S. B. JAIN INSTITUTE OF TECHNOLOGY, MANAGEMENT & RESEARCH, NAGPUR.

Practical No. 4

Aim: Split dataset for Testing and Training a model and print size of it.

Name of Student: Shrutika Pradeep Bagdi

Roll No.: **CS22130**

Semester/Year: **IV/VII**

Academic Session: 2025-2026

Date of Performance: _____

Date of Submission: _____

AIM: Split dataset for Testing and Training a model and print size of it

OBJECTIVE/EXPECTED LEARNING OUTCOME:

- Understanding Training and testing data

HARDWARE AND SOFTWARE REQUIRMENTS:

Hardware Requirement:

Software Requirement:

THEORY:

You are given a dataset – "housing.csv." Load the dataset into a DataFrame. Considering the "median_house_value" column as output/target and the rest of the columns as input, perform the following tasks:

1. Find the number of null values in each column and replace them with the mean values
2. Split the data into test and train fragments using train_test_split() function in 80:20 ratio (80% train, 20% test)
3. Print the size of test and train data (For both input and output)

CODE:

```
#Name: Shrutika Bagdi_CS22130
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split

# Load the Iris dataset
iris = load_iris()

# Features and target
X = iris.data      # All feature columns (sepal length, sepal width, petal length, petal width)
y = iris.target    # Target variable (species)

# Split dataset: 70% training, 30% testing
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Print sizes of training and testing sets
print(f"Training set size: {X_train.shape[0]} samples")
print(f"Testing set size: {X_test.shape[0]} samples")
```

OUTPUT (SCREENSHOT):

```
Training set size: 105 samples
Testing set size: 45 samples
```

N-Grams

Corpus A

Select Corpus

(eos) Can I sit near you (eos) You can sit (eos) Sit near him (eos) I can sit you (eos)

Find Bigram Probabilities

	(eos)	I	you	him	can	near	sit
(eos)	0	0.2	0.2	0	0.2	0	0.2
I	0	0	0	0	0.5	0	0.5
you	0.66	0	0	0	0.33	0	0
him	1	0	0	0	0	0	0
can	0	0.33	0	0	0	0	0.66
near	0	0	0.5	0.5	0	0	0
sit	0.25	0	0.25	0	0	0.5	0

Submit

Take a quiz

Sentence	Probability
I sit you EOS	1
Can you sit near I EOS	2
I can sit EOS	3
You sit EOS	4

Submit

CONCLUSION:

DISCUSSION AND VIVA VOCE:

- Why data is spitted for test and training
- How data is spitted for test and training

REFERENCE:

- *www.w3schools.com*
- *www.tutorialsmade.com*
- *https://www.javatpoint.com/*