Lab-1

Demonstrating various data Preprocessing techniques for

1) Data cleaning: Handling missing values, Handling Categorical data, handling outliers

2) Data Transformations: min-max Scalar / Normalisation, Standard Scalar.

1) Housing dataset

(1) Column information (ii) Statistical summary of columns

(3) unique values in Ocean Proximity column

(4) Columns with missing values.

Code

```
import pandas as pd
housing_df = pd.read_csv('/path/to/housing.csv')
Print ("Column Information:")
Print (housing_df.info())
Print ("\n Statistical Summary:")
Print (housing_df.describe())
Print ("\n unique values in 'Ocean Proximity':")
Print (housing_df['ocean Proximity'].value_counts())
Print ("\n Columns with missing values:")
Print (housing_df.isnull().sum())
```

output

Index (['longitude', 'latitude', 'housing-median-age', 'total-rooms', 'total-bedrooms', 'population', 'households', 'median-income', 'median-house-value', 'ocean-proximity'],

Column Information:

Data columns (total 10 columns):

	~~column~~ ~~longitude~~ columns	~~non-null~~ non-null ~~Count~~ ~~Count~~ Count	Dtype Dtype
0	longitude	20640 non-null	float 64
1	latitude	20640 non-null	float 64
2	housing-median-age	20613 non-null	float 64.
:			
9	ocean-Proximity	20640 non-null	Object.

Statistical summary

	longitude	latitude	housingmedian-age	totalrooms
count	20640.0000	20640.0000	20640.0000	20640.0000
mean	-119.5697	35.6318	28.6394	2635.54
std	2.0035	2.1359	12.5843	2181.53
min	-124.35000	3.25400	18.0000	1447.543
~~max~~ 25%	-121.8000	33.9300	18.100	1447.430
50%	-118.542	34.2600	29.010	2127.250
75%	-118.0100	37.7100	37.200	3148.100
max	-114.3100	41.9500	52.153	39320.010

unique values in 'Ocean Proximity'

ocean_Proximity

1H ocean	9136
INLand	6551
Near ocean	2658
Near Bay	2290
Island	5

~~N~~ dtype : int

Columns with missing values

longitude	0
latitude	0
housing median_age	0
⋮	
Ocean Proximity	0

dtype : int.

Scanned with OKEN Scanner

2) Python code to implement the following data Preprocessing techniques for Diabetes & adult Income dataset.

```
import Pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_csv ('/content/dataset of diabetes.csv')
  df.head(10)

df = pd.read_csv ("/content/adult data 1.csv")
  df.head(10)

Import numpy as np
# introduce some missing values for demonstration.
  df.loc[5, 'AGE'] = np.nan
  df.loc[9, 'BMI'] = np.nan
  df.head(10)
```

Output

	ID	No_Pation	Gender	Age	Urea	Cr	BMI	Class
			F	50.0	4.7	46	NaN	N
0	502	17975			4.5	62	0.17913	N
1	412	47975	M	43.0			0.17916	N
2	327	87656	F	32.0	4.2	72		
3	634	67643	M	NaN	4.7	2.3	0.69	N

i) Which columns in dataset had missing values? How did you handle them?

- Adult Income dataset (adult.csv)

  No columns had missing values

- Diabetes dataset (Dataset of diabetes.csv)

  No columns has missing values

  - Handling method: Since no missing value were found, no imputation was performed

② Which categorical column did you identify in dataset? How did you encode them?

Adult Income Dataset (adult.csv)

→ Categorical columns

→ workclass

→ Education

→ occupation

→ race

→ gender

→ income

→ relationship

Evaluating method: One-hot encoding was applied using pd.get dummies (df, drop_first = True)

→ diabetes Dataset

Categorical column

→ Gender

→ class

encoding One-hot encoding applied pd.get_dummies (df, drop_first = True)

(3) difference b/w min-max Scaling & Standardization

Feature	min max scaling	Standardization
~~type~~ definition	Scales values b/w n fixed range (usually 0 & 1)	centers data on mean (0) with SD 1
formula	$X'_{scaled} = x_{max} - x_{mi\cancel{x}} -$ / x_{min}	X_{stand} $d = \sigma x - \mu$
Effect	Preserves original data distribution but scales it within a fixed range	Changes data distribution to have zero mean & unit variance

(4) When would you use one over other

min max - we need data to be within a fixed range

The dataset doesn't contain extreme outliers

Standardization is used when

- The dataset has ~~anything~~ Varying units, a Gaussian distribution (normal)

- There are significant outliers, as Standardization is less sensitive to them.

Sen
10-03-2024