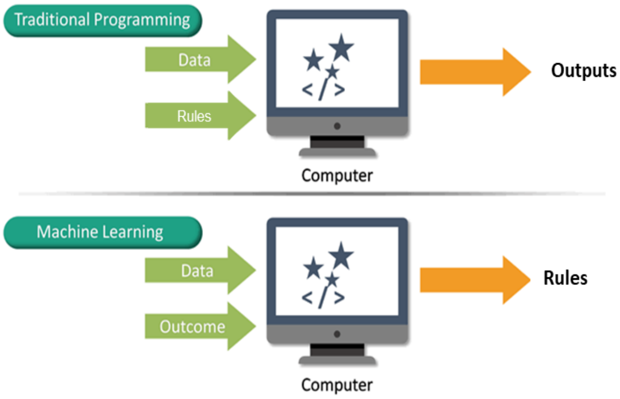
0. Introduction to Machine Learning

0.1. What is Machine Learning?

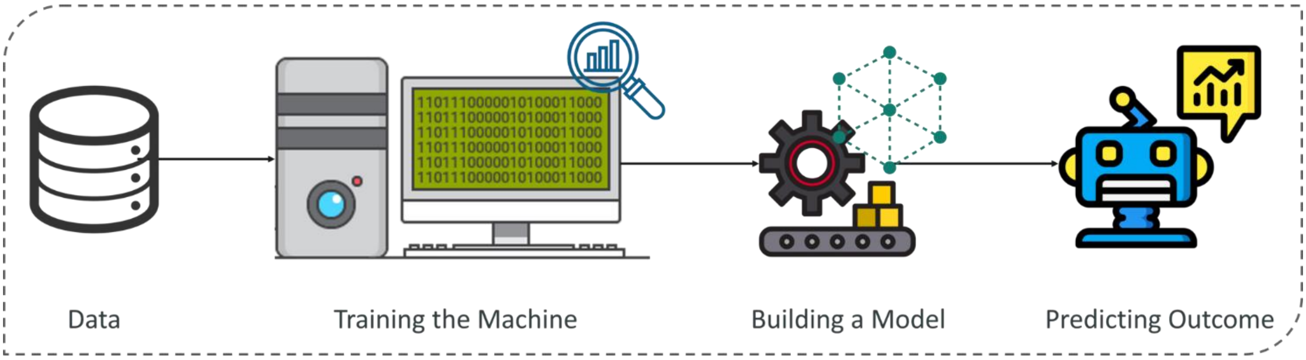
"Machine learning is field of study that gives computers the ability to learn without being explicitly programmed." – 1959, Arthur Samuel, a pioneer in the field of ML. In other words, Machine learning is a form of AI that enables a system to learn from data rather than through explicit programming. Thus, (ML) is a set of algorithms that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.

Traditional Programming vs Machine Learning



Machine learning is a form of AI that enables a system to learn from data rather than through explicit programming. However, machine learning is not a simple process. Machine learning uses a variety of algorithms that iteratively learn from data to improve, describe data, and predict outcomes.

As the algorithms ingest training data, it is then possible to produce more precise models based on that data. A machine learning model is the output generated when you train your machine learning algorithm with data. After training, when you provide a model with an input, you will be given an output. For example, a predictive algorithm will create a predictive model. Then, when you provide the predictive model with data, you will receive a prediction based on the data that trained the model.

