Week 1

Welcome to Machine Learning

Week 1 (index 2): introduction to Machine Learning.

Machine learning is the subfield of computer science that gives "computers the ability to learn without being explicitly programmed."

Let me explain what I mean when I say "**without being explicitly programmed**." Assume that you have a dataset of images of animals such as cats and dogs, and you want to have software or an application that can recognize and differentiate them.

The first thing that you have to do here is interpret the images as a set of feature sets.

For example,

- ✓ does the image show the animal's eyes? If so, what is their size?
- ✓ Does it have ears?
- ✓ What about a tail?
- ✓ How many legs?
- ✓ Does it have wings?

Prior to machine learning, each image would be transformed to a vector of features.

Then, traditionally, we had to write down some rules or methods in order to get computers to be intelligent and detect the animals.

But it was a failure. Why? Well, as you can guess, it needed a lot of rules, highly dependent on the current dataset, and not generalized enough to detect out-of-sample cases.

This is when machine learning entered the scene. Using machine learning, allows us to build a model that looks at all the feature sets, and their corresponding type of animals, and it learns the pattern of each animal.

It is a model built by machine learning algorithms.

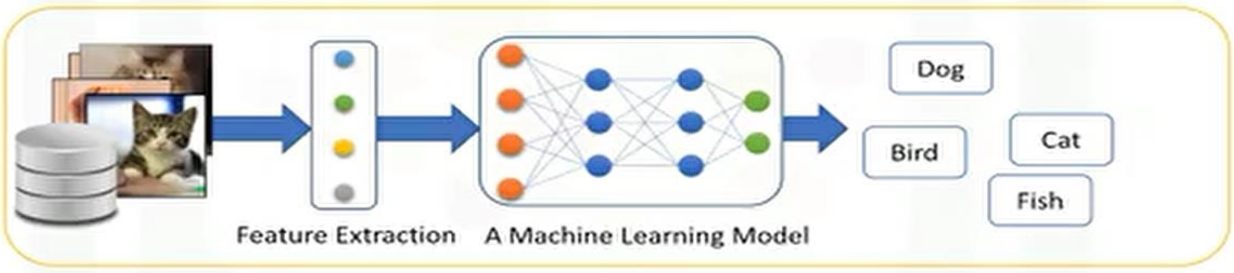
Note: It detects without explicitly being programmed to do so.

In essence, machine learning follows the same process that a 4-year-old child uses to learn, understand, and differentiate animals.

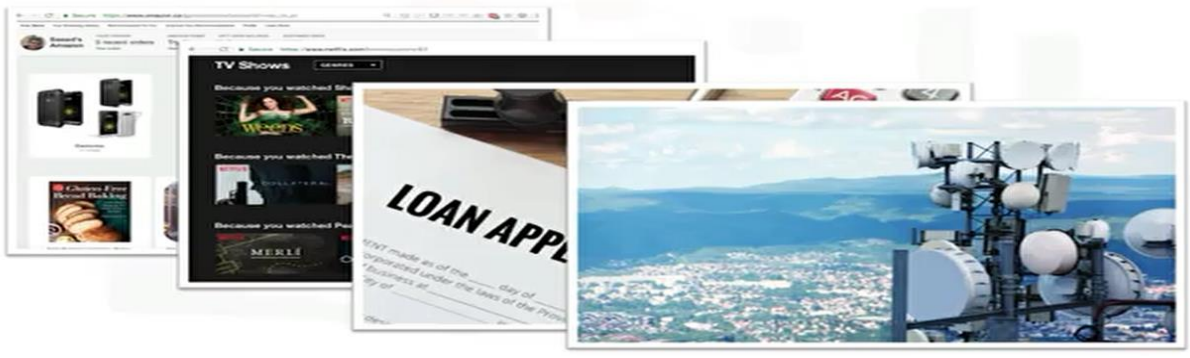
So, machine learning algorithms, inspired by the human learning process, iteratively learn from data, and allow computers to find hidden insights.

These models help us in a variety of tasks, such as object recognition, summarization, recommendation, and so on.

How Machine learning works?



Examples of machine learning



Machine Learning impacts society in a very influential way. Here are some real-life examples.

