PARTE 1. CUESTIONES TIPO TEST

Rodee con un círculo la respuesta correcta.

1. ¿Cómo se denomina la función que optimiza un algoritmo de Machine Learning?
    a. Función Objetivo.
    b. Cross Validation.
    c. Métrica de evaluación.

2. En los equipos de BigData los perfiles técnicos básicos con los que se debería contar son:
    a. Todos los que se puedan considerar como analistas de negocio.
    b. Data Engineer (para preparar todos los accesos a datos) y Data Scientist (para determinar patrones de interés).
    c. Un Business Analyst, un DBA y un Scrum Master al menos.

3. ¿Qué es el Big Data?
    a. Un sistema para el almacenamiento y procesamiento masivo de datos.
    b. Datos muy grandes.
    c. Un lenguaje de programación que se desarrolló en el año 2000.

4. En el enfoque Deep Learning (Aprendizaje Profundo):
    a. Sólo se pueden realizar clasificaciones.
    b. Se pueden realizar clasificaciones y regresiones.
    c. No se pueden realizar regresiones.

5. Los proyectos de Analítica Avanzada, debieran de arrancar de una necesidad de negocio:
    a. Depende de quién desarrolle el proyecto, si lo hace IT sin problema.
    b. No, no hace falta, simplemente con que el proyecto funcione.
    c. Sí, debiera de ser así, de otra forma, es muy dudoso que el proyecto pueda ayudar a la organización a generar valor y a su correcta implantación.

6. ¿Qué falta en esta consulta: "SELECT nombre_cliente, sum(revenue) from orders"?
    a. La sentencia GROUP BY.
    b. La sentencia WHERE.
    c. Ninguna de las anteriores.

7. Qué tipo de variables son, respectivamente, las siguientes: a={ "nombre": "juan" , "apellidos" : "fernandez" } b={"juan","maría","josé"}
    a. un set y un diccionario.
    b. una lista y una tupla.
    c. un diccionario y un set.

8. Selecciona cuál de las siguientes afirmaciones es falsa:
    a. El método .fillna(), aplicado sobre una columna de un dataframe, devuelve un vector con True/False para cada elemento indicando si es o no un nulo.
    b. El método .describe(), aplicado sobre un dataframe, devuelve ciertos estadísticos básicos sobre las columnas numéricas (media, desviación, conteo, cuantiles principales…).
    c. El método .isna(), aplicado sobre una columna de un dataframe, devuelve un vector con True/False para cada elemento indicando si es o no un nulo.

9. Queremos modificar una variable de nuestro dataframe, selecciona el orden verdadero de cada uno de los siguientes métodos atendiendo a su eficiencia computacional:
    a. Vectorización, List Comprehensions, Dataframe.apply(), Iterar sobre el Dataframe.
    b. Iterar sobre el Dataframe, Vectorización, List Comprehensions, Dataframe.apply().
    c. List Comprehensions, Dataframe.apply(), Vectorización, Iterar sobre el DataFrame con .itertuples().

10. Cuál de los siguientes comandos generaría una gráfica correctamente:
    a. plt.plot(df[“paises”], df[“valores”], linewidth = 3, color = “verde oscuro”).
    b. plt.barh(y = df[“paises”], width = df[“valores”]).
    c. plt.scatter(df[“paises”], df[“valores”], alpha = ‘low transparency’).

11. ¿Qué supone más consumo de tiempo y recursos?
    a. Los trabajos previos de revisión de calidad y preparación de data.
    b. Preparar las distintas visualizaciones que se necesitan.
    c. Modelar la narrativa y storytelling.

12. Cuando pensemos en disponer de información en tiempo real...
    a. Hay que pensar cada cuanto un mismo usuario la consultará esperando datos diferentes.
    b. Debemos refrescarla cada vez que disponemos de datos nuevos.
    c. Hay que asegurar que al menos cada hora se actualice.

13. ¿Qué tipo da campo tiene que ser una variable para poder pintar un mapa?
    a. Geográfico.
    b. Decimal.
    c. Entero.

