**Machine Learning Algorithms**

1. **Logistic Regression** (Supervised)

It models the probabilities for classification problems with two possible outcomes. It’s an extension of the linear regression model for classification problems.

It is used to describe data to explain the relationship between one dependant binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Basically works with binary data where the event happens (1) or it doesn’t happen (0) i.e. when the dependant variable (target) is categorical i.e. to predict if an email is spam (1) or (0), whether a tumor is malignant (1) or not (0).

Linear and logistic regression models fail where the relationship between features and outcomes is nonlinear or where features interact with each other. Is being used to predict cervical cancer based on some risk factors.

1. **Decision Tree** (Supervised)

Builds classification or regression models in the form of a tree structure. It breaks down the data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. Splits are formed on a particular variable i.e. gender, class etc. This results in a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf nodes represents a classification or decision. The top most decision node in a tree corresponds to the best predictor called root node. Basically, it follows a set of if-else conditions to visualise the data and classify it according to the conditions. Decision trees can handle both categorical and numerical data. Prior training data classifies coordinates into groups identified by an attribute. They can be used in both a regression and a classification context. This is an example of adaptive basis function models which learns the features directly from the data.

Are used in forecasting i.e. future asset prices or liquidity of certain instruments.

1. **Support Vector Machine (SVM)** (Supervised)

Finds intense application in pattern recognition, data mining and intrusion detection. Uses classification algorithms for two-group classification problems and can be used for regression tasks. SVM model uses sets of labelled training data for each category to categorise new text. It uses a technique called the kernel trick to transform the data and then based on these transformations it finds an optimal boundary between the possible outputs. You can capture much more complex relationships between the datapoints without having to perform difficult transformations. The downside is that the training time is much longer as more complex hence, often referred to as a black box.

It is used for text classification tasks I,e, category assignment, detecting spam and sentiment analysis. Commonly used for image recognition tasks and handwritten digit recognition i.e. postal automation services.

1. **Naïve Bayes** (Supervised)

This classifier is a probabilistic machine learning model used for classification task. This is based on Bayes theorem in that we can find the probability of A happening given that B has occurred. B is the evidence, A is the hypothesis. Presence of one particular feature does not affect the other hence, it is called naïve.

They are mainly used in sentiment analysis, spam filtering, recommendation systems etc. Fast and easy to implement but the biggest disadvantage is the requirement of predictors to be independent.

1. **K-Nearest Neighbours (KNN)** (Supervised)

Used mainly for classification but also for regression predictive problems. It is commonly used for its easy interpretation and low calculation time. This algorithm assumes that similar things are near to each other. This feature similarity is used to predict the values of new datapoints. The new data points will be assigned a value based on how closely it matches the points in the training set. It’s a non-parametric learning algorithm as it doesn’t assume anything about the underlying data. It is also a lazy algorithm as it does not use the training data points to do any generalisation.

It is used for credit ratings as the financial characteristics are compared with similar features to a database. Bank loan approvals i.e. is that person closer in characteristics to people who default on their loans. Handwriting detection, image recognition and video recognition are more advance examples.

1. **Random Forest** (Supervised)

Capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration. The forest it builds consists of multiple decision trees which are merged together to get a more accurate and stable prediction. Instead of searching for the most important feature while spitting the node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Also, it is very easy to measure the relative importance of each feature on the prediction, with the sum of importance is equal to one. This can allow you to possibly drop features as they don’t contribute enough to the prediction process.

It is used in a lot of fields from banking to identify customers more likely to repay their debt on time, healthcare to identify the correct combination of components in medicine and to analyse a patient’s medical history to identify diseases. Another example is e-commerce to determine whether a customer will actually like a product or not.