|  |  |
| --- | --- |
| 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 |
| 27/10/2022 | 16 | 5 hours |
| Activities done during the day:  Dimensionality reduction in Python.   1. What is Dimensionality Reduction?  * We are generating a tremendous amount of data daily. * In fact, 90% of the data in the world has been generated in the last 3-4 years! The numbers are truly mind boggling. * Data generation and collection keeps increasing, visualizing it and drawing inferences becomes more and more challenging. * One of the most common ways of doing visualization is through charts. * Suppose we have 2 variables, Age and Height. We can use a scatter or line plot between Age and Height and visualize their relationship easily:   IMG_256  2. Why is Dimensionality Reduction required?  Here are some of the benefits of applying dimensionality reduction to a dataset:   * Space required to store the data is reduced as the number of dimensions comes down. * Less dimensions lead to less computation/training time. * Some algorithms do not perform well when we have a large dimensions. So reducing these dimensions needs to happen for the algorithm to be useful. * It takes care of multicollinearity by removing redundant features. For example, you have two variables – ‘time spent on treadmill in minutes’ and ‘calories burnt’. * These variables are highly correlated as the more time you spend running on a treadmill, the more calories you will burn. Hence, there is no point in storing both as just one of them does what you require. * It helps in visualizing data. As discussed earlier, it is very difficult to visualize data in higher dimensions so reducing our space to 2D or 3D may allow us to plot and observe patterns more clearly.  1. Common Dimensionality Reduction Techniques   Dimensionality reduction can be done in two different ways:   * By only keeping the most relevant variables from the original dataset (this technique is called feature selection) * By finding a smaller set of new variables, each being a combination of the input variables, containing basically the same information as the input variables (this technique is called dimensionality reduction)   Way for dimensionality reduction :  1. Missing Value Ratio  2. Low Variance Filter  3. High Correlation filter  4. Random Forest  5. Backward Feature Elimination  6.Forward Feature Selection  7. Factor Analysis  8. Principal Component Analysis (PCA)  9. Independent Component Analysis  10. Methods Based on Projections  11. t- Distributed Stochastic Neighbor Embedding (t-SNE)  12. UMAP | | |