1. Ejercicio Práctico 1: Clasificación de Hábitos de Fumadores

Contexto: Te encuentras como analista de datos en el laboratorio "HealthData Lab", un referente en investigación epidemiológica y análisis de datos de salud pública. Tu equipo ha recopilado datos de salud de individuos en un estudio observacional con el fin de identificar patrones relacionados con el hábito de fumar.

El desafío: Utilizar este conjunto de datos para clasificar a los individuos según sean fumadores o no fumadores. Esta clasificación ayudará en el desarrollo de campañas de salud pública y en la comprensión de los impactos del tabaquismo en la salud general.

Objetivo: Basándote en el conjunto de datos proporcionado, tu tarea es clasificar cada registro como fumador o no fumador. La categorización deberá estar fundamentada en un análisis exploratorio de los datos y las características observadas.

2. Instrucciones:

- **Exploración Inicial de Datos (4 punto):**
    - **Ejercicio 1** - Declara el tamaño del dataset. Con respecto a las columnas, identifica qué variables son numéricas y cuáles son categóricas. ¿Cuál es el individuo menos pesado? ¿Y el más alto? Dibuja un histograma de hombres y mujeres atendiendo a si fuman o no.
    - **Ejercicio 2 -** Sobre la variable continua "age" aplica una binarización por umbralización, tomando como umbral la media de las edades, incluyendo esta variable en el Dataframe como "age_bin". Haz un conteo de las dos categorías resultantes en la variable binarizada. Sobre la variable continua "Cholesterol" aplica una agrupación por cuantiles usando percentiles. Incluye esta variable en el dataframe como "Cholesterol_per".
    - **Ejercicio 3 -** Sobre la variable continua "fasting blood sugar" aplica un escalado máximo-mínimo. Dibuja un histograma de la variable original y otro histograma de la variable tras el escalado: ¿qué conclusiones sacas viendo ambos histogramas?
    - **Ejercicio 4** - Aisla la variable discreta "blood_group" en un dataframe que se componga de esa única variable. Genera 3 dataframes diferentes:
        - Uno con variables generadas mediante el método one-hot encoding
        - Otro con variables generadas mediante el método dummy coding
        - Un último con variables generadas mediante el método effect coding
- **Preprocesamiento de Datos (1 punto)**
    - **Ejercicio 5**: En esta sección, continuar con la preparación de nuestro conjunto de datos para el análisis. Basándose en las tareas realizadas anteriormente:
        - **División del conjunto de datos**: Segmentar los datos utilizando la librería scikitlearn con la semilla reproducible "1234", utilizando un 80% de conjunto de entrenamiento y un 20% de conjunto de testeo. Indicar el tamaño de ambos conjuntos.

- **Análisis Exploratorio de Datos (1 punto)**
    - **Ejercicio 6:** Explorar profundamente los datos para obtener insights que guíen la construcción del modelo:
        - **Distribución de variables:** Para la variable "age" aplicar una agrupación por cuantiles utilizando deciles. Para cada grupo calcular el WoE asociado de forma manual. Calcular el IV de la variable discretizada con respecto al target. ¿Crees que es una variable importante de cara a su relación con la variable objetivo (smoking)?
        - **Relaciones entre variables:** Realiza el mismo ejercicio para la variable "Height". Según los resultados obtenidos, ¿cuál crees que tiene un poder predictor más fuerte?

- **Selección de Características (1 punto)**
    - **Ejercicio 7**: Evaluar y seleccionar las características más informativas para el modelo:
        - **Determinación de características relevantes**: Utilizando el atributo ".corr()" del DataFrame en formato pandas enuncia las 3 variables que más se correlen con la variable objetivo. ¿Qué interpretación lógica puedes dar a la correlación obtenida con respecto a dichas variables?

- **Construcción del Modelo (1 punto)**
    - **Ejercicio 8:** Elegir y aplicar el modelo de clasificación adecuado a partir de las opciones vistas en clase:
        - **Entrenamiento:** Capacitar el modelo seleccionado con el conjunto de entrenamiento mediante el algoritmo kNN o el SVM. Justifica tu respuesta.

- **Evaluación del Modelo (1 punto)**
    - **Ejercicio 9:** Evaluar el rendimiento del modelo es clave para entender su efectividad. Se incluye:
        - **Comparación de métricas:** Observar y analizar métricas de rendimiento vistas en clase tales como el *accuracy*, la precisión, *recall*, *F1-score*, AUC-ROC y matriz de confusión**.** ¿Qué conclusiones se pueden extraer de cada una de estás métricas?
        - **Validación cruzada**: Utilizar la validación cruzada para optimizar hiperparámetros y confirmar la estabilidad del modelo. ¿Cuál es la mejor configuración de parámetros?

- **Interpretación de Resultados y Conclusiones (1 punto)**
    - **Ejercicio 10:** Analizar la influencia de cada característica y propuestas de mejora del modelo.
        - Según lo examinado en el modelo, ¿Cuál crees que es la característica más importante? Justifica tu respuesta.
        - ¿Cómo crees que podrías **mejorar** el modelo en futuras iteraciones?

3. Entrega:

- Se espera que completes el NoteBook adjunto al presente documento (clasificación_fumadores.ipynb) abarcando todas las fases mencionadas anteriormente. El informe (Notebook) debe incluir el código en Python empleado, gráficos, resultados y conclusiones.
- Solo debe ser entregado el documento 'clasificación_fumadores.ipynb' con la extensión de nombre y apellidos. Ejemplo: clasificación_fumadores_Juan_Cuesta.ipynb

4. Instrucciones y anotaciones extra:

- En el notebook proporcionado (clasificación_fumadores.ipynb), se detallan de manera precisa las instrucciones y requerimientos para cada ejercicio.
- Es fundamental que todas las respuestas, comentarios y análisis, especialmente aquellos relacionados con los gráficos, estén adecuadamente justificados.
- Esta práctica es de carácter **individual**, por lo que no se permite el trabajo en grupo.
- La fecha límite para la entrega de la práctica es el **23 de febrero**, incluido este día.