A machine learning project may not be linear, but it has a number of well known steps:

1. Define Problem.
2. Prepare Data.
3. Evaluate Algorithms.
4. Improve Results.
5. Present Results.

CLASSIFICATION

Classification, a sub-category of supervised learning, is defined as the process of separating data into distinct categories or classes. These models are built by providing a labeled dataset and making the algorithm learn so that it can predict the class when new data is provided. The most popular classification algorithms are Decision Tree, SVM. We will study these algorithms in the coming tutorials.

REGRESSION

While classification deals with predicting discrete classes, regression is used in predicting continuous numerical valued classes. Regression is also falls under supervised learning generally used to answer “How much?” or “How many?”. Regressions create relationships and correlations between different types of data. Linear Regression is the most common regression algorithm. Regression is the method which measures the average relationship between two or more continuous variables in term of the response variable and feature variables. In other words, regression analysis is to know the nature of the relationship between two or more variables to use it for predicting the most likely value of dependent variables for a given value of independent variables. Linear regression is a mostly used regression algorithm.

Clustering

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

OVERFITTING

When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our dataset. The cause of overfitting is non-parametric and non-linear methods. We use cross-validation to reduce overfitting which allows you to tune hyperparameters with only your original training set. This allows you to keep your test set as truly unseen dataset for selecting your final model