TP5_Aprendizaje

June 27, 2025

1 TP 5: Aprendizaje I

1.1 Fecha de entrega: 02/07/2025 a las 23:59hs.

Condiciones de entrega: el trabajo práctico deberá ser realizado en forma individual. Se deberá subir en la sección del Campus Virtual correspondiente el enlace a Colab. No olvidar configurar el documento para que sea accesible a cualquier persona con el enlace o en su defecto compartirlo con cristian.pacifico@uner.edu.ar y antonella.garcia@uner.edu.ar

Los datasets necesarios los encuentran en la carpeta compartida a inicio del cuatrimestre "**IA-1C2025**".

Se debe presentar el código solución a la consigna y algunas líneas adicionales de código que sirvan para testear la solución presentada. Además, pueden incluir en un campo de texto las respuestas, suposiciones y aclaraciones pertinentes de cada punto.

2 Arboles de decisión.

3 Dataset: MBA

####Utilizar el dataset MBA.csv, que contiene datos sintéticos generados a partir de las estadísticas de la Clase Wharton 2025 de University of Pennsylvania. El objetivo es analizar los datos y construir un modelo que prediga el estado de admisión de los estudiantes. ####Metadatos: * application_id: Identificador único para cada aplicación * género: género del solicitante (masculino, femenino) * internacional: Estudiante internacional (VERDADERO/FALSO) * gpa: Promedio de calificaciones del solicitante (en una escala de 4.0) * Especialidad: Licenciatura (Negocios, STEM, Humanidades) * raza: origen racial del solicitante (por ejemplo, blanco, negro, asiático, hispano, otro / nulo: estudiante internacional) * gmat: Puntuación GMAT del solicitante (800 puntos) * work_exp: Número de años de experiencia laboral (Año) * work_industry: Industria de la experiencia laboral previa del solicitante (por ejemplo, consultoría, finanzas, tecnología, etc.) * Admisión: Estado de admisión (Admitir, Lista de espera, Nulo: Denegar)

1. Carga el dataset "MBA.csv" en un DataFrame "df" y mostrar las primeras 5 filas.

[ ]:
```
import pandas as pd

#incorporar el código correspondiente
```

2. Mostrar la información básica y las estadísticas descriptivas de las variables.

```
[ ]: #incorporar el código correspondiente
```

3.1 Preprocesamiento del DataFrame.

3. Asignar el valor "Denegado" a las filas que contengan valores nulos en la columna "Admisión".

```
[ ]: #incorporar el código correspondiente
```

4. Convertir las variables categóricas a valores numéricos y estandarizar las columnas que sean necesarias para el análisis.

```
[ ]: #incorporar el código correspondiente
```

5. Separar los datos en entrenamiento y testeo, utilizando el 25% de los datos para testeo.

```
[ ]: #incorporar el código correspondiente
```

6. Construir un modelo de árbol de decisión para predecir si un estudiante será admitido o no.

```
[ ]: #incorporar el código correspondiente
```

7. Calcular las métricas que permitan evaluar la precisión del modelo.

```
[ ]: #incorporar el código correspondiente
```

8. A partir de las métricas obtenidas: ¿Qué podemos decir del modelo creado?

```
[ ]: #incorporar el código correspondiente
```

#**Dataset** *car.csv* Se debe crear un modelo de árbol de decisión confiable que sea capaz de ayudar a una empresa a encontrar automóviles que los clientes probablemente comprarán. Se debe construir un modelo de árbol de decisión que clasifique los automóviles como aceptables o no aceptables. El dataset se encuentra disponible en el campus junto a este práctico *car.csv* y se compone de seis características diferentes: compra, mantenimiento, puertas, personas, maletero y seguridad. La variable objetivo clasifica la aceptabilidad de un automóvil determinado. Puede tomar 0 o 1, siendo 1 aceptable.

1. Utilizar el 70% de los datos para entrenamiento y el 30% restante para testeo.

```
[ ]: #incorporar el código correspondiente
```

2. Evaluar la precisión del modelo utilizando **Accuracy**.

```
[ ]: #incorporar el código correspondiente
```

3. A partir de la métrica obtenida: ¿Qué podemos decir del modelo creado?

Introduzca su respuesta aquí...

4 Regresión Logística

5 Dataset: *Breast Cancer.csv*

Utilizar el dataset de cáncer de mama disponible en sklearn para predecir la presencia de cáncer maligno utilizando regresión logística.

1. Utilizar el 20% de los datos para testeo.

```
[ ]: #incorporar el código correspondiente
```

2. Evaluar el rendimiento del modelo utilizando las métricas **Accuracy**, **Precision** y **Recall**.

```
[ ]: #incorporar el código correspondiente
```

3. Obtener la matriz de confusión del modelo.

```
[ ]: #incorporar el código correspondiente
```

6 Dataset: *ClientesEnLinea.csv*

Crear un modelo de Regresión Logística utilizando el dataset *ClientesEnLinea.csv* que cuenta con información de clientes que compran o no ciertos productos en línea para ello contamos con información sobre el género, la edad y el salario estimado, clasificando a los clientes con 0 y 1 si no compró o si compró respectivamente.

1. Utilizar como métrica comparativa el promedio de una validación cruzada K-fold con 5 folds para entrenamiento y testeo.

```
[ ]: #incorporar el código correspondiente
```

2. ¿Cómo se comporta el modelo si consideramos todos los predictores?

Introduzca su respuesta aquí...

3. ¿Qué sucede cuando solo consideramos como predictores Sexo y Edad?

```
[ ]: #incorporar el código correspondiente
```

Introduzca su respuesta aquí...

#***Regresión lineal***

7 Dataset: *articulos_ml.csv*

A partir del dataset *articulos_ml.csv* que se encuentra disponible en el campus y contiene diversas URLs a artículos sobre Machine Learning. Se debe construir un modelo de regresión lineal para predecir cuantas veces será compartido un artículo en redes sociales basándonos en la cantidad de palabras del artículo.

1. Mostrar las columnas disponibles en el dataset *articulos_ml.csv*.

```
[ ]: #incorporar el código correspondiente
```

2. Crear gráficos para visualizar la relación entre las variables del dataset.

```
[ ]: #incorporar el código correspondiente
```

3. Filtrar los artículos que tengan menos de 3500 palabras y una cantidad de compartidos menor a 80,000 para analizar un conjunto más específico de datos.

```
[ ]: #incorporar el código correspondiente
```

4. Utilizar los datos filtrados para generar un modelo de regresión lineal y graficar la relación entre las palabras del artículo y la cantidad de veces que son compartidos.

```
[ ]: #incorporar el código correspondiente
```

5. Utilizar el modelo generado para predecir la cantidad de veces que serán compartidos artículos de 2000, 5000 y 10000 palabras.

```
[ ]: #incorporar el código correspondiente
```

6. Mostrar los coeficientes del modelo.

```
[ ]: #incorporar el código correspondiente
```

7. Evaluar el modelo aplicando las métricas **Error Cuadrático Medio** y **Coeficiente de Determinación (R2)**.

```
[ ]: #incorporar el código correspondiente
```