|  |  |
| --- | --- |
| 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 |
| 21/10/2022 | 10 | 5hours |
| Activities done during the day:  Learned about data pre-processing steps for Machine Learning & Data analytics.  **Why data processing is needed ?**   * Data in the real world is quite dirty and corrupted with inconsistencies, noise, incomplete information, and missing values. * It is aggregated from diversified sources using data mining and warehousing techniques. * The majority of the real-world datasets for machine learning are highly susceptible to be **missing**, **inconsistent**, and **noisy** due to their heterogeneous origin. * Applying data mining algorithms on this noisy data **would not give quality results** as they would fail to identify patterns effectively. Data Processing is, therefore, important to improve the overall data quality. * Quality decisions must be based on quality data. Data Preprocessing is important to get this quality data, without which it would just be a **Garbage In, Garbage Out scenario**.   **How data processing can be done ?**   * Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning. * Steps in data processing  1. Data quality assessment 2. Data cleaning 3. Data transformation 4. Data reduction   Learned in detail about the 4 data processing steps in TCS Ion portal video and in the below reference link.  IMG_256    Reference: <https://monkeylearn.com/blog/data-preprocessing/>  <https://serokell.io/blog/data-preprocessing>  Downloaded a sample data on medical transcriptions and studied the format of data : <https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions> | | |