SANDIPUNIVERSITY, SIJOUL

School of Computer Science& Engineering

Name of Student: **PRN:**
Date of Performance: **Date of Completion:**

ExperimentNo: 01

Aim: Loading and Exploring a Dataset (e.g., Iris, MNIST) Using Pandas and Calculating Summary Statistics

Objective:

- To learn how to load datasets (Iris and MNIST) into Python using the Pandas library.
- To explore the structure and content of the datasets.
- To calculate and interpret summary statistics (e.g., mean, median, standard deviation) for the datasets.

Prerequisites

- **Software**: Python 3.x installed with the following libraries:
  - pandas
  - numpy
  - scikit-learn (for Iris)
  - tensorflow or keras (for MNIST, optional for raw data loading)
- **Hardware**: A computer with sufficient memory (at least 4GB RAM recommended).
- **Skills**:
  - Basic Python programming knowledge.
  - Familiarity with installing Python libraries using pip (e.g., pip install pandas numpy scikit-learn).
- **Dataset Availability**:
  - Iris dataset (built into scikit-learn).
  - MNIST dataset (available via tensorflow.keras.datasets or as raw CSV files).

Theory

- **Pandas**: A powerful Python library for data manipulation and analysis, providing data structures like DataFrames.
- **Iris Dataset**: A classic multivariate dataset with 150 samples of 4 features (sepal length, sepal width, petal length, petal width) and 3 species classes.
- **MNIST Dataset**: A large database of handwritten digits (0–9) with 70,000 images (28x28 pixels each), often used for machine learning.
- **Summary Statistics**: Descriptive measures (e.g., mean, median, min, max) that summarize the central tendency, dispersion, and shape of a dataset's distribution.

Apparatus/Software Requirements

- Python IDE or Jupyter Notebook.
- Internet connection (for initial library installation).

Procedure

Step 1: Set Up the Environment

1. Install required libraries if not already installed:

```
pip install pandas numpy scikit-learn tensorflow
```

2. Launch Jupyter Notebook or your preferred Python environment.

Step 2: Load the Iris Dataset

1. Import necessary libraries.
2. Load the Iris dataset using scikit-learn.
3. Convert it to a Pandas DataFrame for analysis.

Step 3: Explore the Iris Dataset

1. Display the first few rows of the dataset.
2. Check the dataset's shape, column names, and data types.
3. Calculate summary statistics.

Step 4: Load the MNIST Dataset

1. Import the MNIST dataset using tensorflow.keras.datasets.
2. Preprocess the data (e.g., flatten images) and create a DataFrame.
3. Explore the dataset and calculate summary statistics.

Step 5: Analyze and Interpret Results

1. Compare summary statistics between Iris and MNIST.
2. Document observations (e.g., range of values, missing data).

Program Code

Lab Program: Loading and Exploring Datasets with Pandas

Import libraries

```
import pandas as pd
import numpy as np
from sklearn.datasets import load_iris
from tensorflow.keras.datasets import mnist
```

Step 1: Load and Explore Iris Dataset

```
print("=== Exploring Iris Dataset ===")
```

Load Iris dataset

```
iris = load_iris()
iris_df = pd.DataFrame(data=iris.data, columns=iris.feature_names)
iris_df['target'] = iris.target
```

Display first 5 rows

```
print("First 5 rows of Iris dataset:")
print(iris_df.head())
```

Check dataset info

```
print("\nDataset Info:")
print(iris_df.info())
```

Calculate summary statistics

```
print("\nSummary Statistics for Iris dataset:")
print(iris_df.describe())
```

Step 2: Load and Explore MNIST Dataset

```
print("\n=== Exploring MNIST Dataset ===")
```

Load MNIST dataset
(X_train, y_train), (X_test, y_test) = mnist.load_data()

Flatten images and create DataFrame (using first 1000 samples for simplicity)
mnist_df = pd.DataFrame(X_train.reshape(X_train.shape[0], -1)[:1000])
mnist_df['label'] = y_train[:1000]

Display first 5 rows
print("\nFirst 5 rows of MNIST dataset:")
print(mnist_df.head())

Check dataset info
print("\nDataset Info:")
print(mnist_df.info())

Calculate summary statistics (for pixel values)
print("\nSummary Statistics for MNIST dataset (pixel values):")
print(mnist_df.drop(columns=['label']).describe())

Expected Output
Iris Dataset

First 5 rows of Iris dataset:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0
1	4.9	3.0	1.4	0.2	0
2	4.7	3.2	1.3	0.2	0
3	4.6	3.1	1.5	0.2	0
4	5.0	3.6	1.4	0.2	0

Summary Statistics:

- Mean sepal length: ~5.84 cm
- Min petal width: 0.1 cm
- Max petal length: 6.9 cm

MNIST Dataset

- **First 5 Rows** (partial view of flattened 28x28=784 pixels + label):

First 5 rows of MNIST dataset:

	0	1	2	...	783	label
0	0	0	0	...	0	5
1	0	0	0	...	0	0
2	0	0	0	...	0	4
3	0	0	0	...	0	1
4	0	0	0	...	0	9

Summary Statistics (for pixel values, 0–255 range):

- Mean: ~30–40 (varies by sample subset)
- Min: 0
- Max: 255

Observations

- **Iris**: The dataset has 150 rows and 5 columns (4 features + target). All features are numeric with no missing values. Summary statistics show a range of values typical for flower measurements.
- **MNIST**: The dataset has 60,000 training images (subset to 1000 here), with 784 pixel features per image. Pixel values range from 0 (white) to 255 (black), with labels 0–9.

Conclusion

This lab demonstrated how to load and explore the Iris and MNIST datasets using Pandas. Summary statistics provided insights into the data's distribution, preparing the ground for further analysis (e.g., machine learning).

Viva question:

1. What differences do you notice in the summary statistics between Iris and MNIST?
2. How might missing values or outliers affect the summary statistics?
3. Suggest a way to visualize the Iris dataset's features (e.g., scatter plot).

NAMEOFTEACHER:
DATE:
SIGN