|  |  |
| --- | --- |
| Name Of The Student | Vaishnavi G |
| Internship Project Topic | Build a Classification Model for Drug Trials Dataset |
| Name of the Organization | TCS iON |
| Name of the Industry Mentor | Himdweep Walia |
| Name of the Institute | SRM Institute of Science and Technology |

|  |  |  |
| --- | --- | --- |
| Date | Day # | Hours Spent |
| 30/10/2022 | 19 | 5 hours |
| Activities done during the day:  **How to split training and testing data sets in Python?**  Before splitting the data, make sure that the dataset is large enough. Train/Test split works well with large datasets.  **Import the Dataset in a Pandas Dataframe:**   * Pandas DataFrames are a thing of beauty. DataFrames in Python makes the handling of data very user friendly. * You can import large datasets using Pandas and then manipulate them effectively. You can easily import CSV data into a Pandas DataFrame. * Python Pandas module is basically an open-source Python module. It has a wide scope of use in the field of computing, data analysis, statistics, etc. * Pandas module uses the basic functionalities of the NumPy module.   **To import the dataset into a Pandas Dataframe use the following set of lines:**   |  | | --- | | import pandas as pd  housing = pd.read\_csv('path\_to\_dataset') |   This will store the dataset as a DataFrame in the variable ‘housing’.  **To Get the first 5 rows**  After importing a dataset for the first time it is common for data scientists to have a look at the first five rows of the Dataframe. It gives a rough idea of what the data looks like.  **To output the first five rows of the Dataframe, use the following line of code:**   |  | | --- | | housing.head() |   When you run the following line, you will see the output as :  IMG_256  **To Get statistical summary**  To get a statistical summary of your Dataframe you can use the .describe() method provided by pandas.  **The line of code to display the statistical summary is as follows :**   |  | | --- | | housing.describe() |   Running this line of code will give the following output.  IMG_256  **Get a quick description of the data**  To get the quick description of the type of data in the table you can use .info() method provided by Pandas.  **Use the following line of code to get the description :**   |  | | --- | | housing.info() |   The output looks like as shown below :  IMG_256  The output contains a row for each column of the dataset. For each column label you get the count of non null entries and the data-type of the entry.  Knowing the data type of the columns in your dataset allows you to make better judgements when it comes to using the data to train models.  To Get count for each column  You can directly get the count of entries in each column using the .count() method in Pandas.  **You can use this method as shown in the following line of code :**   |  | | --- | | housing.count() |   The output comes out as following:  IMG_256  **Get a Histogram for each column in your dataset**  Pandas allow you to display histograms for each and every column in just one line of code.  **To display histograms use the following line of code :**   |  | | --- | | housing.hist() |   After running the line above, we get the output as :  IMG_256  Data scientists often use histograms to form a better understanding of the data.  Reference:  [https://www.geeksforgeeks.org/how-to-split-a-dataset-into-train-and-test-sets-using-python//](https://www.geeksforgeeks.org/how-to-split-a-dataset-into-train-and-test-sets-using-python/)  <https://builtin.com/data-science/train-test-split/> | | |