Feature Engineering

Raw Data == Bad Data ⟶ Feature Engineering ⟶ Raw Data == Good Data

1. Handling Missing Values

a. MCAR - Missing Completely At Random

b. MAR - Missing At Random (Pattern)

c. MNAR - Missing Not at Random (Missing for a reason)

⟶ Elimination - Loss of Data

⟶ Imputation - Mean, Median, Mode

2. Handling Imbalanced Dataset
    a. UnderSampling → Not Preferrable
    b. OverSampling → Same data is added again and again
    c. SMOTE → Synthetic Minority OverSampling Technique
        i. Within the data it will add the data
        ii. Instead of adding duplicate data into the dataset like Oversampling this will add new data.

3. Outlier Detection & Removal
    a. Using boxplot we are able to find out whether the data has Outliers
    b. Fiver Number Summary

4. Encoding Categorical Values

    a. Onehot Encoding → Using one column it will make more column. Ex: Gender, G_female, G_male.

    b. Label Encoding → Using one column it will make the different data. Ex: Gender, Gender_label: female-0, male-1

    c. Ordinal Encoding → Using the column and giving priority the data is aligned in the column

5. Feature Scaling

    a. To make every data in the proper scale makes the model performance good and it is required step to do in the data preprocessing.

    b. To make the data in one scale using Normalization and Standard Deviation.