Cuestionario: Preprocesado y Aprendizaje Supervisado

1. ¿Cuál es el objetivo principal del preprocesamiento de datos?
    a) Visualizar los datos sin procesarlos.
    b) **Preparar los datos para el análisis posterior.**
    c) Eliminar todos los valores faltantes.
    d) Calcular la media y la mediana de todas las variables.
2. ¿Cuál de las siguientes NO es una tarea común del preprocesamiento?
    a) Tratamiento de valores faltantes.
    b) Detección y tratamiento de outliers.
    c) Estandarización de datos.
    d) **Realizar un análisis de regresión lineal.**
3. ¿Qué medida de centralidad se recomienda utilizar cuando los datos presentan outliers?
    a) Media.
    b) **Mediana.**
    c) Moda.
    d) Rango.
4. ¿Qué método de estandarización se considera el más efectivo según [MC88]?
    a) Z-score.
    b) Escalamiento decimal.
    c) **Estandarización por rangos.**
    d) Ninguno de los anteriores.
5. ¿Cuál es el objetivo de la selección de variables?
    a) **Elegir las variables más relevantes para el análisis.**
    b) Aumentar la dimensionalidad de los datos.
    c) Eliminar todas las variables con valores faltantes.
    d) Convertir variables categóricas en variables continuas.
6. ¿Qué medida de correlación se utiliza para variables cuantitativas con dependencia lineal?
    a) Fisher score.
    b) V de Cramer.
    c) Ganancia de entropía.
    d) **Correlación de Pearson.**
7. Describe dos posibles causas de la aparición de outliers en un conjunto de datos.
    - **Errores en la introducción de datos:** Por ejemplo, escribir un dígito incorrecto.
    - **Variabilidad natural:** La presencia de un valor extremo debido a la naturaleza del fenómeno estudiado.
8. ¿Qué es el principio de parsimonia en estadística?
    - **Apuesta por emplear los modelos más sencillos posibles.**

9. Explica brevemente la técnica de jackknife para la detección de observaciones influyentes.
    - **Consiste en realizar el análisis estadístico sin cada uno de los datos de la muestra, para observar cómo cambia el índice de calidad del modelo.**
10. ¿Qué problema se busca resolver con la ponderación de variables?
    - **Que no todas las variables son igualmente importantes para el análisis.**
11. Los outliers siempre deben ser eliminados del análisis.
    - **Falso.** Depende del análisis y la naturaleza del outlier.
12. La desviación estándar muestral es una medida de dispersión de los datos.
    - **Verdadero.**
13. El coeficiente de correlación de Pearson puede tomar valores entre 0 y 1.
    - **Falso.** Puede tomar valores entre -1 y 1.
14. La estandarización de datos puede ayudar a mejorar la comparabilidad entre variables.
    - **Verdadero.**
15. El método de jackknife se utiliza para calcular la media y la mediana de una muestra.
    - **Falso.** Se usa para detectar observaciones influyentes.
16. La función de densidad de probabilidad (f.d.p) de una variable normal se representa con una campana de Gauss.
    - **Verdadero.**
17. La mediana es una medida de centralidad más robusta que la media en presencia de valores atípicos.
    - **Verdadero.**
18. La varianza muestral se calcula dividiendo la suma de los cuadrados de las diferencias entre cada dato y la media muestral entre el número total de datos (n).
    - **Falso.** Se divide entre n-1.
19. La tipificación de una variable aleatoria normal la transforma en una variable normal estándar con media 0 y varianza 1.
    - **Verdadero.**
20. La suma de dos variables aleatorias normales independientes también sigue una distribución normal.
    - **Verdadero.**
21. Un valor influyente siempre es un outlier, pero no todos los outliers son valores influyentes.
    - **Verdadero.**

