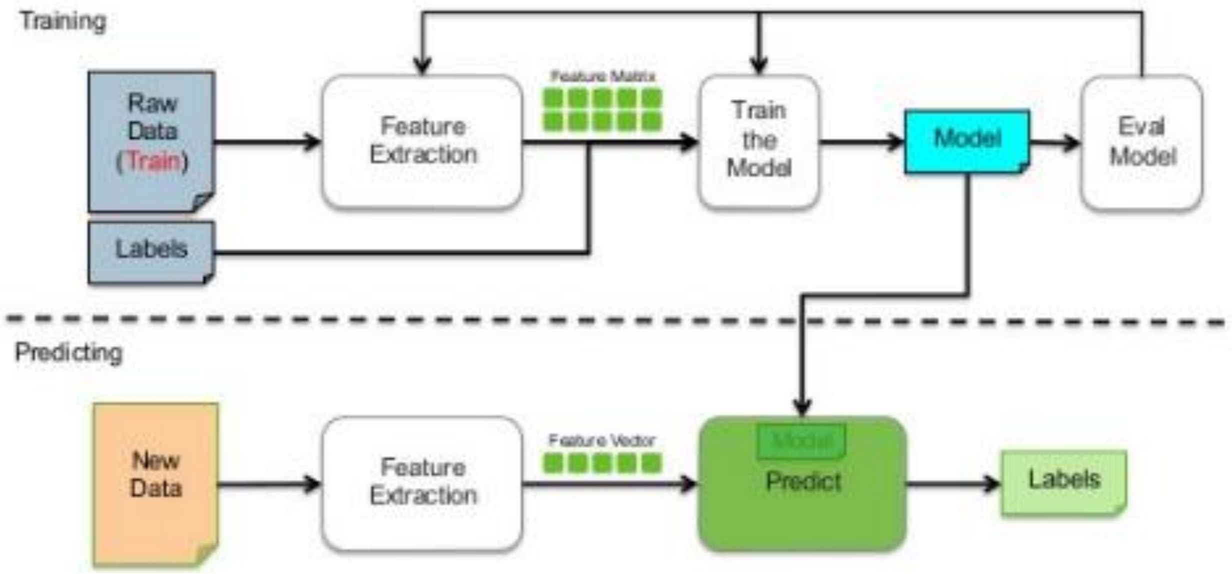
1. Supervised Learning

What Supervised Learning?

Supervised learning is the most common sub-branch of machine learning. Typically, new machine learning practitioners begin their journey with supervised learning algorithms. Supervised machine learning algorithms are designed to learn by example. The name "supervised" learning originates from the idea that training this type of algorithm is like having a teacher supervise the whole process.

When training a supervised learning algorithm, the training data will consist of inputs paired with the correct outputs. During training, the algorithm will search for patterns in the data that correlate with the desired outputs. After training, a supervised learning algorithm will take in new unseen inputs and will determine which label the new inputs will be classified as based on prior training data.

In short: **A supervised learning algorithm learns from labelled training data, helps you to predict outcomes for unforeseen data.**



Flow chart of a Supervised Machine Learning

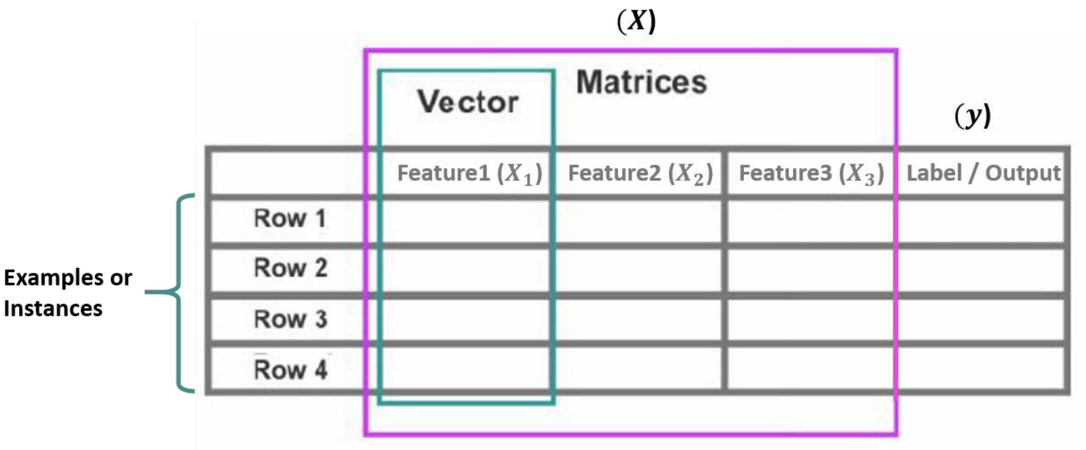
There are two types of supervised machine learning algorithms: Regression and classification. The former predicts continuous value outputs while the latter predicts discrete outputs. For instance, predicting the price of a house in dollars is a regression problem whereas predicting whether a tumor is malignant or benign is a classification problem.

