**National University of Computer and Emerging Sciences**



**Lab Manual 01**

**Data Mining**

Department of Computer Science FAST-NU, Lahore, Pakistan

# Objectives

Data Preprocessing and Exploratory Data Analysis

# Online Python Interpreter

We will be working on [Google Collab](https://colab.research.google.com/) to run Python programs.

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser without any configuration. To do that we need to create a Colab notebook.

Colab notebooks allow you to combine executable code and rich text in a single document, along with images, HTML, LaTeX and more. When you create your own Colab notebooks, they are stored in your Google Drive account. You can easily share your Colab notebooks with co- workers or friends, allowing them to comment on your notebooks or even edit them.

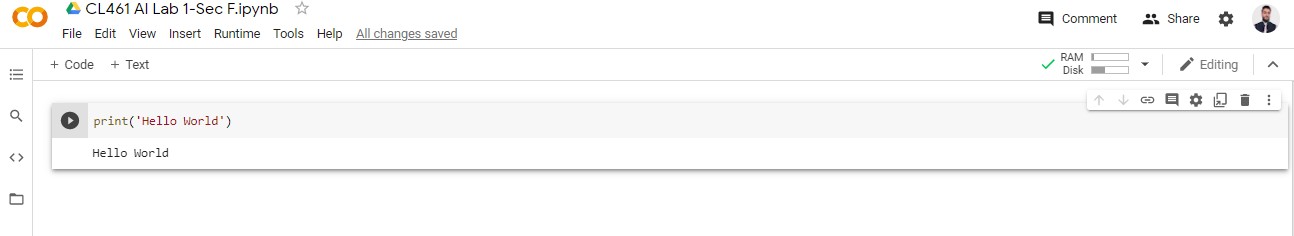
**Colab notebooks are Jupyter notebooks that are hosted by Colab.** The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more. Learn more about Jupyter [here.](https://jupyter.org/)

Follow the instructions given below:

* Visit the following link: <https://colab.research.google.com/>
* Sign in using your nu email id.
* You will be directed to ‘Welcome to Colaboratory’ page.
* Go to **File->New Notebook**.
* A new notebook is created by the name ‘Untitled0.ipynb’.
* Rename the notebook to your roll number.
* Write your Python program by typing the following statement in the first cell

o print(‘Hello World’)

* Execute this cell by clicking the play button on the left of the cell or by pressing Ctrl+Enter.
* You will notice that the notebook will connect to a runtime. RAM and Disk resources are allocated. (Refer to the screenshot below)



# Data Preprocessing Exercise 1:

**Create a Text cell in your Google Colab notebook and answer the following:**

Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

Example: Age in years. Answer: Discrete, quantitative, ratio

(a) Time in terms of AM or PM.

(b) Brightness as measured by a light meter.

(c) Brightness as measured by people’s judgments.

(d) Angles as measured in degrees between 00 and 3600

(e) Bronze, Silver, and Gold medals as awarded at the Olympics.

(f) Height above sea level.

(g) Number of patients in a hospital.

(h) ISBN numbers for books. (Look up the format on the Web.)

(i) Ability to pass light in terms of the following values: opaque, translucent, transparent.

(j) Military rank.

(k) Distance from the center of campus.

(l) Density of a substance in grams per cubic centimeter.

(m) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)

# Exploratory Data Analysis Exercise:

First Upload the given dataset in Google Colab using the following instructions:

## **Step 1: Mount Your Google Drive**

The initial step is to establish a connection between your Google Drive account and your Colab notebook. To achieve this, run the following code:

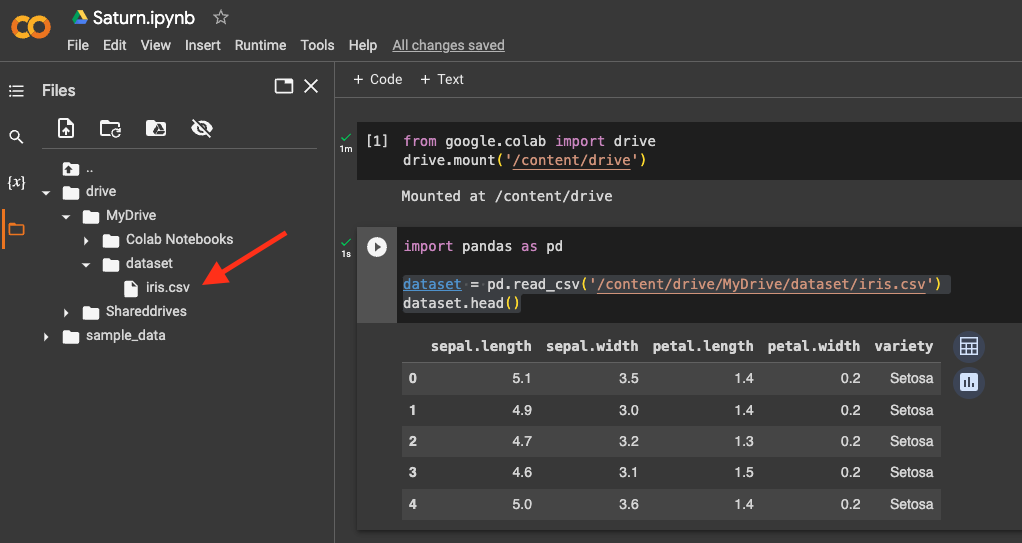
from google.colab import drive

drive.mount('/content/drive')

This action will trigger an authorization request, permitting Colab to access your Google Drive account. Follow the on-screen instructions, and when prompted, enter the provided authorization code. With your Google Drive successfully mounted, you gain access to your files directly within your Colab notebook.

From here, you can import files from Google Drive to Colab by navigating to the file path in the left-hand sidebar of your Colab notebook, as illustrated below.

## **Step 2: Locate the File You Want to Import**

Once your Google Drive is mounted, you can navigate to the file that you want to import to Colab by looking at the panel on the left of the Colab notebook, as shown in the following figure.

## **Step 3: Import the File into Your Colab Notebook**

In this instance, we are importing a CSV file using the [pandas](https://saturncloud.io/glossary/pandas) library. Replace data.csv with the actual file file name. You can use similar code to import other file types like images or text files.

import pandas as pd

df = pd.read\_csv('/content/drive/MyDrive/dataset/data.csv')

df.head()

Now after Importing the dataset, use pandas, matplotlib and seaborn for EDA

1. Classifying the dependent and Independent Variable
2. Report how many examples we have of each feature (use df.shape() df.value\_counts(), and df.dtypes)
3. Report if any attribute has missing value
4. Report mean, median, mode, range, and STD for each attribute
5. Visualize data by creating pair plot, scattered plot, and histrograms, Heatmap Correlation
6. Submission Instructions

Always read the submission instructions carefully.

* Rename your Jupyter notebook to your roll number\_Lab01 and download the notebook as **.ipynb** extension.
* To download the required file, go to **File->Download .ipynb**
* Only submit the **.ipynb** file. DO NOT **zip** or **rar** your submission file
* Submit this file on Google Classroom under the relevant assignment.
* Late submissions will not be accepted.