**Machine learning** is the process of making systems/machines learn and improve by themselves, by being specifically programmed.

The ultimate goal of machine learning is to design [algorithms](https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article" \t "_blank" \o "algorithms) that automatically help a system gather data and use that data to learn more. Systems are expected to look for patterns in the data collected and use them to make vital decisions for themselves.

**Machine learning is a subset of AI** which allows a machine to automatically learn from past data without programming explicitly.

Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns and supports the development of new products. Many of today's leading companies, such as Facebook, Google, and Uber, make machine learning a central part of their operations.

From targeted ads to even cancer cell recognition, machine learning is everywhere. In the real world, there are existing machine learning models capable of tasks like:

* Separating spam from actual emails, as seen in Gmail
* Correcting grammar and spelling mistakes, as seen in auto correct
* Object and image recognition
* Detecting fake news
* Understanding written or spoken words
* Bots on websites that interact with humans, like humans
* Self-driven car

**Types of Machine Learning**

There are three basic types of algorithms: [supervised](https://www.techtarget.com/searchenterpriseai/definition/supervised-learning) learning, [unsupervised](https://www.techtarget.com/whatis/definition/unsupervised-learning) learning, and reinforcement learning. The type of algorithm data scientists choose to use depends on what type of data they want to predict.

* **Supervised learning:** In this type of machine learning, [data scientists](https://www.techtarget.com/searchenterpriseai/definition/data-scientist) supply algorithms with both the inputs and the outputs, and the algorithm must find a method to determine how to arrive at those inputs and outputs. That means it is mapping the inputs to corresponding outputs.

The Supervised learning data comprises: Classification, Regression, and Forecasting.

* **Unsupervised learning:**

Here, the machine learning algorithm studies data to identify patterns. There is no answer key or human operator to provide instruction. Instead, the machine determines the correlations and relationships by analyzing available data.

In an unsupervised learning process, the machine learning algorithm is left to interpret large data sets and address that data accordingly. The algorithm tries to organize that data in some way to describe its structure. This might mean grouping the data into clusters or arranging it in a way that looks more organized.

The unsupervised learning has : **Clustering and** **Dimension reduction**

* **Reinforcement learning:**Data scientists typically use [reinforcement learning](https://www.techtarget.com/searchenterpriseai/definition/reinforcement-learning) to teach a machine to complete a multi-step process for which there are clearly defined rules. By defining the rules, the machine learning algorithm then tries to explore different options and possibilities, monitoring and evaluating each result to determine which one is optimal.

**Choosing the right machine learning algorithm depends on several factors,** including, but not limited to data size, quality, and diversity, as well as what answers businesses want to derive from that data.

Additional considerations include accuracy, training time, parameters, data points, and much more.

Therefore, choosing the right algorithm is both a combination of business need, specification, experimentation, and time available. Even the most experienced data scientists cannot tell you which algorithm will perform the best before experimenting with others.

## Machine Learning Steps

The task of imparting intelligence to machines seems daunting and impossible. But it is easy. It can be broken down into 6 major steps :

**Step 1: Collect Data**

Data is the collection of information.

Types of Data:

QUALITATIVE DATA---nominal (eye color, blood groups, skin color)

---ordinal (agree/disagree, yes/no) data)

QUANTITATIVE DATA---Discrete (countable data like 200 stocks,450 people)

---continuous data (stocks, weights, temperature)

* This step is very important because the quality and quantity of data that you gather will directly determine how good your predictive model can be.
* You may have the information in an existing database or you must create it from scratch.
* It is also common to use the web scraping technique to automatically collect information from various sources.

### **Step 2: Preparing the Data/ Data Preprocessing:**

Sometimes the data we collect needs other forms of adjusting and manipulation. Things like normalization, error correction, and more. These would all happen at the data preparation step.

Other terms for this are Data Wrangling and Data Mugging.

After you have your data, you have to prepare the data.

You can do this by putting together all the data you have and randomizing it. This helps make sure that data is evenly distributed. It cleans the data to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc.

**Step 3: Data Visualization**

Now is a good time to [visualize your data](https://analyticsindiamag.com/how-to-get-started-with-visual-ai-the-new-automl-solution-by-datarobot/) and check if there are correlations between the different characteristics that we obtained

* [Visualize the data](https://www.simplilearn.com/data-visualization-article" \t "_blank" \o "Visualize the data) to understand how it is structured and understand the relationship between various variables and classes present.
* Splitting the cleaned data into two sets - a training set and a testing set. The training set is the set your model learns from. A testing set is used to check the accuracy of your model after training.

You must also separate the data into two groups: one for training and the other for model evaluation which can be divided approximately in a ratio of 80/20 but it can vary depending on the case and the volume of data we have.

At this stage, you can also pre-process your data by normalizing, eliminating duplicates, and making error corrections.

**Step 4: Choose the model**

The next step in our work flow is choosing a model. There are many models that researchers and data scientists have created over the years. Some are very well suited for image data, others for sequences (like text, or music), some for numerical data, and others for text-based data.

This means, there are several models that you can choose according to the objective that you might have: you will use algorithms like supervised([regression](https://analyticsindiamag.com/ann-with-linear-regression/)), unsupervised ([clustering](https://analyticsindiamag.com/comparison-of-k-means-hierarchical-clustering-in-customer-segmentation/))

There are various models to be used depending on the data you are going to process such as images, sound, text, and numerical values.

**Step 5: Train and Test your machine model**

You will need to train the data sets to run smoothly and see an incremental improvement in the prediction rate. Once training is complete, it’s time to see if the model is any good, using **Evaluation/Testing**. This is where that data set that we set aside earlier comes into play. Evaluation allows us to test our model against data that has never been used for training. This metric allows us to see how the model might perform against data that it has not yet seen. This is meant to be representative of how the model might perform in the real world.

A good rule of thumb for a training-evaluation split is somewhere on the order of 80/20 or 70/30.

You will have to check the machine created against your evaluation data set that contains inputs that the model does not know and verify the precision of your already trained model. If the accuracy is less than or equal to 50%, that model will not be useful.

**Step 6: Deployment of the model**

Deployment is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data. It is one of the last stages in the machine learning life cycle and can be one of the most cumbersome.

There are various ways to deploy our model.

1. Cloud Deployment like AWS,Azure etc.
2. Database Deployment in SQL,Mangodb
3. Github
4. Heroku
5. Pythonanywhere

In Conclusion, the tools or software used in each step is as follows:

**Data Collection:**

Software/Tools used are kessel , Stats,Web Scrapping

**Data PreProcessing:**

Software/Tools used are Pandas, Numpy, Math, Stats, ETL Tools, Excel

**Data Visualization:**

Software and Tools are Seaborn, Matplotlib, PowerBI,Tableau

**All three together called**

**Exploratory Data Analysis**

(EDA)

**Model selection:** Software/Tools used are SkiLearn

**Train & Test:** Tools used are SkiLearn

**Deployment:** Software/Tools used are AWS, Flask, Django github, SQL