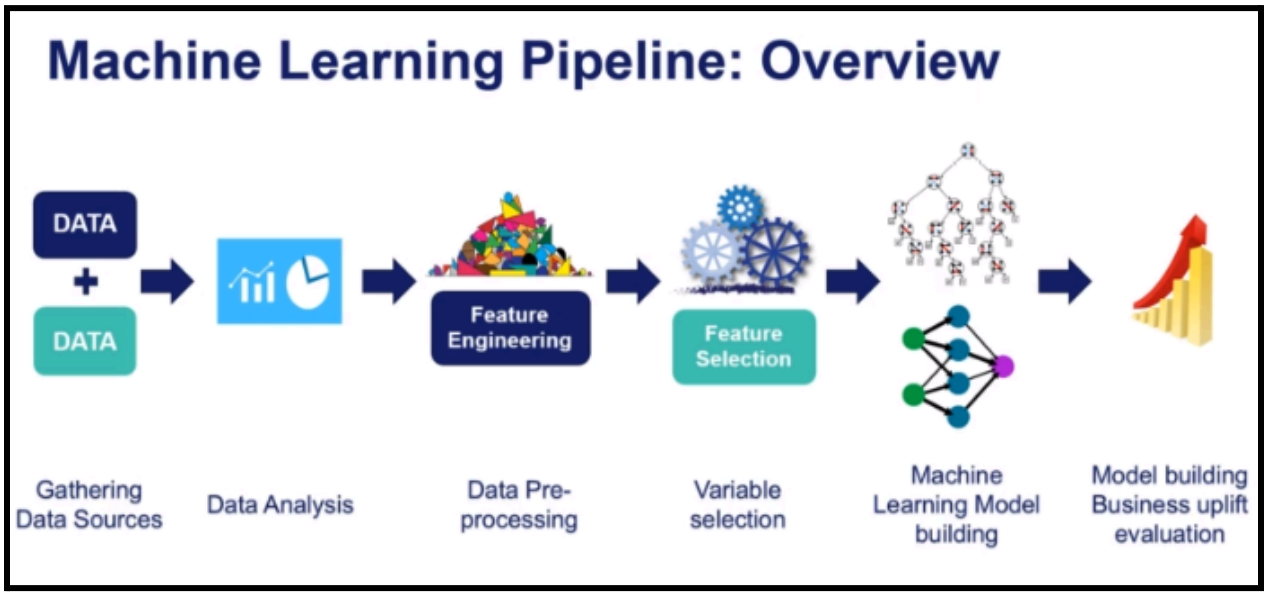
End to end Data Preprocessing Approach



Data Gathering : Web Scrapping, Surveying (Manually),Google form.
Data Analysis : Performing EDA to understand and visualize the relationships between the features (independent and dependent).
Data Cleaning : It is the process of fixing or removing incorrect,corrupted,incorrectly formatted,duplicate or incomplete data within a dataset.
Data Binning : Process of assigning data into right datatype by changing or creating new features.

Feature Engineering:
Feature Engineering is the process of creating new features or transforming existing features to improve the performance of a machine-learning model. It involves selecting relevant information from raw data and transforming it into a format that can be easily understood by a model. The goal is to improve model accuracy by providing more meaningful and relevant information.

Methods to handle Missing Value:

1.Ignore the missing value(delete row)
2.Fill the missing value manually.
3.Global Constant
4.using measure of Central Tendency (Mean,Median,Mode)
5.Measure of Central Tendency for each class.
6.Forward fill
7.Backward fill
8.Most Probable Value.(using ML Algoritms) - Time Consuming.

1.MCAR -Missing Completely at Random.
2.MAR- Missing At Random.
3.NMAR- Not Missing At Random
4.SM-Structure Missing(Reason of value missing is known).

Imputation: Univariate vs Multivarite.

Techniques of Imputation:(Single Imputation)
1.Listwise deletion
2.Mean/Median/Mode
3.Deck Imputation
    3.1:Cold Imputation
    3.2:Hot Imputaion

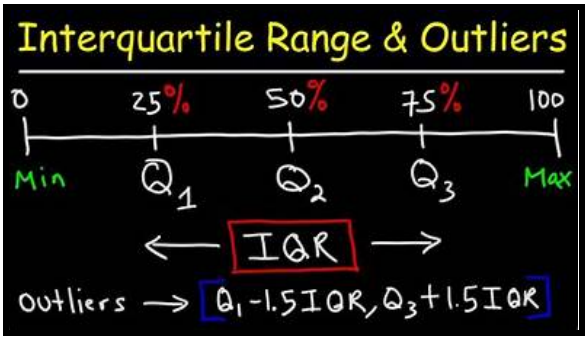
4.Model Based Imputation
    4.1:KNN
    4.2:Expectation Maximization
    4.3:Maximum Likelihood
    4.4:Regression.

