Data

https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand

Fecha de entrega: 25/10/2023

Instrucciones

- Deberán entregar la resolución en un archivo de texto (preferentemente en formato pdf) y la notebook .ipynb (**archivo, no link a colab**)
    - Piense al archivo de texto como un informe: debe contener explicaciones (en español) y gráficos que la acompañen, no código
    - El código se debe encontrar exclusivamente en la notebook
- Tanto el archivo de texto como la notebook deben tener el mismo nombre:
    - icd_2023_**apellido_nombre**
    - Ejemplo: si el alumno Pepe Pompín desea ser evaluado, debe enviar 2 archivos:
        - icd_2023_pompin_pepe.pdf
        - icd_2023_pompin_pepe.ipynb
        - Evitar acentos y mayúsculas en los nombres de los archivos que vayan a enviar

Consignas

1. Exploración
    a. Leer el archivo "hotel_bookings.csv". ¿Qué puede mencionar sobre su estructura y variables?
    b. ¿Cómo es la correlación entre las variables numéricas? Utilice y analice en detalle algún gráfico que sirva para sacar conclusiones sobre la asociación de variables realizando apertura por cancelación de reserva.
    c. Para las distintas categorías dentro de la variable 'deposit_type', analice gráficamente la distribución de la variable 'is_canceled'. ¿Qué puede concluir al respectp? (Extra: puede realizar un test de Chi cuadrado para evaluar la asociación entre estas 2 variables categóricas)
    d. Explore visualmente la relación entre 'lead_time' y 'is_canceled'.
2. Partición de datos
    a. Realice una partición de datos **estratificada** entre entrenamiento y test (80-20%), **usando como semilla su número de documento**
    b. Genere 2 listados de atributos, una lista de Python con los **nombres** de los atributos **predictores numéricos** que considere adecuados y otra lista con los **nombres** de los atributos **predictores categóricos** que considere adecuados
3. Modelos
    a. Ajuste un árbol de clasificación usando los atributos categóricos, realice las transformaciones necesarias (ej: *one-hot-encoding*) para poder realizar el procesamiento de datos.
    b. Elija una métrica de performance para este problema de clasificación y justifique su elección.
    c. Realice búsqueda de hiper-parámetros mediante *cross validation (5 folds) y grid-search* para el árbol de clasificación (con el set de entrenamiento) e indique cómo es la estructura del árbol que emerge como el óptimo según esta búsqueda. ¿Qué performance promedio obtuvo?

    d. Repita los puntos a. y c. pero ahora tenga en cuenta además de los atributos categóricos, los atributos numéricos.
    e. Finalmente realice la búsqueda de hiperparámetros pero esta vez teniendo en cuenta todos los atributos . ¿Cómo es el árbol de mejor performance? ¿Difiere mucho la performance de cross validation respecto a los árboles hallados en los puntos c. y e.?
4. Importancia de features y testing
    a. Grafique el árbol de mejor performance (si es muy grande puede recortar la imagen de manera que se vea la parte de arriba del árbol)
    b. ¿Qué puede decir de la importancia de los features en este árbol?
    c. Finalmente, entrene este árbol con todo el conjunto de datos de entrenamiento y evalúe su performance con los datos de prueba (test). Reporte las métricas de accuracy, recall y precision y grafique la matriz de confusión. Interprete los resultados obtenidos
5. Ensambles
    a. Ajuste un modelo de *Random Forest* (el default de sklearn, sin buscar hiperparámetros) sobre los datos de entrenamiento y evalúe su performance en test. Compare los resultados con los del mejor árbol obtenido.