Universidad Autónoma de San Luis Potosí
Facultad de ingeniería
Inteligencia Artificial Aplicada
Practica: 4
Nombre Práctica: Preprocesamiento de Datos
Nombre del Alumno: Manuel Ramírez Galván

Fecha: 19/02/2025

Procedimiento

4.1.- Inicie Jupyter Notebooks y abra los notebooks "introducción"
4.2.- Siga las instrucciones en los notebooks para explorar los conceptos básicos de preprocesamiento de datos.
4.3.- Aplicar las técnicas de preprocesamiento de datos al dataset “Social_Network_Ads”

Resultados

```
import pandas as pd
import numpy as np

data = pd.read_csv("C:/Users/HUAWEI/Desktop/
data.head()
```

✓ 0.0s

	Gender	Age	EstimatedSalary	Purchased
0	Male	19	19000	False
1	Male	35	20000	False
2	Female	26	43000	False
3	Female	27	57000	False
4	Male	19	76000	False

Imagen 1.- Carga y Visualización de Datos

```
data.columns
```

✓ 0.0s

```
Index(['Gender', 'Age', 'EstimatedSalary', 'Purchased'], dtype='object')
```

Imagen 2.- Visualización de las columnas

```
data.shape
```

✓ 0.0s

```
(99, 4)
```

Imagen 3.- Tamaño de la tabla

```
data.dtypes
```
✓ 0.0s

```
Gender              object
Age                  int64
EstimatedSalary      int64
Purchased             bool
dtype: object
```

Imagen 4.- Tipos de datos de cada columna

```
data.describe().transpose()
```
✓ 0.0s

	count	mean	std	min	25%	50%	75%	max
Age	99.0	30.282828	8.230159	18.0	25.0	28.0	33.5	59.0
EstimatedSalary	99.0	57616.161616	33344.126268	15000.0	27000.0	52000.0	81500.0	150000.0

Imagen 5.- Descripción de los datos

```
data.isnull().sum()
```
✓ 0.0s

```
Gender              0
Age                 0
EstimatedSalary     0
Purchased           0
dtype: int64
```

Imagen 6.- Verificar si faltan datos

```
df = pd.get_dummies(data, columns = ['Gender'], drop_first = True, dtype=int)
df.head()
```
✓ 0.0s

	Age	EstimatedSalary	Purchased	Gender_Male
0	19	19000	False	1
1	35	20000	False	1
2	26	43000	False	0
3	27	57000	False	0
4	19	76000	False	1

Imagen 7.- Conversión de variables y agregar Columna

```
df = pd.get_dummies(df, columns = ['Purchased'], drop_first = True, dtype=int)
df.head()
```
✓ 0.0s

	Age	EstimatedSalary	Gender_Male	Purchased_True
0	19	19000	1	0
1	35	20000	1	0
2	26	43000	0	0
3	27	57000	0	0
4	19	76000	1	0

Imagen 8.- Conversión de variable binaria y agregar Columna

```
df[['Age', 'EstimatedSalary']] = df[['Age', 'EstimatedSalary']].agg(np.log)
df.head()
```

✓ 0.0s

	Age	EstimatedSalary	Gender_Male	Purchased_True
0	2.944439	9.852194	1	0
1	3.555348	9.903488	1	0
2	3.258097	10.668955	0	0
3	3.295837	10.950807	0	0
4	2.944439	11.238489	1	0

Imagen 9.- Transformación logarítmica

```
df = df.drop(columns=['Purchased_True'])
df.head()
```

✓ 0.0s

	Age	EstimatedSalary	Gender_Male
0	2.944439	9.852194	1
1	3.555348	9.903488	1
2	3.258097	10.668955	0
3	3.295837	10.950807	0
4	2.944439	11.238489	1

Imagen 10.- Eliminación de Columna

```
import matplotlib.pyplot as plt
import seaborn as sns
```

Imagen 11.- Agregar más librerías

```
sns.set(style="whitegrid")
```

Imagen 12.- Mejora visualización de graficos

```
fig, ax = plt.subplots(1, 2, figsize=(15, 5))
sns.histplot(df['Age'], kde=True, ax=ax[0])
ax[0].set_title('Distribucion de la Edad')
sns.histplot(df['EstimatedSalary'], kde=True, ax=ax[1])
ax[1].set_title('Distribucion del Salario Estimado')
```

```
Text(0.5, 1.0, 'Distribucion del Salario Estimado')
```

