Date = 3/03/25

Lab-1

Data Browsing tab

1.) Load csv file

Import pandas as pd

⇒ file = pd.read_csv("housing.csv")

2.) Information of all col^n's

⇒ file1 = file.info()

3.) Statistical info of all numerical columns

⇒ file.describe()

4.) Count of unique labels. values for "ocean_proximity".

⇒ file['ocean_proximity'].value_count()

5.) To count attribute value having missing values count 70

⇒ missing_value = df.isnull().sum()

count = missing_value[missing_value 70]

print(count)

3-3-25

After Execution,

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

Ans ⇒ No column in both the dataset had any missing value.

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

For Adult Dataset,

Ans ⇒ Income column was categorical. Handled using ordinal encoder with <=50K & >50K

Work class column was handled using one-hot encoding.

→ Diabetes dataset:

~~No column had~~

Gender and class were categorical.

Gender → ordinal encoding ['M', 'F']

Class → one-hot encoding

Q.3.)

Ans ⇒

Q.3.) What is the difference between Min-Max Scaling and Standardization? When would you use one over the other?

Ans ⇒

Min Max Scaling

$$X_i' = \frac{X_i - \text{Min}}{\text{Max} - \text{Min}}$$

Range between [0, 1]

ideal for distance-based Model but can distort data with extreme outliers.

Standardization:

$$X_i' = \frac{X_i - \text{mean}}{\text{std}}$$

range = [-1, 1]

Works well for normally distributed data but sensitive to outliers.