Laboratorio 5.
Validación Cruzada

OBJETIVOS:

El objetivo de este laboratorio es aplicar técnicas de minería de datos en un conjunto de datos real, utilizando Python y sus librerías más comunes como Numpy, Pandas, y Matplotlib. Para esto deben emplearse distintos modelos de aprendizaje automático y técnicas de preprocesamiento de datos, enfatizando el uso de la Validación Cruzada para evaluar el desempeño de los modelos. Este ejercicio tiene como fin que los estudiantes tomen decisiones informadas sobre qué modelos son más adecuados para el problema dado, cómo ajustarlos, y cómo validar su efectividad.

DESCRIPCIÓN DEL CONJUNTO DE DATOS

Se utilizará el conjunto de datos de calidad de vinos disponibles en el UCI Machine Learning Repository. Este conjunto de datos contiene características físico-químicas de vinos (como acidez, azúcar, alcohol, etc.) junto con una calificación de calidad. El reto consiste en predecir la calidad del vino basándose en estas características.

URL del conjunto de datos:

Wine Quality Dataset (https://archive.ics.uci.edu/ml/datasets/wine+quality)

EJERCICIOS

1. Exploración y Preprocesamiento de Datos (15 puntos)

- Cargar el conjunto de datos utilizando Pandas.
- Realizar un análisis exploratorio de datos (EDA) básico: estadísticas descriptivas, distribución de variables, etc.
- Preprocesar los datos: limpieza, tratamiento de valores faltantes, normalización/estandarización si es necesario.

2. División del Conjunto de Datos (10 puntos)

- Dividir el conjunto de datos en entrenamiento y prueba utilizando una proporción 80-20.

3. Selección y Entrenamiento de Modelos (30 puntos)

- Entrenar al menos tres modelos diferentes seleccionados de los temas vistos en clases. Por ejemplo, Regresión Logística, KNN, y Random Forest.
- Para cada modelo, realizar un ajuste de hiperparámetros básico.
- Utilizar técnicas de reducción de dimensionalidad como PCA si se considera necesario.

4. Validación Cruzada (25 puntos)

- Implementar validación cruzada en cada modelo para evaluar su desempeño.
- Comparar los resultados de la validación cruzada entre los modelos para decidir cuál tiene mejor desempeño.

5. Interpretación de Resultados y Conclusión (20 puntos)

- Interpretar los resultados obtenidos.
- Escribir una conclusión sobre qué modelo funcionó mejor y por qué.
- Incluir visualizaciones de datos y resultados que apoyen las conclusiones.
- Sugerir posibles mejoras o pasos futuros para investigaciones adicionales.

Entrega:

Deben entregar un informe que incluya:

- Código fuente utilizado para el análisis.

- Visualizaciones generadas.

- Una discusión de los resultados de la validación cruzada.

- Reflexión final sobre el laboratorio y los insights obtenidos.

Puede integrarse todo en un Notebook ipynb!

EVALUACIÓN

Total: 100 puntos

Exploración y Preprocesamiento de Datos: 15 puntos

División del Conjunto de Datos: 10 puntos

Selección y Entrenamiento de Modelos: 30 puntos

Validación Cruzada: 25 puntos

Interpretación de Resultados y Conclusión: 20 puntos

Notas Adicionales

- Trabajar en parejas. Es importante que se registren en uno de los grupos configurados para este laboratorio. De no hacerlo, no tendrán nota.
- Es esencial hacer un uso adecuado de las bibliotecas de Python mencionadas y seguir las buenas prácticas de programación.
- Se valorará la creatividad en la presentación de resultados y el análisis crítico de los modelos utilizados.