

Trabajo Práctico Integrador Grupal

Aplicaciones de Inteligencia de Datos

Objetivo

El objetivo de este Trabajo Práctico Integrador es que los estudiantes integren y apliquen los conocimientos adquiridos a lo largo de la Maestría mediante el análisis de un caso real de Inteligencia de Datos. Se espera que puedan realizar un proceso completo que incluya análisis exploratorio, preprocesamiento de datos y el desarrollo de modelos de machine learning adecuados para abordar el problema planteado.

Metodología y Entrega del Trabajo Práctico

- El trabajo práctico será grupal, con un máximo de 5 estudiantes por grupo.
- La entrega del trabajo se realizará por medio del Campus Virtual.
- Se debe adjuntar un informe en pdf y el Notebook correspondiente.
- No hay restricciones en el uso de librerías de programación.
- El nombre de los archivos debe contener los apellidos de los integrantes del equipo.
- Cada grupo defenderá el trabajo realizado mediante la presentación de un Power Point.

Enunciado

La asignatura dispondrá de una base de datos para realizar el trabajo.

Consideraciones importantes a tener en cuenta sobre el dataset

- La variable objetivo del problema es “**Radiacion_Global**” (columna AB).
- Las variables “**Heliofania_Efectiva**” (columna O) y “**Heliofania_Relativa**” (columna P) **no deben ser utilizadas** en ninguna etapa del análisis ni del modelado.

Parte I – EDA y Preprocesamiento:

1. Realizar un preprocesamiento básico y una limpieza inicial del dataset, respondiendo al menos a las siguientes preguntas:
 - a. ¿Cuántos registros y variables contiene el dataset?
 - b. ¿Qué tipo de variables están presentes?
 - c. ¿Existen valores faltantes, duplicados o inconsistencias?
 - d. ¿Se identifican valores atípicos relevantes?

2. Resumir estadísticamente los datos: obtener las principales medidas de posición y de dispersión de cada variable.
3. Generar visualizaciones que permitan realizar un análisis univariado, bivariado y multivariado, justificando brevemente la elección de cada tipo de gráfico.
4. Realizar un preprocesamiento y limpieza avanzada de datos.
5. Elaborar conclusiones parciales que sinteticen los principales hallazgos del análisis exploratorio y las decisiones de preprocesamiento adoptadas.

(Al finalizar la Parte I se debe contar con un dataset listo para ser utilizado con algoritmos de aprendizaje automático).

Parte II – Modelado, Evaluación e Interpretabilidad:

1. Selección de modelos:

Seleccionar al menos dos algoritmos de *machine learning* adecuados para el problema planteado y justificar su elección.

2. Entrenamiento, validación y optimización de modelos:

- a. Realizar el entrenamiento de los modelos seleccionados, incorporando técnicas de validación cruzada (*cross-validation*) y la optimización de hiperparámetros mediante algún método de búsqueda (por ejemplo, *Grid Search*, *Random Search* u otros).
- b. Describir el esquema de validación utilizado y los principales hiperparámetros optimizados.

3. Evaluación y análisis de resultados:

- a. Evaluar y comparar el desempeño de los modelos utilizando métricas adecuadas para el problema.
- b. Visualizar y analizar el *output* de los modelos, incluyendo comparaciones entre valores reales y predichos sobre el conjunto de test.
- c. Detallar y representar gráficamente la “confianza” o incertidumbre asociada a las predicciones, cuando corresponda.

4. Interpretabilidad de los modelos:

Realizar un análisis de interpretabilidad de los modelos entrenados, identificando las variables más relevantes y discutiendo su impacto en las predicciones.

5. Conclusiones:

Elaborar conclusiones integradoras a partir de los resultados obtenidos, destacando fortalezas, limitaciones y posibles mejoras.