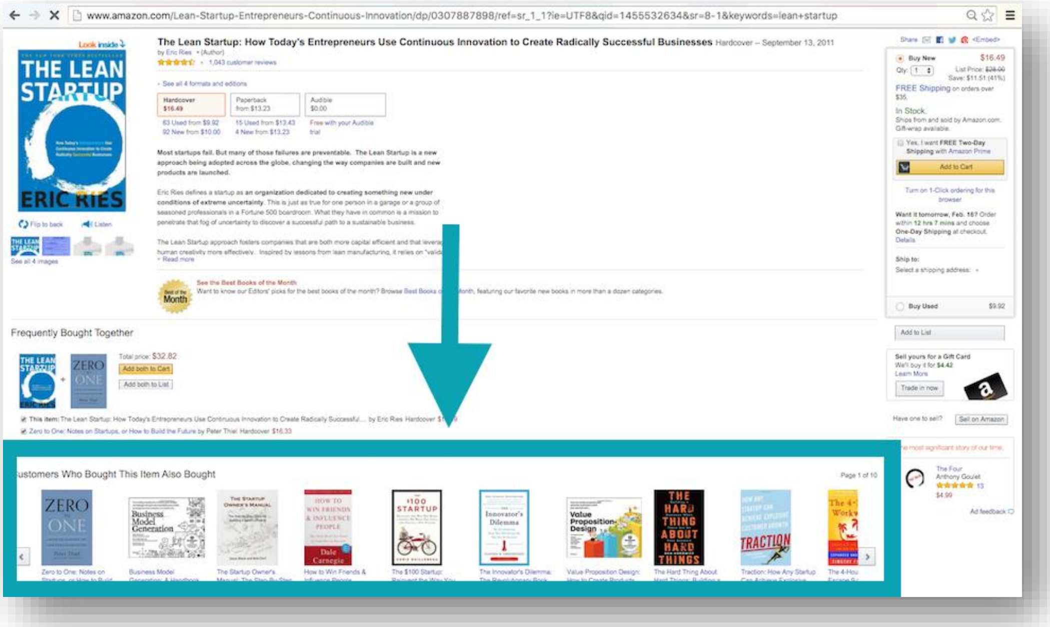


Thus, Machine learning involves designing of efficient and accurate prediction algorithms. Since the success of a learning algorithm depends on the data used, machine learning is inherently related to data analysis and statistics. More generally, learning techniques are data-driven methods combining fundamental concepts in computer science with ideas from statistics, probability and optimization.

The main practical objectives of machine learning consist of generating accurate predictions for unseen items and of designing efficient and robust algorithms to produce these predictions.

You likely interact with machine learning applications without realizing.

For example, when you visit an e-commerce site and start viewing products and reading reviews, you're likely presented with other, similar products that you may find interesting. These recommendations aren't hard coded by an army of developers. The suggestions are served to the site via a machine learning model.



www.amazon.com/Lean-Startup-Entrepreneurs-Continuous-Innovation/dp/0307887898/ref=sr_1_1?ie=UTF8&qid=1455532634&sr=8-1&keywords=lean+startup

The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses Hardcover – September 13, 2011

by Eric Ries (Author)

1,043 customer reviews

See all 4 formats and editions

Hardcover	Paperback	Audible
$16.49	from $13.23	$0.00
63 Used from $9.92 92 New from $10.00	15 Used from $13.43 4 New from $13.23	Free with your Audible trial

Most startups fail. But many of those failures are preventable. The Lean Startup is a new approach being adopted across the globe, changing the way companies are built and new products are launched.

Eric Ries defines a startup as **an organization dedicated to creating something new under conditions of extreme uncertainty**. This is just as true for one person in a garage or a group of seasoned professionals in a Fortune 500 boardroom. What they have in common is a mission to penetrate that fog of uncertainty to discover a successful path to a sustainable business.

The Lean Startup approach fosters companies that are both more capital efficient and that leverag human creativity more effectively. Inspired by lessons from lean manufacturing, it relies on "valid

Read more

See the Best Books of the Month

Want to know our Editors' picks for the best books of the month? Browse Best Books o Month, featuring our favorite new books in more than a dozen categories.

Frequently Bought Together

Total price: $32.82

This item: The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful... by Eric Ries Hardcover

Zero to One: Notes on Startups, or How to Build the Future by Peter Thiel Hardcover $16.33

ustomers Who Bought This Item Also Bought

Page 1 of 10

Buy New $16.49

Qty: 1

List Price: ~~$28.00~~

Save: $11.51 (41%)

FREE Shipping on orders over $35.

In Stock.

Ships from and sold by Amazon.com. Gift-wrap available.

□ Yes, I want **FREE Two-Day Shipping** with Amazon Prime

Turn on 1-Click ordering for this browser

Want it tomorrow, Feb. 16? Order within **12 hrs 7 mins** and choose **One-Day Shipping** at checkout. Details

Ship to:

Select a shipping address:

○ Buy Used $9.92

Sell yours for a Gift Card

We'll buy it for **$4.42**

Learn More

Have one to sell?

The Four

Anthony Goulet

13

$4.99

The model ingests your browsing history along with other shoppers' browsing and purchasing data in order to present other similar products that you may want to purchase.

