|  |  |
| --- | --- |
| 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 |
| 24/11/2022 | 44 | 5 hours |
| Activities done during the day:  **Project Hands on - Train your model with training data and test the results using testing set**   |  | | --- | | # splitting data into train and validation  x\_train, x\_test, y\_train, y\_test = train\_test\_split(df, target, test\_size=.3) |   Once the data is divided into attributes and labels, the final preprocessing step is to divide data into training and test sets. Luckily, the model\_selection library of the Scikit-Learn library contains the train\_test\_split method that allows us to seamlessly divide data into training and test sets.  Splitting data into train, test, and validation sets is a repetitive task.   |  | | --- | | from sklearn.model\_selection import train\_test\_split |   **Training Set**   * It is the set of data that is used to train and make the model learn the hidden features/patterns in the data. * In each epoch, the same training data is fed to the neural network architecture repeatedly, and the model continues to learn the features of the data. * The training set should have a diversified set of inputs so that the model is trained in all scenarios and can predict any unseen data sample that may appear in the future.   **The Validation Set**   * The validation set is a set of data, separate from the training set, that is used to validate our model performance during training. * This validation process gives information that helps us tune the model’s hyperparameters and configurations accordingly. It is like a critic telling us whether the training is moving in the right direction or not. * The model is trained on the training set, and, simultaneously, the model evaluation is performed on the validation set after every epoch. * The main idea of splitting the dataset into a validation set is to prevent our model from overfitting i.e., the model becomes really good at classifying the samples in the training set but cannot generalize and make accurate classifications on the data it has not seen before.   **The Test Set**   * The test set is a separate set of data used to test the model after completing the training. * It provides an unbiased final model performance metric in terms of accuracy, precision, etc.   **Data splitting**   * Data splitting is when data is divided into two or more subsets. Typically, with a two-part split, one part is used to evaluate or test the data and the other to train the model. * Data splitting is an important aspect of data science, particularly for creating models based on data. This technique helps ensure the creation of data models and processes that use data models -- such as machine learning -- are accurate.   Split the dataset by using the function train\_test\_split(). you need to pass 3 parameters features, target, and test\_set size. Additionally, you can use random\_state to select records randomly.  **Code:**   |  | | --- | |  | | | |