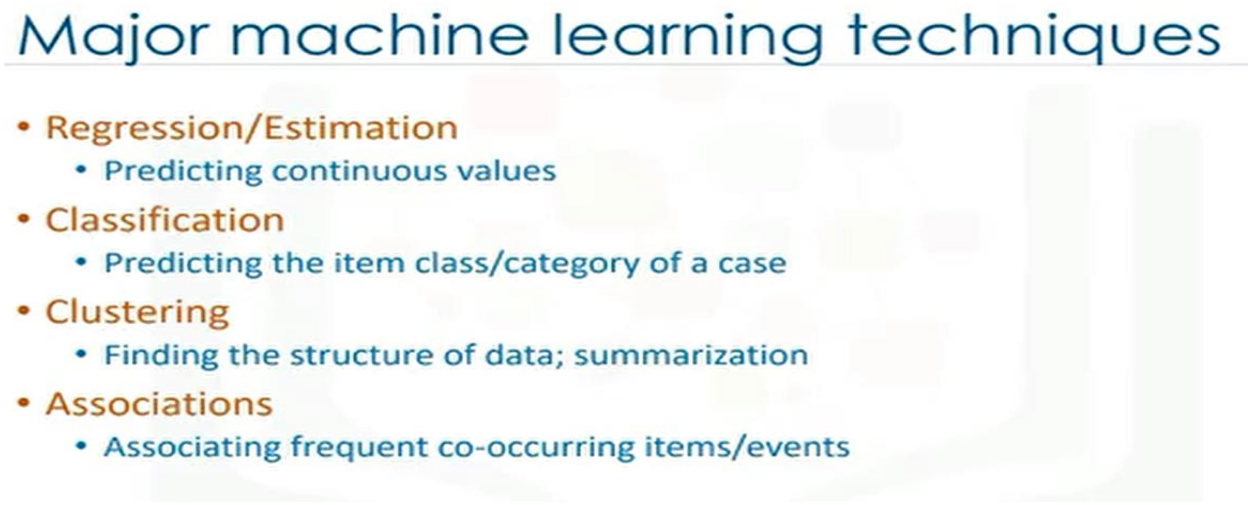
- First, how do you think Netflix and Amazon recommend videos, movies, and TV shows to its users?
  They use Machine Learning to produce suggestions that you might enjoy!
  This is similar to how your friends might recommend a television show to you, based on their knowledge of the types of shows you like to watch.
- How do you think banks make a decision when approving a loan application?
  They use machine learning to predict the probability of default for each applicant, and then approve or refuse the loan application based on that probability.
- Telecommunication companies use their customers' demographic data to segment them, or predict if they will unsubscribe from their company the next month.

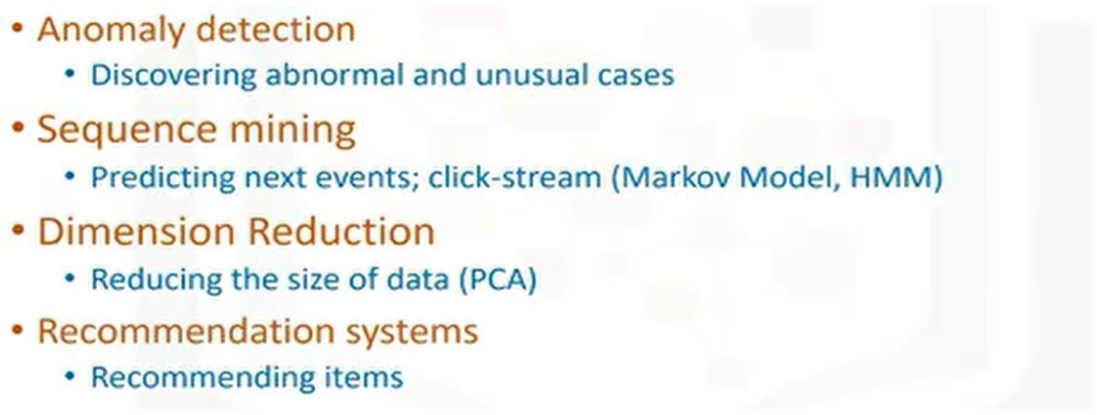
There are many other applications of machine learning that we see every day in our daily life, such as

- ✓ chatbots,
- ✓ logging into our phones or even computer games using face recognition.

Each of these use different machine learning techniques and algorithms.

