1. Explain the term machine learning, and how does it work? Explain two machine learning applications in the business world. What are some of the ethical concerns that machine learning applications could raise?

Machine learning is one of the most exciting technologies that one would have ever come across. As is evident from the name, it gives the computer that which makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

Today, companies are using Machine Learning to improve business decisions, increase productivity, detect disease, forecast weather, and do many more things. With the exponential growth of technology, we not only need better tools to understand the data we currently have, but we also need to prepare ourselves for the data we will have. To achieve this goal we need to build intelligent machines. We can write a program to do simple things. But most of the time, Hardwiring Intelligence in it is difficult. The best way to do it is to have some way for machines to learn things themselves. A mechanism for learning – if a machine can learn from input then it does the hard work for us. This is where [Machine Learning](https://www.geeksforgeeks.org/machine-learning/) comes into action. Some of the most common examples are:

* Image Recognition
* Speech Recognition
* Recommender Systems
* Fraud Detection
* Self Driving Cars
* Medical Diagnosis
* Stock Market Trading
* Virtual Try On

**Image Recognition**

Image Recognition is one of the reasons behind the boom one could have experienced in the field of [Deep Learning](https://www.geeksforgeeks.org/deep-learning-tutorial/). The task which started from classification between cats and dog images has now evolved up to the level of Face Recognition and real-world use cases based on that like employee attendance tracking.

Also, [image recognition](https://www.geeksforgeeks.org/image-recognition-with-mobilenet/) has helped revolutionized the healthcare industry by employing smart systems in disease recognition and diagnosis methodologies.

**Speech Recognition**

Speech Recognition based smart systems like Alexa and Siri have certainly come across and used to communicate with them. In the backend, these systems are based basically on Speech Recognition systems. These systems are designed such that they can convert voice instructions into text.

One more application of the Speech recognition that we can encounter in our day-to-day life is that of performing Google searches just by speaking to it.

**Recommender Systems**

As our world has digitalized more and more approximately every tech giants try to provide customized services to its users. This application is possible just because of the [recommender systems](https://www.geeksforgeeks.org/data-mining-and-recommender-systems/) which can analyze a user’s preferences and search history and based on that they can recommend content or services to them.

An example of these services is very common for example youtube. It recommends new videos and content based on the user’s past search patterns. Netflix recommends movies and series based on the interest provided by users when someone creates an account for the very first time.

**Fraud Detection**

In today’s world, most things have been digitalized varying from buying toothbrushes or making transactions of millions of dollars everything is accessible and easy to use. But with this process of digitization cases of [fraudulent transactions](https://www.geeksforgeeks.org/online-payment-fraud-detection-using-machine-learning-in-python/) and fraudulent activities have increased. Identifying them is not that easy but machine learning systems are very efficient in these tasks.

Due to these applications only whenever the system detects red flags in a user’s activity than a suitable notification be provided to the administrator so, that these cases can be monitored properly for any spam or fraud activities.

**Self Driving Cars**

It would have been assumed that there is certainly some ghost who is driving a car if we ever saw a car being driven without a driver but all thanks to machine learning and deep learning that in today’s world, this is possible and not a story from some fictional book. Even though the algorithms and tech stack behind these technologies are highly advanced but at the core it is machine learning which has made these applications possible.

The most common example of this use case is that of the Tesla cars which are well-tested and proven for autonomous driving.

**Medical Diagnosis**

If you are a machine learning practitioner or even if you are a student then you must have heard about projects like [breast cancer Classification](https://www.geeksforgeeks.org/ml-kaggle-breast-cancer-wisconsin-diagnosis-using-knn/), [Parkinson’s Disease Classification](https://www.geeksforgeeks.org/parkinson-disease-prediction-using-machine-learning-python/), [Pneumonia detection](https://www.geeksforgeeks.org/pneumonia-detection-using-deep-learning/), and many more health-related tasks which are performed by machine learning models with more than 90% of accuracy.

Not even in the field of disease diagnosis in human beings but they work perfectly fine for plant disease-related tasks whether it is to predict the type of disease it is or to detect whether some disease is going to occur in the future.

**Stock Market Trading**

Stock Market has remained a hot topic among working professionals and even students because if you have sufficient knowledge of the markets and the forces which drives them then you can make fortune in this domain. Attempts have been made to create intelligent systems which can predict future price [trends](https://www.geeksforgeeks.org/what-is-a-trend-in-time-series/) and market value as well.

The ethical implications of artificial intelligence (AI) and machine learning technologies are vast and multifaceted. Here are some of the key areas of concern:

**Bias and Fairness**

* **Algorithmic Bias**: AI systems can perpetuate and even amplify existing biases if they are trained on biased data. This can lead to unfair treatment of certain groups based on race, gender, socioeconomic status, etc.
* **Transparency**: The decision-making process of AI systems can be opaque, making it difficult to understand and challenge decisions, particularly those that affect people's lives significantly.

**Privacy and Security**

* **Data Privacy**: AI systems often require large amounts of data, raising concerns about how this data is collected, stored, and used. There is a risk of infringing on individuals' privacy rights.
* **Surveillance**: The use of AI for surveillance purposes can lead to intrusive monitoring and tracking of individuals, potentially eroding civil liberties.

**Accountability and Responsibility**

* **Accountability**: Determining who is accountable for the actions of AI systems can be challenging. Is it the developers, the operators, or the AI itself?
* **Autonomous Systems**: As AI systems gain more autonomy, ensuring they act in ways that are aligned with human values and ethics becomes crucial.

**Employment and Economic Impact**

* **Job Displacement**: AI and automation have the potential to displace a significant number of jobs, leading to economic inequality and social disruption.
* **Economic Power**: The benefits of AI might be concentrated in the hands of a few large corporations, exacerbating economic disparities.