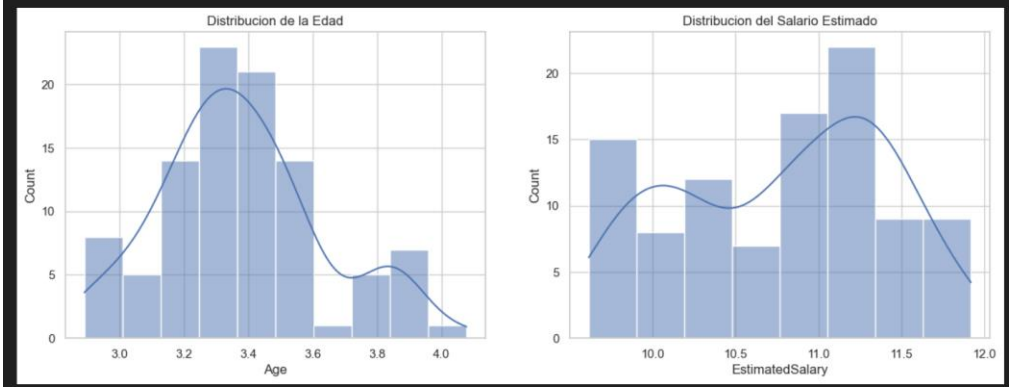

Imagen 13.- Graficas de las Distribuciones de Edad y Salario

```
plt.figure(figsize=(10, 5))
sns.scatterplot(data=df, x='Age', y='EstimatedSalary', hue='Gender_Male', palette='viridis')
plt.title('Relacion entre Edad y Salario Estimado')
```

```
Text(0.5, 1.0, 'Relacion entre Edad y Salario Estimado')
```

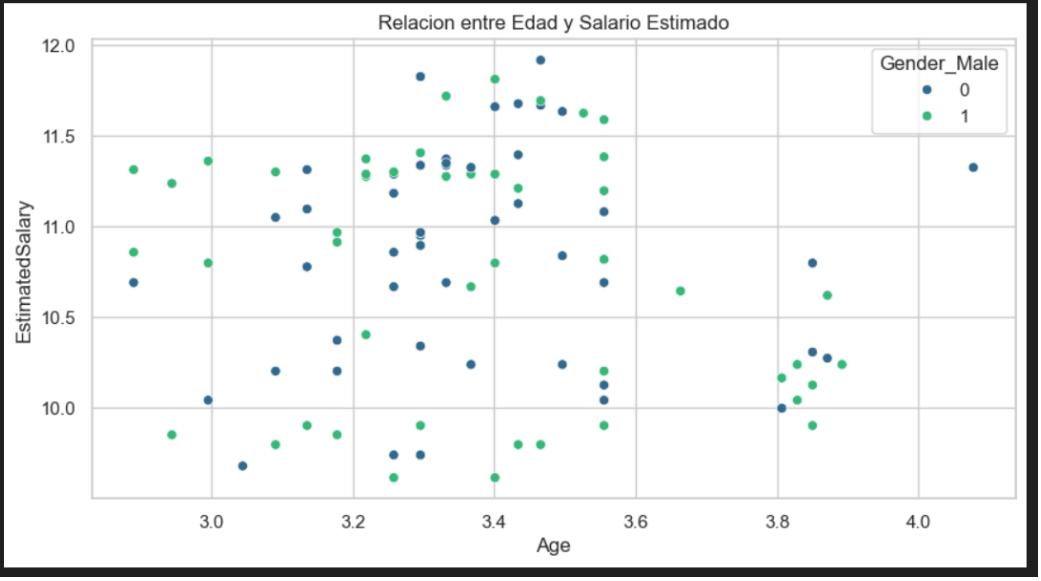

Imagen 14.- Grafico de Dispersión de las Distribuciones

```
corr = df.corr(numeric_only=True)
sns.heatmap(corr, annot=True)
```

```
<Axes: >
```