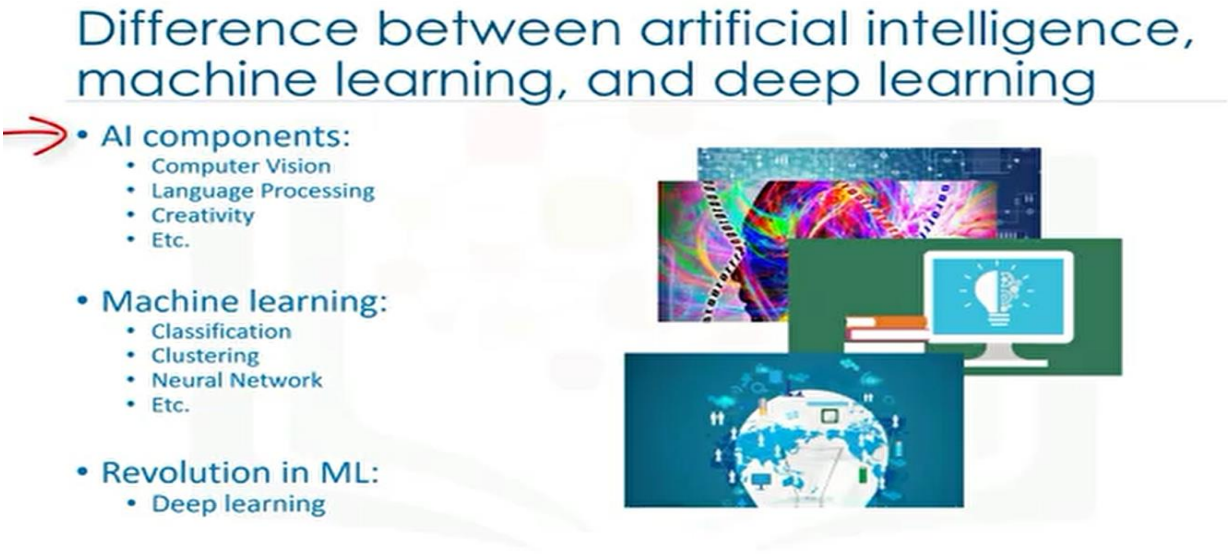
let's quickly examine a few of the more popular techniques.





1. **Regression/Estimation technique** is used for predicting a continuous value.
   For example,
   predicting things like the price of a house based on its characteristics, or to estimate the Co2 emission from a car's engine.
2. A **Classification technique** is used for Predicting the class or category of a case.
   for example,
   if a cell is benign or malignant, or whether or not a customer will churn.
3. **Clustering groups** of similar cases,
   for example,
   can find similar patients, or can be used for customer segmentation in the banking field.
4. **Association technique** is used for finding items or events that often co-occur,
   for example,
   grocery items that are usually bought together by a particular customer.
5. **Anomaly detection** is used to discover abnormal and unusual cases,
   for example,
   it is used for credit card fraud detection. Sequence mining is used for predicting the next event, for instance, the click-stream in websites.
6. **Dimension reduction** is used to reduce the size of data.
7. **Recommendation systems**, this associates people's preferences with others who have similar tastes, and recommends new items to them, such as books or movies.

What is the difference between these buzzwords that we keep hearing these days, such as Artificial intelligence (or AI), Machine Learning and Deep Learning?



- AI tries to make computers intelligent in order to mimic the cognitive functions of humans.

  So, Artificial Intelligence is a general field with a broad scope including: Computer Vision,
  - ✓ Language Processing,
  - ✓ Creativity, and
  - ✓ Summarization.

- Machine Learning is the branch of AI that covers the statistical part of artificial intelligence.

  It teaches the computer to solve problems by looking at hundreds or thousands of examples, learning from them, and then using that experience to solve the same problem in new situations.

- Deep Learning is a very special field of Machine Learning where computers can actually learn and make intelligent decisions on their own.

  Deep learning involves a deeper level of automation in comparison with most machine learning algorithms.

Week 1 (index 3): Python for machine learning

Python is a popular and powerful general purpose programming language that recently emerged as the preferred language among data scientists.

We can write your machine-learning algorithms using Python, and it works very well. However, there are a lot of modules and libraries already implemented in Python, that can make your life much easier.

- NumPy which is a math library to work with N-dimensional arrays in Python.
  It enables you to do computation efficiently and effectively.
  It is better than regular Python because of its amazing capabilities.
  For example,
  for working with arrays, dictionaries, functions, datatypes and working with images you need to know NumPy.