Applications

There are several scenarios and several real-world tasks where Machine Learning algorithms have been successfully deployed, including:

- Text or document classification, e.g., spam detection;
- Natural language processing, e.g., morphological analysis, part-of-speech tagging,
- Speech recognition, speech synthesis, speaker verification;
- Optical character recognition (OCR);
- Computer vision tasks, e.g., image recognition, face detection;
- Fraud detection (credit card, telephone);
- Unassisted vehicle control (robots, navigation);
- Medical diagnosis, e.g., Tumour / Cancer Detection;
- Recommendation systems e.g., Product recommendations in online shopping platforms

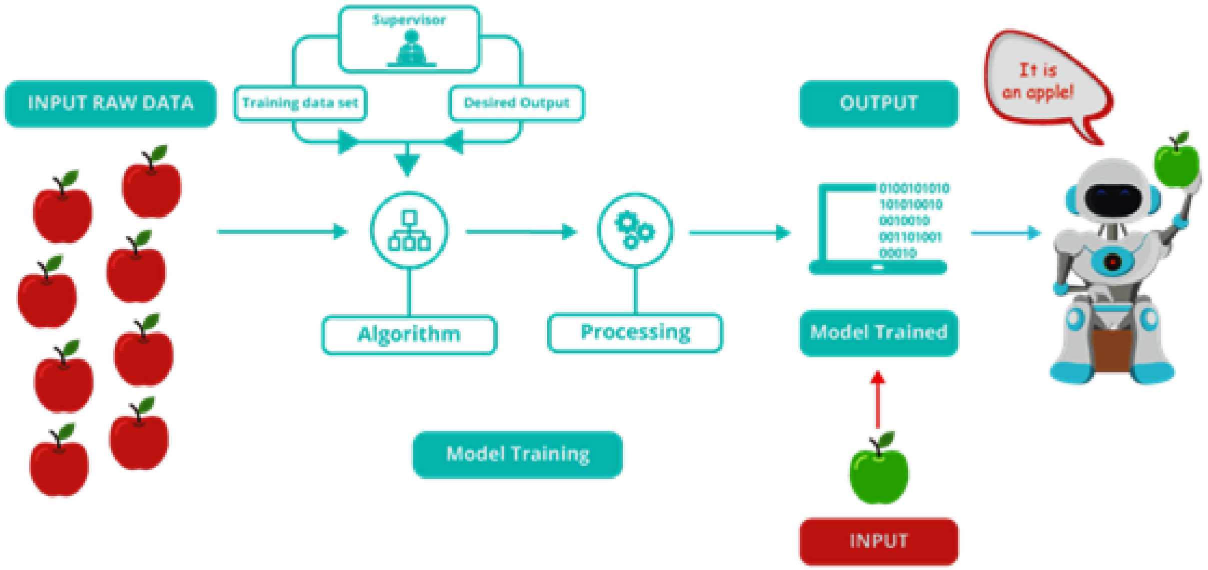
0.2. Approaches to Machine Learning

We next briefly describe common machine learning approaches. These differ in the types of training data available to us.

Supervised learning:

The learning algorithm receives a set of labelled examples as training data and makes predictions for all unseen points. This is the most common scenario associated with classification, regression problems.

The purpose of this method is for the algorithm to be able to "learn" by comparing its actual output with the "taught" outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabelled data. This is precisely why supervised learning methods are extensively used in predictive analytics where the main objective is to predict some response for some input data that's typically fed into a trained supervised ML model.



Supervised learning methods or algorithms include learning algorithms that take in data samples (known as training data) and associated outputs (known as labels or responses) with each data sample during the model training process. The main objective is to learn a mapping or association between input data samples x and their corresponding outputs y based on multiple training data instances. This learned knowledge can then be used in the future to predict an output y′ for any new input data sample x′ which was previously unknown or unseen during the model training process. These methods are termed as supervised because the model learns on data samples where the desired output responses/labels are already known beforehand in the training phase.

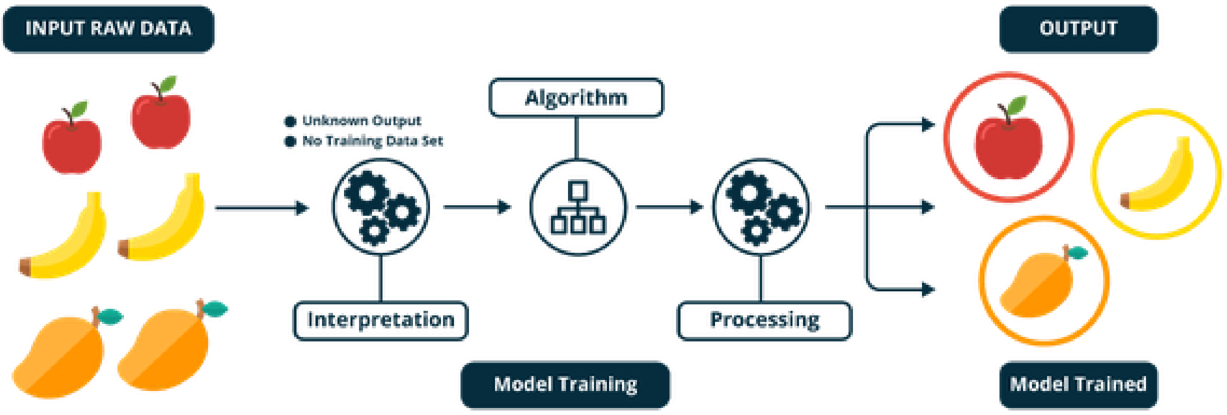
Some of the most popular supervised learning algorithms are: Linear Regression, Logistic Regression, Decision Trees and Random Forests, K-Nearest Neighbours etc.