14. ¿En Power BI cuál es la diferencia entre campo calculado y medida?
    a. Ninguna, los 2 funcionan igual.
    b. El campo calculado se mostrará como un campo más de una tabla mientras que la medida no. También el campo se calculará siempre que se actualicen los datos mientras que la medida solo se calculará cuando sea llamado.
    c. La única diferencia está en el nombre.

15. ¿Qué función DAX se usa para traducir un código a icono?
    a. MEAN()
    b. MAX()
    c. UNICHAR()

16. ¿Dónde nunca evaluaremos un modelo?
    a. En el conjunto de datos de evaluación.
    b. En el conjunto de datos de test.
    c. En el conjunto de datos de entrenamiento.

17. A la hora de buscar la mejor partición en un nodo, un modelo de árbol de decisión de clasificación multiclase selecciona:
    a. Aquella que genera el menor RMSE.
    b. Aquella que genera el mayor RMSE.
    c. Aquella que genera la menor media ponderada del Gini de los nodos resultantes.

18. Cuando nos encontramos con outliers en nuestro caso de estudio lo que debemos hacer es…
    a. Tras analizar su impacto, eliminarlos siempre pues nos introducen ruido y pueden dañar el modelo y métricas.
    b. Tras analizar su impacto, llevar a cabo un análisis de los mismos y decidir si es correcto que permanezcan en el estudio o por el contrario eliminarlos al ser identificados como ruido o elementos no significativos.
    c. Tras analizar su impacto, siempre debemos mantenerlos y añadir su información a nuestros análisis.

19. ¿Cuál de las siguientes afirmaciones sobre Random Forests es FALSA?
    a. El método de agregación de los árboles de un Random Forest también se conoce como prunning.
    b. La predicción final es una media o la moda de los n árboles que componen el Random Forest dependiendo de si hablamos de una regresión o una clasificación respectivamente.
    c. Son modelos estocásticos, es decir, no deterministas, es decir, el proceso de entrenamiento no tiene por qué dar lugar a un mismo modelo si se repite varias veces.

20. La técnica PCA:
    a. No tiene problemas para tratar la no linealidad de los datos.
    b. Crea nuevas dimensiones combinando las variables originales de forma que maximiza la señal en cada una de esas nuevas dimensiones.
    c. Tiene en cuenta la variable target.

PARTE 2. CUESTIONES CORTAS

Responda a las siguientes cuestiones en la misma hoja que se plantean.

CUESTIÓN 1

Tenemos dos tablas en una base de datos: "empleados" y "ventas". La tabla "empleados" tiene los siguientes campos:

- id_empleado: ID único del empleado
- nombre: nombre del empleado
- cargo: cargo del empleado en la empresa
- salario: salario mensual del empleado

La tabla "ventas" tiene los siguientes campos:

- id_venta: ID único de la venta
- id_empleado: ID del empleado que realizó la venta
- fecha: fecha en la que se realizó la venta
- monto: monto total de la venta

Necesitamos escribir una consulta SQL que nos muestre el nombre del empleado, el cargo del empleado, el monto total de ventas realizadas por el empleado en el año 2021 y el porcentaje que representa ese monto sobre el total de ventas del año 2021. Además, queremos que los resultados estén ordenados de mayor a menor por el monto total de ventas realizadas por el empleado.

CUESTIÓN 2

Estamos preparando este dataset para el entrenamiento de modelos de machine learning.

	A	B	C	y
0	Cat_A	10.0	10000	1
1	Cat_A	12.0	11000	0
2	Cat_B	30.0	9000	0
3	Cat_A	9.0	8000	1
4	Cat_B	NaN	11500	1

En esta muestra de 5 filas puedes ver que:

- A es una variable categórica binaria
- La escala de las variables numéricas B y C
- La presencia de nulos en la Columna B

Indica los pasos que seguirías para preparar este dataset para el entrenamiento de modelos. No es necesario que escribas código para indicar estos pasos, aunque puedes hacerlo si lo prefieres.

