Apuntes

SIMULACRO DE EXAMEN 01

Análisis de Datos Masivos para el Negocio

Duración: 2 horas **Material permitido**: RStudio, apuntes, chuleta impresa o Word **Entrega**: Respuestas escritas + archivo .R adjunto **Datasets**: `diabetes.csv` y `Mall_Customers.csv`

INSTRUCCIONES: Se te proporcionan dos bases de datos. Responde a cada pregunta con el código R necesario y una breve interpretación del resultado. Recuerda que se evalúa el planteamiento y la interpretación, no solo que el código funcione.

PARTE 1: DATASET DIABETES (Clasificación)

Trabajarás con el archivo `diabetes.csv` que contiene datos clínicos de pacientes para predecir si desarrollan diabetes (variable `Outcome`: 1 = diabetes, 0 = no diabetes).

Variables: Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome

---

Pregunta 1 (Carga y exploración) — 0.5 pts

Carga el dataset `diabetes.csv` en R. ¿Cuántas filas y columnas tiene? ¿Qué tipo de dato tiene cada variable?

Pregunta 2 (Nulos y anomalías) — 0.5 pts

¿Existen valores nulos (NA) en el dataset? Ejecuta el código necesario para comprobarlo.

Pregunta 3 (Trampa de calidad de datos) — 1 pt

Observa las variables Glucose, BloodPressure, SkinThickness, Insulin y BMI. ¿Tiene sentido que alguna de estas variables tenga valor 0? ¿Qué significan esos ceros en realidad? ¿Cuántos registros con valor 0 hay en cada una de esas variables?

Pregunta 4 (Limpieza) — 0.5 pts

Elimina las filas donde Glucose o BMI sean 0 (son valores imposibles clínicamente). ¿Cuántos registros quedan tras la limpieza?

Pregunta 5 (Estadística descriptiva) — 0.5 pts

Calcula la media y mediana de Glucose para los pacientes CON diabetes y para los pacientes SIN diabetes. ¿Qué observas? Interpreta la diferencia en lenguaje de negocio (como si se lo explicaras a un médico).

Pregunta 6 (Filtro y conteo) — 0.5 pts

¿Cuántas mujeres mayores de 40 años tienen diabetes (Outcome = 1)? ¿Qué porcentaje representan del total de mujeres mayores de 40?

Pregunta 7 (Tablas de frecuencia) — 0.5 pts

Crea una tabla de frecuencias y una tabla de proporciones de la variable Outcome. ¿Cuál es la probabilidad (proporción) de tener diabetes en este dataset?

Pregunta 8 (Crear variable binaria) — 0.5 pts

Crea una nueva columna llamada `Obesidad` que valga 1 si BMI >= 30 y 0 en caso contrario. ¿Cuántos pacientes son obesos?

Pregunta 9 (Correlación) — 0.5 pts

Calcula la matriz de correlación entre todas las variables numéricas. ¿Qué variable tiene la correlación más alta con Outcome? ¿Es fuerte o débil? Interpreta el resultado.

Pregunta 10 (Preparación del modelo — Conversión a factor) — 0.5 pts

Convierte la variable Outcome a factor. ¿Por qué es necesario este paso antes de crear un modelo de clasificación?

Pregunta 11 (División train/test) — 0.5 pts

Divide el dataset en conjunto de entrenamiento (80%) y testeo (20%) usando `createDataPartition`. Recuerda fijar la semilla. ¿Cuántas filas tiene cada conjunto?

Pregunta 12 (Regresión logística) — 1 pt

Crea un modelo de regresión logística para predecir Outcome usando todas las demás variables. ¿Qué variables son significativas (p-valor < 0.05)? Interpreta el coeficiente de Glucose: si la glucosa aumenta en 1 unidad, ¿qué ocurre con la probabilidad de diabetes?

Pregunta 13 (Predicción con regresión logística) — 1 pt

Usa el modelo logístico para predecir sobre los datos de test. Recuerda que la regresión logística da probabilidades: convierte las probabilidades a 0/1 usando un umbral de 0.5. Calcula la matriz de confusión. ¿Cuál es el Accuracy del modelo?

Pregunta 14 (Árbol de decisión) — 1 pt

Crea un árbol de decisión con `rpart` para predecir Outcome. Dibuja el árbol con `rpart.plot`. ¿Cuál es la variable más importante según el árbol (la que está en la raíz)?

Pregunta 15 (Comparación de modelos) — 1 pt

Calcula la matriz de confusión del árbol de decisión sobre los datos de test. Compara el Accuracy de la regresión logística y del árbol de decisión. ¿Qué modelo recomendarías al hospital y por qué? Comenta también la Sensibilidad: en un contexto médico, ¿qué es más grave, un falso positivo o un falso negativo?

PARTE 2: DATASET MALL CUSTOMERS (Clustering)

Trabajarás con el archivo `Mall_Customers.csv` que contiene datos de clientes de un centro comercial.

Variables: CustomerID, Gender, Age, Annual Income (k$), Spending Score (1-100)

Pregunta 16 (Carga y exploración) — 0.5 pts

Carga el dataset. ¿Cuántas filas y columnas tiene? ¿Hay valores nulos?

Pregunta 17 (Selección de variables para clustering) — 0.5 pts

Para hacer clustering, ¿qué variables seleccionarías? Justifica por qué descartas alguna. Recuerda que clustering necesita variables numéricas.

Pregunta 18 (Número óptimo de clusters) — 1 pt

Determina el número óptimo de clusters usando NbClust (si se cuelga, usa el método del codo como alternativa). ¿Cuántos grupos recomienda? Muestra el código y el resultado.

Pregunta 19 (K-means) — 1 pt

Ejecuta el algoritmo kmeans con el número de clusters que obtuviste. Muestra los centros de cada grupo. Interpreta cada grupo en lenguaje de negocio (como si le explicaras al director del centro comercial qué tipo de clientes tiene).

Pregunta 20 (Pregunta de negocio) — 1 pt

El director del centro comercial quiere lanzar una campaña de fidelización dirigida al grupo de clientes más valioso. Basándote en los resultados del clustering, ¿a qué grupo le recomendarías dirigir la campaña? ¿Qué características tiene ese grupo? ¿Qué tipo de acciones de marketing sugerirías?

RESUMEN DE PUNTUACIÓN

Bloque	Preguntas	Puntos
Carga y exploración	1, 2, 16	1.5
Calidad y limpieza	3, 4	1.5

Bloque	Preguntas	Puntos
Descriptivo y filtros	5, 6, 7, 8	2.0
Correlación	9	0.5
Preparación modelo	10, 11	1.0
Regresión logística	12, 13	2.0
Árbol de decisión	14, 15	2.0
Clustering	17, 18, 19	2.5
Interpretación negocio	20	1.0
TOTAL	**20**	**14 pts (ajustar a 10)**

Nota: La puntuación se escala a 10. Cada pregunta tiene una parte de código y una parte de interpretación. El código sin interpretación puntúa la mitad.