|  |  |
| --- | --- |
| 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 |
| 23/10/2022 | 12 | 6 hours |
| Activities done during the day:  Learned about data Cleaning With Pandas and NumPy ( Commands and Usage).  **What is data cleaning ?**   * When working with multiple data sources, there are many chances for data to be incorrect, duplicated, or mislabeled. If data is wrong, outcomes and algorithms are unreliable, even though they may look correct. * Data cleaning is the process of changing or eliminating garbage, incorrect, duplicate, corrupted, or incomplete data in a dataset.   **Data cleaning cycle**   * It is the method of analyzing, distinguishing, and correcting untidy, raw data. Data cleaning involves filling in missing values, distinguish and fix errors present in the dataset. Whereas the techniques used for data cleaning might vary in step with different types of datasets, the following are standard steps to map out data cleaning:   IMG_256  **Data cleaning with Pandas**   * Pandas stand for “Python Data Analysis Library” . * Pandas is the popular Python library that is mainly used for data processing purposes like cleaning, manipulation, and analysis. * It consists of classes to read, process, and write CSV data files. * There are numerous Data cleaning tools present but, the Pandas library provides a really fast and efficient way to manage and explore data. * It does that by providing us with Series and DataFrames, which help us not only to represent data efficiently but also manipulate it in various ways.  |  | | --- | | #importing module  import pandas as pd |  * To import the dataset we use the **read\_csv()** function of pandas and store it in the DataFrame named as data. * As the dataset is in tabular format, when working with tabular data in Pandas it will be automatically converted in a **DataFrame**. * DataFrame is a two-dimensional, mutable data structure in Python. It is a combination of rows and columns like an excel sheet.  |  | | --- | | #importing the dataset by reading the csv file  data = pd.read\_csv('Iris.csv')  #displaying the first five rows of dataset  print(data.head()) |  * The head() function is a built-in function in pandas for the dataframe used to display the rows of the dataset. * We can specify the number of rows by giving the number within the parenthesis. By default, it displays the first five rows of the dataset. * If we want to see the last five rows of the dataset we use the tail()function of the dataframe like this:  |  | | --- | | #displayinf last five rows of dataset  data.tail() |   **Rebuild Missing Data**  To find and fill the missing data in the dataset we will use another function. There are 4 ways to find the null values if present in the dataset.   1. Using isnull() function:   This function provides the boolean value for the complete dataset to know if any null value is present or not.   |  | | --- | | data.isnull() |  1. Using isna() function:  |  | | --- | | data.isna() |   This is the same as the isnull() function. Ans provides the same output.   1. Using isna().any()  |  | | --- | | data.isna().any() |   This function also gives a boolean value if any null value is present or not, but it gives results column-wise, not in tabular format.   1. Using isna(). sum()  |  | | --- | | data.isna().sum() |   This function gives the sum of the null values preset in the dataset column-wise.   1. Using isna().any().sum()  |  | | --- | | data.isna().any().sum() |   This function gives output in a single value if any null is present or not.  If there are any null value s preset we can fill those places with any other value using the fillna() function of DataFrame.Following is the syntax of fillna() function:   |  | | --- | | DataFrame\_name.fillna(value=None, method=None, axis=None, inplace=False, limit=None, downcast=None) |   This function will fill NA/NaN or 0 values in place of null spaces.  **Standardization and Normalization**   * Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. * Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.   **De-Duplicate**   * De-Duplicate means remove all duplicate values. There is no need for duplicate values in data analysis. These values only affect the accuracy and efficiency of the analysis result. * To find duplicate values in the dataset we will use a simple dataframe function i.e. duplicated().  |  | | --- | | data.duplicated() |   **Export Dataset**  This is the last step of the data cleaning process. After performing all the above operations, the data is transformed into clean the dataset and it is ready to export for the next process in Data Science or Data Analysis.  Reference:  #1 - <https://realpython.com/python-data-cleaning-numpy-pandas/>  #2 - <https://www.analyticsvidhya.com/blog/2021/06/data-cleaning-using-pandas/>  #3 - <https://colab.research.google.com/notebooks/mlcc/intro_to_pandas.ipynb> | | |