BASES DEL TERCER ENTREGABLE: ALGORITMOS DE MACHINE LEARNING

Objetivo

El objetivo de este último entregable/proyecto final es el de desarrollar alguno de los modelos de aprendizaje automático (supervisados: regresión lineal y clasificación, o no supervisados: clustering), aplicando los conceptos de machine learning y análisis de datos aprendidos en clase. Los alumnos y alumnas deben demostrar habilidades para preparar, analizar y modelar datos, interpretando adecuadamente los resultados para generar conclusiones claras.

Conjuntos de Datos

En la carpeta adjunta llamada “Datos”, se ofrecen seis conjuntos de datos en formato csv o Excel. Cada alumno deberá elegir una técnica de modelado de datos (regresión lineal, clasificación o clustering) y, una vez escogida la técnica, deberá seleccionar un único conjunto de datos que sea de su preferencia. Si no os sentís cómodos con ninguno de estos datasets, siempre podéis trabajar con un conjunto de datos que sea de vuestro gusto.

Los conjuntos de datos que se ofertan son los siguientes:

Modelos de Regresión Lineal

- *Precio de viviendas*: un dataset con variables sobre tamaño, ubicación, y antigüedad de viviendas, junto con el precio de venta como variable objetivo.

- *Consumo de energía*: contiene registros de consumo eléctrico diario en distintos sectores industriales, con variables como temperatura, humedad, y demanda estacional.

Modelos de Clasificación

- *Detección de fraude financiero*: un dataset con transacciones bancarias, con atributos como monto, tipo de transacción, y origen. La variable objetivo es si la transacción es fraudulenta.

- *Diagnóstico médico*: datos de pacientes con distintas mediciones (edad, presión arterial, colesterol) y una variable objetivo que indica si el paciente tiene una enfermedad.

Modelos de Clustering

- *Segmentación de clientes*: incluye datos de clientes de un comercio (edad, ingresos, historial de compras), ideal para agrupar en segmentos de consumo.

- *Agrupación de ciudades por indicadores demográficos:* un dataset con ciudades y variables como densidad de población, ingresos medios, y tasa de crecimiento, adecuado para identificar patrones demográficos.

Requisitos Técnicos

Para la realización de este proyecto, los requisitos técnicos mínimos (según el algoritmo escogido) serán los siguientes:

1. Análisis exploratorio de los datos y preprocesamiento:

1.1. Carga de datos y exploración inicial

- Importar el conjunto de datos.
- Mostrar las primeras filas y el tamaño del dataset para entender la estructura básica.
- Inspeccionar los tipos de datos de cada columna y los principales estadísticos para las distintas variables (numéricas y categóricas).

1.2. Detección de valores nulos y duplicados

- Identificar valores nulos por columna y evaluar su tratamiento (eliminar, imputar o dejar según el contexto).
- Identificar registros duplicados y decidir si deben eliminarse o si aportan valor al análisis.

1.3. Detección y tratamiento de outliers

- Utilizar métodos estadísticos (como el rango intercuartílico o z-score) y/o visuales (boxplots, scatterplots) para identificar outliers en las variables numéricas.
- Decidir cómo tratarlos: eliminarlos, transformarlos o dejar algunos si son relevantes para el análisis.

1.4. Transformación de variables categóricas y numéricas

- Convertir variables categóricas en formato numérico usando técnicas como One-Hot Encoding o Label Encoding, según la naturaleza de la variable y el modelo a utilizar.
- Normalización y Escalado de Variables Numéricas: aplicar alguna de las técnicas estudiadas teniendo en cuenta la distribución de los datos, la sensibilidad ante valores extremos y la escala de los datos.

1.5. Análisis visual de distribuciones y correlaciones (opcional)

- Crear gráficos de distribución, como histogramas y boxplots, para explorar la forma y dispersión de los datos.
- Generar gráficos de dispersión entre variables numéricas clave para observar relaciones.
- Generar un heatmap de correlación para identificar relaciones fuertes o débiles entre variables, lo cual puede orientar futuras selecciones de características.

1.6. Resumen de insights clave

- Extraer y resumir los insights más importantes del análisis visual y estadístico, enfocándose en patrones, relaciones y posibles factores que influyan en la variable objetivo o en los resultados esperados del análisis.

2. Algoritmos de Machine Learning:

2.1. Modelos de regresión lineal

- Entrenar un modelo de regresión lineal, interpretar coeficientes y evaluar con métricas como RMSE y R^2.

2.2. Modelos de clasificación

- Entrenar un modelo de clasificación (como árboles de decisión, random forest o regresión logística) y evaluar usando métricas de precisión, recall, y F1-score.

2.3. Técnicas de clustering

- Aplicar métodos de clustering (k-means, jerárquico o DBSCAN) y evaluar con la métrica de silhouette.

3. Interpretación y Conclusiones:

- Interpretar los resultados del modelo y discutir posibles mejoras.
- Explicar el uso práctico del modelo para la toma de decisiones en el contexto del dataset.

Entrega

El proyecto deberá ser entregado en formato de Jupyter Notebook (.ipynb) con todas las celdas ejecutadas y comentarios que expliquen el código. Dicho código, tendrá que seguir las pautas de buenas prácticas vistas durante las clases (nombres de variables con sentido, limpieza y orden, etcétera).

Cuando tengáis que desarrollar un párrafo para explicar lo que vais haciendo, utilizad celdas de markdown en lugar de poner comentarios en las celdas de código.

La fecha límite de entrega es el domingo 19 de noviembre de 2024.

El entregable se remitirá a la siguiente dirección de correo electrónico: eduardo.pastor@thepower.education indicando en el asunto del mail vuestro nombre y apellidos seguido del texto “- ENTREGABLE 3”.

En el correo que enviéis, debéis adjuntar una carpeta que contenga:

- El Jupyter Notebook (archivo.ipynb).
- Los conjuntos de datos.
- Archivo de soporte de funciones y/o clases (opcional).
- Todo aquel archivo que consideréis necesario para la comprensión de vuestro ejercicio: Word, PDF, imágenes, etcétera (opcional).

Evaluación

El entregable será calificado como apto o no apto y evaluado en función de los siguientes criterios:

- Correcta implementación de las herramientas y técnicas vistas en clase.
- Estructura y claridad del código (uso adecuado de comentarios, nombres de variables, etcétera).
- Interpretación y análisis de los resultados.
- Creatividad y profundidad en el análisis de datos.