EX.NO: 1	Study of Machine Learning Tools and Working with Dataset
DATE: 25.03.2022	

AIM:
To study and implement machine learning tools and working with datasets.

1. **What is a dataset? Also create a sample dataset as csv file. Check the properties of a csv file**.

A dataset in machine learning is, quite simply, a collection of data pieces that can be treated by a computer as a single unit for analytic and prediction purposes. This means that the data collected should be made uniform and understandable for a machine that doesn't see data the same way as humans do

```
import csv
Head = ['football', 'volleyball', 'cricket', 'basketball']
Data = [ ['20EUAI005', '20EUAI010', '20EUAI011', '20EUAI006'],
      ['20EUAI016', '20EUAI003', '20EUAI007', '20EUAI022'],
      ['20EUAI0020', '20EUAI018', '20EUAI025', '20EUAI030'],
      ['20EUAI026', '20EUAI015', '20EUAI033', '20EUAI037'],
      ['20EUAI044', '20EUAI040', '20EUAI029', '20EUAI035'] ]

with open('Score.csv', 'w', encoding='UTF8') as f:
  writer = csv.writer(f)
  writer.writerow(Head)
  writer.writerows(Data)

with open('Score.csv','rt') as f:
  data = csv.reader(f)
  for row in data:
    print(row)
```

THE PROPERTIES OF A CSV FILE:
A comma-separated values (CSV) file is a delimited text file that uses a comma to separate values. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format.

2. Perform a self-study on the python Packages used in Machine Learning.

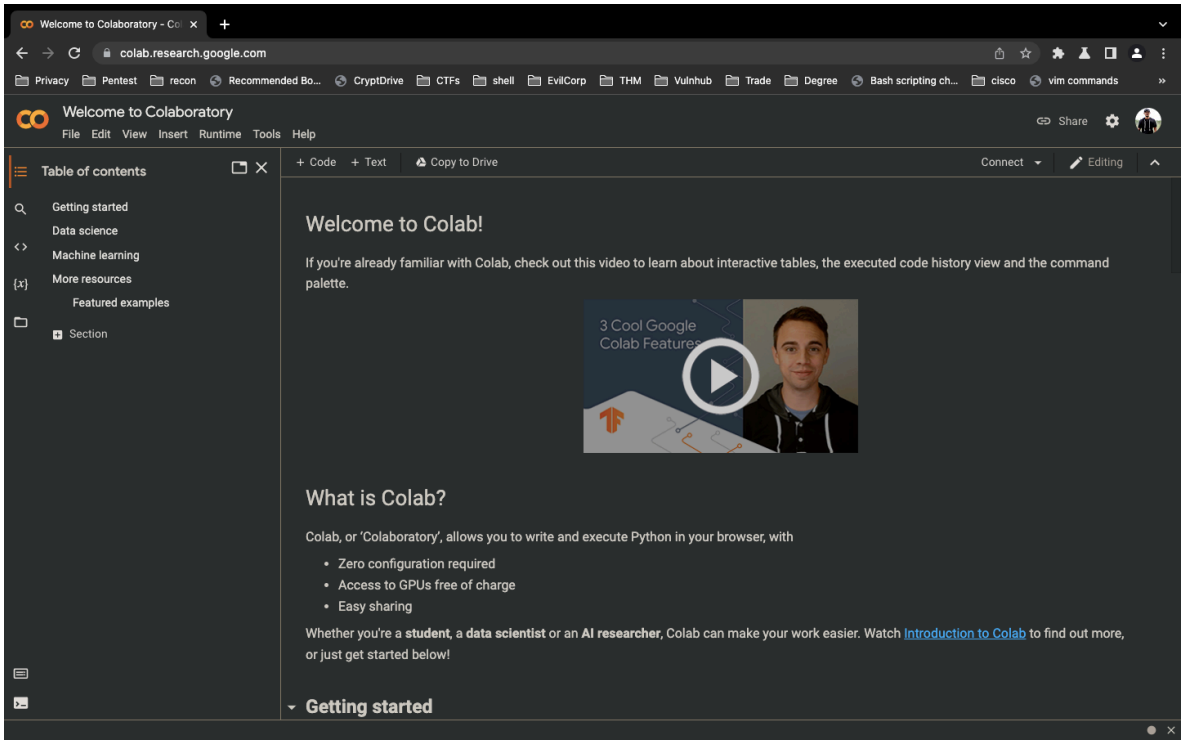
a. Tensorflow
    i. Tensorflow is a open source library which is used for Machine learning and Artificial intelligence which is used in python and java programming languages. It is focused on deep neural networks.
b. Natural Language ToolKit (NLTK)
    i. NLTK is the widely used library for Text Classification and Natural Language Processing.
c. Sci-kit
    i. Scikit-learn is mostly focused on various data modeling concepts like regression, classification, clustering, model selections
d. Keras
    i. Keras provides a Python interface of Tensorflow Library especially focused on AI neural networks.
e. Pytorch
    i. The main focus of the library is only on developing and training deep learning models only.
f. MLpack
    i. The main emphasis while developing this library was on making it a fast, scalable, and easy-to-understand as well as an easy-to-use library so that even a coder new to programming can understand and use it without any problem
g. OpenCv
    i. OpenCV is an open-source platform dedicated to computer vision and image processing

