Training Day-66 Report:

Data Transformation in Machine Learning:-

Data transformation is the most important step in a machine learning pipeline which includes modifying the raw data and converting it into a better format sothat it can be more suitable for analysis and model training purposes. In data transformation, we usually deal with issues such as noise, missing values, outliers, and non-normality.

Different Data Transformation Technique:-

Data transformation in machine learning involves a lot of techniques, let's discuss8 of the major techniques that we can apply to data to better fit our model and produce better results in the prediction process.
The choice of data transformation technique depends on the characteristics of thedata and the machine learning algorithm that we intend to use on the data. Here are the mentioned techniques discussed in details.

Handling Missing Data:-

Most of the times the data that is received from different sources miss some of the values in it, if we train our model on this data the model might behave differently or even produce error while training. Therefore, handling missing databecomes an important aspect to consider while transforming the data, there are different techniques through which we can handle the missing data which can help us improve our model performance. Let's discuss some of the techniques in details here:

- **Removing the Missing Data**: We can delete the rows or columns whichare having missing data. This is significant only when a small number
  of data is missing, if there's a large value of missing data points in our dataset we must consider some other technique otherwise deleting the rows or columns with large number of missing values will change the way our model performs since it might cause the model to train on less data. We can drop the rows which contain the missing values using the dropna method in pandas if the data we have is stored in a pandas dataframe.

- **Imputation:** In this technique we remove the missing values by filling the missing values positions with some other value, for example we canfill in the missing values with the mean of the different values from the same column which is in the same category or data type as of the

missing value. The most common types of imputation methods includemean, median, mode imputation. We can also fill in the missing valueswith a constant value that we want to be present in the data instead of the missing value. Imputation is also implanted within the sklearn library, we can impute different missing values with the help of KNNImputer(K-Nearest Neighbors) which is a part of sklearn.impute.

- **Forward Fill or Backward Fill**: Usually in time series analysis, wherethe data is produced after a constant time, if some data goes missing wecan replace the missing value with forward fill or the backward fill options, The forward fill method fills the missing value with the previous non missing value whereas backward fill method fills the missing value with the next non missing value to the missing value.

- **Interpolation**: Missing data can also be handled by interpolation technique, it involves predicting the missing values based on observed values in the dataset. There are multiple interpolation methods, the choice of the method is based on the data that we have. Most commonlyused interpolation is linear interpolation which assumes there is a linear relationship between observed values and missing data points, this
method predicts the missing value by fitting a straight line between twoadjacent non-missing points.