DHRUV GARG 16 BCE 1190

PRE-PROCESSING OF THE DATASET

LAB 1

DATASET CHOSEN: Absenteeism at work (https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work)

The first step to implementing machine learning on any dataset is to **understand and clean the dataset** – in other words, perform pre-processing. In the process of understanding the dataset, we might find some faults in the data (**inconsistent measures, missing values**) – which will affect the algorithm implementation.

My first step was to import the pandas library and **read the ".csv" file using "read_csv"**. On reading, all the attribute values for a row went into the same column. This was found using the "shape" attribute of the dataframe. I observed that the delimiter used in the csv file was semi-colon(;) and not comma(,). Thus, all values were incorrectly going into a single column. To correct this import of csv data, I added a " sep = ';' " argument in the read_csv command. Now when the first 5 rows of "absent_data" (dataframe) were printed using absent_data.head(), the columns and rows were correctly printed. I used the "describe", "shape", "index" and "columns" attributes to understand the dataset.

Next, was an important step – **to identify and correct null valued rows in the dataset**. This was done using "isnull().sum()" on the absent_data dataframe. Sequentially, all the columns with number of null valued rows were printed. It was noted that there were no null values in the dataset.

We know that machine learning algorithms are basically mathematical operations on the dataset. Hence, **math cannot be done on categorical (string/character) valued attributes.** We must convert them into numerical values. In this dataset, all values were already converted into numerical values by the dataset donator. But *for the sake of learning, 'Day of the week' attribute was changed from numerical to categorical and back to numerical value.* I tried using LabelEncoder() method of sklearn, but was unsuccessful. The encoder assigned numerical values to the days uniquely but randomly. This would change the meaning of the digits representing days in the original dataset.

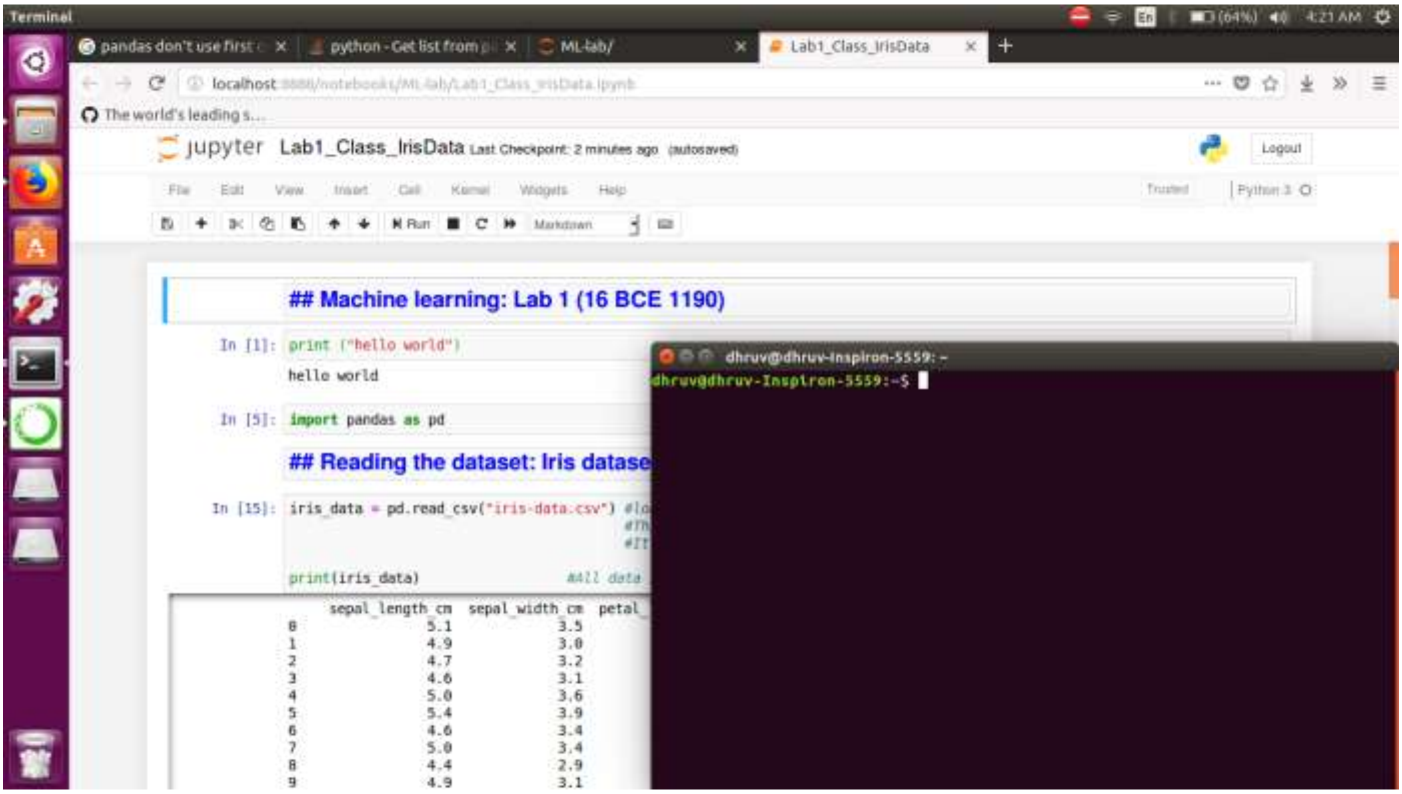
Next, the **data** in the dataset **was scaled using the sklearn's StandardScaler.** Scaling of variables needs to be done so that no attribute gets higher/lower weight just because of the values it stores. Also, using the scaled values, computation of the algorithm is faster and so is its convergence.