3. **Give due importance to the following and see what the following modules and functions do:**

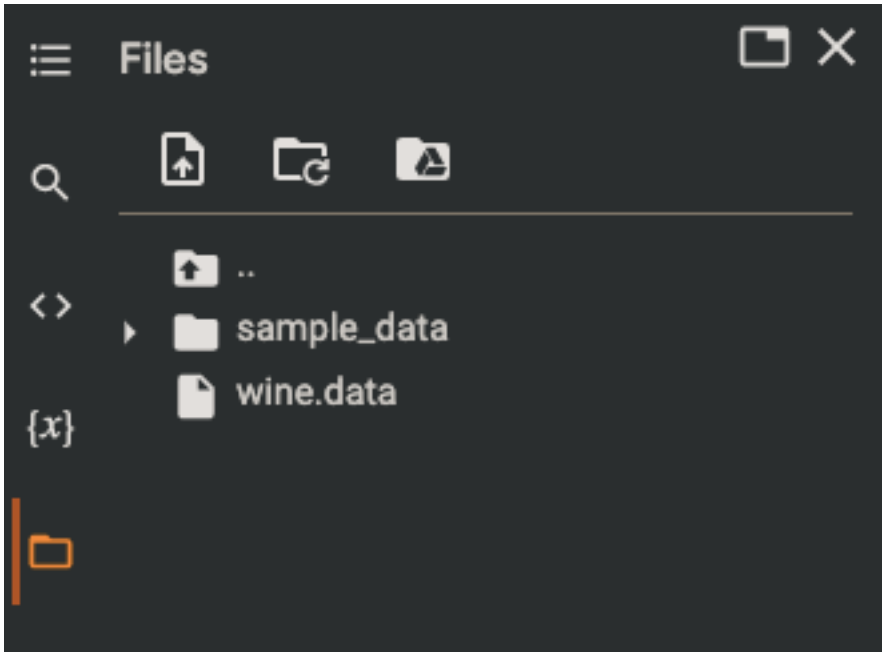
   a. Scipy
      i. SciPy is a free and open-source Python library used for scientific computing and technical computing. SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers and other tasks common in science and engineering.
   b. Numpy
      i. NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
   c. Pandas
      i. pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series
   d. Sklearn
      i. Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines.
   e. Matplotlib
      i. Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

4. Explore Google Colab and Load a dataset and see how it works.

a. This is how Google Colab looks like.



b. Upload the csv file on the left menu.



c. Write the python code in a code block and execute it using the play button on the left top corner of the code block cell.

```python
import csv
Head = ['football', 'volleyball', 'cricket', 'basketball']
Data = [ ['20EUAI005', '20EUAI010', '20EUAI011', '20EUAI006'],
         ['20EUAI016', '20EUAI003', '20EUAI007', '20EUAI022'],
         ['20EUAI0020', '20EUAI018', '20EUAI025', '20EUAI030'],
         ['20EUAI026', '20EUAI015', '20EUAI033', '20EUAI037'],
         ['20EUAI044', '20EUAI040', '20EUAI029', '20EUAI035'] ]

with open('Score.csv', 'w', encoding='UTF8') as f:
  writer = csv.writer(f)
  writer.writerow(Head)
  writer.writerows(Data)

with open('Score.csv','rt') as f:
  data = csv.reader(f)
  for row in data:
    print(row)
```

d. The output will be displayed below each cell block.

```
['football', 'volleyball', 'cricket', 'basketball']
['20EUAI005', '20EUAI010', '20EUAI011', '20EUAI006']
['20EUAI016', '20EUAI003', '20EUAI007', '20EUAI022']
['20EUAI0020', '20EUAI018', '20EUAI025', '20EUAI030']
['20EUAI026', '20EUAI015', '20EUAI033', '20EUAI037']
['20EUAI044', '20EUAI040', '20EUAI029', '20EUAI035']
```

5. **Create a dataset with the following features:**

Height(in cms)	Weight(in Kgs)
158	48
159	62
160	72
162	71
157	70
164	61
166	82
172	71
161	86
165	65

a. Create the dataset in csv file and upload the file in Google Colab. Write the code and get the output.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

data = pd.read_csv("example_dataset.csv")
data.head(10)
```

	Height (in cms)	Weight (in kgs)
0	158	48
1	159	62
2	160	72
3	162	71
4	157	70
5	164	61
6	166	82
7	172	71
8	161	86
9	165	65

b. Print some details about the csv file (dataset).

