**(ML) AI Section 4: Analyze Data - Understanding models used for data analysis**

**Main Learning Goal**

The main goal of this unit is to understand how you train Machine Learning models to analyze data. This unit also looks at how different ML models can be used in order to and aid data-based decision making using financial information as a real-world example.

**Focus Question: How can you train ML models to analyze data?**

**Background**

**Supervised learning** is a machine learning approach that’s defined by its use of labeled datasets. These datasets are designed to train or “supervise” algorithms into classifying data or predicting outcomes accurately. Using labeled inputs and outputs, the model can measure its accuracy and learn over time.

Supervised learning can be separated into two types of problems when data mining: **regression** and **classification**:

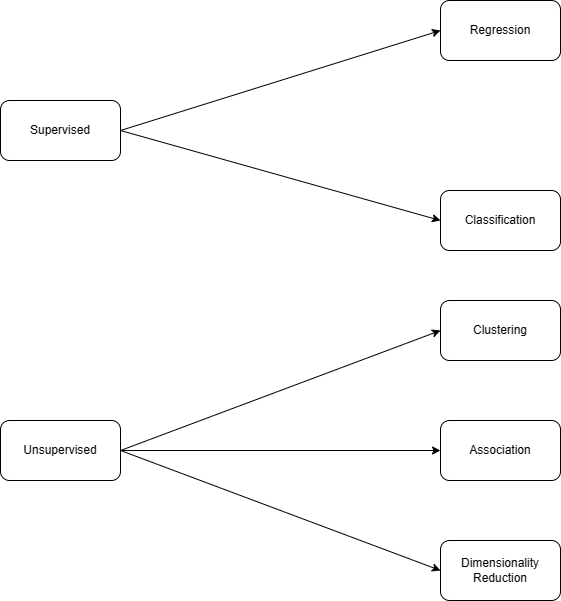
* **Regression** - uses an algorithm to understand the relationship between dependent and independent variables. You looked at one kind of regression in a previous lesson (linear regression with the application of trendlines) but there are many different kinds. Regression models are helpful for predicting numerical values based on different data points, such as sales revenue projections for a given business. Some popular regression algorithms are linear regression, logistic regression and polynomial regression.
* **Classification** - uses an algorithm to accurately assign test data into specific categories, such as separating apples from oranges. Or, in the real world, supervised learning algorithms can be used to classify spam in a separate folder from your inbox.    
  Linear classifiers, support vector machines, decision trees and random forest are all common types of classification algorithms.

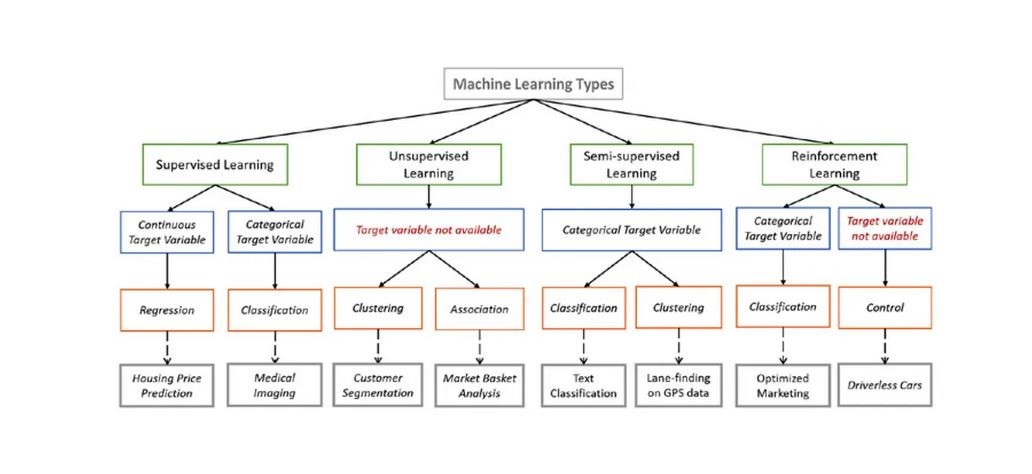
**Unsupervised learning** uses machine learning algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns in data without the need for human intervention (hence, they are “unsupervised”).

Unsupervised learning models are used for three main tasks: **clustering, association** and **dimensionality reduction**.

* **Clustering** - a data mining technique for grouping unlabeled data based on their similarities or differences. For example, K-means clustering algorithms assign similar data points into groups, where the K value represents the size of the grouping and granularity. This technique is helpful for market segmentation, image compression, etc.
* **Association** - another type of unsupervised learning method that uses different rules to find relationships between variables in a given dataset. These methods are frequently used for market basket analysis and recommendation engines, along the lines of “Customers Who Bought This Item Also Bought” recommendations.
* **Dimensionality reduction** - a learning technique used when the number of features (or dimensions) in a given dataset is too high. It reduces the number of data inputs to a manageable size while also preserving the data integrity. Often, this technique is used in the preprocessing data stage, such as when autoencoders remove noise from visual data to improve picture quality.

The main distinction between the two approaches is the use of **labeled datasets**. To put it simply, supervised learning uses **l**abeled input and output data, while an unsupervised learning algorithm does not.





**Energy Use and Requirements for Machine Learning**

These models can use large amounts of energy and computing power. They also require specialized hardware including  Specialized Graphics Processing Units (**GPU**), Central Processing Units (**CPU**),  Coprocessors and AI accelerators costing large amounts of money. For more on the carbon footprint of large AI systems see this [article .](https://hellofuture.orange.com/en/keeping-watch-on-the-carbon-footprint-of-machine-learning/)(pdf version:[What is the carbon footprint of machine learning for AI).](https://ufl.instructure.com/courses/487632/files/80371917?wrap=1)

In the early days, the Cloud was the only way to provide sufficient computing resources for AI workloads. Hosted platforms to deploy AI models include Microsoft Azure, Viso Suite, Hugging Face, ChatGPT, Google Collab, or Amazon SageMaker.  In recent years,  Edge Computing made it possible to deploy models to the network edge (Edge AI). Edge computing is computing that takes place at or near the physical location of either the user or the source of the data.

Not all data is compared to "something" and not all data can be fitted to some mathematical algorithm. Sometimes data is used to recognize something like an image of a sports star or something. While there is generally always some math involved in the algorithms, AI can use and analyze data using many different methods.