Estadística

Histograma
Población/Muestra
Nivel de confianza y el rango
Correlación 0 y dependencia. Correlación lineal

Análisis
Proceso EDA

- Comprender el problema el ámbito
- Formular hipótesis
- Datos
- Univariante
- Bivariante
- …
- Conclusiones

Hipótesis
Univariante
Bivariante
Multivariante
Con dos variables categóricas cómo verías si

SQL

Sesgos

- De muestreo
- Tunel visión

ML
Tipos Supervisado/No-Supervisado/Por refuerzo
Cómo funciona un Regresor lineal
Regresiones
Clasificación
Precisión/Recall

Deep Learning
Qué es una red neuronal?
Qué es una red convolucional?
Una **red neuronal convolucional** (o CNN, por sus siglas en inglés: *Convolutional Neural Network*) es un tipo de red neuronal especialmente diseñada para procesar datos que tienen una estructura de grilla, como imágenes. Son ampliamente utilizadas en tareas de visión por computadora, como clasificación de imágenes, detección de objetos y segmentación de imágenes.

Conceptuales
Ciencia de Datos
Machine Learning

Deep Learning

Estadística

1. **Pregunta:** ¿Cuál es la diferencia entre población y muestra?
   Respuesta: Una población incluye todos los elementos de interés, mientras que una muestra es un subconjunto representativo de la población.
2. **Pregunta:** ¿Qué es el p-valor y cómo se interpreta?
   Respuesta: Es la probabilidad de obtener resultados iguales o más extremos bajo la hipótesis nula. Un p-valor bajo sugiere que la hipótesis nula debe rechazarse.
3. **Pregunta:** ¿Qué es un intervalo de confianza?
   Respuesta: Un rango estimado que contiene el parámetro poblacional con un nivel de confianza especificado (e.g., 95%).
4. **Pregunta:** ¿Qué es una variable categórica y cómo se analiza?
   Respuesta: Es una variable que toma valores discretos o etiquetas. Se analiza con tablas de frecuencia, proporciones y gráficos de barras.
5. **Pregunta:** ¿Qué representa la desviación estándar?
   Respuesta: Es una medida de dispersión que indica cuánto se desvían los datos de la media.
6. **Pregunta:** ¿Qué diferencia hay entre correlación y regresión?
   Respuesta:
   - La correlación mide la relación y dirección entre dos variables.
   - La regresión predice valores de una variable basada en otra.
7. **Pregunta:** ¿Qué significa que dos eventos sean mutuamente excluyentes?
   Respuesta: Que no pueden ocurrir al mismo tiempo. La probabilidad de ambos es 0.
8. **Pregunta:** ¿Qué es un test de hipótesis?
   Respuesta: Es una técnica para evaluar si una afirmación sobre un parámetro poblacional es consistente con los datos.
9. **Pregunta:** ¿Qué es un estadístico de prueba?
   Respuesta: Es un valor calculado a partir de los datos que se compara con una distribución para decidir si rechazar la hipótesis nula.
10. **Pregunta:** ¿Qué es la regresión logística y para qué se usa?
    Respuesta: Es un modelo usado para predecir variables categóricas (e.g., clasificación binaria).
11. **Pregunta:** ¿Qué es la ley de los grandes números?
    Respuesta: Establece que a medida que aumenta el tamaño de la muestra, la media muestral se acerca a la media poblacional.
12. **Pregunta:** ¿Qué es la probabilidad condicional?
    Respuesta: Es la probabilidad de que ocurra un evento dado que otro ha ocurrido.
13. **Pregunta:** ¿Qué es el coeficiente de determinación (R^2)?
    Respuesta: Indica qué porcentaje de la variabilidad en la variable dependiente está explicado por las independientes.
14. **Pregunta:** ¿Qué es una distribución normal?
    Respuesta: Una distribución simétrica en forma de campana, donde la media, mediana y moda coinciden.