5.Prior Knowledge.

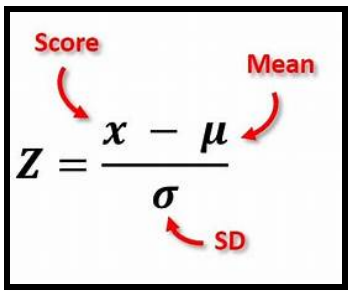
Outlier Detection and Removing:

1. Using Boxplot or Distribution Plot(use seaborn library)
2. IQR(inter -quartile range)

    In descriptive statistics, the interquartile range (IQR) is a measure of statistical dispersion, which is the spread of the data. The IQR may also be called the midspread, middle 50%, fourth spread, or H-spread. It is defined as the difference between the 75th and 25th percentiles of the data.



3. Z -score



Feature Encoding:
Encoding is the technique of transforming categorical value into numerical.

1. One hot Encoding(used for nominal data)(Creates a sparse matrix)
2. Label Encoding(used for nominal data)(like mapping any value)
3. Ordinal Encoding (used for ordinal data like - Rank)

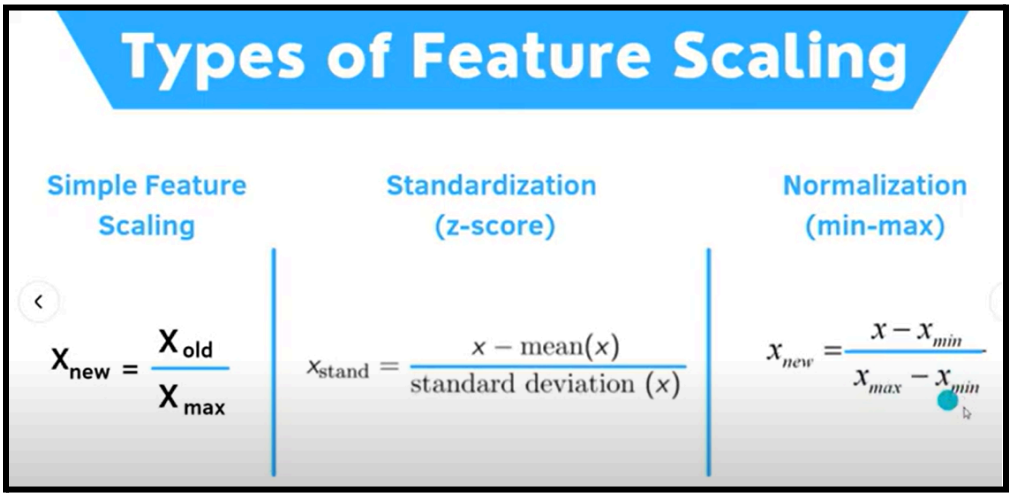
Nominal data : Different Categories have no relation or inter dependencies between them, like Gender,Married/Unmarried.

Ordinal data : Different Categories having relation or inter dependencies between them, like 1st ,2nd 3rd student of a class, Categories different Qualifications.

Feature Scaling and Transformation:
Feature scaling is a crucial step in the feature transformation process that ensures all features are on a similar scale. It is the process that normalizes the range of input columns and makes it useful for further visualization and machine learning model training.

There are three different techniques of Feature Scaling:
1.Simple Feature Scaling
2.Stadardization(z-score)(used in all cases)
3.Normalization(Min-Max)

Here the transform method will scale test data with same scaler which we have obtained during fit method on training data.So that our data remain constant.



When to use which Tranformation:

It is very difficult to say when to use use which transformation.Its depends on the type of probelm .

1. For Distance-Based Algorithm like KNN,Clustering,SVM data -scaling is very important,
2. On the other hand non-distance-based algorithm s like Naive-Bayes or Tree based Algorithms Scaling is not so important.
3. Nomalization scaled data between 0 to 1. Standardization makes data with mean_value =0 and standard deviation = 1.
4. If datasets Features values range have bigger difference among them Normalization could be use.
5. If dataset contains a lots of outliers Standardization could be use.
6. Overall Standardization do a better work on the dataset.

Handling the duplicate values:
Just drop the duplicate rows of the dataset.