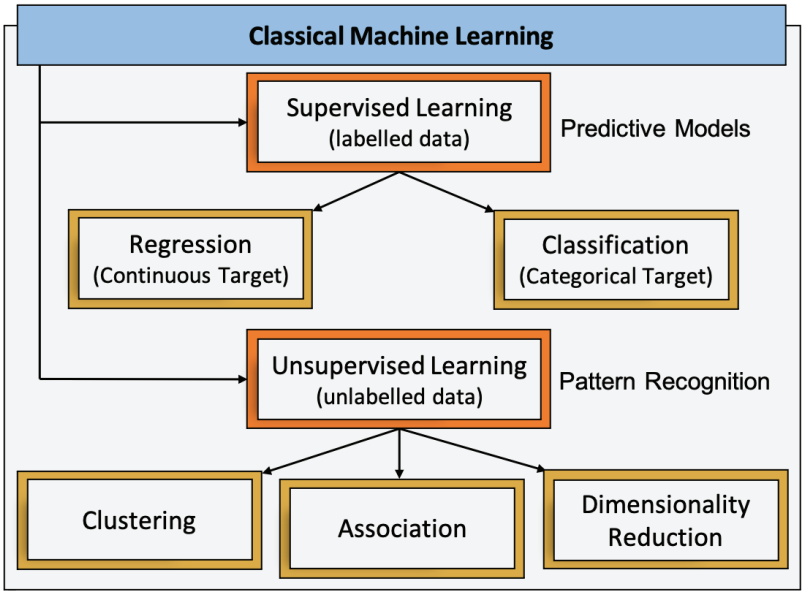
CENGN

# MACHINE LEARNING WITH PYTHON TAKEAWAY

## What Is Machine Learning?

Machine learning refers to algorithms that use statistical reasoning to find patterns within massive amounts of data, to uncover associations and to make predictions.

Machine learning has two basic paradigms: supervised and unsupervised learning.




## Supervised Learning

Aims to learn the underlying relationship between the inputs and outputs based on historical data. It is often used to predict outcomes using labelled data to train a model. Supervised learning has two main algorithm types:

**Regression:** An algorithm typically used to predict or forecast an outcome variable in terms of a set of input variables (predictors). For example, predicting house price based on features like location, number of bedrooms, lot area, etc.

**Classification:** An algorithm that can predict the class of an unknown entity (whether belonging to a set of pre-defined classes). For example: spam/not spam, pass/fail, benign/malignant, etc.

The table below summarizes key points about common supervised learning models:

| | | |
|---|---|---|
| Regression | Linear Regression | Identify relationships between quantitative data typically used to predict or forecast an outcome variable in terms of a set of input variables. |
| | Regression Trees | Decision trees work in a sequential process. Data is evaluated and split multiple times according to certain cutoff values in the predictor variables. |
| Classification | Logistic Regression | Uses the logistic function to model the probabilities for a classification problem with two possible outcomes such as pass/fail, spam/no spam, healthy/sick, etc. |
| | Naive Bayes | Uses the Bayes' theorem to predict the probability of an observation belonging to a particular class. It assumes all features are uncorrelated. |
| | SVM | SVMs find a line, curve, or manifold that optimally separates the classes from each other. SVMs rely on the kernel to specify the decision boundary (linear/non-linear). |

CENGN

# MACHINE LEARNING WITH PYTHON TAKEAWAY

## Unsupervised Learning

Aims to infer some natural structure present within the data by searching for patterns and grouping similar observations. Unsupervised learning has three main algorithm types:

**Clustering**: An algorithm that finds subgroups or clusters of similar observations in a given data. For example, segmenting customers into groups based on their salary, age, and purchasing behavior.

**Dimensionality reduction**: An algorithm that aims to reduce the dimensions in a dataset that has a high number of correlated features. For example, datasets of images are highly dimensional, thus will require reducing its dimensions prior to modeling.

**Association learning**: An algorithm that identifies items that appear together or are otherwise associated. For example, Amazon's frequently bought together feature suggests items for purchase based on purchasing habits of other customers.

The table below summarizes key points about common unsupervised learning models:

| | | |
|---|---|---|
| Clustering | K-Means | A simple distance-based clustering approach for partitioning a dataset into a pre-defined number of clusters (K). It aims to minimize the within-cluster variation. |
| | Hierarchical Clustering | This algorithm produces a dendrogram - an interpretable tree-based representation of the observations. It does not require us pre-specifying the number of clusters . |
| | Mixture Models | A clustering algorithm that can discover complex patterns, it can handle data with clusters of different shapes easily. |
| Association Learning | Apriori | A simple association learning iterative algorithm that starts out with lists of each individual item and then builds up lists one item at a time. |
| | ECLAT | A more efficient association learning algorithm. |
| Dimensionality Reduction | PCA | Dimensionality reduction techniques attempt to reduce the number of dimensions in large datasets while retaining much of the variation in the dataset. |
| | t-SNE | PCA and t-SNE are two examples. |

## Evaluating a Model's Performance

After building a model, its performance can be evaluated using metrics like the below:

**Confusion Matrix**: Visually depicts whether the model is correctly labeling the observations.
**Accuracy**: Identifies how often a model makes the right predictions.
**Precision**: A measure of how many positives classified by a model are really positives.
**Recall**: A measure of how many positives were correctly classified as positive.
**F1-Score:** Is the harmonic mean of precision and recall.
**MSE**: Is the mean of the square of the errors. It is used for evaluating regression models.

Accuracy, precision, recall, and F1-score are used for evaluating classification models.