- SciPy is a collection of numerical algorithms and domain specific toolboxes, including signal processing, optimization, statistics and much more.

  SciPy is a good library for scientific and high-performance computation.

- Matplotlib is a very popular plotting package that provides 2D plotting, as well as 3D plotting.

Basic knowledge about these three packages which are built on top of Python, is a good asset for data scientists who want to work with real-world problems.

data analysis with Python course

- Pandas' library is a very high-level Python library that provides high performance easy to use data structures.

  It has many functions for data importing, manipulation and analysis.

  In particular, it offers data structures and operations for manipulating numerical tables and timeseries.

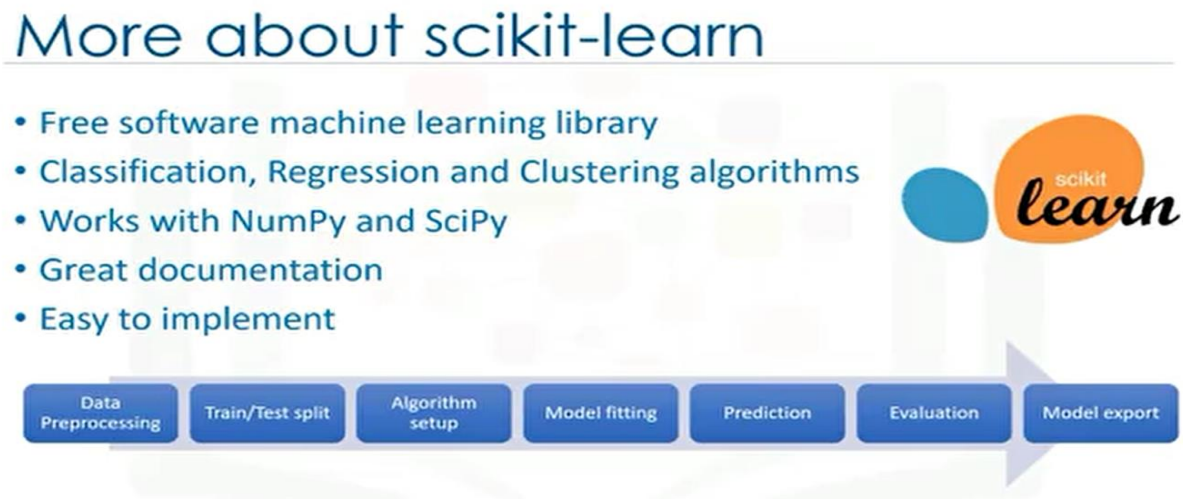
- SciKit Learn is a collection of algorithms and tools for machine learning.

  which is our focus here and which you'll learn to use within this course.

  As we'll be using SciKit Learn quite a bit in the labs, let me explain more about it and show you why it is so popular among data scientists.

  SciKit Learn is a free Machine Learning Library for the Python programming language.

  It has most of the classification, regression and clustering algorithms, and it's designed to work with a Python numerical and scientific library: NumPy and SciPy.



Also, it includes very good documentation. On top of that, implementing machine learning models with SciKit Learn is really easy with a few lines of Python code.

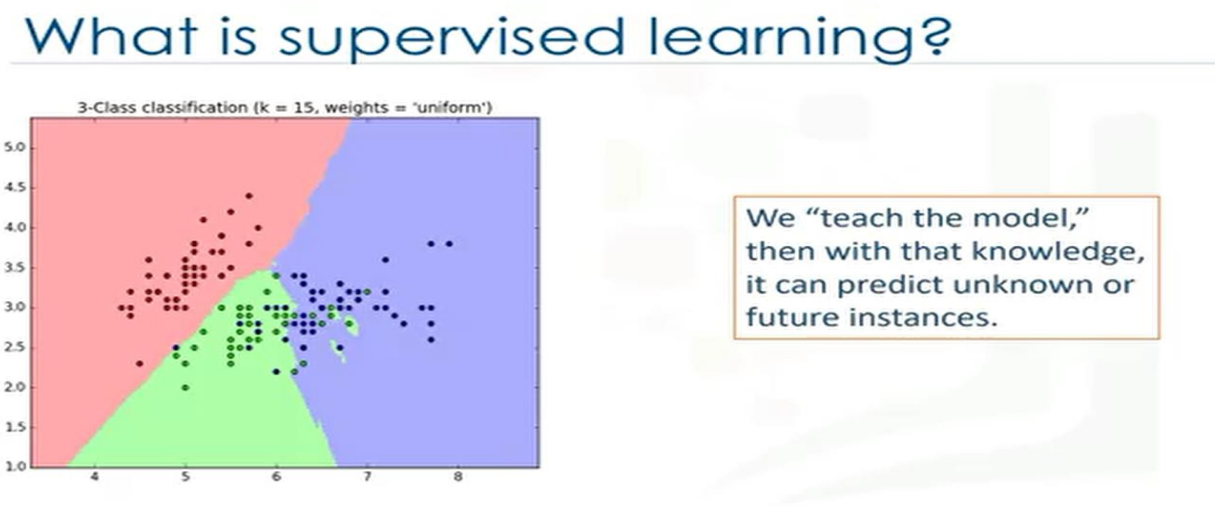
Most of the tasks that need to be done in a machine learning pipeline are implemented already in Scikit Learn including pre-processing of data, feature selection, feature extraction, train test splitting, defining the algorithms, fitting models, tuning parameters, prediction, evaluation, and exporting the model.

