|  |  |
| --- | --- |
| 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 |
| 26/10/2022 | 15 | 5 hours |
| Activities done during the day:  Data reduction in data analysis.  **What is data reduction ?**  Data reduction is a method of reducing the size of original data so that it may be represented in a much smaller space. While reducing data, data reduction techniques preserve data integrity  **Why data reduction is needed ?**  When data analysis and mining is done on a huge amount of data then it takes a very long time to process, which makes it impractical and infeasible. It can reduce the processing time for data analysis, data reduction techniques are used to obtain a reduced representation of the dataset that is much smaller in volume by maintaining the integrity of the original data. By reducing the data, the efficiency of the data mining process is improved which produces the same analytical results.  There are various strategies for data reduction which are as follows −   1. **Data cube aggregation** − In this method, where aggregation operations are used to the data in the construction of a data cube. These data include the All Electronics sales per quarter, for the years 2002 to 2004. It is interested in the annual sales (total per year), rather than the total per quarter. Thus the data can be aggregated so that the resulting data summarize the total sales per year instead of per quarter. The resulting data set is smaller in volume, without loss of data essential for the analysis task.   IMG_256   1. **Attribute subset selection** − In this method, where irrelevant, weakly relevant, or redundant attributes or dimensions can be discovered and deleted. Data sets for analysis can include hundreds of attributes, some of which can be irrelevant to the mining task or redundant. For instance, if the task is to arrange customers as to whether or not they are likely to purchase a popular new CD at All Electronics when notified of a sale, attributes such as the customer’s telephone number are likely to be irrelevant, unlike attributes such as age or music\_taste. 2. **Dimensionality reduction** − Encoding mechanisms are used to reduce the data set size. In dimensionality reduction, data encoding or transformations are applied to obtain a reduced or “compressed” representation of the original data. If the original data can be reconstructed from the compressed data without any loss of information, the data reduction is called lossless.   IMG_256   1. **Numerosity reduction** − The data are restored or predicted by alternative, smaller data representations including parametric models (which are required to save only the model parameters rather than the actual data) or nonparametric methods including clustering, sampling, and the use of histograms. 2. **Discretization and concept hierarchy generation** − In this method, where raw data values for attributes are replaced by ranges or higher conceptual levels. Data discretization is a form of numerosity reduction that is very beneficial for the automatic production of concept hierarchies. Discretization and concept hierarchy generation are dynamic tools for data mining, in that they enable the mining of data at various levels of abstraction. | | |