```
print(f"Shape of the data: {data.shape}")
print(f"Type of the data: {type(data)}")
print(f"Size of the data: {data.size}")
```

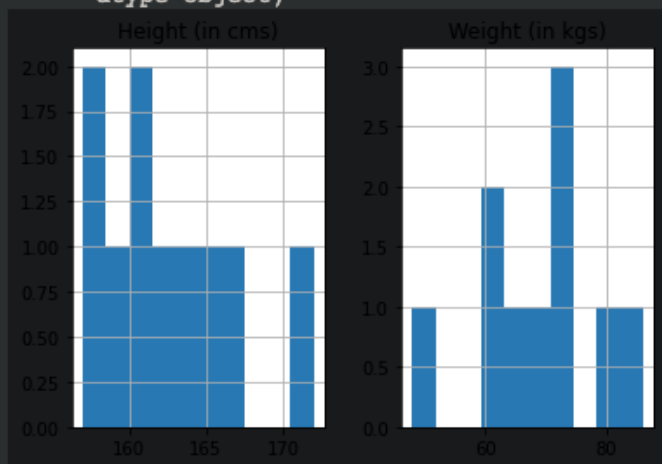
```
Shape of the data: (10, 2)
Type of the data: <class 'pandas.core.frame.DataFrame'>
Size of the data: 20
```

6. Identify the instances and plot a graph using these dimensions using matplotlib

Plotting histogram for the given csv file (dataset)

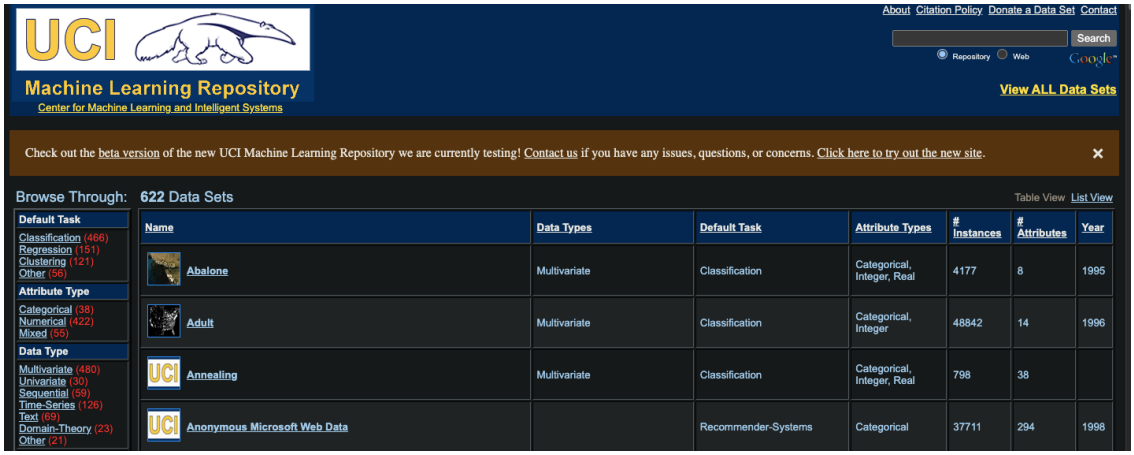
```
[6] data.hist()
```

