Date 3/3/25
Page

Lab 1

To demonstrate various data pre-processing techniques for a given dataset.

Python code of "housing.csv" file

```
import pandas as pd
# i) To load csv file into data frame
filename = "housing.csv"
df = pd.read_csv(filename)

# ii) To display information of all columns
print("Columns:")
print(df.info())

# iii) Statistical information of all numerical
print(df.describe())

# iv) Count of unique labels in Ocean proximity
print(df['ocean_proximity'].value_counts())

# v) Attributes or columns having missing
value greater than zero

print(df.isnull().sum()[df.isnull().
sum() > 0])
```

Date __/__/__
Page _____

Python code to implement the following data preprocessing techniques for

① Diabetes

Data preprocessing techniques:
1. Data cleaning: Handling missing values, Handling categorical data, handling outliers
2. Data transformation: min-max scaler / Normalization: Standard scaler

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.preprocessing import LabelEncoder
data = pd.read_csv('/Dataset of Diabetes.csv')

1. Data cleaning
print("Missing values in each column:")
print(data.isnull().sum())
numerical_columns = ['Urea', 'Cr', 'HbA1c',
    'Chol', 'TG', 'HDL', 'LDL', 'VLDL', 'BMI']
for col in numerical_columns:
    data[col] = data[col].fillna(data[col].median())

data.drop_duplicates(inplace=True)

print("Missing values after handling:")
print(data[numerical_columns].isnull().sum())
print("Shape after removing duplicates:",
                    data.shape)
```

Date __/__/__
Page _____

Dataset of diabetes.csv

1. Which columns in the dataset had missing values? How did you handle them?

* In this dataset, there were no explicit missing values. This was check when the dataframe were loaded to pandas data.isnull().
* But missing values may exists in numerical columns
* The numerical missing values were handles by filling it with median value of respective columns.
* In categorical columns, no missing values were explicitly handled.

2. Which categorical columns did you identify in the dataset? How did you encode them?

* The categorical columns were identified as Gender and Class, where Gender contains F & M value & Class contains N, P & Y
* This value were encoded using label Encoder(), where
  Gender F → 0, M → 1
  Class N → 0, P → 1 & Y → 2

Date __/__/__
Page _____

3. What is the difference between min max scaling & standardization? When would you use one of the other?

min-max:

Transforms features to a fixed range [0,1].

Formula: $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$

Preserves the shape of the dataset

Standardization:

Transforms features to have a mean of 0 & standard deviation of 1.

Use:

minmax is used during uniform scaling [0,1]

Standardization is used for normally distributed data.

⇒ Adult income dataset.csv

1. Which column in the dataset had missing values? How did you handle them

columns with missing values:
workclass, occupation, native-country

2. Which categorical column did you identify? How to encode them

Date __/__/__
Page _____

categorical columns: workclass, education, marital-status, occupation, relationship, race, gender, native-country, income.
use label encoder!

3. what is the difference between min max scaling and standardization?

→ min max scale data to a fixed range typically [0,1]
used when needed bounded values.

→ standardization transforms data to mean 0 & standard deviation 1.