CUESTIÓN 3

Explica en tus términos cómo se produce el proceso de un entrenamiento de un árbol de decisión de clasificación. Puedes usar para tus explicaciones (si lo necesitas) el dataset de la cuestión anterior.

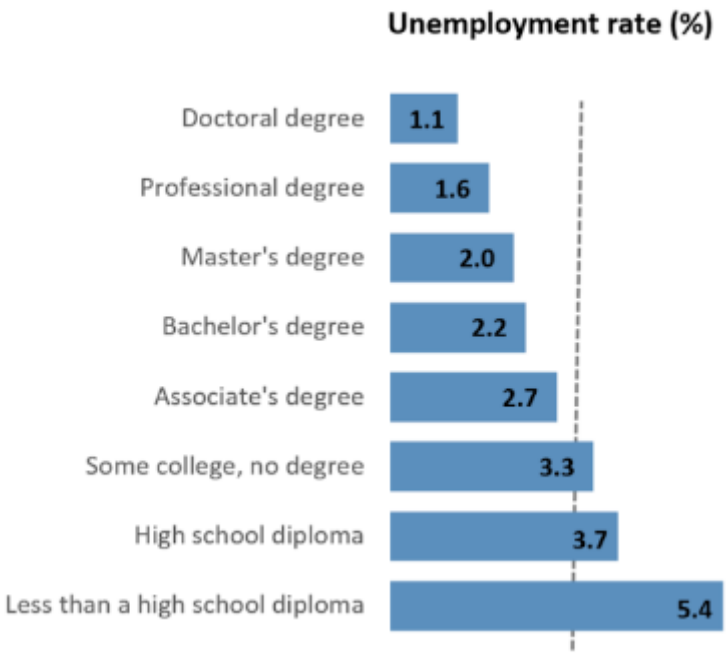
CUESTIÓN 4

¿Qué es el gobierno del dato? Describe brevemente sus principios, beneficios y roles.

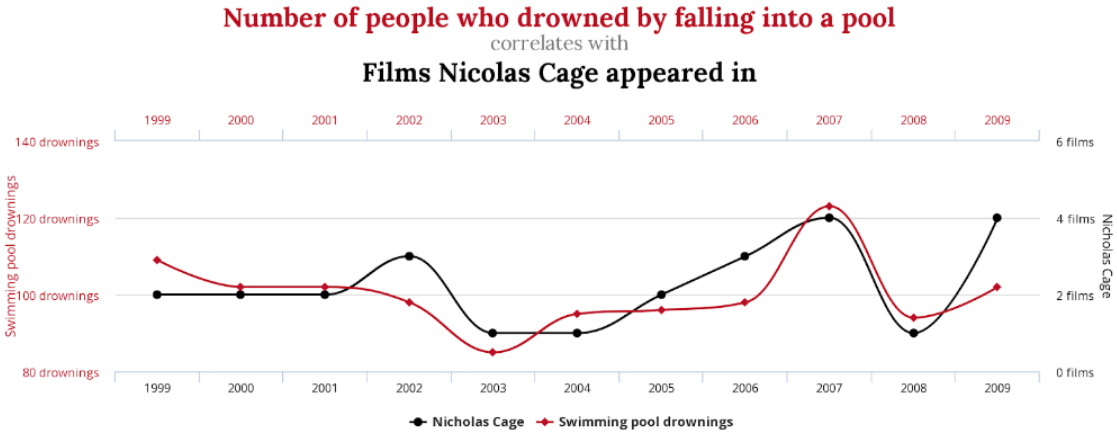
CUESTIÓN 5:

La afirmación de "correlación no implica causalidad" es de sobra conocida, pero entonces, ¿qué origina la correlación cuando no hay una relación causa efecto entre las variables? Observa la asociación entre los siguientes pares de variables y explica qué puede haber detrás de sus respectivas correlaciones (se valorará positivamente que incluyas referencias a coeficientes de correlación y confianza estadística en tus respuestas).

- a) Nivel educativo y tasa de desempleo



- b) Ahogamientos en cierto estado de EEUU y número de películas protagonizadas por Nicolas Cage



- c) venta de helados y muertes por ataques de tiburones