Unsupervised learning:

This learning algorithm exclusively receives unlabelled training data, and makes predictions for all unseen points. Since in general no labelled example is available the learning algorithm is left to find commonalities among its input data.

For example, a business may wish to group its customers into distinct categories based on their purchasing behaviour without knowing in advance what these categories maybe. Clustering and dimensionality reduction are example of unsupervised learning problems.

The goal of unsupervised learning may be as straightforward as discovering hidden patterns and structures within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify raw data.



Unsupervised learning is more concerned with trying to extract meaningful insights or information from data rather than trying to predict some outcome based on previously available supervised training data. There is more uncertainty in the results of unsupervised learning but you can also gain a lot of information from these models that was previously unavailable to view just by looking at the raw data.

Some of the most popular unsupervised learning algorithms are: K-means clustering, Hierarchical Clustering, Principal Component Analysis etc.

Reinforcement Learning:

Reinforcement learning is a behavioural learning model. Reinforcement learning differs from other types of supervised learning because the system isn’t trained with the sample data set. Rather, the system learns through trial and error.

The learning system, called an agent in this context, can observe the environment, select and perform actions, and get rewards in return or penalties in the form of negative rewards. Therefore, a sequence of successful decisions will result in the process being “reinforced” because to get the most reward over time it best solves the problem at hand.

One of the most common applications of reinforcement learning is in robotics or game playing. Reinforcement learning is also the algorithm that is being used for self-driving cars.

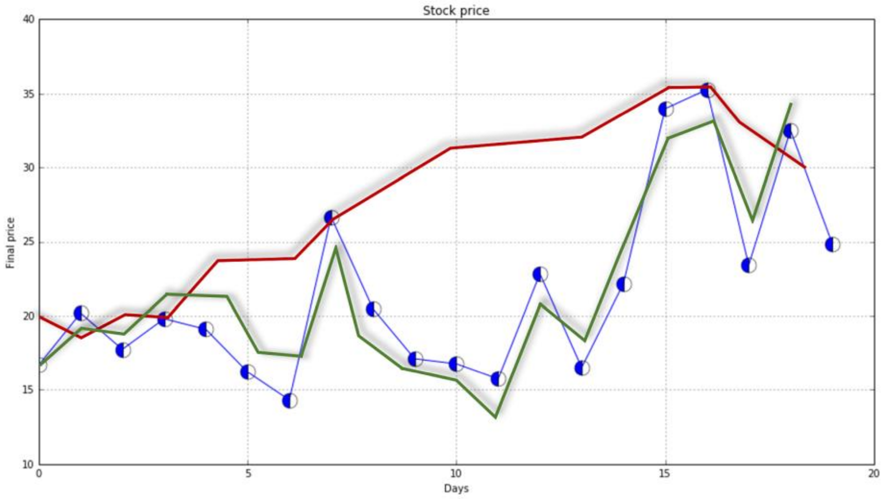
This list is by no means comprehensive, and learning algorithms are applied to new applications every day.

0.3. Types of machine learning algorithms:

Selecting the right algorithm is part science and part art. Two data scientists tasked with solving the same business challenge may choose different algorithms to approach the same problem. However, understanding different classes of machine learning algorithms helps data scientists identify the best types of algorithms. Some major classes of Machine Learning problems are:

- **Regression:** *Predict a real value for each item*.

In this type of problem, the output is a continuous quantity. So, for example, if you want to predict the sales of a product given the advertisement expenses, it is a Regression problem.