You have to split your dataset into train and test sets to train your model and then test the model's accuracy separately. SciKit Learn can split arrays or matrices into random train and test subsets for you in one line of code.

Week 1 (index 4): supervised algorithms versus unsupervised algorithms

Supervise, means to observe, and direct the execution of a task, project, or activity.

Obviously, we aren't going to be supervising a person, instead will be supervising a machine learning model that might be able to produce classification regions like we see here.

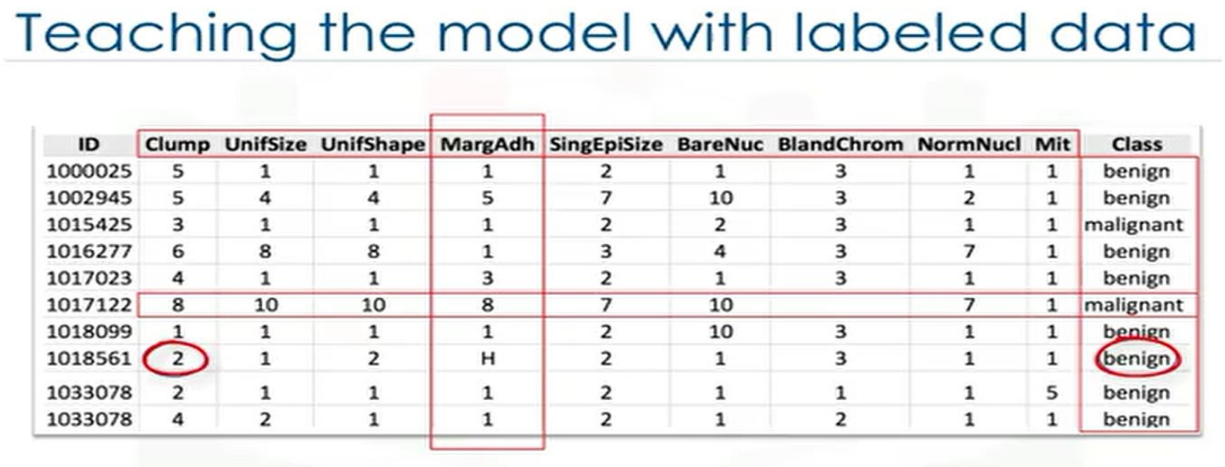


Q: How do we supervise a machine learning model?

- ✓ We do this by teaching the model, that is we load the model with knowledge so that we can have it predict future instances.

Q: But this leads to the next question which is, how exactly do we teach a model?

- ✓ We teach the model by training it with some data from a labeled dataset.



ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

clumps are uniformity of cell size, uniformity of cell shape, marginal adhesion and so on are called attributes.
The columns are called features which include the data.

If you plot this data, and look at a single data point on a plot, it'll have all of these attributes that would make a row on this chart also referred to as an observation.

