Minería de Datos

Solemne 1: Requerimientos del Proyecto 1

---

1. Contexto

¿Alguna vez te has preguntado cuál es la mejor época del año para reservar una habitación de hotel? ¿O la duración óptima de la estancia para obtener la mejor tarifa diaria? ¿Qué pasaría si quisieras predecir si es probable que un hotel reciba o no un número desproporcionadamente alto de solicitudes especiales?

En este proyecto utilizaremos un conjunto de datos de reservas de hotel que puede ayudar a explorar las preguntas anteriores mediante la analítica predictiva o descriptiva, aunque **el objetivo de esta actividad será practicar el proceso de preparación de los datos con miras a obtener una vista analítica adecuada.**

El conjunto de datos y su descripción se encuentra en el siguiente enlace (se adjunta a este documento el archivo hotel_bookings.csv):

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

Este conjunto de datos contiene información de reserva para un hotel urbano y un hotel resort, e incluye información como: cuándo se realizó la reserva, duración de la estancia, número de adultos, niños y/o bebés, y número de plazas de aparcamiento disponibles, entre otras cosas. Toda la información de identificación personal ha sido eliminada de los datos.

2. Informe

Los grupos de proyecto deberán realizar un informe que desarrolle los siguientes contenidos:

a. **Portada** con el nombre del proyecto, y nombres y RUT de los integrantes del grupo.

b. **Catálogo de datos**. Se debe describir la siguiente información con respecto a los datos que se están analizando:

- ¿Quién es el propietario o responsable de la fuente de datos?
- ¿Dónde se encuentra ubicada la fuente de datos (URL de acceso)?
- ¿Tiene algún costo?
- ¿En qué tipo de archivo se encuentran los datos?
- ¿Cuántos registros tiene?

- ¿Cuántos atributos tiene?
- ¿Cuál es el significado de cada atributo?
- ¿Qué tamaño tiene en Mega o Giga Bytes?
- ¿Tiene algún requisito de privacidad?

c. **Verificación de correspondencia** lógica entre variables (atributos) y los tipos de datos que almacenan.

- Tipos de datos de cada variable (atributo).
- Análisis de la correspondencia lógica entre variables y tipos.

d. **Interpretación preliminar y visualización**. Codificar, ejecutar y explicar los resultados de las instrucciones del lenguaje Python para:

- Importar librerías Python.
- Cargar los datos en un dataframe.
- Determinar la cantidad total de registros y de variables (atributos).
- Determinar los tipos de datos por cada variable (atributo).
- Determinar valores estadísticos básicos para cinco (5) variables numéricas: media, moda, varianza, desviación estándar, etc.
- Generar histogramas para las cinco (5) variables numéricas.
- Realizar un análisis preliminar sobre la simetría de la distribución de los datos de una (1) variable numérica (indicar si presenta una distribución simétrica, asimetría positiva o asimetría negativa). Este item no requiere una instrucción Python, pero si requiere indicar los resultados del análisis.
- Determinar valores estadísticos básicos para cinco (5) variables categóricas: cantidad de valores por atributo, valor máximo (moda), frecuencia del valor máximo, etc.
- Generar gráficas de barras para las cinco (5) variables categóricas.
- Realizar un análisis preliminar sobre los valores mínimos y máximos de los atributos de una (1) variable categórica.
- Generar la gráfica de correlación de todas las variables e identificar posibles correlaciones existentes en pares de variables.
- Elegir dos variables numéricas (por ejemplo: lead_time y arrival_date_month) y generar un gráfico que muestre la correlación de las mismas. Interpretar el resultado producto de la visualización.

- Generar un gráfico que visualice la cantidad de Reservas asociadas al tipo de hotel, si es Resort o City. Responder a la pregunta: ¿qué tipo de hotel tiene más reservas?

e. **Preparación de los datos.** Codificar, ejecutar y explicar los resultados de las instrucciones del lenguaje Python para:

- Detectar y eliminar datos duplicados:
  - Identificar todos los registros duplicados de la base de datos.
  - Eliminar todos los registros duplicados de la base de datos.
- Detectar y tratar datos faltantes:
  - Identificar todas las variables (columnas) con datos faltantes.
  - Elegir una (1) variable numérica que presente datos faltantes y eliminar los registros (filas) que contengan dichos datos.
  - Elegir una (1) variable numérica que presente datos faltantes para realizar un análisis de simetría y determinar si es mejor la media o la mediana para sustituirlos. Sustituir los datos faltantes por la media o la mediana dependiendo del análisis de simetría de la variable.
  - Elegir tres (3) variables categóricas que presenten datos faltantes y sustituir éstos por la moda.
- Detectar y tratar datos atípicos:
  - Detectar los datos atípicos de la variable StaysInWeekendNights utilizando la visualización a través de un Diagrama de Caja y Bigotes. Eliminar los registros que contengan datos atípicos.
  - Elegir tres (3) variables numéricas diferentes a la variable StaysInWeekendNights y utilizar el método LOF (Local Outlier Factor) para detectar registros con valores atípicos. En caso de existir estos registros, eliminarlos de la base de datos.

f. **Enlace al video explicativo del proyecto.**

Nota importante: se sugiere la elaboración de un notebook de Google Colab que integre la explicación del paso a paso de los análisis junto con las instrucciones Python implementadas. Esto ayudaría considerablemente a una ordenada y detallada explicación de lo realizado y a la elaboración del video correspondiente.

3. Video

3.1. Contenidos del Video:

i. Presentación del equipo y del proyecto: señale el nombre del proyecto y los nombres de cada uno de los integrantes del equipo.
ii. Explicación del Catálogo de Datos.
iii. Verificación de correspondencia lógica entre variables y los tipos de datos que almacenan.
iv. Explicación de la interpretación preliminar y visualización de datos.
v. Explicación del proceso de preparación de los datos.
vi. Conclusiones. Resuma los resultados obtenidos en el proyecto explicando: conocimientos adquiridos, principales dificultades que el grupo enfrentó y cómo pudieron resolverlas.

3.2. Condiciones del Video:

i. El video a desarrollar debe tener una duración mínima de 5 y máxima de 60 minutos.
ii. Todos los estudiantes del grupo deben participar en el video. Al momento de participar se les debe escuchar su voz y ver al menos el rostro. No olvide que su participación en el video debe aportar a la explicación del contenido.
iii. El video debe preparar material para contestar cada uno de los contenidos solicitados.
iv. La calidad del video debe considerar el uso de imágenes claras y nítidas, sonido apropiado y contenidos que aporten claridad de la explicación. Además, debe estar organizado para ayudar a comprender los contenidos solicitados.
v. El video puede ser desarrollado con las herramientas que al grupo de estudiantes les parezcan apropiadas.
vi. El grupo de estudiantes entregará un enlace al video para su revisión en línea, y éste deberá presentarse en un formato que no amerite una instalación de software especial.
vii. Se recomienda el uso de la plataforma Zoom institucional para la grabación de este video.

4. Aspectos administrativos:

- Modalidad: en grupos de mínimo dos (2) y máximo cinco (5) estudiantes.
- Lugar: Sección de “Evaluaciones” del aula virtual habilitado para tal fin.
- Entregable: Informe en formato Word o PDF.
- Fecha tope: miércoles 24-04-2024, 23:59 horas.

Docente Ismael Moreno Flores
i.morenoflores@uandresbello.edu