2017 con las reservas de Enero 1 de 2018, las de Agosto 25 de 2017 con Agosto 25 de 2018, etc… lo anterior lo hicimos buscando conservar y sumar aquellas épocas de mayores reservas. Cabe aclarar que el Dataset fue "Synthetically-Generated" por lo que cada mes consta de 31 días. A continuación, hicimos 2 copias de la hoja completa original: Una donde organizábamos los valores por cada columna de la forma Fill 1 (ver Imagen 2) y otra donde los organizábamos de la forma Fill 2 (ver Imagen 3).

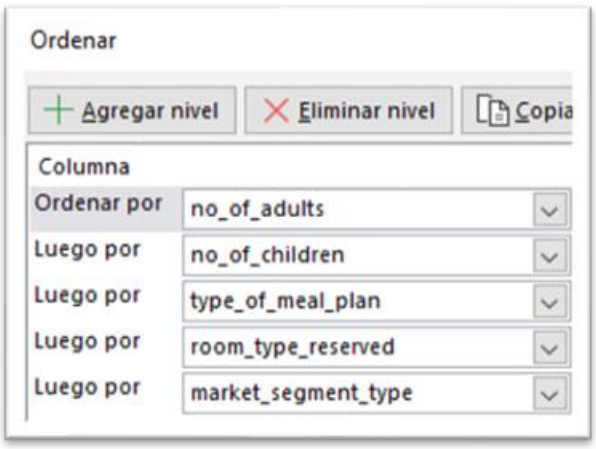

Imagen 2. Forma Fill 1

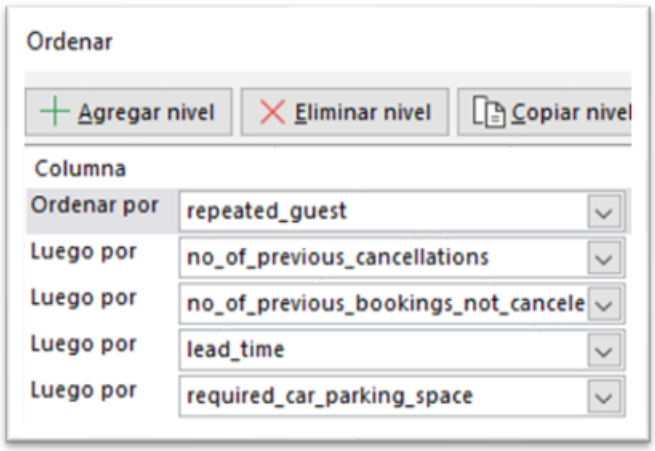

Imagen 3. Forma Fill 2

Para llenar los datos simplemente seleccionamos las columnas "arrival_year", "arrival_month" y "arrival_date" (aquellas con datos faltantes) y por medio de la función de Excel *Buscar y Seleccionar – Ir a Especial – Celdas en blanco* llenamos la información con los datos de la celda inmediatamente superior (esto para cada una de las dos hojas, Fill 1 y Fill 2)

Finalmente llevando a cabo el mismo procedimiento realizado con la hoja original (tabla dinámica y suma de reservas de cada fecha), pudimos graficar las 3 tablas y comparar los datos originales con las dos formas de completar la información (Fill 1 y Fill 2). Esperábamos que la cantidad total aumentara (pues se crearían 1500 valores nuevos), pero tuvimos cuidado de seleccionar aquella que se ajustara más a la original.

A continuación en la Imagen 4 verán los datos graficados de el documento original (con datos faltantes) en color azul, luego se verán los datos Fill 1 en naranja y Fill 2 en gris. Será claro que las variaciones de la forma Fill 1 (naranja) son muy superiores y "radicales" comparadas con la forma de llenado Fill 2 que está en color gris.

Finalmente, decidimos trabajar con el documento ordenado y llenado de la forma Fill 2, lo que quiere decir que las columnas “repeated_guest”, “no_of_previous_cancellations”, “no_of_previous_bookings_not_canceled”, “lead_time” y “required_car_parking_space” están altamente relacionadas con las fechas de las reservas (que eran los datos que debíamos completar).