15. **Pregunta:** ¿Cómo se detecta heterocedasticidad en un modelo de regresión?
    Respuesta: Analizando los residuos mediante gráficos o pruebas como la de Breusch-Pagan.

Exploratory Data Analysis (EDA)

1. **Pregunta:** ¿Cómo lidias con datos faltantes?
   Respuesta: Eliminando filas/columnas, imputando valores o usando modelos predictivos.
2. **Pregunta:** ¿Qué son los outliers y cómo los detectas?
   Respuesta: Son valores extremos que se detectan con gráficos como boxplots o con métricas como el IQR.
3. **Pregunta:** ¿Qué son las estadísticas descriptivas?
   Respuesta: Medidas que resumen datos, como media, mediana, moda, varianza y percentiles.
4. **Pregunta:** ¿Qué es un heatmap y cuándo se usa?
   Respuesta: Es un gráfico de matriz coloreado que representa correlaciones o relaciones entre variables.
5. **Pregunta:** ¿Qué técnicas usarías para analizar la distribución de una variable continua?
   Respuesta: Histogramas, boxplots, gráficos de densidad y medidas de tendencia central y dispersión.
6. **Pregunta:** ¿Por qué es importante analizar la correlación entre variables?
   Respuesta: Para entender relaciones y evitar multicolinealidad en modelos predictivos.
7. **Pregunta:** ¿Qué pasos sigues para limpiar un conjunto de datos?
   Respuesta: Manejar datos faltantes, eliminar duplicados, corregir formatos, detectar outliers y normalizar datos.
8. **Pregunta:** ¿Qué es la transformación de datos?
   Respuesta: Cambiar la escala o aplicar funciones (e.g., logaritmos) para mejorar la interpretación o modelado.
9. **Pregunta:** ¿Cómo evalúas la simetría de una distribución?
   Respuesta: Usando medidas como el sesgo (skewness) o visualizaciones como histogramas.
10. **Pregunta:** ¿Qué significa que una variable tenga kurtosis alta?
    Respuesta: Que tiene colas más pesadas o extremos más frecuentes que una distribución normal.
11. **Pregunta:** ¿Qué significa escalar datos?
    Respuesta: Transformar los valores para que estén en el mismo rango, usando normalización o estandarización.
12. **Pregunta:** ¿Qué es un scatterplot y qué información ofrece?
    Respuesta: Un gráfico de dispersión que muestra la relación entre dos variables numéricas.
13. **Pregunta:** ¿Qué son las dimensiones de los datos?
    Respuesta: El número de variables o características en el conjunto de datos.

14. **Pregunta:** ¿Cómo reduces la dimensionalidad de un dataset?
    Respuesta: Usando técnicas como PCA (Análisis de Componentes Principales) o seleccionando características.
15. **Pregunta:** ¿Qué herramientas visuales usas para analizar datos categóricos?
    Respuesta: Tablas de frecuencia, gráficos de barras, diagramas de pastel y gráficos de mosaico.

