**1. EDA [Conversion of raw data into useful data]**

a. Identify numerical and categorical features

b. Identify missing values and visualize them

c. Identify outliers (boxplot)

d. Cleaning the raw data as required

**2. Handling the missing values**

a. Mean, median, mode etc.

b. Random Sample Imputation

c. Capturing NAN values with a new feature

d. End of Distribution imputation

e. Arbitrary Value Imputation

*Categorical Missing Values*

a. Frequent Category Imputation + Adding a variable to capture NaN

b. Suppose if you have more frequent categories, we just replace NaN with a new category (e.g.: replace NaN with 'Missing')

**3. Handling imbalanced dataset**

a. Under Sampling

b. Over Sampling

**4. Treating the outliers**

**5. Scaling down the data and transformation**

a. Standardization, Normalization

b. Scaling to Minimum And Maximum values

c. Scaling To Median And Quantiles

d. Guassian Transformation, Logarithmic Transformation, Reciprocal Transformation, Square Root Transformation, Exponential Transformation, Box Cox Transformation

**6. Converting categorical features into numerical features**

**7. Feature Selection:**

a. Correlation

b. K Neighbours

c. Chisquare

d. Genetic Algorithm

e. Feature importance (Eg: Extra tree Classifier)