```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f07b1ea2f90>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f07b1e80ed0>]],
      dtype=object)
```

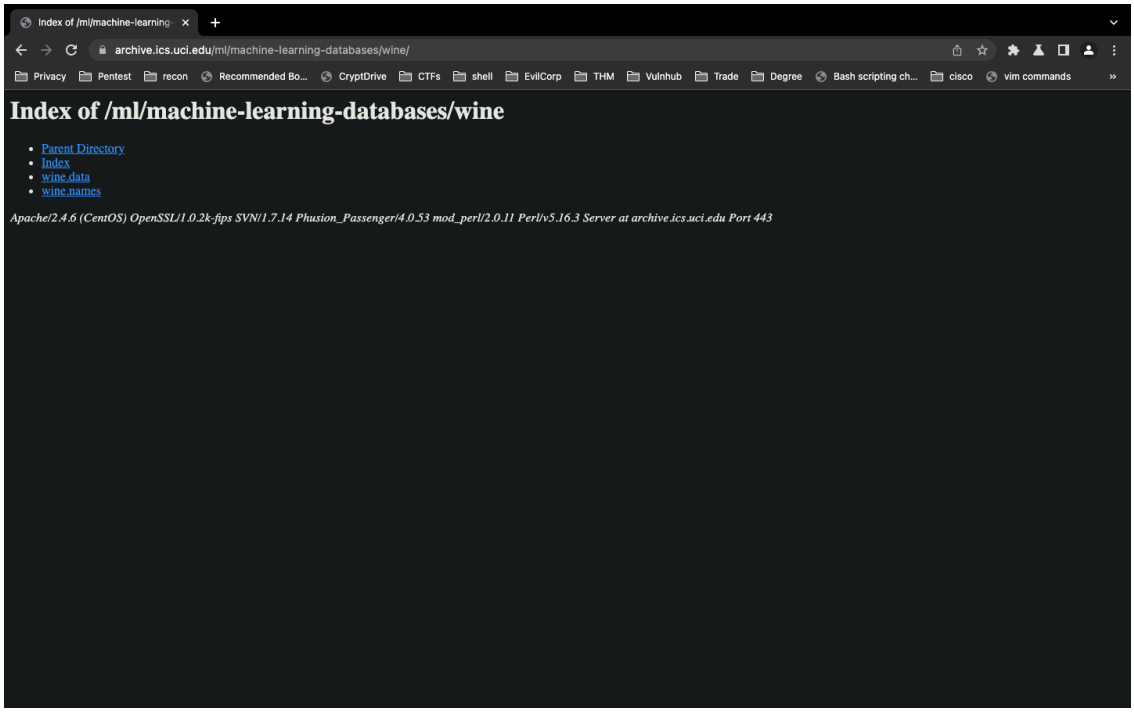


7. **Explore more about importing datasets from UCI repositories and perform the same.**

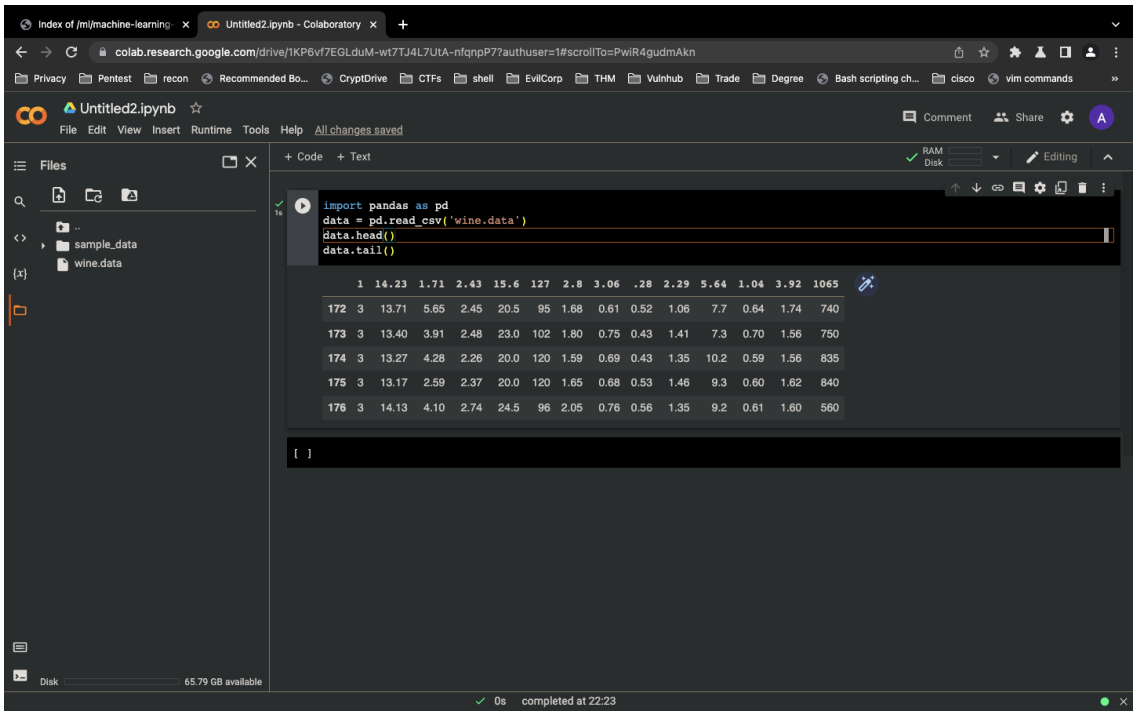
a. Search for the required datasets from UCI repository.



b. Download the required dataset.



c. Upload and run the python code to get the output.



RESULT:

We have successfully studied and implemented machine learning tools and working with datasets.