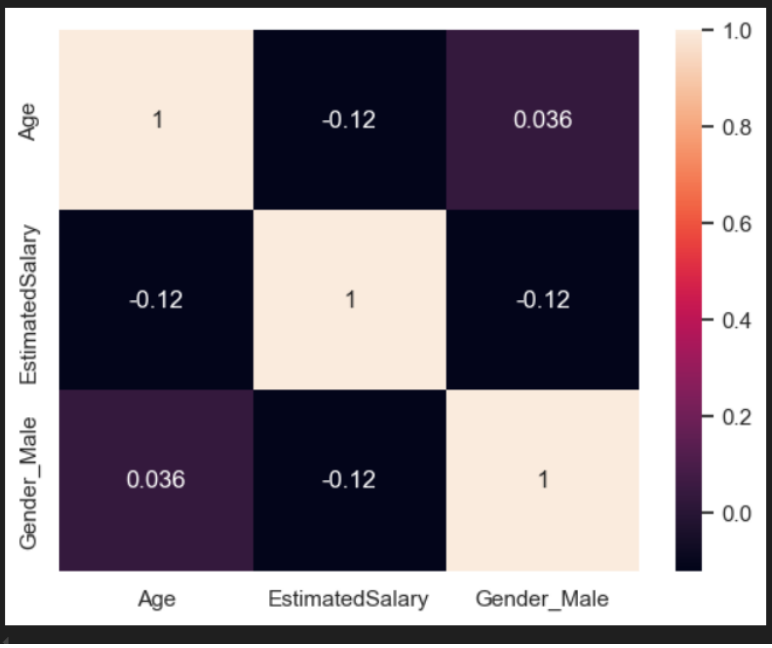

Imagen 15.- Obtención de Matriz de Correlación

Comprensión

1. **¿Cuál es la importancia del preprocesamiento de datos en el análisis de datos y el aprendizaje automático?**

Esta etapa garantiza que los datos sean limpios, estructurados y adecuados para su análisis, ya que, sin esto, se pueden producir resultados incorrectos, inconsistentes o ineficientes.

Corrige valores erróneos, evita sesgos y errores en el modelo, acelera y optimiza el procesamiento, convierte datos categóricos en valores numéricos, hace normalización.

2. **Mencione al menos tres técnicas de preprocesamiento de datos y explique su función.**

Limpieza de datos: Elimina errores, valores nulos, duplicados o inconsistencias en los datos para evitar sesgos en los análisis o modelos.

Normalización y escalado: Ajusta los valores de las variables numéricas a una escala común para evitar que una variable domine sobre las demás.

Codificación de Datos Categóricos: Convierte variables categóricas en valores numéricos que pueden ser procesados por modelos de Machine Learning.

3. **¿Qué son los datos faltantes y como se pueden manejar durante el preprocesamiento?**

Son valores ausentes en un conjunto de datos, lo que puede afectar en el análisis y rendimiento de los modelos. Pueden ocurrir por errores en la recopilación, fallas, problemas en bases de datos o datos incompletos.

Se eliminan filas o columnas con valores nulos si la cantidad de datos perdidos es significativa y afecta el análisis.

Se remplazan los valores faltantes con estimaciones como la media, mediana, moda o interpolaciones.

4. **¿Qué son los valores atípicos y como se pueden detectar y tratar?**

Son valores que desvían significativamente el resto de los valores en un conjunto de datos, esto por errores de medición, problemas en la recolección de datos, etc.

Se puede detectar por el Método de Cuartiles por identificación de valores extremos.

Método de Desviación Estándar donde los valores fuera de 3 desviaciones estándar de la media suelen considerarse atípicos.

Además del Método de Boxplot o Diagrama de Cajas para una inspección visual.

Para tratarlos se utiliza la Eliminación de Outliers, cuando los valores atípicos son errores o datos irrelevantes.

También por la Transformación de Datos al aplicar algoritmos o escalado para reducir el impacto de los valores atípicos.

Además, por la Sustitución por Mediana o Media al remplazar los valores atípicos por valores más representativos.

5. ¿Cuál es la importancia de la codificación de variables categóricas?

Muchos modelos requieren variables numéricas en lugar de categóricas. La codificación de variables categóricas convierte estos valores en un formato numérico para que los algoritmos puedan procesarlos correctamente.

Hace que los datos sean compatibles con modelos de Machine Learning, mejora la precisión del modelo, facilita el análisis de datos, optimiza el rendimiento del modelo.

Conclusiones

El preprocesamiento de datos es la etapa mas importante para el análisis de datos, ya que se asegura que los datos estén limpios, estructurados y sean los adecuados para su uso en modelos de Machine Learning y análisis. Sin este preprocesamiento, los modelos pueden producir malos resultados y erroneos.