TALLER N°1 MACHINE LEARNING – Exploratory Data Analysis (EDA)

A. Objetivo General:

- Desarrollar habilidades de análisis de datos y comprensión de las relaciones entre diferentes variables en un conjunto de datos reales, aplicando técnicas de EDA y preparando el terreno para futuros análisis de Machine Learning.

B. Objetivos Específicos:

1. **Comprender las Distribuciones que presentan los Datos**: Analizar la distribución de diferentes variables en el conjunto de datos, identificar valores atípicos y entender las relaciones entre diferentes variables. Esto incluirá la visualización de los datos utilizando gráficos apropiados.
2. **Verificación de Hipótesis**: Formular y verificar hipótesis sobre relaciones entre diferentes variables en el conjunto de datos, por ejemplo, la relación entre el consumo de alcohol y el rendimiento académico, la relación entre la estructura familiar y la salud de los estudiantes, etc.
3. **Preparación de Datos para Modelado**: Preparar los datos para futuros análisis de Machine Learning. Esto incluirá la limpieza de los datos, la gestión de valores perdidos y atípicos, y la transformación de variables categóricas.

C. Historia de los Datos

En La Serena, hay una universidad que ha notado un cambio en el comportamiento y rendimiento de sus estudiantes mechones durante los últimos años. Aunque no tienen datos históricos sobre el comportamiento de los estudiantes, han decidido comenzar a recopilar datos sobre varios aspectos de la vida de los estudiantes, como el rendimiento académico, el consumo de alcohol, la salud, las relaciones sociales y la estructura familiar. La universidad ha recopilado estos datos durante un año y ahora quiere analizarlos para entender mejor la situación y tomar medidas para mejorar la vida de los estudiantes.

Objetivos del proyecto para la Universidad

El objetivo principal del proyecto es entender la relación entre el consumo de alcohol y varios otros factores de la vida de los estudiantes. Ustedes como asesores de la Universidad deben realizar un análisis exploratorio de datos (EDA) para entender la distribución de los datos, buscar patrones, verificar hipótesis y, finalmente, crear un **modelo intuitivo** para predecir ciertos el nivel de consumo de alcohol de los estudiantes (NO DEBE USAR algoritmos de ML).

D. Tareas Específicas (QUE SE REQUIERE EN ESTE TALLER)

d.1) Análisis descriptivo: Se debe comenzar con un análisis descriptivo de los datos para esto Desarrolle estadísticas descriptivas (tendencia central, percentiles y dispersión) y desarrolle gráficos coherentes con las variables. **Interprete los resultados.** (14 puntos)

d.2) Verificación de hipótesis: La universidad tiene varias hipótesis que necesita verificar (confirmar o rechazar) con análisis basados en los datos. Estas hipótesis son:

1. El rendimiento académico está relacionado con el consumo de alcohol.
2. El rendimiento académico está relacionado con el estado de salud de los estudiantes.
3. El rendimiento académico está relacionado con la estructura familiar de los estudiantes.

4. El consumo de alcohol está relacionado con el estado de salud de los estudiantes.
5. El consumo de alcohol está relacionado con la estructura familiar de los estudiantes.
6. El consumo de alcohol está relacionado con el tiempo que los estudiantes pasan con amigos.
7. El estado de salud de los estudiantes está relacionado con la estructura familiar.
8. El estado de salud de los estudiantes está relacionado con el tiempo que pasan con amigos.
9. La estructura familiar de los estudiantes está relacionada con el tiempo que pasan con amigos.
10. La estructura familiar de los estudiantes está relacionada con la inasistencia a clases.
11. El tiempo que los estudiantes pasan con amigos está relacionado con la inasistencia a clases.
12. El rendimiento académico está relacionado con la inasistencia a clases.
13. El rendimiento académico está relacionado con el tiempo que los estudiantes pasan con amigos.
14. El consumo de alcohol está relacionado con la inasistencia a clases.

- Para verificar (confirmar o rechazar) cada una de las hipótesis desarrolle un análisis con datos que debe seleccionar e interpretar los resultados del análisis. NO SE debe LIMITAR sólo a relaciones lineales. (14 puntos).
- Desarrolle 7 hipótesis con sus correspondientes análisis que permitan verificar (confirmar o rechazar) cada una de ellas e interprete resultado. NO SE debe LIMITAR sólo a relaciones lineales. (7 puntos)

d.3) Creación de modelo: Basándose en el análisis anterior, ustedes como asesores de la universidad debe crear un modelo (que este basado en su intuición) para predecir el consumo de alcohol de un estudiante basado factores que ustedes deberán seleccionar o crear a partir de las variables disponibles. Y provea una **medida de desempeño o calidad de su modelo** (14+7=21 puntos).