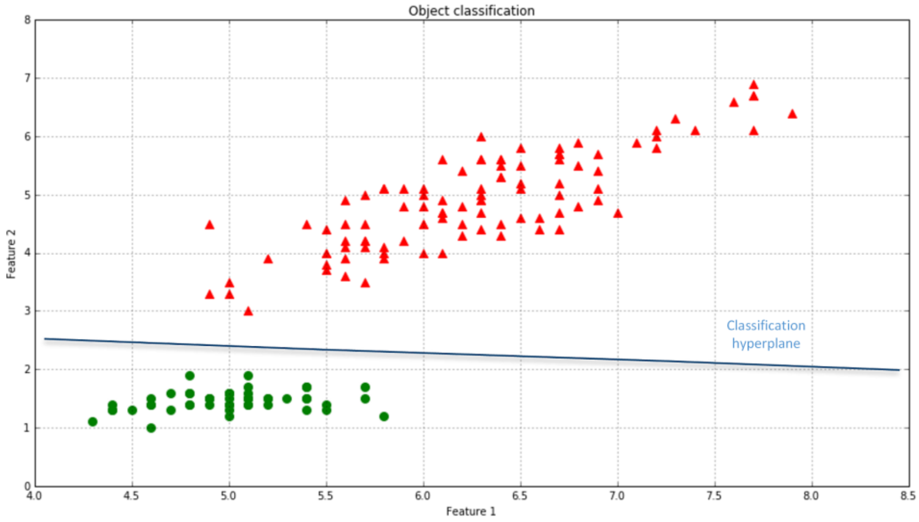


The main objective is value estimation. Regression based methods are trained on input data samples having output responses that are continuous numeric values unlike classification.

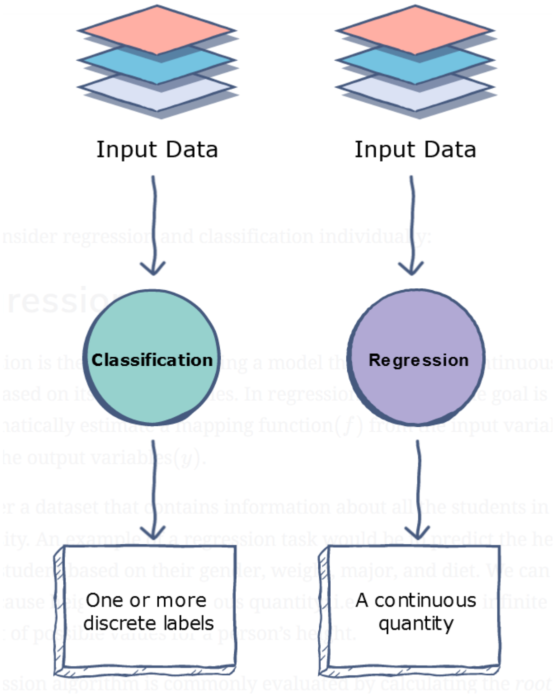
- **Classification:** *Assign a category to each item*.

In this type, the output is a categorical value. For example, Classifying emails into two classes, spam and non-spam is a classification problem. Another example is of image classification, where the model may assign items with categories such as landscape, portrait, or animal.

Output here are also known as classes or class labels which are categorical in nature meaning they are unordered and discrete values. Thus, each output response belongs to a specific discrete class or category.



Difference between a Classification and a Regression Problem:

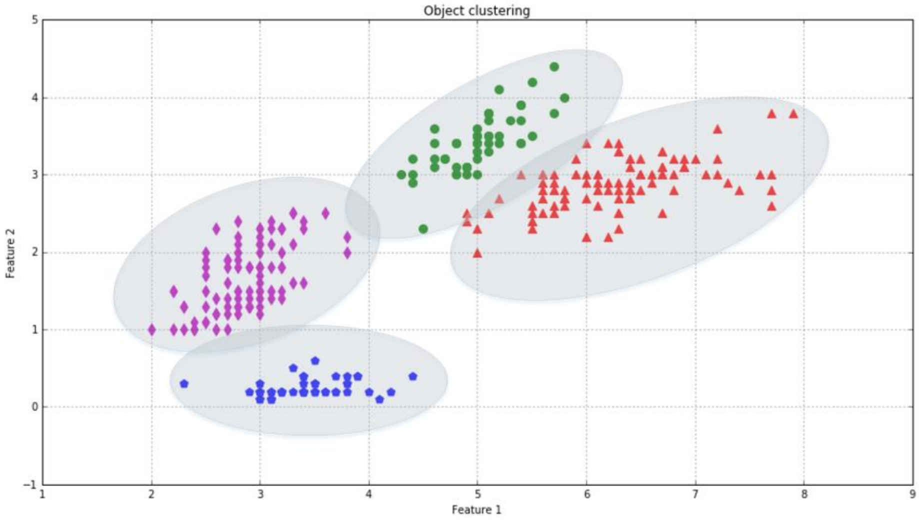


- **Clustering:** *Partition items into homogeneous regions*.

Clustering is often performed to find patterns of similarity and relationships among data samples in our dataset and then cluster these samples into various groups, such that each group or cluster of data samples has some similarity, based on the inherent attributes or features.

