Decision trees, especially *classification and regression trees* can be used to quickly predict either categorical or numeric outcomes. Classification trees are used when the response or target variable is categorical in nature. Regression trees are needed when the response variable is numeric or continuous.

When to use

Classification trees: used to separate the dataset into classes belonging to the response variable. Usually the response variable has two classes: Yes or No (1 or 0). Thus classification trees are used when the response or target variable is categorical in nature.

Regression trees are needed when the response variable is numeric or continuous. For example, the predicted price of a consumer good. Thus regression trees are applicable for *prediction* type of problems as opposed to *classification*.

How they work

In a standard classification tree, the idea is to split the dataset based on homogeneity of data. For example we have two variables: age and weight that predict if a person is going to sign up for a gym membership or not. In our training data if it showed that 90% of the people who are older than 40 signed up, we split the data here and age becomes a top node in the tree. We can almost say that this split has made the data "90% pure

In a regression tree the idea is this: since the target variable does not have classes, we fit a regression model to the target variable using each of the independent variables. Then for each independent variable, the data is split at several split points. At each split point, the "error" between the predicted value and the actual values is squared to get a "Sum of Squared Errors (SSE)". The split point errors across the variables are compared and the variable/point yielding the lowest SSE is chosen as the root node/split point. This process is recursively continued.