**Ethical Use of AI**

* **Purpose of AI**: Ethical considerations about the purposes for which AI is used are crucial. For instance, the use of AI in military applications or in generating deepfake videos raises significant ethical questions.
* **Human-AI Interaction**: As AI systems become more integrated into daily life, ensuring that they enhance rather than detract from human well-being is important.

**Long-term Impact and Existential Risk**

* **Superintelligent AI**: There are concerns about the long-term impact of creating superintelligent AI systems that could surpass human intelligence and control.
* **Control and Governance**: Developing frameworks for the control and governance of AI technologies is essential to ensure they are used for the common good and do not pose existential risks.

Addressing these ethical implications requires a multidisciplinary approach involving ethicists, technologists, policymakers, and the public to create robust guidelines and regulations that promote the responsible development and use of AI technologies.

2. Describe the process of human learning:

i. Under the supervision of experts

ii. With the assistance of experts in an indirect manner

iii. Self-education

Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Supervised learning and unsupervised learning are two main types of[machine learning](https://www.geeksforgeeks.org/machine-learning/).

In [supervised learning](https://www.geeksforgeeks.org/supervised-machine-learning/), the machine is trained on a set of labeled data, which means that the input data is paired with the desired output. The machine then learns to predict the output for new input data. Supervised learning is often used for tasks such as classification, regression, and object detection.

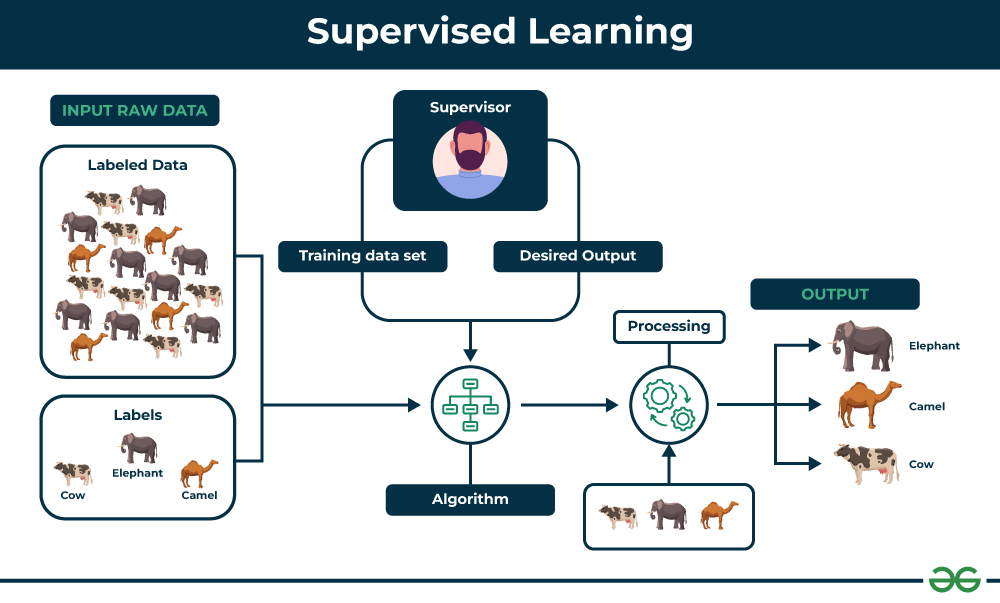
In unsupervised learning, the machine is trained on a set of unlabeled data, which means that the input data is not paired with the desired output. The machine then learns to find patterns and relationships in the data. Unsupervised learning is often used for tasks such as [clustering](https://www.geeksforgeeks.org/clustering-in-machine-learning/), dimensionality reduction, and anomaly detection.

## ****What is Supervised learning?****

Supervised learning is a type of [machine learning algorithm](https://www.geeksforgeeks.org/machine-learning-algorithms/) that learns from labeled data. Labeled data is data that has been tagged with a correct answer or classification.

Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Supervised learning is when we teach or train the machine using data that is well-labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.

For example, a labeled dataset of images of Elephant, Camel and Cow would have each image tagged with either “Elephant” , “Camel”or “Cow.”



**Key Points:**

* Supervised learning involves training a machine from labeled data.
* Labeled data consists of examples with the correct answer or classification.
* The machine learns the relationship between inputs (fruit images) and outputs (fruit labels).
* The trained machine can then make predictions on new, unlabeled data.

**Example:**

Let’s say you have a fruit basket that you want to identify. The machine would first analyze the image to extract features such as its shape, color, and texture. Then, it would compare these features to the features of the fruits it has already learned about. If the new image’s features are most similar to those of an apple, the machine would predict that the fruit is an apple.

**For instance**, suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all the different fruits one by one like this:

* If the shape of the object is rounded and has a depression at the top, is red in color, then it will be labeled as –**Apple**.
* If the shape of the object is a long curving cylinder having Green-Yellow color, then it will be labeled as –**Banana**.

Now suppose after training the data, you have given a new separate fruit, say Banana from the basket, and asked to identify it.

Since the machine has already learned the things from previous data and this time has to use it wisely. It will first classify the fruit with its shape and color and would confirm the fruit name as BANANA and put it in the Banana category. Thus the machine learns the things from training data(basket containing fruits) and then applies the knowledge to test data(new fruit).

## Types of Supervised Learning

Supervised learning is classified into two categories of algorithms:

* [**Regression**:](https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/) A regression problem is when the output variable is a real value, such as “dollars” or “weight”.
* [**Classification**:](https://www.geeksforgeeks.org/getting-started-with-classification/) A classification problem is when the output variable is a category, such as “Red” or “blue” , “disease” or “no disease”.

Supervised learning deals with or learns with “labeled” data. This implies that some data is already tagged with the correct answer.

### ****1- Regression****

Regression is a type of supervised learning that is used to predict continuous values, such as house prices, stock prices, or customer churn. Regression algorithms learn a function that maps from the input features to the output value.