This type of problem involves assigning the input into two or more clusters based on feature similarity. For example, clustering customers into similar groups based on their interests, spending etc.

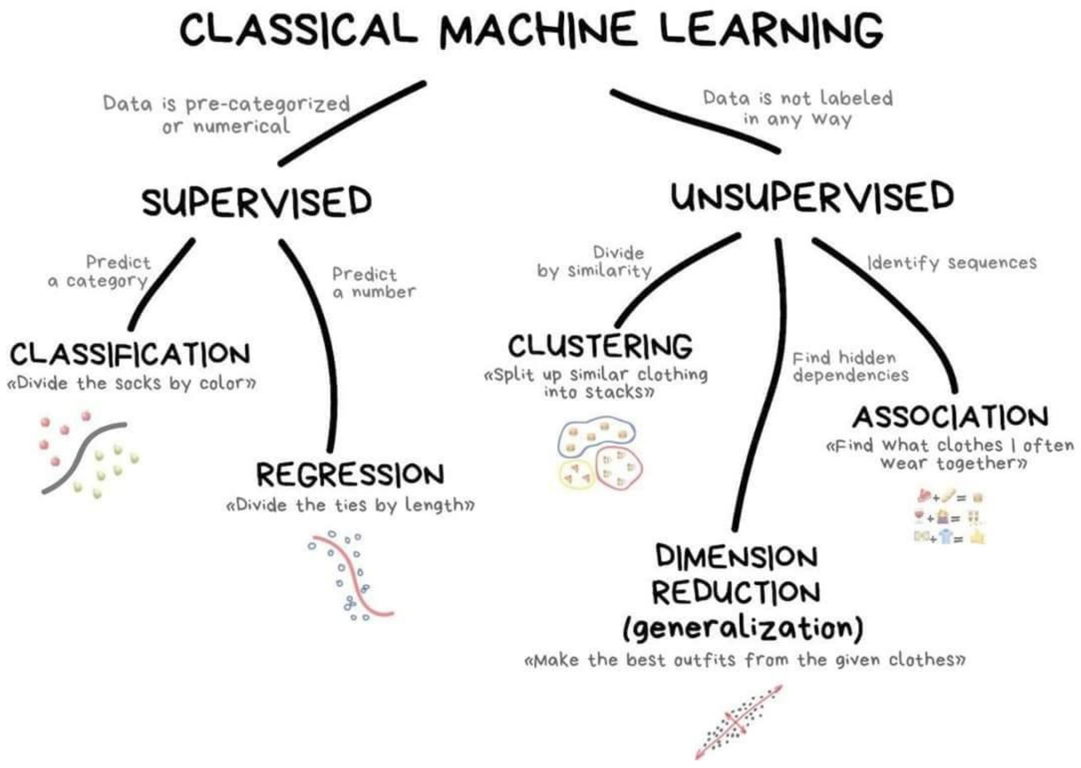
The main objective is always to create clusters such that elements of each cluster are near each other and far apart from elements of other clusters.



- **Dimensionality Reduction:** *Transformation into lower dimension*

These algorithms transform an initial representation of items into a lower-dimensional representation of these items while preserving some properties of the initial representation.

The goal is to simplify the data without losing too much information. One way to do this is to merge several correlated features into one. For example, a car's mileage may be very correlated with its age, so the dimensionality reduction algorithm will merge them into one feature that represents the car's wear and tear. This is called feature extraction.


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
Divide by similarity
Identify sequences
Find hidden dependencies
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

What do you think of the following problems?

Problem Statement 1: Study a bank credit dataset and make a decision about whether to approve the loan of an applicant based on his socio-economic profile.

Problem Statement 2: To study the House Sales dataset and build a Machine Learning model that predicts the house pricing index.

0.4. Machine Learning Pipeline

The objective of Machine Learning, data mining, or artificial intelligence is to make our lives easier, automate tasks, and take better decisions. The best way to solve a real-world Machine Learning or analytics problem is to use a Machine Learning pipeline starting from getting your data to transforming it into information and insights using Machine Learning algorithms and techniques. A Machine Learning pipeline will mainly consist of elements related to data retrieval and extraction, preparation, modelling, evaluation, and deployment.

The major steps in the pipeline are briefly mentioned here:

1. Data retrieval: This is mainly data collection, extraction, and acquisition from various data sources and data stores.

2. Data preparation: In this step, we pre-process the data, clean it, wrangle it, and manipulate it as needed. Initial exploratory data analysis is also carried out. Next steps involved extracting, engineering, and selecting features/attributes from the data.

   (i) Data processing and wrangling: Mainly concerned with data processing, cleaning, munging, wrangling and performing initial descriptive and exploratory data analysis.

