KDD Process:

1. Data Cleaning – check typos, check missing value (dropna)
2. Data Integration – tight coupling(Extraction, Transformation..)
3. Data Selection – select relevant data
4. Data Transformation – to stabilize variance, ex. Normalization
5. Data Mining – method to extract data patterns
6. Pattern Evaluation – identify interesting patterns.
7. Knowledge presentation – visualization and knowledge representation

Data Reduction: ex. Dimensionally Reduction(PCA)

Principle Component Analysis:

Create one or more index variables from a larger set of measured variables. It does this using a linear combination (basically a weighted average) of a set of variables. The created index variables are called components.

This model can be set up as a simple equation: $C = w_1(Y_1) + w_2(Y_2) + w_3(Y_3) + w_4(Y_4)$

Factor Analysis: is based on a formal model predicting observed variables from theoretical latent (hidden) factors.

Linear Regression: is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).

Logistic regression: is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Comparison:

In linear regression, the outcome (dependent variable) is continuous. It can have any one of an infinite number of possible values.

In logistic regression, the outcome (dependent variable) has only a limited number of possible values.

Why applying PCA: Low-variance data often, but not always, has little predictive power, so removing low-variance dimensions of your dataset can be an effective way of improving predictor running time.

Random Forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction of the individual trees.

AdaBoost: It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers.

KNN: Predictions are made for a new data point by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances. For regression problems, this might be the mean output variable, for classification problems this might be the mode (or most common) class value.

Evaluating Model Accuracy:

The goal of the ML model is to learn patterns that generalize well for unseen data instead of just memorizing the data that it was shown during training. Once you have a model, it is important to check if your model is performing well on unseen examples that you have not used for training the model. To do this, you use the model to predict the answer on the evaluation dataset (held out data) and then compare the predicted target to the actual answer