Some common [regression algorithms](https://www.geeksforgeeks.org/types-of-regression-techniques/) include:

* Linear Regression
* Polynomial Regression
* Support Vector Machine Regression
* Decision Tree Regression
* Random Forest Regression

### ****2- Classification****

Classification is a type of supervised learning that is used to predict categorical values, such as whether a customer will churn or not, whether an email is spam or not, or whether a medical image shows a tumor or not. Classification algorithms learn a function that maps from the input features to a probability distribution over the output classes.

Some common[classification algorithms](https://www.geeksforgeeks.org/top-6-machine-learning-algorithms-for-classification/) include:

* Logistic Regression
* Support Vector Machines
* Decision Trees
* Random Forests
* Naive Baye

### Evaluating Supervised Learning Models

Evaluating supervised learning models is an important step in ensuring that the model is accurate and generalizable. There are a number of different [metrics](https://www.geeksforgeeks.org/metrics-for-machine-learning-model/) that can be used to evaluate supervised learning models, but some of the most common ones include:

#### For Regression

* **Mean Squared Error (MSE):** MSE measures the average squared difference between the predicted values and the actual values. Lower MSE values indicate better model performance.
* **Root Mean Squared Error (RMSE):** RMSE is the square root of MSE, representing the standard deviation of the prediction errors. Similar to MSE, lower RMSE values indicate better model performance.
* **Mean Absolute Error (MAE):** MAE measures the average absolute difference between the predicted values and the actual values. It is less sensitive to outliers compared to MSE or RMSE.
* **R-squared (Coefficient of Determination):** R-squared measures the proportion of the variance in the target variable that is explained by the model. Higher R-squared values indicate better model fit.

#### For Classification

* **Accuracy:** Accuracy is the percentage of predictions that the model makes correctly. It is calculated by dividing the number of correct predictions by the total number of predictions.
* **Precision:** Precision is the percentage of positive predictions that the model makes that are actually correct. It is calculated by dividing the number of true positives by the total number of positive predictions.
* **Recall:** Recall is the percentage of all positive examples that the model correctly identifies. It is calculated by dividing the number of true positives by the total number of positive examples.
* **F1 score:** The F1 score is a weighted average of precision and recall. It is calculated by taking the harmonic mean of precision and recall.
* **Confusion matrix:** A confusion matrix is a table that shows the number of predictions for each class, along with the actual class labels. It can be used to visualize the performance of the model and identify areas where the model is struggling.

### ****Applications of Supervised learning****

Supervised learning can be used to solve a wide variety of problems, including:

* **Spam filtering:** Supervised learning algorithms can be trained to identify and classify spam emails based on their content, helping users avoid unwanted messages.
* **Image classification:** Supervised learning can automatically classify images into different categories, such as animals, objects, or scenes, facilitating tasks like image search, content moderation, and image-based product recommendations.
* **Medical diagnosis:** Supervised learning can assist in medical diagnosis by analyzing patient data, such as medical images, test results, and patient history, to identify patterns that suggest specific diseases or conditions.
* **Fraud detection:** Supervised learning models can analyze financial transactions and identify patterns that indicate fraudulent activity, helping financial institutions prevent fraud and protect their customers.
* **Natural language processing (NLP):** Supervised learning plays a crucial role in NLP tasks, including sentiment analysis, machine translation, and text summarization, enabling machines to understand and process human language effectively.

### ****Advantages of Supervised learning****

* Supervised learning allows collecting data and produces data output from previous experiences.
* Helps to optimize performance criteria with the help of experience.
* Supervised machine learning helps to solve various types of real-world computation problems.
* It performs classification and regression tasks.
* It allows estimating or mapping the result to a new sample.
* We have complete control over choosing the number of classes we want in the training data.

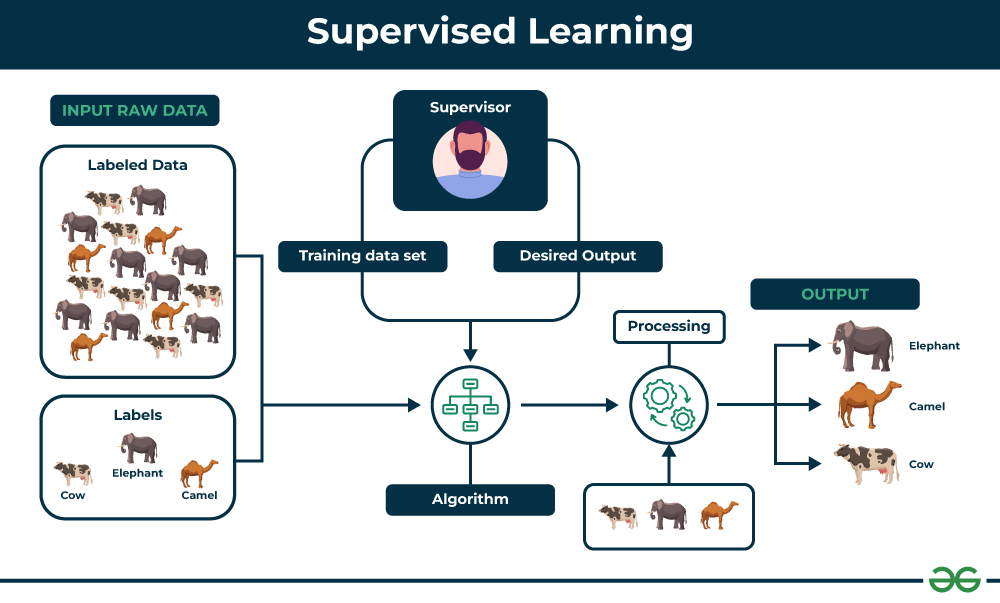
### ****Disadvantages of Supervised learning****

* Classifying big data can be challenging.
* Training for supervised learning needs a lot of computation time. So, it requires a lot of time.
* Supervised learning cannot handle all complex tasks in Machine Learning.
* Computation time is vast for supervised learning.
* It requires a labelled data set.
* It requires a training process.