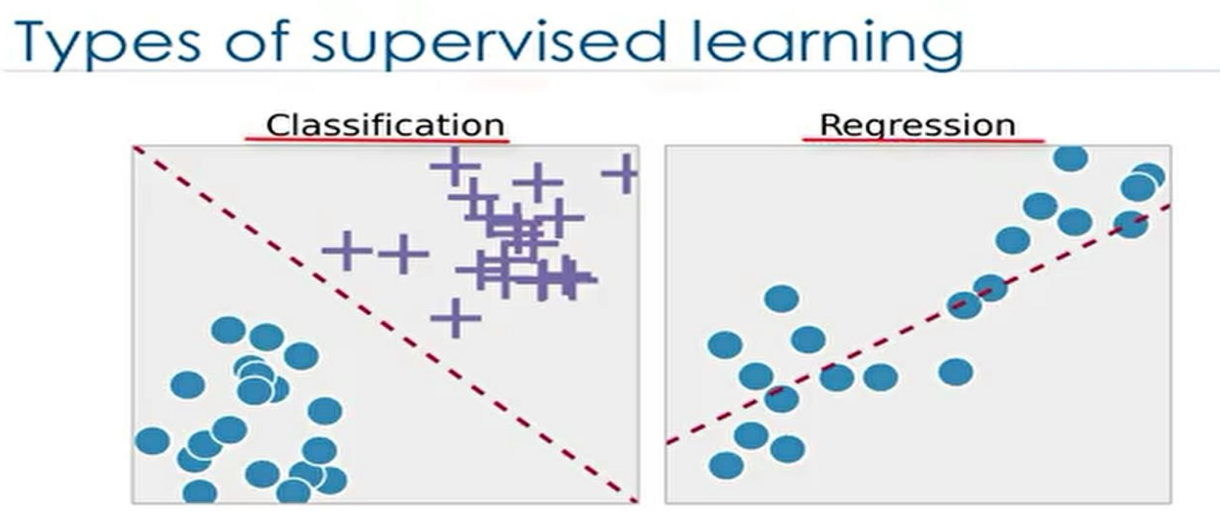
Looking directly at the value of the data, you can have two kinds.

- ✓ The first is numerical. When dealing with machine learning, the most commonly used data is numeric.
- ✓ The second is categorical, that is its non-numeric because it contains characters rather than numbers.
  In this case, it's categorical because this dataset is made for classification.

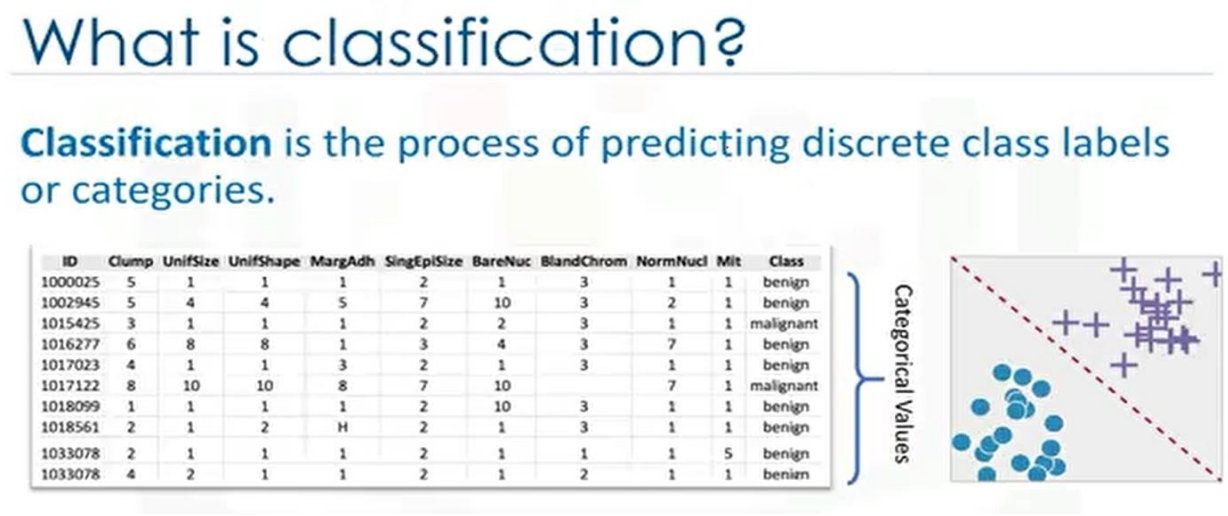
There are two types of supervised learning techniques.

