Universidad Autónoma de San Luis Potosí
Facultad de ingeniería
Inteligencia Artificial Aplicada
Practica 4
Preprocesamiento de Datos
Ana Sofía Medina Martínez

Fecha 18/09/2024

Objetivo

Que el estudiante adquiera conocimientos sobre las técnicas de preprocesamiento de datos utilizadas en el análisis de datos y el aprendizaje automático.

Procedimiento

4.1.- Inicie Jupyter Notebooks y abra los notebooks "introduccion"

4.2.- Siga las instrucciones en los notebooks para explorar los conceptos básicos de prepocesamiento de datos.

4.3.- Aplicar las técnicas de preprocesamiento de datos al dataset “Social_Network_Ads”

Resultados

Aplicar las técnicas al dataset “Social_Network_Ads”

```
[ ] import pandas as pd
    social = pd.read_csv('/content/datasets/Social_Network_Ads.csv')
    social.head()
```

	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

Next steps: Generate code with social | View recommended plots | New interactive sheet

```
[ ] social.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 99 entries, 0 to 98
Data columns (total 4 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   Gender           99 non-null     object
 1   Age              99 non-null     int64
 2   EstimatedSalary  99 non-null     int64
 3   Purchased        99 non-null     bool
dtypes: bool(1), int64(2), object(1)
memory usage: 2.5+ KB
```

```
[ ] social['Gender'] = social['Gender'].str.strip().str.capitalize().map({'Male': 1, 'Female': 0})
    social.head()
```

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

```
[ ] social['Purchased'] = social['Purchased'].replace({False: 0, True: 1})
    social.head()
```

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

Next steps: Generate code with social | View recommended plots | New interactive sheet

```
[ ] social.describe()
```

	Gender	Age	EstimatedSalary	Purchased
count	99.000000	99.000000	99.000000	99.000000
mean	0.505051	30.282828	57616.161616	0.191919
std	0.502519	8.230159	33344.126268	0.395814
min	0.000000	18.000000	15000.000000	0.000000
25%	0.000000	25.000000	27000.000000	0.000000
50%	1.000000	28.000000	52000.000000	0.000000
75%	1.000000	33.500000	81500.000000	0.000000
max	1.000000	59.000000	150000.000000	1.000000

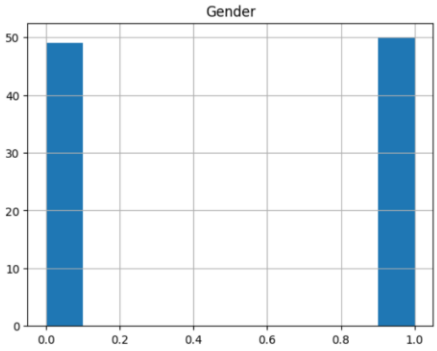
```
[ ] social.var()
```

	0
Gender	2.525253e-01
Age	6.773552e+01
EstimatedSalary	1.111831e+09
Purchased	1.566687e-01

dtype: float64

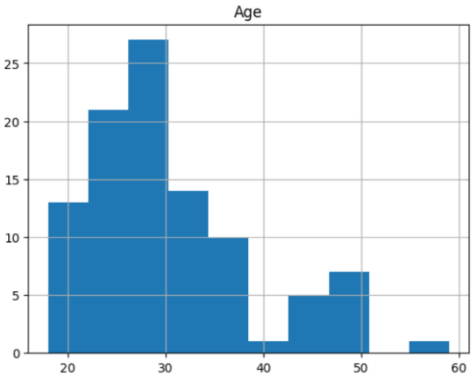
```
[ ] social.hist('Gender')
```

```
array([[<Axes: title={'center': 'Gender'}>]], dtype=object)
```



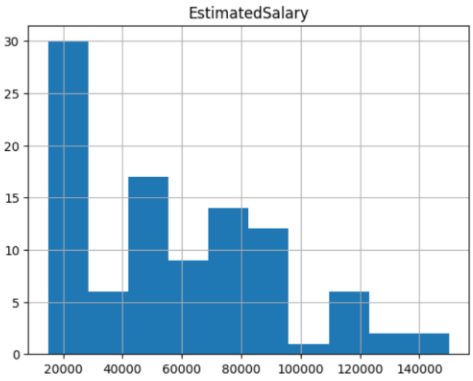

```
social.hist('Age')
```

```
array([[<Axes: title={'center': 'Age'}>]], dtype=object)
```



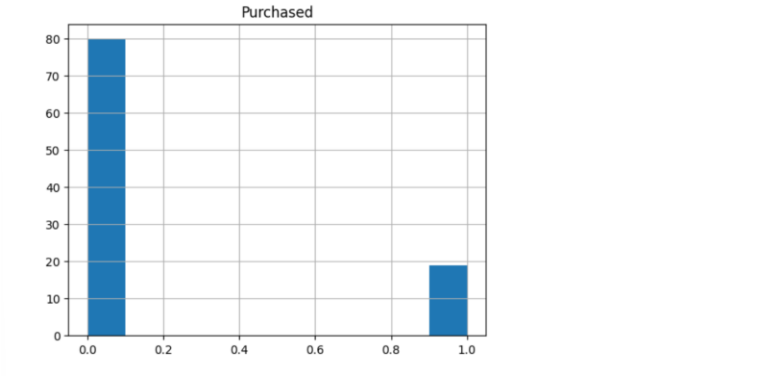

```
[ ] social.hist('EstimatedSalary')
```

```
array([[<Axes: title={'center': 'EstimatedSalary'}>]], dtype=object)
```



```
[ ] social.hist('Purchased')
```

```
array([[<Axes: title={'center': 'Purchased'}>]], dtype=object)
```



Comprensión

1. **¿Cuál es la importancia del preprocesamiento de datos en el análisis de datos y el aprendizaje automático?**
   Mejora la calidad y consistencia de los datos, lo que optimiza el rendimiento de los modelos de aprendizaje automático.

2. **Mencione al menos tres técnicas de preprocesamiento de datos y explique su función.**
   - Normalización: Escalar datos en un rango específico.
   - Imputación de valores faltantes: Reemplazar datos faltantes con la media, mediana, etc.
   - Codificación categórica: Convertir variables categóricas en números.

3. **¿Qué son los datos faltantes y cómo se pueden manejar durante el preprocesamiento?**
   Se pueden eliminar o imputar con la media, mediana, moda, etc.

4. **¿Qué son los valores atípicos y cómo se pueden detectar y tratar?**
   Son datos extremos que se pueden detectar con gráficos y z-score; se eliminan o ajustan según el caso.

5. **¿Cuál es la importancia de la codificación de variables categóricas?**
   Permite que los algoritmos manejen datos no numéricos transformándolos en números.

Conclusiones

El preprocesamiento de datos es esencial para mejorar la calidad y precisión de los modelos de análisis y aprendizaje automático, ya que permite manejar datos faltantes, valores atípicos y variables categóricas de manera efectiva, asegurando que los algoritmos funcionen de manera óptima.