3. Provide a few examples of various types of machine learning.

### 1. Supervised Machine Learning

[Supervised learning](https://www.geeksforgeeks.org/supervised-machine-learning/) is defined as when a model gets trained on a **“Labelled Dataset”**. Labelled datasets have both input and output parameters. In **Supervised Learning** algorithms learn to map points between inputs and correct outputs. It has both training and validation datasets labelled.



*Supervised Learning*

Let’s understand it with the help of an example.

**Example:**Consider a scenario where you have to build an image classifier to differentiate between cats and dogs. If you feed the datasets of dogs and cats labelled images to the algorithm, the machine will learn to classify between a dog or a cat from these labeled images. When we input new dog or cat images that it has never seen before, it will use the learned algorithms and predict whether it is a dog or a cat. This is how **supervised learning** works, and this is particularly an image classification.

There are two main categories of supervised learning that are mentioned below:

* [Classification](https://www.geeksforgeeks.org/getting-started-with-classification/)
* [Regression](https://www.geeksforgeeks.org/types-of-regression-techniques/)

#### Classification

[**Classification**](https://www.geeksforgeeks.org/getting-started-with-classification/)deals with predicting **categorical** target variables, which represent discrete classes or labels. For instance, classifying emails as spam or not spam, or predicting whether a patient has a high risk of heart disease. Classification algorithms learn to map the input features to one of the predefined classes.

Here are some classification algorithms:

* [**Logistic Regression**](https://www.geeksforgeeks.org/understanding-logistic-regression/)
* [**Support Vector Machine**](https://www.geeksforgeeks.org/support-vector-machine-algorithm/)
* [**Random Forest**](https://www.geeksforgeeks.org/random-forest-regression-in-python/)
* [**Decision Tree**](https://www.geeksforgeeks.org/decision-tree/)
* [**K-Nearest Neighbors (KNN)**](https://www.geeksforgeeks.org/k-nearest-neighbours/)
* [**Naive Bayes**](https://www.geeksforgeeks.org/naive-bayes-classifiers/)

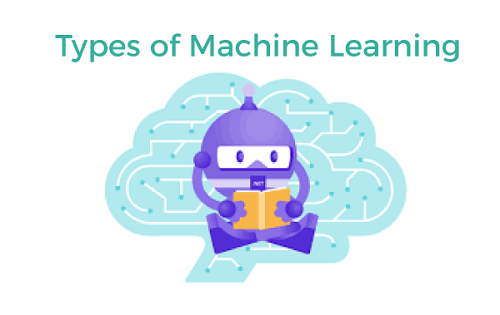
#### Regression

[**Regression**](https://www.geeksforgeeks.org/regression-classification-supervised-machine-learning/), on the other hand, deals with predicting **continuous** target variables, which represent numerical values. For example, predicting the price of a house based on its size, location, and amenities, or forecasting the sales of a product. Regression algorithms learn to map the input features to a continuous numerical value.

4. Examine the various forms of machine learning.

# Types of Machine Learning

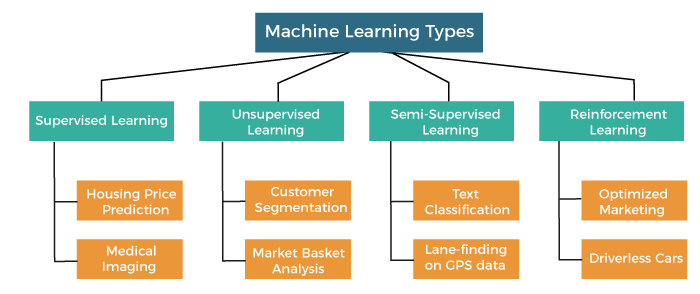
**Machine learning is a subset of AI, which enables the machine to automatically learn from data, improve performance from past experiences, and make predictions**. Machine learning contains a set of algorithms that work on a huge amount of data. Data is fed to these algorithms to train them, and on the basis of training, they build the model & perform a specific task.



These ML algorithms help to solve different business problems like Regression, Classification, Forecasting, Clustering, and Associations, etc.

Based on the methods and way of learning, machine learning is divided into mainly four types, which are:

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Semi-Supervised Machine Learning
4. Reinforcement Learning



In this topic, we will provide a detailed description of the types of Machine Learning along with their respective algorithms:

## 1. Supervised Machine Learning

As its name suggests, [Supervised machine learning](https://www.javatpoint.com/supervised-machine-learning) is based on supervision. It means in the supervised learning technique, we train the machines using the "labelled" dataset, and based on the training, the machine predicts the output. Here, the labelled data specifies that some of the inputs are already mapped to the output. More preciously, we can say; first, we train the machine with the input and corresponding output, and then we ask the machine to predict the output using the test dataset.

Let's understand supervised learning with an example. Suppose we have an input dataset of cats and dog images. So, first, we will provide the training to the machine to understand the images, such as the **shape & size of the tail of cat and dog, Shape of eyes, colour, height (dogs are taller, cats are smaller), etc.** After completion of training, we input the picture of a cat and ask the machine to identify the object and predict the output. Now, the machine is well trained, so it will check all the features of the object, such as height, shape, colour, eyes, ears, tail, etc., and find that it's a cat. So, it will put it in the Cat category. This is the process of how the machine identifies the objects in Supervised Learning.

**The main goal of the supervised learning technique is to map the input variable(x) with the output variable(y).** Some real-world applications of supervised learning are **Risk Assessment, Fraud Detection, Spam filtering,** etc.

### Categories of Supervised Machine Learning

Supervised machine learning can be classified into two types of problems, which are given below:

* **Classification**
* **Regression**

### a) Classification

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "**Yes" or No, Male or Female, Red or Blue, etc**. The classification algorithms predict the categories present in the dataset. Some real-world examples of classification algorithms are **Spam Detection, Email filtering, etc.**

Some popular classification algorithms are given below:

* **Random Forest Algorithm**
* **Decision Tree Algorithm**
* **Logistic Regression Algorithm**
* **Support Vector Machine Algorithm**

### b) Regression

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.

