Introducción a la Inteligencia Artificial

Trabajo Práctico Integrador

Resumen

En este trabajo se espera que el alumno presente el resultado del análisis completo del conjunto de datos elegido. Para ello, se espera que el alumno aplique los conceptos adquiridos en el desarrollo de este curso.
El presente trabajo se realizará, presentará, expondrá y evaluará de manera grupal, en grupos de a lo sumo 4 personas.

Objetivo del Trabajo

El objetivo de este trabajo práctico es que los estudiantes, organizados en grupos, apliquen las técnicas vistas en el curso para realizar un análisis completo de un conjunto de datos.

Alcance

El análisis debe incluir desde la exploración y limpieza de datos hasta el uso de técnicas de aprendizaje supervisado y no supervisado, culminando con el uso de modelos de aprendizaje profundo. Cada grupo deberá entregar un informe escrito y realizar una presentación oral defendiendo sus hallazgos y conclusiones.

Fases del Trabajo

1. Selección del Dataset:

En la carpeta “Datasets para el TP Integrador” encontrará los datasets enunciados a continuación:

ID	Nombre	Tamaño (Mb)
1	breast_cancer	0,12
2	car_prices	88
3	creditcard	150,8
4	globalTerrorism	162,8
5	Landslides-after-rain	0,442
6	online_shoppers-intention	1,1
7	police_killings	0,129
8	stock_data	28,1
9	telco	1,9
10	winequality	0,341

Los datasets que tienen más de 50MB no se presentan dentro del campus (por política del Campus). Los mismos pueden hallarse en una carpeta compartida en One Drive: Datasets para TP IntegradorDatasets para TP Integrador

2. Análisis Exploratorio de Datos (EDA)

Se debe realizar un primer abordaje al conjunto de datos seleccionado:

- **Descripción del dataset**: Resumen de las variables disponibles, identificación de variables numéricas, categóricas y la variable objetivo (si aplica).
- **Visualización de datos**: Generación de gráficos descriptivos que permitan entender las relaciones y tendencias entre variables.
- **Análisis de datos faltantes**: Identificación de valores nulos o inconsistentes y estrategias para tratarlos (imputación, eliminación, etc.).
- **Análisis de outliers**: Identificación y tratamiento de valores atípicos.

3. Proceso ETL (Extracción, Transformación y Carga)

- **Extracción y limpieza de datos**: Transformación del dataset para asegurar su integridad, normalización de variables numéricas, codificación de variables categóricas, etc.
- **Generación de nuevas variables (feature engineering)**: Creación de nuevas características que puedan mejorar el rendimiento de los modelos, basadas en las variables existentes.

4. Análisis Supervisado

Los estudiantes deben aplicar técnicas de aprendizaje supervisado sobre el dataset procesado:

- **Selección de modelos**: Prueba de varios modelos de aprendizaje supervisado (e.g., regresión logística, random forest, etc.).
- **Entrenamiento y validación**: Evaluación de los modelos mediante validación cruzada o partición en conjunto de entrenamiento y prueba.
- **Métricas de evaluación**: Presentación de métricas relevantes para la tarea (e.g., accuracy, F1-score, etc.).

5. Análisis No Supervisado

El análisis no supervisado debe incluir:

- **Reducción de dimensionalidad**: Aplicación de técnicas como Análisis de correlación, entropía, PCA (Análisis de Componentes Principales).
- **Clusterización**: Uso de algoritmos de clustering (K-Means, a-priori, ente otros) para agrupar los datos y analizar patrones ocultos.

6. Uso de Herramientas de Aprendizaje Profundo

- Utilización de modelos Transformers sobre el conjunto de datos.
- Evaluación de resultados.

7. Informe

El informe deberá incluir:

- **Introducción**: Breve resumen, objetivos, alcance del trabajo y del dataset elegido.
- **Marco teórico:** Una breve descripción de cada una de las técnicas que utilizará en el análisis (con referencias bibliográficas).

- **Desarrollo técnico: Contiene**
    - **Análisis Exploratorio**: Resultados obtenidos en la fase de análisis exploratorio de datos (EDA).
    - **Proceso ETL**: Descripción de los pasos de limpieza, transformación y generación de nuevas variables.
    - **Modelos supervisados**: Detalles de los modelos supervisados probados, justificación de la elección final y métricas de rendimiento.
    - **Modelos no supervisados**: Resultados obtenidos de la clusterización y análisis no supervisado.
    - **Aprendizaje profundo**: Explicación del modelo usado y sus resultados.
- **Resultados y hallazgos:** Presente resumidamente los hallazgos obtenidos
- **Conclusiones**: Reflexión final sobre el proceso, limitaciones del análisis y sugerencias para futuros trabajos.
- **Referencias Bibliográficas:** En formato APA

8. Presentación y Defensa Oral

- **Duración**: 10-15 minutos por grupo.
- **Contenido**: Exposición de los resultados más relevantes, con énfasis en el análisis, los modelos utilizados y las conclusiones extraídas.
- **Defensa**: Los estudiantes deben estar preparados para responder preguntas sobre los métodos aplicados y sus hallazgos.

Pautas de Evaluación

- **Informe** (50%): Claridad, rigor técnico, y estructura.
- **Presentación oral** (30%): Capacidad de comunicación, claridad de la exposición (individual), y defensa de los resultados. Uso de lenguaje técnico.
- **Resultados obtenidos** (20%): Calidad de las conclusiones, validez de los modelos y profundidad del análisis.