d.4) Recomendaciones: Basándose en su análisis, como asesores deben hacer recomendaciones a la universidad sobre cómo pueden mejorar la situación. (7 puntos)

E. Dataset para el Modelamiento ML

Los datos que se utilizarán para este proyecto son los datos recopilados por la universidad en sus encuestas anuales a los estudiantes. Estos datos incluyen información sobre el consumo de alcohol de los estudiantes, su rendimiento académico, su salud, sus relaciones sociales, etc. El dataset para este Taller-N°1 sobre EDA se denomina "estudialcohol.csv", el que representa una muestra de datos compilados en las ciudades de La Serena y Valparaíso sobre estudiantes de primeros años universitarios que tienen hábitos alcohólicos.

OBS: los nombres de las variables no usan tilde ni "ñ" a propósito, con el fin de ser usadas sin inconvenientes en Python/Jupyter notebook:

Las características del dataset son las siguientes:

1. **ciudad**: Ciudad de residencia del estudiante
2. **genero**: Género del estudiante (M = Masculino, F = Femenino)
3. **edad**: Edad del estudiante
4. **direccion**: Tipo de dirección del estudiante (URB = Urbana, RUR = Rural)
5. **famtam**: Tamaño de la familia del estudiante (MI3 = Menor o igual a 3 integrantes, MQ3 = más de 3 integrantes en la familia)
6. **estadoP**: Estado parental (SEP = Separados, JUN = Juntos)
7. **eduM**: Nivel educativo de la madre (0: Ninguna, 1: E.Básica, 2: E.Media, 3: Técnica, 4:Universitaria)

8. **eduP**: Nivel educativo del padre (0: Ninguna, 1: E.Básica, 2: E.Media, 3: Técnica, 4:Universitaria)
9. **trabM**: Trabajo de la madre
10. **trabP**: Trabajo del padre
11. **razon**: Razón de elección de la escuela
12. **asistioNMM**: Asistió al nivel medio mayor (preescolar)
13. **internet**: Dispone de acceso a internet en casa
14. **tutor**: Tutor asignado
15. **tpoviaje**: Tiempo de viaje (total) a la universidad
16. **tpoestudio**: Tiempo dedicado al estudio
17. **areprobadas**: Número de asignaturas reprobadas
18. **apoyoextra**: Recibe algún apoyo extra de la universidad
19. **apoyofam**: Recibe apoyo familiar en los estudios
20. **extrapaga**: Recibe clases particulares pagadas
21. **activextra**: Realiza actividades extracurriculares
22. **qsuperior**: Quería seguir con estudios superiores (está convencido que son útiles)
23. **pololea**: Tiene pololo/a
24. **qfamrel**: Calidad de las relaciones familiares
25. **tpolibre**: Tiempo libre tras la escuela
26. **saleamigos**: Sale con amigos los fines de semana
27. **salud**: Percepción de salud del estudiante.
28. **inasistencia**: Número de inasistencias
29. **G1**: Nota en el primer periodo
30. **G2**: Nota en el segundo periodo
31. **G3**: Nota en el tercer periodo
32. **nivelalco**: Nivel de consumo de alcohol.

F. Entregables

Para lo anterior se pide entrega, con la estructura del nombre indicado:

1. Archivo **EDA-GRUPO-Nnn.ipynb** con todo el código y los análisis. (donde nn=n° del grupo asignado)
2. Documento en pdf (**EDA-GRUPO-Nnn.pdf)** con las siguientes secciones:
   i) El análisis descriptivo de los datos, gráficos e interpretación de lo todo lo graficado y observado.
   ii) La validación de las hipótesis (confirmación | rechazo) con sus respectivos análisis gráfico que la valida y su interpretación.
   iii) La descripción y explicación del modelo predictivo intuitivo.
   iv) Las Recomendaciones a la Universidad con todo lo que analizó.
3. Buena documentación y presentación de su código en Jupiter (Python) [**7 puntos**]
   i) Justificación del uso de cada librería
   ii) Nivel de documentación del código. (Orden y claridad del markdown usado para documentar el código)
4. Formalidad del documento pdf (tapa con grupo e integrantes, asignatura, tabla de contenidos, cabecera sobria y pie de página) [**7 puntos**]

$$\textbf{NOTA FINAL} = \frac{(ptje\ d.1 + ptje\ d.2 + ptje\ d.3 + ptje\ d.4 + ptje\ f.3 + ptje\ f.4)}{11}$$

BUEN DESARROLLO DE PROYECTO !!