Some popular Regression algorithms are given below:

* **Simple Linear Regression Algorithm**
* **Multivariate Regression Algorithm**
* **Decision Tree Algorithm**
* **Lasso Regression**

### Advantages and Disadvantages of Supervised Learning

**Advantages:**

* Since supervised learning work with the labelled dataset so we can have an exact idea about the classes of objects.
* These algorithms are helpful in predicting the output on the basis of prior experience.

**Disadvantages:**

* These algorithms are not able to solve complex tasks.
* It may predict the wrong output if the test data is different from the training data.
* It requires lots of computational time to train the algorithm.

### Applications of Supervised Learning

Some common applications of Supervised Learning are given below:

* **Image Segmentation:**  
  Supervised Learning algorithms are used in image segmentation. In this process, image classification is performed on different image data with pre-defined labels.
* **Medical Diagnosis:**  
  Supervised algorithms are also used in the medical field for diagnosis purposes. It is done by using medical images and past labelled data with labels for disease conditions. With such a process, the machine can identify a disease for the new patients.
* **Fraud Detection -** Supervised Learning classification algorithms are used for identifying fraud transactions, fraud customers, etc. It is done by using historic data to identify the patterns that can lead to possible fraud.
* **Spam detection -** In spam detection & filtering, classification algorithms are used. These algorithms classify an email as spam or not spam. The spam emails are sent to the spam folder.
* **Speech Recognition -** Supervised learning algorithms are also used in speech recognition. The algorithm is trained with voice data, and various identifications can be done using the same, such as voice-activated passwords, voice commands, etc.

## 2. Unsupervised Machine Learning

[Unsupervised learnin](https://www.javatpoint.com/unsupervised-machine-learning)g is different from the Supervised learning technique; as its name suggests, there is no need for supervision. It means, in unsupervised machine learning, the machine is trained using the unlabeled dataset, and the machine predicts the output without any supervision.

In unsupervised learning, the models are trained with the data that is neither classified nor labelled, and the model acts on that data without any supervision.

**The main aim of the unsupervised learning algorithm is to group or categories the unsorted dataset according to the similarities, patterns, and differences.** Machines are instructed to find the hidden patterns from the input dataset.

Let's take an example to understand it more preciously; suppose there is a basket of fruit images, and we input it into the machine learning model. The images are totally unknown to the model, and the task of the machine is to find the patterns and categories of the objects.

So, now the machine will discover its patterns and differences, such as colour difference, shape difference, and predict the output when it is tested with the test dataset.

### Categories of Unsupervised Machine Learning

Unsupervised Learning can be further classified into two types, which are given below:

* **Clustering**
* **Association**

### 1) Clustering

The clustering technique is used when we want to find the inherent groups from the data. It is a way to group the objects into a cluster such that the objects with the most similarities remain in one group and have fewer or no similarities with the objects of other groups. An example of the clustering algorithm is grouping the customers by their purchasing behaviour.

Some of the popular clustering algorithms are given below:

* **K-Means Clustering algorithm**
* **Mean-shift algorithm**
* **DBSCAN Algorithm**
* **Principal Component Analysis**
* **Independent Component Analysis**

### 2) Association

Association rule learning is an unsupervised learning technique, which finds interesting relations among variables within a large dataset. The main aim of this learning algorithm is to find the dependency of one data item on another data item and map those variables accordingly so that it can generate maximum profit. This algorithm is mainly applied in **Market Basket analysis, Web usage mining, continuous production**, etc.

Some popular algorithms of Association rule learning are **Apriori Algorithm, Eclat, FP-growth algorithm.**

### Advantages and Disadvantages of Unsupervised Learning Algorithm

**Advantages:**

* These algorithms can be used for complicated tasks compared to the supervised ones because these algorithms work on the unlabeled dataset.
* Unsupervised algorithms are preferable for various tasks as getting the unlabeled dataset is easier as compared to the labelled dataset.

**Disadvantages:**

* The output of an unsupervised algorithm can be less accurate as the dataset is not labelled, and algorithms are not trained with the exact output in prior.
* Working with Unsupervised learning is more difficult as it works with the unlabelled dataset that does not map with the output.

### Applications of Unsupervised Learning

* **Network Analysis:** Unsupervised learning is used for identifying plagiarism and copyright in document network analysis of text data for scholarly articles.
* **Recommendation Systems:** Recommendation systems widely use unsupervised learning techniques for building recommendation applications for different web applications and e-commerce websites.
* **Anomaly Detection:** Anomaly detection is a popular application of unsupervised learning, which can identify unusual data points within the dataset. It is used to discover fraudulent transactions.
* **Singular Value Decomposition:** Singular Value Decomposition or SVD is used to extract particular information from the database. For example, extracting information of each user located at a particular location.

## 3. Semi-Supervised Learning

**Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning**. It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms and uses the combination of labelled and unlabeled datasets during the training period.

**A**lthough Semi-supervised learning is the middle ground between supervised and unsupervised learning and operates on the data that consists of a few labels, it mostly consists of unlabeled data. As labels are costly, but for corporate purposes, they may have few labels. It is completely different from supervised and unsupervised learning as they are based on the presence & absence of labels.

**To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced**. The main aim of [semi-supervised learning](https://www.javatpoint.com/semi-supervised-learning) is to effectively use all the available data, rather than only labelled data like in supervised learning. Initially, similar data is clustered along with an unsupervised learning algorithm, and further, it helps to label the unlabeled data into labelled data. It is because labelled data is a comparatively more expensive acquisition than unlabeled data.

We can imagine these algorithms with an example. Supervised learning is where a student is under the supervision of an instructor at home and college. Further, if that student is self-analysing the same concept without any help from the instructor, it comes under unsupervised learning. Under semi-supervised learning, the student has to revise himself after analyzing the same concept under the guidance of an instructor at college.

