ML ASSIGNMENT 4

Q1) data validation, data insertion, data preprocessing, feature extraction, training model, testing model, deployment.

Q2) onetime data, streaming and realtime data (sensors), batch wise data, various formats(csv, excel, json etc), string data, numerical data, categorical data etc.

Q3) numeric means numbers (integers, real numbers etc), category means categories and need not be numbers (like A,B,C, etc categories). Feature selection means extracting features while dimensionality reduction means reducing dimension of features.

Q4) histogram gives probability of outcome in a given subcategory while subcategories are arranged in order and are of equal intervals. Scatter plot directly plots points using the coordinates.

Q5) investigation of data is essential to understand use case, to remove unnecessary data, to preprocess data, to choose right models and parameters and improve accuracy and understanding of model etc. quantitative data can be explored to understand its attributes like mean, variance etc while qualitative data is explored in terms of it relevance in usage.

Q6) various histogram shapes - gaussian, skewed, double peaked etc.

Entire dataset (features) is divided into bins of equal length starting from minimum point to maximum point arranged in order on x axis.. The probability of occurrence in these bins in reflected on y axis.

Q7) remove data with outliers, replace outliers with mean or mode etc.

Q8) mean, median and mode. When data is skewed, mean varies significantly from mode.

Q9) scatter plots one variable on x axis and another variable on y axis which allows to examine bivariate relationships. It is possible by looking for huge deviations in plot away from general trend.

Q10) categories of one variable are represented in rows while categories of other variable are represented in columns in cross tabs.