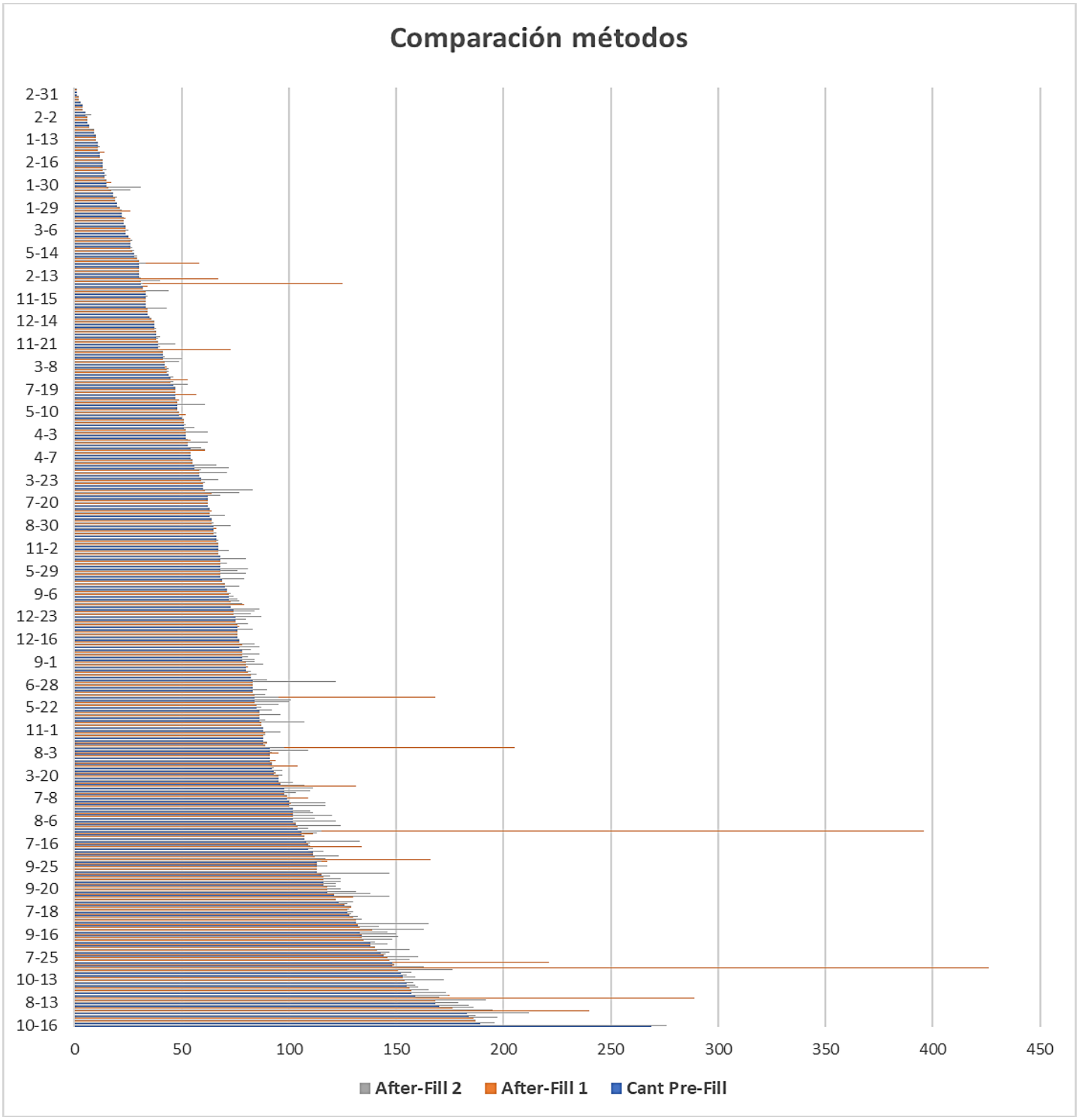

Imagen 4. Comparación Pre-Fill, After-Fill 1 y After-Fill 2