### Advantages and disadvantages of Semi-supervised Learning

**Advantages:**

* It is simple and easy to understand the algorithm.
* It is highly efficient.
* It is used to solve drawbacks of Supervised and Unsupervised Learning algorithms.

**Disadvantages:**

* Iterations results may not be stable.
* We cannot apply these algorithms to network-level data.
* Accuracy is low.

## 4. Reinforcement Learning

**Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance.** Agent gets rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agent is to maximize the rewards.

In reinforcement learning, there is no labelled data like supervised learning, and agents learn from their experiences only.

The [reinforcement learning](https://www.javatpoint.com/reinforcement-learning) process is similar to a human being; for example, a child learns various things by experiences in his day-to-day life. An example of reinforcement learning is to play a game, where the Game is the environment, moves of an agent at each step define states, and the goal of the agent is to get a high score. Agent receives feedback in terms of punishment and rewards.

Due to its way of working, reinforcement learning is employed in different fields such as **Game theory, Operation Research, Information theory, multi-agent systems.**

A reinforcement learning problem can be formalized using **Markov Decision Process(MDP).** In MDP, the agent constantly interacts with the environment and performs actions; at each action, the environment responds and generates a new state.

### Categories of Reinforcement Learning

Reinforcement learning is categorized mainly into two types of methods/algorithms:

* **Positive Reinforcement Learning:** Positive reinforcement learning specifies increasing the tendency that the required behaviour would occur again by adding something. It enhances the strength of the behaviour of the agent and positively impacts it.
* **Negative Reinforcement Learning:** Negative reinforcement learning works exactly opposite to the positive RL. It increases the tendency that the specific behaviour would occur again by avoiding the negative condition.

### Real-world Use cases of Reinforcement Learning

* **Video Games:**  
  RL algorithms are much popular in gaming applications. It is used to gain super-human performance. Some popular games that use RL algorithms are **AlphaGO** and **AlphaGO Zero**.
* **Resource Management:**  
  The "Resource Management with Deep Reinforcement Learning" paper showed that how to use RL in computer to automatically learn and schedule resources to wait for different jobs in order to minimize average job slowdown.
* **Robotics:**  
  RL is widely being used in Robotics applications. Robots are used in the industrial and manufacturing area, and these robots are made more powerful with reinforcement learning. There are different industries that have their vision of building intelligent robots using AI and Machine learning technology.
* **Text Mining**  
  Text-mining, one of the great applications of NLP, is now being implemented with the help of Reinforcement Learning by Salesforce company.

### Advantages and Disadvantages of Reinforcement Learning

**Advantages**

* It helps in solving complex real-world problems which are difficult to be solved by general techniques.
* The learning model of RL is similar to the learning of human beings; hence most accurate results can be found.
* Helps in achieving long term results.

**Disadvantage**

* RL algorithms are not preferred for simple problems.
* RL algorithms require huge data and computations.
* Too much reinforcement learning can lead to an overload of states which can weaken the results.

5. Can you explain what a well-posed learning problem is? Explain the main characteristics that must be present to identify a learning problem properly.

**Well Posed Learning Problem –** A computer program is said to learn from experience E in context to some task T and some performance measure P, if its performance on T, as was measured by P, upgrades with experience E.

Any problem can be segregated as well-posed learning problem if it has three traits –

* Task
* Performance Measure
* Experience

**Certain examples that efficiently defines the**well-posed**learning problem are –**

**1. To better filter emails as spam or not**

* Task – Classifying emails as spam or not
* Performance Measure – The fraction of emails accurately classified as spam or not spam
* Experience – Observing you label emails as spam or not spam

**2. A checkers learning problem**

* Task – Playing checkers game
* Performance Measure – percent of games won against opposer
* Experience**–** playing implementation games against itself

**3. Handwriting Recognition Problem**

* Task – Acknowledging handwritten words within portrayal
* Performance Measure – percent of words accurately classified
* Experience – a directory of handwritten words with given classifications

**4. A Robot Driving Problem**

* Task – driving on public four-lane highways using sight scanners
* Performance Measure – average distance progressed before a fallacy
* Experience – order of images and steering instructions noted down while observing a human driver

**5. Fruit Prediction Problem**

* Task – forecasting different fruits for recognition
* Performance Measure – able to predict maximum variety of fruits
* Experience – training machine with the largest datasets of fruits images

**6. Face Recognition Problem**

* Task – predicting different types of faces
* Performance Measure – able to predict maximum types of faces
* Experience – training machine with maximum amount of datasets of different face images

**7. Automatic Translation of documents**

* Task – translating one type of language used in a document to other language
* Performance Measure – able to convert one language to other efficiently
* Experience – training machine with a large dataset of different types of languages

6. Is machine learning capable of solving all problems? Give a detailed explanation of your answer.

## Real-World Problems that Machine Learning can solve

### 1. Recommending Products after Collecting Previous Data

Recommendation systems are one of the most common machine learning use cases in day-to-day life. These systems are used mainly by search engines like Google and Bing and the top eCommerce platforms like Amazon and eBay.

The ML integrated systems show a list of recommended products individually for each of their consumers. These suggestions are based on data like previous purchases, wish lists, searches, clicks, inquiries, and browsing history. This data is fed to a comprehensive ML algorithm to strike the user at the right moment and enhance their customer engagement.

### 2. Works as the Best Image and Video Recognition Tool

If you have come across features like face recognition, text detection, object detection, and landmark detection, it is because of the integration of deep learning in machine learning. When ML algorithms are trained with innovative deep learning frameworks, they can quickly identify and classify objects and make things easier for a non-native person.

MNL can also be used to determine handwritten text by segmenting a piece of writing into smaller images, each containing a single character.