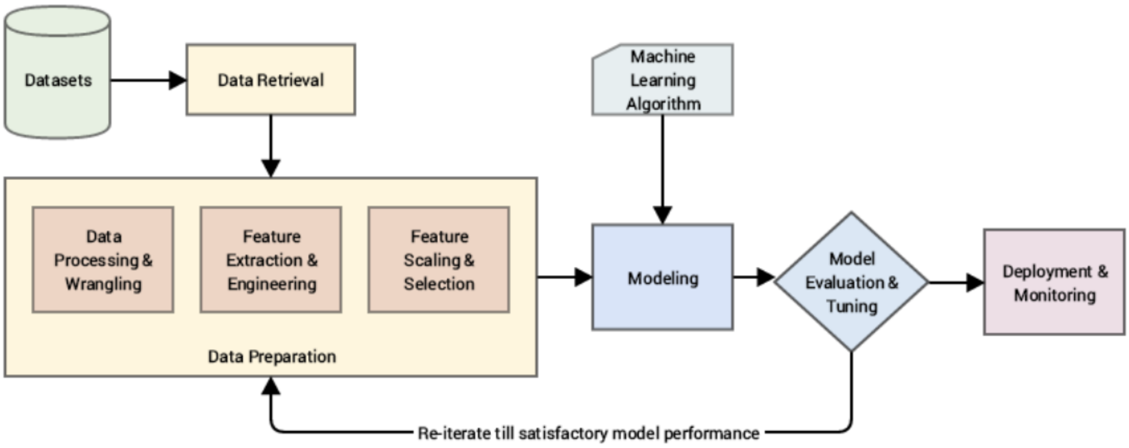
   (ii) Feature extraction and engineering: Here, we extract important features or attributes from the raw data and even create or engineer new features from existing features.

   (iii) Feature scaling and selection: Data features often need to be normalized and scaled to prevent Machine Learning algorithms from getting biased. Besides this, often we need to select a subset of all available features based on feature importance and quality. This process is known as feature selection.

3. Modelling: In the process of modelling, we usually feed the data features to a Machine Learning method or algorithm and train the model, typically to optimize a specific cost function in most cases with the objective of reducing errors and generalizing the representations learned from the data.

4. Model evaluation and tuning: Built models are evaluated and tested on validation datasets and, based on metrics like accuracy, F1 score, and others, the model performance is evaluated. Models have various parameters that are tuned in a process called hyper-parameter optimization to get models with the best and optimal results.

5. Deployment and monitoring: Selected models are deployed in production and are constantly monitored based on their predictions and results.

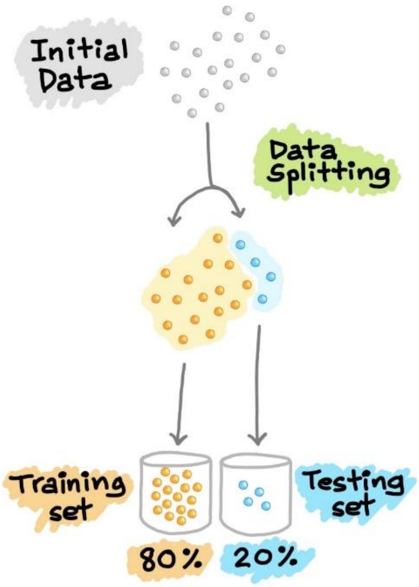

Datasets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

To sum it up, take a look at the above figure. A Machine Learning process begins by feeding the machine lots of data, by using this data the machine is trained to detect hidden insights and trends. These insights are then used to build a Machine Learning Model by using an algorithm in order to solve a problem.

0.5. Measuring Success:

• Training and Testing Data

Before we can apply our model to real-world data, we need to know whether it actually works i.e., whether we should trust its predictions.



Unfortunately, we cannot use the data we used to build the model to evaluate it. This is because our model can always simply remember the whole training set, and will therefore always predict the correct label for any point in the training set. This "remembering" does not indicate to us whether our model will generalize well (in other words, whether it will also perform well on new data).

To assess the model's performance, we show it new data (data that it hasn't seen before) for which we have labels. This is usually done by splitting the labelled data we have collected into two parts. One part of the data is used to build our machine learning model, and is called the training data or training set. The rest of the data will be used to assess how well the model works; this is called the test data.

