Trabajo Práctico #2 - Clasificación de Noticias

Fecha de entrega: lunes 28 de Abril

Objetivo

Desarrollar un modelo de aprendizaje automático que clasifique con precisión los artículos de noticias según su contenido. Este ejercicio demostrará la importancia del preprocesamiento de datos, la extracción de características y la evaluación de modelos en tareas de clasificación de texto.

Descripción

El trabajo se centra en desarrollar una comprensión integral y aplicar el procesamiento de datos, la extracción de características, el aprendizaje no supervisado (utilizando el algoritmo k-means), el aprendizaje supervisado (a través de la clasificación con árboles de decisión) y la evaluación del modelo dentro del contexto de una tarea práctica de aprendizaje automático.

Se comenzará con datos en bruto, que debe ser limpiado y preprocesado para garantizar su idoneidad para el análisis. Esto incluye el manejo de valores faltantes, la normalización de datos y la codificación de variables categóricas. Luego, extraerán características significativas de los datos procesados, empleando técnicas como **TF-IDF** para datos de texto o **PCA** para datos numéricos, con el fin de reducir la dimensionalidad y mejorar el rendimiento del modelo.

El trabajo se divide en dos tareas de aprendizaje:

1. **Aprendizaje no supervisado**, donde se debe utilizar el algoritmo **k-means** para identificar agrupaciones inherentes dentro de los datos sin respuestas pre etiquetadas.

2. **Aprendizaje supervisado**, donde se debe aplicar un **clasificador con árboles de decisión** para predecir resultados basados en las características extraídas.

Finalmente, se deberá evaluar sus modelos utilizando métricas adecuadas:

- **Para tareas no supervisadas**: la métrica de **SSD** para evaluar la calidad de las agrupaciones.
- **Para tareas supervisadas**: precisión (*accuracy*), precisión positiva (*precisión*), exhaustividad (*recall*) y puntaje F1 (*F1-score*).

Recursos

- **BBC News Dataset**: Un conjunto de datos de noticias de la BBC en formato CSV con su categoría correspondiente.

Entregables

Un notebook de *Google Collab* con el modelo desarrollado, junto con las pruebas. Estas mismas deberán contener las métricas mencionadas anteriormente.