### 3. Your Virtual Assistant

A virtual assistant, which is also very common as an AI assistant, is an application program that comprehends natural language voice commands and finishes the tasks for the users like searching the web, booking an appointment, etc. If you have also asked Google Assistant in your android phones to wake you up at 5 AM or asked Siri on your iPhones for directions to the nearest restaurant, then ML has also made your life easier.

Some principal personal assistants or smart assistants available in the market are Siri, Google Assistants, Alexa, Echo, and Google mini. These assistants can help you look for information by voice commands or answer your questions by searching your query on the web.

### 4. Ingenious Gaming Using ML

With the advancement in technologies, we can improve the graphics of the games and give them a mind at the same time. Lately, if you have been facing difficulties beating the bot in a chess game, then ML might take it over. Today’s games not only simply analyze your moves but are also learning how to play the game better than you by practicing numerous times. Now using your mind against such an intelligent system will surely give you brains and make you smarter at the same time.

### 5. Devising Superior Health Care Methods

Even hospitals are utilizing machine learning to cure and treat patients. Thanks to our wearable devices, doctors can get accurate data on our health from anywhere in the world and suggest an aid if they find something helpful. The integration of ML in some essential tools can quickly provide real-time insights and combine with the explosion of computing power.

It can help doctors to diagnose critically faster and more accurately. Not only this, AI is assisting in the development of new medications and treatments, predicting harmful reactions at the early stages, and working towards finding a way to lower the costs of healthcare for providers and patients.

### 6. Protecting Environment in the Most Impactful Way

Aforesaid, possibilities are endless with ML, and it’s just the beginning. Recently IBM’s Green Horizon Project was acknowledged by experts worldwide as it accurately predicted weather and pollution forecasts. We can use it to save and predict natural forecasts with the expertise of the professionals from Singapore by our side. It is helping city planners to run every kind of scenario just by feeding previous data to their ML algorithm to find ways for minimum environmental impact.

### 7. Real-Time Dynamic Pricing

You might have already encountered this scenario while booking a flight ticket to travel on Christmas or booking a cab at peak hours. You will notice a big gap between the regular pricing and pricing at that particular time. So, in these scenarios, the ML and data analysis techniques are helping businesses to get to know more about their users. It answers two critical questions.

First, how are customers reacting to surge prices? And second, whether they are looking for customers because of surge pricing? The integration of AI and ML helps the businesses and the users as it helps determine when customers are looking for the best promotional and discounted prices.

### 8. Innovations in the Finance Sector Including Stock Market

The functioning of the finance sector is about to change in the upcoming years completely. Thanks to technologies like mobile app development and machine learning, the stock market is at its all-time high.

Thanks to AI, deep learning, and machine learning, it has become easy for users to predict the market price by feeding it the previous data. It will allow traders to make better and steady decisions which means less financial loss. Not only this, the machine learning-based anomaly detection models can easily monitor your every transaction request and alert you of any kind of suspicious activity.

### 9. Commute Predictions Using Machine Learning

Almost everyone uses GPS services while driving. A programmed GPS helps us in finding the proper navigation to our destination. But it’s the integration of ML and its features like congestion analysis GPS that helps us by telling us the path to avoid traffic and reach our destination on time. It saves the data like our location and velocities of the vehicles on the same path to determine the traffic and let us know whether it would be the right path to follow.

### 10. Online Video OTT Streaming Applications

The pioneers of online video streaming services are Netflix and Amazon Prime; both of them combined have killed the traditional way of watching television. But how were they able to keep the customer engaged on their platforms? First by offering impressive content, secondly by getting personalized with the users.

They were able to capture mass audiences using machine learning. At the right time, they integrated ML in their program and fed it the user’s data like day and time they watch content, type of content they like to watch, browsing pattern, whether they like to watch trailers before they watch a movie or a show, etc. They are using practical machine learning frameworks to engage their audience by providing quality streaming service right to their homes.

7. What are the various methods and technologies for solving machine learning problems? Any two of them should be defined in detail.

The ten methods described offer an overview — and a foundation you can build on as you hone your machine learning knowledge and skill:

1. **Regression**
2. **Classification**
3. **Clustering**
4. **Dimensionality Reduction**
5. **Ensemble Methods**
6. **Neural Nets and Deep Learning**
7. **Transfer Learning**
8. **Reinforcement Learning**
9. **Natural Language Processing**
10. **Word Embeddings**

One last thing before we jump in. Let’s distinguish between two general categories of machine learning: supervised and unsupervised. We apply **supervised** ML techniques when we have a piece of data that we want to predict or explain. We do so by using previous data of inputs and outputs to predict an output based on a new input. For example, you could use supervised ML techniques to help a service business that wants to predict the number of new users who will sign up for the service next month. By contrast, **unsupervised** ML looks at ways to relate and group data points without the use of a target variable to predict. In other words, it evaluates data in terms of traits and uses the traits to form clusters of items that are similar to one another. For example, you could use unsupervised learning techniques to help a retailer that wants to segment products with similar characteristics — without having to specify in advance which characteristics to use.

# Regression

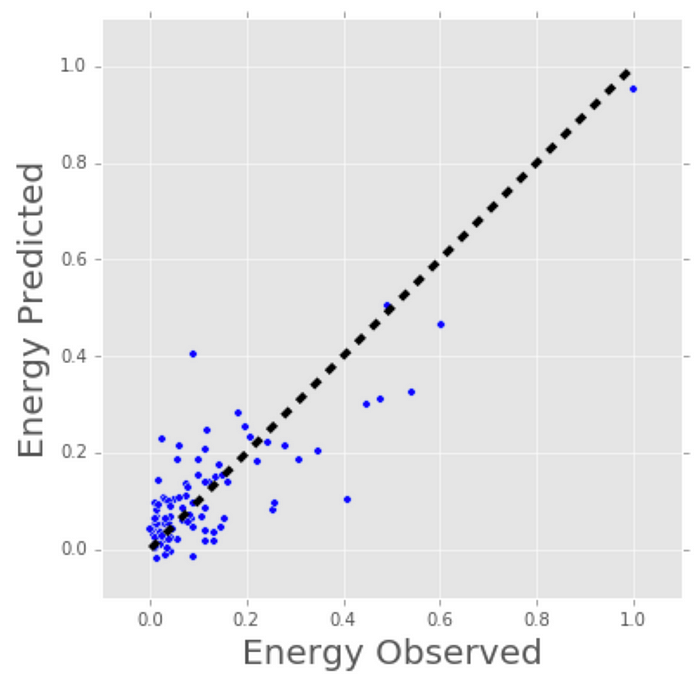
Regression methods fall within the category of supervised ML. They help to predict or explain a particular numerical value based on a set of prior data, for example predicting the price of a property based on previous pricing data for similar properties.

