**Department of Electrical Engineering**

**Faculty Member:** LE Munadi Sial **Date:** 4-Oct-2023

**Semester:** 7th **Group:**

# CS471 Machine Learning

**Lab 4: Introduction to Pandas and MatplotLib**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **PLO4 - CLO4** | **PLO4 -CLO4** | **PLO5 -CLO5** | **PLO8 -CLO6** | **PLO9 -CLO7** |
| **Name** | **Reg. No** | **Viva /Quiz / Lab Performance** | **Analysis of data in Lab Report** | **Modern Tool Usage** | **Ethics** | **Individual and Team Work** |
|  |  | **5 Marks** | **5 Marks** | **5 Marks** | **5 Marks** | **5 Marks** |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

## Introduction

This laboratory exercise is focused on handling and visualizing datasets for machine learning purposes. In any machine learning task, we are working with data. For dataset handling, we use the Pandas library which can load .csv files into a data frame. During machine learning, we also need to make plots. For this, we make use of the PyPlot submodule in the MatplotLib library.

## Objectives

The following are the main objectives of this lab:

* Implement data structures such as lists and dictionaries in python
* Create, alter and loop through lists
* Use slicing to access range of items in a list
* Utilize various list methods such as append, insert, extend, remove, pop etc
* Create and implement a dictionary
* Create Numpy arrays and perform matrix operations and broadcasting
* Use Scipy for minimization, scarce matrices and iterpolation

## Lab Conduct

* Respect faculty and peers through speech and actions
* The lab faculty will be available to assist the students. In case some aspect of the lab experiment is not understood, the students are advised to seek help from the faculty.
* In the tasks, there are commented lines such as #YOUR CODE STARTS HERE# where you have to provide the code. You must put the code between the #START and #END parts of these commented lines. Do NOT remove the commented lines.
* Use the tab key to provide the indentation in python.
* Upon completing the lab, you must delete the manual from the lab computer

**Theory**

Pandas (panel data) is a library that can load tabular data from .csv files and store into a NumPy compatible table known as a “Pandas Data Frame”. Each column in a data frame is of a “Pandas Series” type. Aside from loading datasets, pandas also enables us to perform basic mean, mode, median operations as well as clean up incomplete or duplicate data.

MatplotLib is another library focused on data visualization. It contains many functions for displaying plots, subplots, scatter plots etc. Line plots are used widely for monitoring training accuracies and losses. Scatter plots are used mainly for modeling the feature space of the dataset.

A brief summary of the list functions in python is provided below:

**append(I)** append item I to the end of the list

**insert(i, I)** insert item I at i position of the list

**extend(L)** extend/concatenate a second list L

**remove(I)** remove a specified item I from a list

**pop(i)** remove item at specific index i in the list

**count(I)** return total number of a specific item I from a list

**index(I)** return index of first occurrence of a specific item I

**reverse** reverse the items of the list

For this lab, you will be provided with 2 dataset files in .csv format which you will need for the tasks. Additionally, for the final task, you will need to arrange your own dataset by downloading it from the internet. You will need to import pandas and matplotlib.pyplot for the given tasks.

**Lab Task 1 – Pandas Series and Dataframes \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

1. Create a Pandas series using a dictionary and display the output.
2. Create a Pandas dataframe using a dictionary and display the output.

Provide all of the codes and screenshots of the final outputs.

***### TASK 1 CODE STARTS HERE ###***

*### TASK 1 CODE ENDS HERE ###*

***### TASK 1 OUTPUT SCREENSHOT STARTS HERE ###***

*### TASK 1 OUTPUT SCREENSHOT ENDS HERE ###*

**Lab Task 2 – CSV Files \_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Load dataset 1 into a dataframe and perform the following

1. Print the dataset using the head and tail functions
2. Print any 3 rows from the dataset
3. Print any 5 elements from the dataset
4. Use the mean, mode and median functions for each column in the dataset

Provide all of the codes and screenshots of the final output.

***### TASK 2 CODE STARTS HERE ###***

*### TASK 2 CODE ENDS HERE ###*

***### TASK 2 OUTPUT SCREENSHOTS START HERE ###***

*### TASK 2 OUTPUT SCREENSHOTS END HERE ###*

**Lab Task 3 – Dataset Cleaning \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Load dataset 2 into a dataframe.

1. Write code to remove the incomplete rows from the dataset
2. Write code to remove the duplicated rows from the dataset
3. Save the cleaned dataset into a dataframe. You need to attach this cleaned dataset file (renamed to lab5\_task3.csv) in your lab submission.

***### TASK 3 CODE STARTS HERE ###***

*### TASK 3 CODE ENDS HERE ###*

***### TASK 3 OUTPUT SCREENSHOT STARTS HERE ###***

*### TASK 3 OUTPUT SCREENSHOT ENDS HERE ###*

**Lab Task 4 – Line and Scatter Plots \_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

For this task, you will need to use datasets 1 and 2. You will also require the matplotlib.pyplot module for plotting. Perform the following.

1. Make line plots of the following equations for x = 1 to 100. You will need to make use of NumPy arrays for this part.
   1. y = 2x + 1
   2. y = 3x2
   3. y = cos(x)
2. Load dataset 1 and make a scatter plot (axes x1 and x2)
3. Load the cleaned dataset 2 and make a scatter plot (axes x1 and x2). You need to use markers for the labels (y) such that 0 corresponds to a red circle and 1 corresponds to a blue square. The label y is the “inland” column. For x1 and x2, choose any 2 columns from the dataset and also mention the columns that you are using.
4. Load the cleaned dataset 2 and make a 3-D scatter plot between any three features in the dataset (axes x1, x2, x3)

***### TASK 4 CODE STARTS HERE ###***

*### TASK 4 CODE ENDS HERE ###*

***### TASK 4 OUTPUT SCREENSHOT STARTS HERE ###***

*### TASK 4 OUTPUT SCREENSHOT ENDS HERE ###*

**Lab Task 5 – Dataset Batches \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Load the cleaned dataset 2 into as a dataframe. For this task, you will divide the dataset into 10 batches. For each individual batch, calculate the mean, mode and median for any 3 feature columns. Finally, make line plots showing the batch number on the x-axis and the mean, mode and median on the y-axis.

***### TASK 5 CODE STARTS HERE ###***

*### TASK* ***5*** *CODE ENDS HERE ###*

***### TASK 5 OUTPUT SCREENSHOT STARTS HERE ###***

*### TASK 5 OUTPUT SCREENSHOT ENDS HERE ###*

**Lab Task 6 – Your Own Dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Download your own CSV dataset from the internet (e.g. Kaggle). Your dataset must have at least 500 rows and at least 2 feature columns. Your dataset must also have a labels column with classification data (0/1). Make a scatter plot between the feature axes and show the labels with different markers. Provide all of the codes and screenshots of the plots. You will also need to submit the downloaded dataset with your report (renamed as lab4\_task6.csv). Note that no two submitted datasets must be exactly the same.

***### TASK 6 CODE STARTS HERE ###***

*### TASK 6 CODE ENDS HERE ###*

***### TASK 6 SCREENSHOTS START HERE ###***

*### TASK 6 SCREENSHOTS END HERE ###*