Pre-Processing for Machine Learning Projects:

Data Pre-processing:

Data Pre-processing is a process of converting the raw data into cleaned text. In general pre-processing refers to the transformations applied to the data before feeding into the algorithm. It is merely transforming raw data into an understandable format. It is strongly recommended to understand the Data Description where we can explore the definitions and contents of all the variables before working on ML Algorithm. We have to remember that the quality of input decides the quality of the output. Approximately data pre-processing takes 70% of the total project time which indicates how much pre-processing stage is essential for any ML project. The output from the pre-processing is the final data set.

The methods for data pre-processing are organized into different categories.

Data Integration:

Data Integration is to carefully merge data from multiple sources which helps to reduce and avoid redundancies in the data.

Data Cleaning:

Data Cleaning or Data cleansing is to clean the data by imputing missing values, smoothing noisy data, and identifying or removing outliers. In general, the missing values are found due to collection error or data is corrupted.

The missing data can be filled by using various techniques.

Removing the entire rows which contain missing data or replacing the data with any corresponding value.

Mean / Mode / Median imputation is one of the most frequently used methods. Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable.

We can also use any of the predictive model technique to impute values and choosing the model is purely based on the business goal. We can use KNN imputation which replaces the missing values with most similar to the respective missing values by using distance as a measurement. KNN Imputation is used most often as it is based on the Neighbourhood criteria.

Data Reduction:

Data Reduction strategies are used to reduce the dataset regarding volume without losing its originality.

Dimensionality reduction will reduce the no of attribute or variables whereas numerosity reduction will reduce the data volume by smaller forms of data representations.

Principle component analysis is one of the widely used dimensionality reduction technique which reduces the number of attributes in the data by projecting the data from its original high-dimensional space into a lower-dimensional space.

Data Transformation:

Data Transformation is the process of converting data or information from one format to another which helps to find the patterns or to improve the efficiency of the model or accuracy of the model.

Real world data is collected from different sources in different ways and given data might is on different scales.

Normalization is one of the data transformation techniques where attribute data is scaled to fall within a small range and make all the attributes equal. Column normalization or feature normalization used to transform or compress the data in the range between 0 and 1.

Column standardization or feature standardization used a lot in practice which transforms or compress the data such that their mean is 0 and the standard deviation is 1. Here we are compressing or expanding the data points in the hypercube to make our standard deviation for any feature is 1.

One hot encoding is the technique which is used to convert categorical variables to numerical variables.

Data Aggregation:

Data aggregation is a pre-processing technique to aggregate the data required for the analysis.

For example, the sales data is presented for each day, and we would like to analyse by each month. Here we combine each day are grouped by month to find the sales trend. It is used to improve the stability of the data and to understand the granularity.

Data Discretization:

Data Discretization is one of the other pre-processing techniques which reduces the no. of continuous variable values by grouping them into some bins or intervals.

In simple words, putting the values into buckets by making them discrete.

For example, a set of values like all the different set of students age, are organized in different buckets like young-age, mature-age and old-age.