Machine Learning

1. **Pregunta:** ¿Cuál es la diferencia entre aprendizaje supervisado y no supervisado?
   Respuesta: Supervisado usa etiquetas para entrenar, mientras que no supervisado busca patrones en datos no etiquetados.
2. **Pregunta:** ¿Qué es el underfitting y cómo se soluciona?
   Respuesta: Es cuando un modelo no aprende lo suficiente de los datos. Puede solucionarse aumentando la complejidad del modelo o recolectando más datos.
3. **Pregunta:** ¿Qué es la validación cruzada?
   Respuesta: Es una técnica para evaluar el rendimiento de un modelo dividiendo los datos en subconjuntos de entrenamiento y prueba.
4. **Pregunta:** ¿Qué métricas usas para evaluar un modelo de clasificación?
   Respuesta: Precisión, recall, F1-score y matriz de confusión.
5. **Pregunta:** ¿Qué es un hiperparámetro y cómo se ajusta?
   Respuesta: Es un parámetro del modelo definido antes del entrenamiento. Se ajusta usando técnicas como grid search o random search.
6. **Pregunta:** ¿Qué diferencia hay entre regresión lineal y regresión polinómica?
   Respuesta: La regresión lineal asume una relación lineal, mientras que la polinómica incluye términos no lineales (e.g., cuadrados, cúbicos).
7. **Pregunta:** ¿Qué es la regularización y para qué sirve?
   Respuesta: Es una técnica que penaliza la complejidad del modelo para evitar sobreajuste.
8. **Pregunta:** ¿Cuál es la diferencia entre un árbol de decisión y un bosque aleatorio?
   Respuesta: Un árbol de decisión es un modelo único, mientras que un bosque aleatorio combina múltiples árboles para mejorar la precisión.
9. **Pregunta:** ¿Qué es el aprendizaje por refuerzo?
   Respuesta: Es un tipo de aprendizaje donde un agente toma decisiones para maximizar recompensas en un entorno.
10. **Pregunta:** ¿Qué es la función de pérdida en machine learning?
    Respuesta: Es una métrica que evalúa qué tan bien se ajusta un modelo a los datos.
11. **Pregunta:** ¿Qué es el balanceo de clases y por qué es importante?
    Respuesta: Ajustar las proporciones entre clases en datos desequilibrados para evitar sesgos en el modelo.
12. **Pregunta:** ¿Qué es un modelo de clustering?
    Respuesta: Un modelo que agrupa datos en subconjuntos basados en similitudes.
13. **Pregunta:** ¿Qué es un pipeline en machine learning?
    Respuesta: Una secuencia de pasos automatizados para preprocesar datos y entrenar un modelo.

14. **Pregunta:** ¿Cómo diferencias entre features importantes y no importantes?
    Respuesta: Usando técnicas como importancia de características o selección basada en modelos (e.g., Lasso).
15. **Pregunta:** ¿Qué es un modelo ensamble y para qué se usa?
    Respuesta: Combina varios modelos para mejorar la precisión y reducir errores.

La **precisión** y el **recall** son métricas comunes para evaluar el desempeño de un modelo, especialmente en problemas de clasificación. Estas métricas son útiles para entender cómo un modelo maneja las clases positivas y negativas, y son especialmente importantes en conjuntos de datos desbalanceados.

1. Precisión (Precision):

Indica qué proporción de las predicciones positivas realizadas por el modelo son realmente correctas. En otras palabras, mide la calidad de las predicciones positivas.

Precisio´n=Verdaderos Positivos (TP)Verdaderos Positivos (TP)+Falsos Positivos (FP)\text{Precisión} = \frac{\text{Verdaderos Positivos (TP)}}{\text{Verdaderos Positivos (TP)} + \text{Falsos Positivos (FP)}}Precisio´n=Verdaderos Positivos (TP)+Falsos Positivos (FP)Verdaderos Positivos (TP)

- **TP (True Positives)**: Casos positivos correctamente clasificados.
- **FP (False Positives)**: Casos negativos que el modelo clasificó como positivos erróneamente.

Interpretación:

- Alta precisión significa que la mayoría de las predicciones positivas son correctas, pero no garantiza que el modelo detecte todos los casos positivos.

2. Recall (Sensibilidad o Tasa de Verdaderos Positivos):

Indica qué proporción de los casos positivos reales fueron detectados por el modelo. Mide la capacidad del modelo para encontrar todos los positivos.

Recall=Verdaderos Positivos (TP)Verdaderos Positivos (TP)+Falsos Negativos (FN)\text{Recall} = \frac{\text{Verdaderos Positivos (TP)}}{\text{Verdaderos Positivos (TP)} + \text{Falsos Negativos (FN)}}Recall=Verdaderos Positivos (TP)+Falsos Negativos (FN)Verdaderos Positivos (TP)

- **FN (False Negatives)**: Casos positivos reales que el modelo no detectó.

Interpretación:

- Alto recall significa que el modelo detecta la mayoría de los casos positivos, pero no garantiza que todas las predicciones positivas sean correctas.