• Generalization, Overfitting, and Underfitting

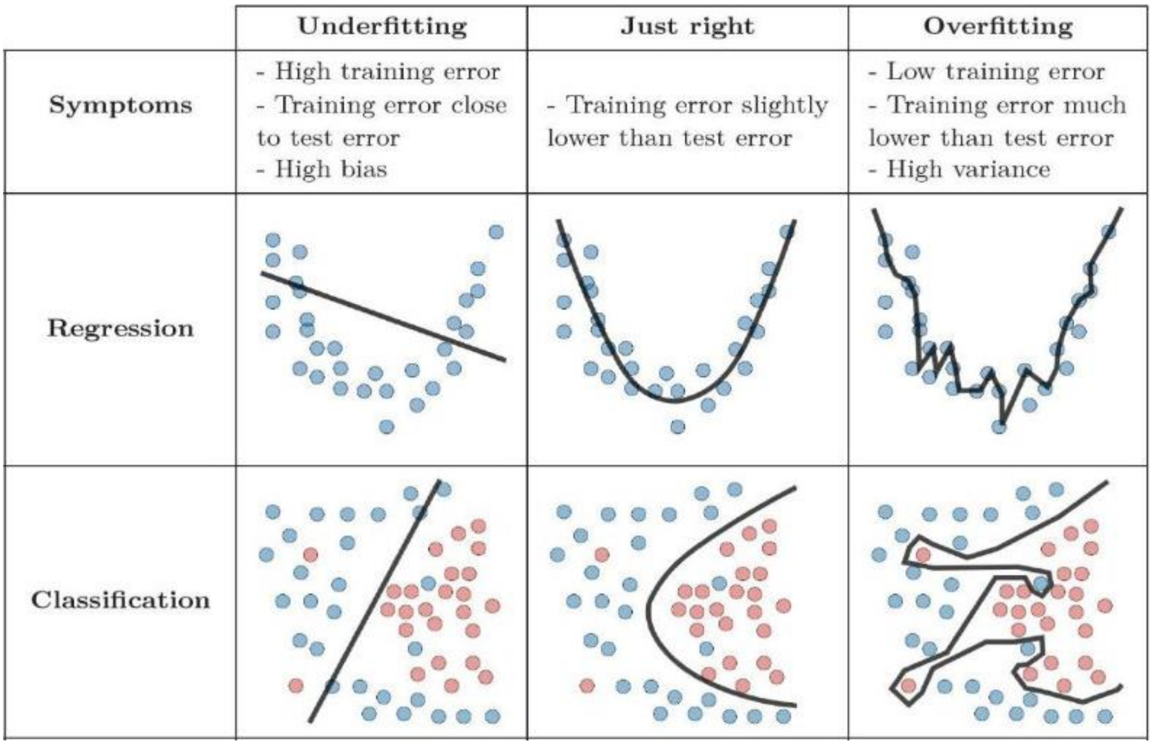
In machine learning, we want to build a model on the training data and then be able to make accurate predictions on new, unseen data that has the same characteristics as the training set that we used. If a model is able to make accurate predictions on unseen data, we say it is able to generalize from the training set to the test set. We want to build a model that is able to generalize as accurately as possible.

Usually we build a model in such a way that it can make accurate predictions on the training set. If the training and test sets have enough in common, we expect the model to also be accurate on the test set. However, there are some cases where this can go wrong. For example, if we allow ourselves to build very complex models, we can always be as accurate as we like on the training set.

Building a model that is too complex for the amount of information we have is called **Overfitting**. Overfitting occurs when you fit a model too closely to the particularities of the training set and obtain a model that works well on the training set but is not able to generalize to new data.

On the other hand, if your model is too simple then you might not be able to capture all the aspects of and variability in the data, and your model will do badly even on the training set. Choosing too simple a model is called **Underfitting**.

The more complex we allow our model to be, the better we will be able to predict on the training data. However, if our model becomes too complex, we start focusing too much on each individual data point in our training set, and the model will not generalize well to new data. There is a sweet spot in between that will yield the best generalization performance. This is the model we want to find.



	Underfitting	Just right	Overfitting
Symptoms	- High training error - Training error close to test error - High bias	- Training error slightly lower than test error	- Low training error - Training error much lower than test error - High variance
Regression			
Classification			

0.6. Essential Libraries and Tools

- Jupyter Notebook
- NumPy
- Pandas
- SciPy
- Matplotlib
- Scikit-learn

Scikit-learn is an open source project. Scikit-learn depends on two other Python packages, NumPy and SciPy. For plotting and interactive development, you should also install matplotlib, IPython, and the Jupyter Notebook.

The scikit-learn project is constantly being developed and improved, and it has a very active user community. It contains a number of state-of-the-art machine learning algorithms, as well as comprehensive documentation about each algorithm. scikit-learn is a very popular tool, and the most prominent Python library for machine learning.