Primera entrega - Introducción a la Inteligencia Artificial: ¿Se cancelará la reserva del hotel?

Integrantes.

Paola Andrea Posada Restrepo <paola.posada1@udea.edu.co>

Stiven Guerra Chaverra<stiven.guerra@udea.edu.co>

Sebastián Gómez Ramírez <sebastian.gomez35@udea.edu.co>

1. **Descripción del problema predictivo a resolver**

Tras buscar entre múltiples datasets que se adecuarán al presente proyecto se decidió abordar un problema de aprendizaje supervisado de clasificación, el cual busca determinar si la reserva realizada en un hotel va a ser cancelada o no; esta clasificación se realiza utilizando información relacionada a una reserva o a la evolución de la misma. El problema abordado puede contribuir a diferentes hoteles para tener en cuenta una posible cancelación y así considerar el uso del espacio reservado.

2. **Descripción del dataset a utilizar**

Para el desarrollo del ejercicio mencionado se hará uso del dataset *Hotel Booking* tomado de Kaggle, el cual cuenta con 119.391 datos reales recolectados de un Hotel de ciudad y un Hotel Resort entre las fechas 01 de Julio de 2015 y 31 de Agosto del 2017. Para el caso de uso se realizó una selección solo para el año 2015, trabajando con un total de 21996 para facilitar el procesamiento y trabajo con el conjunto de datos. Este dataset contiene múltiples datos respecto a cada reserva realizada, las cuales poseen más de 30 características entre las que destacan la fecha de la reserva, cantidad de niños, cantidad de bebés, país de proveniencia, cantidad de cambios en la reserva, si ha tenido una reserva cancelada previamente, entre otros. También vale la pena destacar varios datos categóricos como el tipo de cliente, tipo de habitación, el tipo del depósito, etc. Finalmente, se cuenta también con el dato de si la reserva fue tomada o cancelada.

Se realizó una búsqueda de los datos faltantes y afortunadamente para nuestro modelo y se encontró que el dataset presenta datos nulos en 3 columnas: país, compañía y agente; con 133, 20691 y 3099 datos nulos respectivamente.

3. **las métricas de desempeño requeridas**

Durante el análisis de la base de datos, se encuentra un desbalance significativo entre las dos clases, debido a que la clase 1 (que representa una cancelación en la reserva) cuenta con 8154 datos y la clase 0 (que representa una NO cancelación en la reserva) cuenta con 13854 datos; por lo tanto se busca una métrica de desempeño que no cause sesgo en el resultado final por no tener en cuenta la clase 1 (que representa una minoría). Por lo que se decidió usar la métrica Accuracy, F1, Recall y Precision

4. **Primer criterio sobre cuál sería el desempeño deseable en producción.**

Como desempeño deseable, se buscan valores superiores a 0.6 en las métricas de Accuracy, F1, Recall y Precision