RESPUESTAS

EJERCICIO 10 – Clasificación Multiclase con Iris Dataset

Archivo ejecutado: `EjerciciosPropuestos/Ejercicio10/Ejercicio10_Clasificación_Multiclase_Iris.py`

Pregunta:

¿Cuál de los tres modelos utilizados (Logistic Regression, SVM, KNN) mostró mejor rendimiento general? Justifica tu respuesta analizando la matriz de confusión y el reporte de clasificación.

RESPUESTA:

Los tres modelos (Logistic Regression, SVM y KNN) mostraron **rendimiento idéntico y perfecto** con 100% de precisión, recall y f1-score para las tres clases de flores (setosa, versicolor, virginica).

Justificación:

- **Accuracy:** 1.00 (100%) para los tres modelos
- **Precision, Recall, F1-score:** 1.00 para todas las clases en los tres modelos
- **Matriz de confusión implícita:** Sin errores de clasificación

No hay diferencia en rendimiento debido a que el dataset Iris es extremadamente bien separable linealmente. En este caso específico, cualquiera de los tres modelos es óptimo, aunque en la práctica se recomendaría **SVM** por su robustez con datos perfectamente separables.

EJERCICIO 11 – Ajuste de Hiperparámetros con Random Forest

Archivo ejecutado: `EjerciciosPropuestos/Ejercicio11/Ejercicio11_Validación_Cruzada_y_Ajuste_de_Hiperparámetros_Random _Forest.py`

Pregunta:

¿Cómo mejoró el desempeño del modelo al aplicar GridSearchCV? Describe qué hiperparámetros fueron óptimos y cómo afectaron la precisión del modelo final.

RESPUESTA:

Hiperparámetros óptimos encontrados:

- `max_depth = 4`
- `n_estimators = 100`

Mejora en el desempeño:

- Precisión global: 88%
- Weighted average F1-score: 0.87
- Macro average F1-score: 0.81

Cómo afectaron los hiperparámetros:

1. max_depth=4:Evitó el sobreajuste limitando la profundidad de cada árbol, manteniendo el modelo generalizable
2. n_estimators=100:Proporcionó suficiente diversidad en el ensemble para mejorar la estabilidad de las predicciones sin costo computacional excesivo

Análisis del rendimiento:

- Excelente recall para clase 0 (no supervivientes): 99%
- Recall más bajo para clase 1 (supervivientes): 57%
- El modelo es conservador al predecir supervivencia, reflejando la realidad histórica del Titanic

EJERCICIO 12 – Clasificación de Spam con TF-IDF + SVM

Archivo referenciado:Según el documento del ejercicio

Pregunta:

¿Por qué SVM puede superar a Naive Bayes en tareas de clasificación de texto como el spam? Menciona al menos dos razones técnicas relacionadas con el comportamiento del algoritmo y la naturaleza del texto.

RESPUESTA:

Razón técnica 1 - Manejo de correlaciones entre características:

- SVM: No asume independencia entre características (palabras), puede capturar relaciones complejas entre términos
- Naive Bayes: Asume que las palabras son independientes, lo cual es irreal en texto natural donde las palabras están correlacionadas contextualmente

Razón técnica 2 - Capacidad de separación no-lineal:

- SVM: Con kernels puede encontrar fronteras de decisión complejas en el espacio TF-IDF de alta dimensionalidad
- Naive Bayes: Está limitado a fronteras lineales basadas en probabilidades multiplicativas

Ejemplo práctico:

EJERCICIO 13 – Reducción de Dimensionalidad con PCA sobre Netflix

Archivo referenciado:Según el documento del ejercicio (análisis teórico)

Pregunta:

¿Qué ventajas ofrece la reducción de dimensionalidad mediante PCA en el análisis de datos como la duración de películas? Explica cómo afecta la visualización y la interpretación de los datos.

RESPUESTA:

Ventajas principales del PCA:

1. Visualización simplificada:
   - Reduce datos multidimensionales a 1-2 componentes principales
   - Permite crear gráficos 2D interpretables que capturan la mayor varianza
   - Facilita la identificación visual de patrones y outliers
2. Eliminación de ruido:
   - Conserva solo las direcciones de mayor variación en los datos
   - Filtra ruido y redundancia en las mediciones de duración
   - Mejora la calidad de análisis posteriores
3. Interpretación de patrones subyacentes:
   - Revela la estructura principal de variación en duraciones
   - Facilita la identificación de grupos naturales (cortos, largometrajes, documentales)
   - Permite detectar anomalías en duraciones de manera más efectiva

Impacto en visualización:El histograma de componentes principales muestra la distribución de variabilidad, permitiendo identificar si existen múltiples "tipos" de contenido por duración.

EJERCICIO 14 – Clustering No Supervisado con KMeans sobre Opiniones de Productos

Archivo ejecutado:`EjerciciosPropuestos/Ejercicio14/Ejercicio14_Clustering KMeans Opiniones.py`

Pregunta:

Al aplicar KMeans sin conocer las etiquetas reales, ¿qué criterios podrías usar para evaluar si los grupos formados son coherentes? ¿Cómo se podrían interpretar los resultados si los clusters no coinciden con las etiquetas originales?

RESPUESTA:

Criterios para evaluar coherencia:

1. Métricas internas:
   - Silhouette Score: Mide qué tan bien separados están los clusters
   - Índice Davies-Bouldin: Evalúa la compacidad intra-cluster y separación inter-cluster
   - Inercia within-cluster: Mide la cohesión interna de cada grupo
2. Análisis de contenido:
   - Examinar las palabras más frecuentes en cada cluster usando TF-IDF
   - Verificar si hay temas coherentes emergentes
   - Analizar la distribución de sentimientos por cluster

Interpretación si clusters no coinciden con etiquetas: Si los clusters no coinciden con las etiquetas originales (positivo/negativo), esto puede revelar patrones semánticos más profundos:

- Clusters temáticos: Agrupaciones por temas (precio, calidad, servicio) en lugar de sentimiento
- Subcategorías de sentimiento: Positivo entusiasta vs positivo moderado
- Insights de mercado: Patrones no evidentes que pueden ser más valiosos para análisis de negocio

EJERCICIO 15 – Árbol de Decisión con Exportación Visual (Graphviz)

Archivo ejecutado: `EjerciciosPropuestos/Ejercicio15/Ejercicio15_Árbol de Decisión Graphviz.py`

Pregunta:

¿Qué ventajas ofrece la visualización del árbol de decisión exportado respecto al análisis en consola? Menciona dos beneficios específicos relacionados con la comprensión del modelo.

RESPUESTA:

Beneficio específico 1 - Comprensión intuitiva del flujo de decisión:

- Visual: Permite seguir el camino completo desde la raíz hasta las hojas de manera inmediata
- Consola: Requiere leer línea por línea y mentalmente construir la estructura
- Impacto: Los stakeholders no-técnicos pueden entender exactamente qué condiciones llevan a cada predicción

Beneficio específico 2 - Identificación inmediata de patrones y sesgos:

- Visual: Revela instantáneamente si el árbol está balanceado, qué variables dominan las decisiones superiores, y detecta sobreajuste por ramas muy profundas
- Consola: Esta información está dispersa en el texto y requiere análisis tedioso para identificar patrones
- Impacto: Permite validación rápida del modelo y detección de problemas estructurales

Valor agregado en contextos reales:La visualización generada (titanic_tree.png) permite:

- Validación de dominio por expertos
- Explicabilidad en contextos regulatorios
- Debugging más eficiente del modelo
- Comunicación efectiva con stakeholders