Lastly, using the **seaborn** and **matplotlib** libraries, a **correlation plot was plotted between the various attributes of the dataset.** This helps us to know if there is a strong positive/negative correlation between two variables. Although the plot did not reveal any surprising insight about the dataset, it does lead us to a starting point for future analysis on the dataset.

SCREENSHOTS

DHRUV GARG 16 BCE 1190

Iris-dataset


Terminal
pandas don't use first c
python - Get list from p
ML-lab/
Lab1_Class_IrisData
localhost:8888/notebooks/ML-lab/Lab1_Class_IrisData.ipynb
The world's leading s...
jupyter Lab1_Class_IrisData Last Checkpoint: 2 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Markdown
4:21 AM
dhruv@dhruv-Inspiron-5559: ~

Machine learning: Lab 1 (16 BCE 1190)

```
In [1]: print ("hello world")

hello world

In [5]: import pandas as pd
```

Reading the dataset: Iris datase

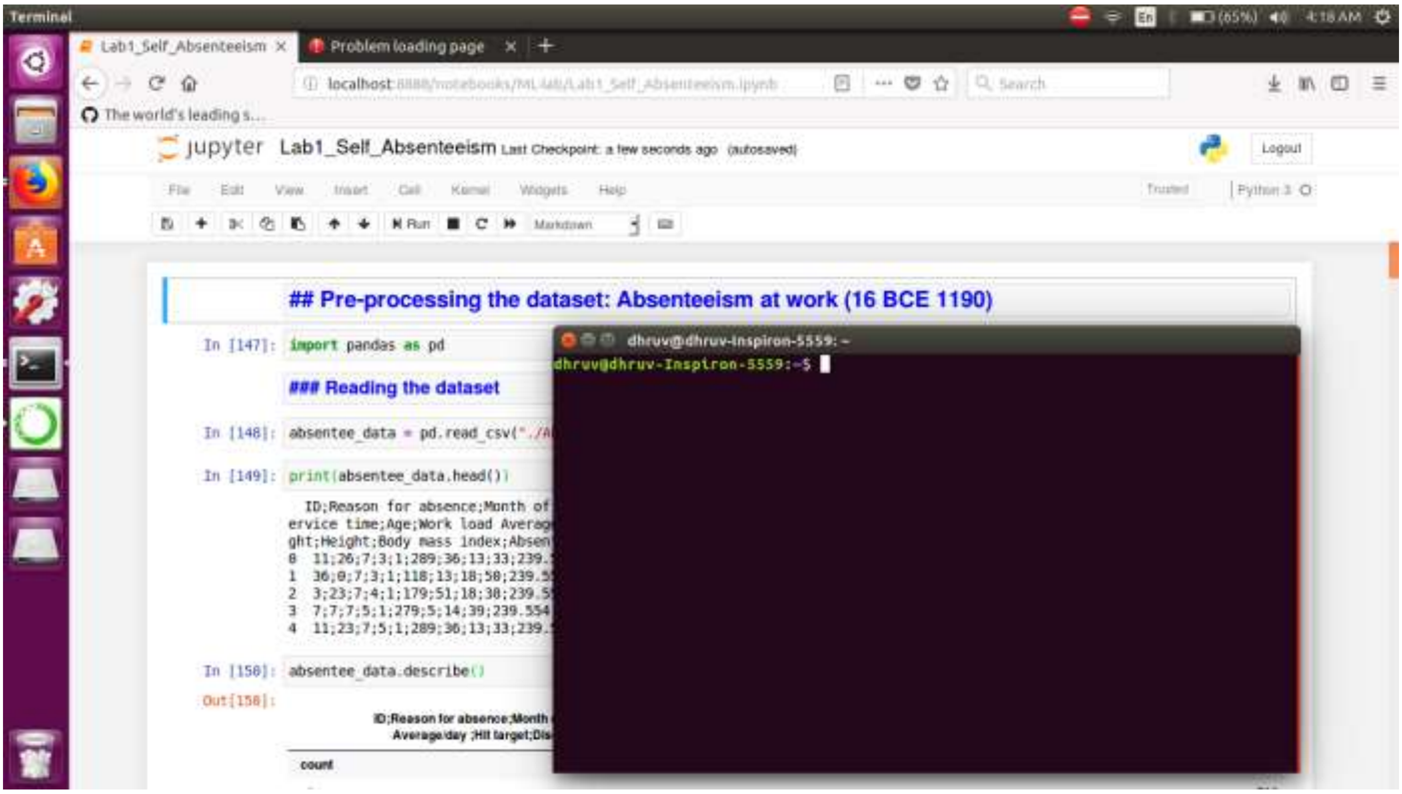
```
In [15]: iris_data = pd.read_csv("iris-data.csv") #lo
                                                  #Th
                                                  #It

         print(iris_data)                   #All data

             sepal_length_cm  sepal_width_cm  petal_
         0              5.1             3.5
         1              4.9             3.0
         2              4.7             3.2
         3              4.6             3.1