They are:

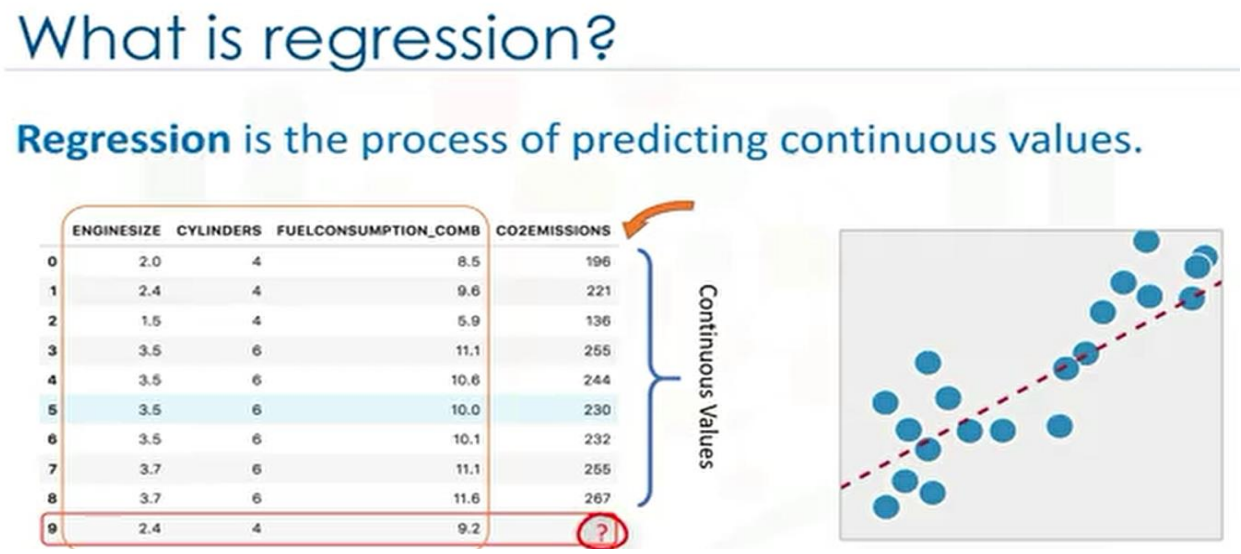
1. Classification
2. Regression.



Classification: classification is the process of predicting a discrete class label, or category.



Regression: regression is the process of predicting a continuous value as opposed to predicting a categorical value in classification.



Look at this dataset. It is related to CO2 emissions of different cars.

It includes; engine size, cylinders, fuel consumption, and CO2 emission of various models of automobiles.

Given this dataset, you can use regression to predict the CO2 emission of a new car by using other fields such as engine size, or number of cylinders.

Unsupervised

Q: what do you think unsupervised learning means?

- ✓ unsupervised learning is exactly as it sounds. We do not supervise the model, but we let the model work on its own to discover information that may not be visible to the human eye.

  It means, the unsupervised algorithm trains on the dataset, and draws conclusions on unlabeled data.

Generally speaking, unsupervised learning has more difficult algorithms than supervised learning since we know little to no information about the data, or the outcomes that are to be expected.

Dimension reduction, density estimation, market basket analysis, and clustering are the most widely used unsupervised machine learning techniques.

- ✓ **Dimensionality reduction**, and/or feature selection, play a large role in this by reducing redundant features to make the classification easier.
- ✓ **Market basket** analysis is a modeling technique based upon the theory that if you buy a certain group of items, you're more likely to buy another group of items.
- ✓ **Density estimation** is a very simple concept that is mostly used to explore the data to find some structure within it.
- ✓ **Clustering** is considered to be one of the most popular unsupervised machine learning techniques used for grouping data points, or objects that are somehow similar.
  Cluster analysis has many applications in different domains, whether it be a bank's desire to segment his customers based on certain characteristics, or helping an individual to organize in-group his, or her favorite types of music.
  Generally speaking, though, clustering is used mostly for discovering structure, summarization, and anomaly detection.

So, to recap, the biggest difference between supervised and unsupervised learning.

- ✓ supervised learning deals with labeled data while unsupervised learning deals with unlabeled data.
- ✓ In supervised learning, we have machine learning algorithms for classification and regression. In unsupervised learning, we have methods such as clustering.
- ✓ In comparison to supervised learning, unsupervised learning has fewer models and fewer evaluation methods that can be used to ensure that the outcome of the model is accurate.
- ✓ As such, unsupervised learning creates a less controllable environment as the machine is creating outcomes for us.