Machine learning algorithms

1. **Linear Regression:** To understand the working functionality of Linear Regression, imagine how you would arrange random logs of wood in increasing order of their weight. There is a catch; however – you cannot weigh each log. You have to guess its weight just by looking at the height and girth of the log (visual analysis) and arranging them using a combination of these visible parameters. This is what linear regression in machine learning is like. In this process, a relationship is established between independent and dependent variables by fitting them to a line. This line is known as the regression line and is represented by a linear equation Y= a *X + b.
2. **Logistic Regression:** Logistic Regression is used to estimate discrete values (usually binary values like 0/1) from a set of independent variables. It helps predict the probability of an event by fitting data to a logit function. It is also called logit regression. These methods listed below are often used to help improve logistic regression models:
   - include interaction terms
   - eliminate features
   - regularize techniques
   - use a non-linear model
3. **Decision Tree:** Decision Tree algorithm in machine learning is one of the most popular algorithm in use today; this is a supervised learning algorithm that is used for classifying problems. It works well in classifying both categorical and continuous dependent variables. This algorithm divides the population into two or more homogeneous sets based on the most significant attributes/ independent variables.
4. **SVM (Support Vector Machine) Algorithm:** SVM algorithm is a method of a classification algorithm in which you plot raw data as points in an n-dimensional space (where n is the number of features you have). The value of each feature is then tied to a particular coordinate, making it easy to classify the data. Lines called classifiers can be used to split the data and plot them on a graph.
5. **Naïve Bayes Algorithm:** A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Even if these features are related to each other, a Naive Bayes classifier would consider all of these properties independently when calculating the probability of a particular outcome. A Naive Bayesian model is easy to build and useful for massive datasets. It's simple and is known to outperform even highly sophisticated classification methods.

6. **KNN (K- Nearest Neighbors) Algorithm:** This algorithm can be applied to both classification and regression problems. Apparently, within the Data Science industry, it's more widely used to solve classification problems. It's a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbors. The case is then assigned to the class with which it has the most in common. A distance function performs this measurement. KNN can be easily understood by comparing it to real life. For example, if you want information about a person, it makes sense to talk to his or her friends and colleagues! Things to consider before selecting K Nearest Neighbours Algorithm:
   - KNN is computationally expensive
   - Variables should be normalized, or else higher range variables can bias the algorithm
   - Data still needs to be pre-processed.
7. **K-Means:** It is an unsupervised learning algorithm that solves clustering problems. Data sets are classified into a particular number of clusters (let's call that number K) in such a way that all the data points within a cluster are homogenous and heterogeneous from the data in other clusters. How K-means forms clusters:
   - The K-means algorithm picks k number of points, called centroids, for each cluster.
   - Each data point forms a cluster with the closest centroids, i.e., K clusters.
   - It now creates new centroids based on the existing cluster members.
   - With these new centroids, the closest distance for each data point is determined. This process is repeated until the centroids do not change.
8. **Random Forest Algorithm:** A collective of decision trees is called a Random Forest. To classify a new object based on its attributes, each tree is classified, and the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest). Each tree is planted & grown as follows:
   - If the number of cases in the training set is N, then a sample of N cases is taken at random. This sample will be the training set for growing the tree.
   - If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M, and the best split on this m is used to split the node. The value of m is held constant during this process.
   - Each tree is grown to the most substantial extent possible. There is no pruning.
9. **Dimensionality Reduction Algorithms:** In today's world, vast amounts of data are being stored and analyzed by corporates, government agencies, and research organizations. As a data scientist, you know that this raw data contains a lot of information - the challenge is to identify significant patterns and variables. Dimensionality reduction algorithms like Decision Tree, Factor Analysis, Missing Value Ratio, and Random Forest can help you find relevant details.

10. **Gradient Boosting Algorithm and AdaBoosting Algorithm:** Gradient Boosting Algorithm and AdaBoosting Algorithm are boosting algorithms used when massive loads of data have to be handled to make predictions with high accuracy. Boosting is an ensemble learning algorithm that combines the predictive power of several base estimators to improve robustness. In short, it combines multiple weak or average predictors to build a strong predictor. These boosting algorithms always work well in data science competitions like Kaggle, AV Hackathon, CrowdAnalytix. These are the most preferred machine learning algorithms today. Use them, along with Python and R Codes, to achieve accurate outcomes.