         4              5.0             3.6
         5              5.4             3.9
         6              4.6             3.4
         7              5.0             3.4
         8              4.4             2.9
         9              4.9             3.1
```

```
dhruv@dhruv-Inspiron-5559:~$
```

Chosen dataset – Absenteeism at work


Terminal
Lab1_Self_Absenteeism
Problem loading page
localhost:8888/notebooks/ML-lab/Lab1_Self_Absenteeism.ipynb
Search
The world's leading s...
jupyter Lab1_Self_Absenteeism Last Checkpoint: a few seconds ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Markdown
4:18 AM
dhruv@dhruv-Inspiron-5559: ~

Pre-processing the dataset: Absenteeism at work (16 BCE 1190)

```
In [147]: import pandas as pd
```

Reading the dataset

```
In [148]: absentee_data = pd.read_csv("./A

In [149]: print(absentee_data.head())

            ID;Reason for absence;Month of
          ervice time;Age;Work load Averag
          ght;Height;Body mass index;Absen
          0  11;26;7;3;1;289;36;13;33;239.
          1  36;0;7;3;1;118;13;18;50;239.5
          2  3;23;7;4;1;179;51;18;38;239.5
          3  7;7;7;5;1;279;5;14;39;239.554
          4  11;23;7;5;1;289;36;13;33;239.

In [150]: absentee_data.describe()

Out[150]:

                    ID;Reason for absence;Month
                      Average/day ;Hit target;Dis

          count
```

```
dhruv@dhruv-Inspiron-5559:~$
```