|  |  |
| --- | --- |
| 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 |
| 24/10/2022 | 13 | 5 hours |
| Activities done during the day:  Learned about data transformation process in ML and data analysis.  **What is data transformation ?**   * Data transformation is used when data needs to be converted to match that of the destination system. This can occur at two places of the data pipeline. First, organizations with on-site data storage use an extract, transform, load, with the data transformation taking place during the middle ‘transform’ step. * Organizations today mostly use cloud-based data warehouses because they can scale their computing and storage resources in seconds. Cloud based organizations, with this huge scalability available, can skip the ETL process. Instead, they use a transformation process that converts the data as the raw data is uploaded, a process called extract, load, and transform. The process of data transformation can be handled manually, automated or a combination of both. * Transformation is an essential step in many processes, such as data integration, migration, warehousing and wrangling. The process of data transformation can be:  1. Constructive, where data is added, copied or replicated 2. Destructive, where records and fields are deleted 3. Aesthetic, where certain values are standardized, or 4. Structural, which includes columns being renamed, moved and combined   **Stages in Data Transformation:**  1. Discovery  The first step is to identify and understand data in its original source format with the help of data profiling tools. Finding all the sources and data types that need to be transformed. This step helps in understanding how the data needs to be transformed to fit into the desired format.  2. Mapping  The transformation is planned during the data mapping phase. This includes determining the current structure, and the consequent transformation that is required, then mapping the data to understand at a basic level, the way individual fields would be modified, joined or aggregated.  3. Code Generation  The code, which is required to run the transformation process, is created in this step using a data transformation platform or tool.  4. Execution  The data is finally converted into the selected format with the help of the code. The data is extracted from the source(s), which can vary from structured to streaming, telemetry to log files. Next, transformations are carried out on data, such as aggregation, format conversion or merging, as planned in the mapping stage. The transformed data is then sent to the destination system which could be a dataset or a data warehouse.  Reference:  #1 - <https://www.tibco.com/reference-center/what-is-data-transformation>  #2 - <https://towardsdatascience.com/data-transformation-and-feature-engineering-e3c7dfbb4899> | | |