22. El rango intercuartílico se utiliza para detectar outliers.
    - **Verdadero.**
23. La estandarización de datos se utiliza para convertir todas las variables a una misma escala.
    - **Verdadero.**
24. El preprocesamiento de datos es un paso opcional en el análisis de datos.
    - **Falso.** Es un paso fundamental para preparar los datos.
25. El objetivo del preprocesamiento de datos es obtener modelos más precisos y fáciles de interpretar.
    - **Verdadero.**
26. ¿Cuál es el objetivo principal del aprendizaje supervisado?
    a) Agrupar datos sin etiquetar en diferentes clústeres.
    b) **Inferir un modelo de predicción a partir de datos etiquetados.**
    c) Encontrar patrones ocultos en datos sin información previa sobre la salida.
    d) Optimizar un agente para que tome decisiones en un entorno.
27. ¿En qué se diferencia la clasificación de la regresión en el contexto del aprendizaje supervisado?
    a) La clasificación predice valores continuos, mientras que la regresión predice clases discretas.
    b) **La clasificación predice clases discretas, mientras que la regresión predice valores continuos.**
    c) La clasificación utiliza algoritmos basados en árboles, mientras que la regresión utiliza algoritmos lineales.
    d) La clasificación se utiliza para datos etiquetados, mientras que la regresión se utiliza para datos sin etiquetar.
28. ¿Cuál de los siguientes es un ejemplo de problema de clasificación?
    a) Predecir el precio de una casa.
    b) Predecir la temperatura para mañana.
    c) **Predecir si un correo electrónico es spam o no.**
    d) Predecir el número de clientes que visitarán una tienda.
29. ¿Qué es el sobreajuste en el aprendizaje supervisado?
    a) **Cuando el modelo se ajusta demasiado a los datos de entrenamiento y no generaliza bien a datos nuevos.**
    b) Cuando el modelo no se ajusta lo suficiente a los datos de entrenamiento y tiene un alto error.
    c) Cuando el modelo se entrena con un conjunto de datos demasiado pequeño.
    d) Cuando el modelo utiliza demasiadas variables predictoras.

30. ¿Cuál de las siguientes técnicas se utiliza para evaluar el rendimiento de un modelo en aprendizaje supervisado?
    a) Clustering.
    b) Reducción de dimensionalidad.
    c) **Validación cruzada.**
    d) Análisis de componentes principales.
31. ¿Qué tipo de modelo generativo se utiliza en el clasificador Naive Bayes Gaussiano?
    a) **Distribución gaussiana.**
    b) Distribución multinomial.
    c) Distribución de Bernoulli.
    d) Distribución uniforme.
32. ¿Cuál de las siguientes métricas se utiliza comúnmente para evaluar modelos de clasificación binaria?
    a) Error cuadrático medio.
    b) **Precisión**.
    c) Coeficiente de determinación R^2.
    d) Error absoluto medio.

33. Describe dos ventajas del clasificador Naive Bayes.

    - **Es extremadamente rápido tanto para el entrenamiento como para la predicción.**
    - **Suele ser muy fácil de interpretar.**

34. ¿Cuál es la principal limitación del algoritmo kNN?

    - **El alto coste computacional, especialmente con grandes conjuntos de datos.**

35. ¿Qué son los datos etiquetados en aprendizaje supervisado?

    - **Son datos que incluyen tanto las variables de entrada como la salida o etiqueta correspondiente a cada instancia.**

36. El algoritmo kNN es un ejemplo de aprendizaje perezoso.

    - **Verdadero.**

37. La regresión lineal siempre es la mejor opción para modelar relaciones entre variables.

    - **Falso.** Depende de la naturaleza de la relación entre las variables.

38. La validación cruzada es una técnica que ayuda a prevenir el sobreajuste.

    - **Verdadero.**

39. El clasificador Naive Bayes asume que todas las variables predictoras son independientes entre sí.

    - **Verdadero.**

40. El algoritmo kNN puede utilizarse tanto para clasificación como para regresión.

    - **Verdadero.**

41. La "maldición de la dimensionalidad" beneficia al rendimiento de los algoritmos de aprendizaje automático.
    - **Falso.** La "maldición de la dimensionalidad" afecta negativamente al rendimiento, especialmente en algoritmos basados en distancias como kNN.
42. Es recomendable ponderar el voto en función de la distancia en kNN cuando se busca reducir el impacto del ruido.
    - **Verdadero.**
43. La matriz de confusión permite evaluar la precisión de un modelo de clasificación.
    - **Verdadero.**
44. La curva ROC se usa para analizar el rendimiento de los modelos de regresión.
    - **Falso.** Se usa para analizar el rendimiento de los modelos de clasificación.
45. En la validación cruzada k-fold, el modelo se entrena y evalúa una sola vez.
    - **Falso.** Se entrena y evalúa k veces.
46. El aprendizaje supervisado siempre es más efectivo que el aprendizaje no supervisado.
    - **Falso.** Depende del problema y la disponibilidad de datos etiquetados.
47. El teorema de Bayes es la base de los clasificadores Naive Bayes.
    - **Verdadero.**
48. En un problema de regresión, el objetivo es predecir una clase o categoría.
    - **Falso.** El objetivo es predecir un valor numérico continuo.
49. El preprocesamiento de datos es fundamental para el aprendizaje supervisado.
    - **Verdadero.** Aunque no se menciona explícitamente en las fuentes, el preprocesamiento de datos es una etapa importante para preparar los datos para el aprendizaje supervisado.
50. La selección de variables no influye en el rendimiento de un modelo de aprendizaje supervisado.
    - **Falso.** Una buena selección de variables puede mejorar la precisión y la interpretabilidad del modelo.