|  |  |
| --- | --- |
| 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 |
| 29/10/2022 | 18 | 5 hours |
| Activities done during the day:  **What is Training Dataset?**  The training data is the biggest (in -size) subset of the original dataset, which is used to train or fit the machine learning model.  **What is Test Dataset?**   * Once we train the model with the training dataset, it's time to test the model with the test dataset. * This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset. * The test dataset is another subset of original data, which is independent of the training dataset.   **The testing data should:**   * Represent or part of the original dataset. * It should be large enough to give meaningful predictions.   **Need of Splitting dataset into Train and Test set**  Splitting the dataset into train and test sets is one of the important parts of data pre-processing, as by doing so, we can improve the performance of our model and hence give better predictability.  IMG_256  Therefore, if we train and test the model with two different datasets, then it will decrease the performance of the model. Hence it is important to split a dataset into two parts, i.e., train and test set.  For splitting the dataset, we can use the train\_test\_split function of scikit-learn.  The bellow line of code can be used to split dataset:   |  | | --- | | from sklearn.model\_selection import train\_test\_split  x\_train, x\_test, y\_train, y\_test= train\_test\_split(x, y, test\_size= 0.2, random\_state=0) |  * we have imported the train\_test\_split function from the sklearn library. * we have used four variables, which are * x\_train: It is used to represent features for the training data * x\_test: It is used to represent features for testing data * y\_train: It is used to represent dependent variables for training data * y\_test: It is used to represent independent variable for testing data * In the train\_test\_split() function, we have passed four parameters. Which first two are for arrays of data, and test\_size is for specifying the size of the test set. The test\_size may be .5, .3, or .2, which tells the dividing ratio of training and testing sets. * The last parameter, random\_state, is used to set a seed for a random generator so that you always get the same result, and the most used value for this is 42.   **Overfitting and Underfitting issues**   * Overfitting and underfitting are the most common problems that occur in the Machine Learning model. * A model can be said as overfitted when it performs quite well with the training dataset but does not generalize well with the new or unseen dataset. * The issue of overfitting occurs when the model tries to cover all the data points and hence starts caching noises present in the data. * Due to this, it can't generalize well to the new dataset. Because of these issues, the accuracy and efficiency of the model degrade. Generally, the complex model has a high chance of overfitting. * There are various ways by which we can avoid overfitting in the model, such as Using the Cross-Validation method, early stopping the training, or by regularization, etc. * On the other hand, the model is said to be under-fitted when it is not able to capture the underlying trend of the data. It means the model shows poor performance even with the training dataset. * In most cases, underfitting issues occur when the model is not perfectly suitable for the problem that we are trying to solve. * To avoid the overfitting issue, we can either increase the training time of the model or increase the number of features in the dataset.   **How do training and testing data work in Machine Learning?**  Machine Learning algorithms enable the machines to make predictions and solve problems on the basis of past observations or experiences. These experiences or observations an algorithm can take from the training data, which is fed to it. Further, one of the great things about ML algorithms is that they can learn and improve over time on their own, as they are trained with the relevant training data.  Once the model is trained enough with the relevant training data, it is tested with the test data.  We can understand the whole process of training and testing in three steps, which are as follows:   * **Feed:** Firstly, we need to train the model by feeding it with training input data. * **Define:** Now, training data is tagged with the corresponding outputs (in Supervised Learning), and the model transforms the training data into text vectors or a number of data features. * **Test:** In the last step, we test the model by feeding it with the test data/unseen dataset. This step ensures that the model is trained efficiently and can generalize well.   The above process is explained using a flowchart given below:  IMG_256  Reference: <https://www.w3schools.com/python/python_ml_train_test.asp/>  <https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_training_test_data.htm/> | | |