RESPUESTAS

RESPUESTAS A LOS EJERCICIOS 10-15

EJERCICIO 10 - Clasificación Multiclase Iris Dataset

Pregunta: ¿Cuál de los tres modelos (Logistic Regression, SVM, KNN) mostró mejor rendimiento?

Respuesta: Los tres modelos tuvieron rendimiento idéntico con 100% de precisión en todas las métricas. No hay diferencia porque el dataset Iris es perfectamente separable. Cualquiera de los tres modelos funciona igual de bien en este caso.

EJERCICIO 11 - Random Forest con GridSearchCV

Pregunta: ¿Cómo mejoró GridSearchCV el modelo y qué hiperparámetros fueron óptimos?

Respuesta: Los mejores parámetros fueron max_depth=4 y n_estimators=100, logrando 88% de precisión.

- max_depth=4 evitó el sobreajuste limitando la profundidad de los árboles
- n_estimators=100 dio suficiente diversidad sin costo computacional excesivo
- El modelo es conservador: predice bien los no supervivientes (99%) pero menos los supervivientes (57%)

EJERCICIO 12 - Detección de Spam con SVM

Pregunta: ¿Por qué SVM supera a Naive Bayes en clasificación de spam?

Respuesta: Los resultados muestran que SVM alcanzó 98% de precisión en spam vs 74% en análisis de sentimientos. Dos razones técnicas:

1. **Patrones estructurados:** SVM aprende secuencias como "FREE money NOW" que son típicas de spam. Naive Bayes trata las palabras independientemente y pierde el contexto.
2. **Datasets desbalanceados:** SVM maneja bien el desbalance (1453 ham vs 219 spam) porque busca el hiperplano óptimo sin sesgo por cantidad. Naive Bayes se sesga hacia la clase mayoritaria.

EJERCICIO 13 - PCA en Netflix

Pregunta: ¿Qué ventajas ofrece PCA para analizar duración de películas?

Respuesta: PCA ofrece tres ventajas principales:

1. **Visualización simplificada:** Convierte datos complejos en gráficos 2D fáciles de interpretar
2. **Eliminación de ruido:** Filtra información irrelevante y conserva solo los patrones importantes
3. **Detección de grupos:** Identifica categorías naturales como cortos, largometrajes y documentales

EJERCICIO 14 - Clustering KMeans

Pregunta: ¿Cómo evaluar si los clusters son coherentes sin etiquetas reales?

Respuesta: Se pueden usar dos enfoques:

Métricas internas: Silhouette Score, índice Davies-Bouldin e inercia para medir qué tan bien separados están los grupos.

Análisis de contenido: Examinar las palabras más frecuentes en cada cluster para ver si hay temas coherentes.

Si los clusters no coinciden con las etiquetas originales: Esto puede revelar patrones más profundos, como agrupaciones por temas (precio, calidad, servicio) en lugar de solo sentimiento positivo/negativo.

EJERCICIO 15 - Visualización de Árbol de Decisión

Pregunta: ¿Qué ventajas tiene la visualización gráfica vs análisis en consola?

Respuesta: Dos beneficios específicos:

1. **Comprensión intuitiva:** La visualización permite seguir todo el flujo de decisión de un vistazo. En consola hay que leer línea por línea y armar mentalmente la estructura.
2. **Detección de patrones:** Se ven inmediatamente problemas como desbalance del árbol, variables dominantes o sobreajuste. En consola esta información está dispersa y es difícil de detectar.

La visualización es especialmente útil para explicar el modelo a personas no técnicas y validar que las decisiones tienen sentido.