Basic Terminology:

Before we proceed, lets understand the basic terminology used in any Machine Learning Algorithm.

- **Algorithm:** A Machine Learning algorithm is a set of rules and statistical techniques used to learn patterns from data and draw significant information from it. It is the logic behind a Machine Learning model. An example of a Machine Learning algorithm is the Linear Regression algorithm.

- **Dataset:** A dataset contains data organized in rows and columns. In each column is a feature. A feature is also known as a variable, a dimension or an attribute—but they all mean the same thing. Each individual row represents a single observation of a given feature/variable. Rows are sometimes referred to as an instance or example.



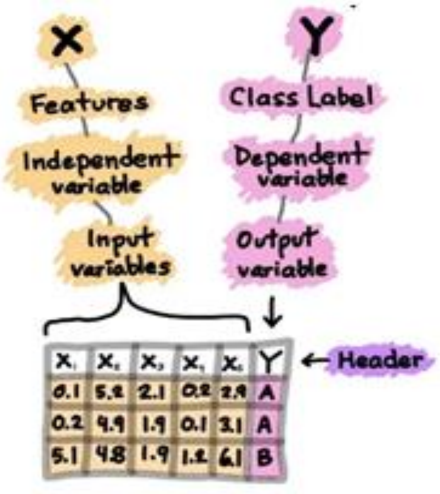
	Feature1 (X_1)	Feature2 (X_2)	Feature3 (X_3)	Label / Output
Row 1				
Row 2				
Row 3				
Row 4				

Each column is known as a vector. Vectors store your X and y values and multiple vectors (columns) are commonly referred to as matrices. In the case of supervised learning, y will already exist in your dataset and be used to identify patterns in relation to independent variables (X).

- **Examples:** Items or instances of data used for learning or evaluation.

- **Features:** The set of attributes, often represented as a vector (or matrix), associated to an example.

- **Labels:** Values or categories assigned to examples. In classification problems, examples are assigned specific categories.



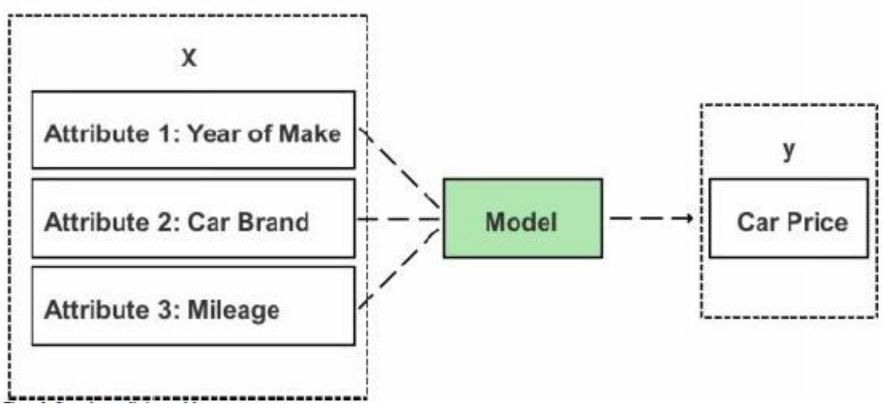
X_1	X_2	X_3	X_4	X_5	Y
0.1	5.2	2.1	0.2	2.9	A
0.2	4.9	1.9	0.1	3.1	A
5.1	4.8	1.9	1.2	6.1	B

- **Training sample:** Examples used to train a learning algorithm. The training data helps the model to identify key trends and patterns essential to predict the output.

- **Test sample:** Examples used to evaluate the performance of a learning algorithm. The test sample is separate from the training and validation data and is not made available in the learning stage. After the model is trained, it must be tested to evaluate how accurately it can predict an outcome. This is done by the testing data set.

As the first branch of machine learning, supervised learning concentrates on learning patterns through connecting the relationship between variables and known outcomes and working with labeled datasets. Supervised learning works by feeding the machine sample data with various features (represented as "X") and the correct value output of the data (represented as "y"). The fact that the output and feature values are known qualifies the dataset as "labeled." The algorithm then deciphers patterns that exist in the data and creates a model that can reproduce the same underlying rules with new data.

For instance, to predict the market rate for the purchase of a used car, a supervised algorithm can formulate predictions by analyzing the relationship between car attributes (including the year of make, car brand, mileage, etc.) and the selling price of other cars sold based on historical data. Given that the supervised algorithm knows the final price of other cards sold, it can then work backward to determine the relationship between the characteristics of the car and its value.



After the machine deciphers the rules and patterns of the data, it creates what is known as a model: an algorithmic equation for producing an outcome with new data based on the rules derived from the training data. Once the model is prepared, it can be applied to new data and tested for accuracy. After the model has passed both the training and test data stages, it is ready to be applied and used in the real world.

Another simple example is to create a model for predicting house values where y is the actual house price and X are the variables that impact y, such as land size, location, and the number of rooms. Through supervised learning, we can create a rule to predict y (house value) based on the given values of various variables (X).

Examples of supervised learning algorithms include regression analysis, decision trees, k-nearest neighbours, etc.