

## MINERÍA DE DATOS

### GUIA DE REFORZAMIENTO

**Asignatura:** Minería de Datos

**Objetivo:** Reforzar conceptos teóricos y prácticos necesarios para rendir la evaluación con éxito.

**Preguntas de desarrollo (debe elaborar un ejemplo ficticio para cada concepto, al igual que lo visto en clases)**

1. Explique por qué es fundamental evaluar la calidad de los resultados obtenidos en un modelo de minería de datos.
2. Describa la relación entre la calidad del dato y la validez del modelo predictivo.
3. ¿Cómo influyen la exactitud, completitud, consistencia y actualidad en la calidad de un dataset?
4. Analice el impacto que pueden tener los datos mal categorizados dentro de un proceso de minería de datos.
5. Explique qué es la Ingeniería de Entrada y por qué es un paso clave en la preparación de datos.
6. ¿Por qué es importante definir correctamente el rol de cada variable (Feature, Target, Meta)?
7. Describa el propósito de la Ingeniería de Salida dentro del flujo de minería de datos.
8. Fundamente la importancia de seleccionar atributos relevantes antes de entrenar un modelo.
9. Compare dos métodos para seleccionar atributos y explique en qué situaciones cada uno es más apropiado.
10. Explique el concepto de discretización y sus ventajas dentro de la construcción de modelos.
11. Mencione dos riesgos de discretizar de manera incorrecta un atributo numérico.
12. ¿Qué ventajas ofrece usar aprendizaje no supervisado en un conjunto de datos sin variable objetivo?
13. Compare brevemente k-Means y Clustering Jerárquico en términos de objetivo y uso.
14. Desde una perspectiva teórica, analice cómo el bootstrapping ayuda a evaluar la estabilidad de un modelo.
15. Explique la diferencia conceptual entre el proceso de entrenamiento y el proceso de testeo.
16. ¿Por qué no es suficiente evaluar un modelo solo con un Train/Test Split? Fundamente.
17. Describa el funcionamiento general de la validación cruzada y por qué reduce el sesgo.
18. Analice el valor de métricas como precisión, recall y F1-score dentro de un problema de clasificación.
19. ¿En qué situaciones es más útil analizar AUC que Accuracy? Fundamente.
20. Explique cómo se interpreta una matriz de confusión y qué tipo de información aporta.
21. Fundamente la importancia de comparar varios modelos bajo las mismas condiciones experimentales.
22. ¿Por qué algunos algoritmos ofrecen predicción de probabilidades y otros no? Explique su utilidad.
23. Analice qué se puede concluir cuando dos modelos obtienen el mismo Accuracy pero diferente AUC.
24. Explique los conceptos MAE, RMSE y  $R^2$  en modelos de predicción numérica.

25. Desde el punto de vista del modelamiento, ¿qué implica la existencia de una brecha entre el resultado esperado y el obtenido?
26. Fundamente el rol de la calibración al evaluar modelos probabilísticos.
27. Describa un proceso iterativo de mejora de modelo a partir de sus brechas de desempeño.
28. Analice cómo la creatividad e innovación pueden surgir del descubrimiento de patrones inesperados.
29. ¿Por qué es importante la ética en el análisis, comunicación e interpretación de resultados?
30. Describa un escenario donde manipular métricas sería antiético y explique por qué.

### **CASO – Satisfacción de Usuarios en la Plataforma SaludPlus**

SaludPlus es una plataforma digital que permite a los usuarios obtener citas médicas, revisar resultados de exámenes y acceder a información de salud general. Durante los últimos meses, la gerencia ha notado un descenso progresivo en la valoración positiva del servicio.

El área de análisis de datos inició un proyecto de minería de datos para comprender los factores que influyen en la satisfacción. Se recopiló información de uso general como:

- Frecuencia semanal de acceso
- Tipo de servicio utilizado (exámenes, consulta, información)
- Tiempo promedio de respuesta de la plataforma
- Nivel de satisfacción (de 1 a 5) → Variable objetivo
- Reportes de errores o fallas experimentadas
- Perfil del usuario (edad y ocupación)

Antes de entrenar modelos predictivos, el equipo decidió realizar un proceso completo de análisis de calidad de datos y selección de atributos, con el fin de comprender mejor qué variables influyen en la satisfacción.

### **PREGUNTAS DEL CASO**

1. ¿Por qué es importante analizar la calidad de los datos antes de construir un modelo de satisfacción en SaludPlus?
2. ¿Qué beneficio aporta identificar cuáles atributos influyen más en el nivel de satisfacción?
3. ¿Cómo podría el análisis de patrones ayudar a mejorar la plataforma SaludPlus?