The simplest method is linear regression where we use the mathematical equation of the line (**y = m \* x + b**) to model a data set. We train a linear regression model with many data pairs **(x, y)** by calculating the position and slope of a line that minimizes the total distance between all of the data points and the line. In other words, we calculate the slope (**m**) and the y-intercept (**b**) for a line that best approximates the observations in the data.

Let’s consider a more a concrete example of linear regression. I once used a linear regression to predict the energy consumption (in kWh) of certain buildings by gathering together the age of the building, number of stories, square feet and the number of plugged wall equipment. Since there were more than one input (age, square feet, etc…), I used a multi-variable linear regression. The principle was the same as a simple one-to-one linear regression, but in this case the “line” I created occurred in multi-dimensional space based on the number of variables.

The plot below shows how well the linear regression model fit the actual energy consumption of building. Now imagine that you have access to the characteristics of a building (age, square feet, etc…) but you don’t know the energy consumption. In this case, we can use the fitted line to approximate the energy consumption of the particular building.

Note that you can also use linear regression to estimate the weight of each factor that contributes to the final prediction of consumed energy. For example, once you have a formula, you can determine whether age, size, or height is most important.



Linear Regression Model Estimates of Building’s Energy Consumption (kWh).

Regression techniques run the gamut from simple (like linear regression) to complex (like regularized linear regression, polynomial regression, decision trees and random forest regressions, neural nets, among others). But don’t get bogged down: start by studying simple linear regression, master the techniques, and move on from there.

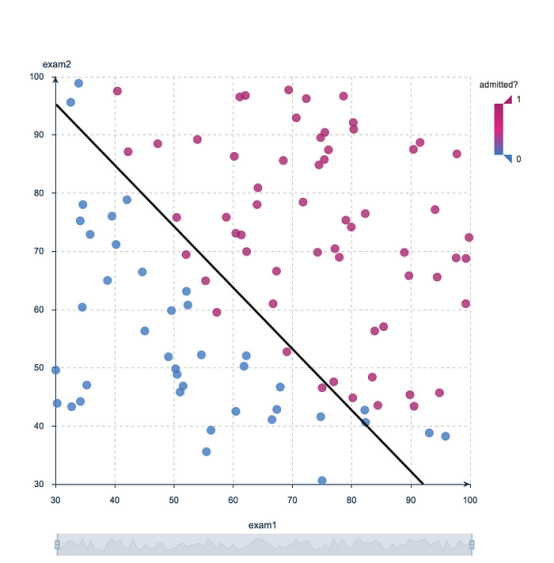
# Classification

Another class of supervised ML, classification methods predict or explain a class value. For example, they can help predict whether or not an online customer will buy a product. The output can be yes or no: buyer or not buyer. But classification methods aren’t limited to two classes. For example, a classification method could help to assess whether a given image contains a car or a truck. In this case, the output will be 3 different values: 1) the image contains a car, 2) the image contains a truck, or 3) the image contains neither a car nor a truck.

The simplest classification algorithm is logistic regression — which makes it sounds like a regression method, but it’s not. Logistic regression estimates the probability of an occurrence of an event based on one or more inputs.

For instance, a logistic regression can take as inputs two exam scores for a student in order to estimate the probability that the student will get admitted to a particular college. Because the estimate is a probability, the output is a number between 0 and 1, where 1 represents complete certainty. For the student, if the estimated probability is greater than 0.5, then we predict that he or she will be admitted. If the estimated probabiliy is less than 0.5, we predict the he or she will be refused.

The chart below plots the scores of previous students along with whether they were admitted. Logistic regression allows us to draw a line that represents the decision boundary.



Logistic Regression Decision Boundary: Admitted to College or Not?

Because logistic regression is the simplest classification model, it’s a good place to start for classification. As you progress, you can dive into non-linear classifiers such as decision trees, random forests, support vector machines, and neural nets, among others.

# Clustering

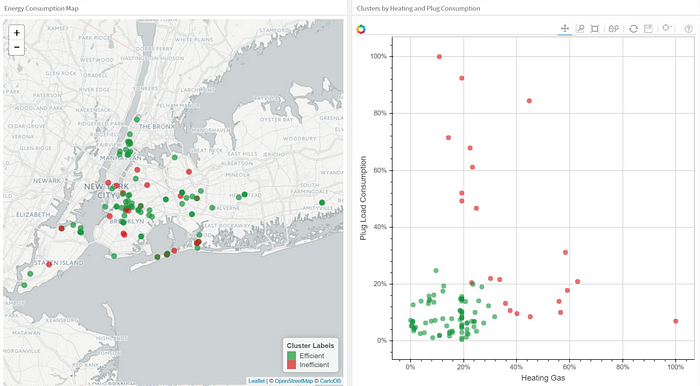
With clustering methods, we get into the category of unsupervised ML because their goal is to group or cluster observations that have similar characteristics. Clustering methods don’t use output information for training, but instead let the algorithm define the output. In clustering methods, we can only use visualizations to inspect the quality of the solution.

The most popular clustering method is K-Means, where “K” represents the number of clusters that the user chooses to create. (Note that there are various techniques for choosing the value of K, such as the elbow method.)

Roughly, what K-Means does with the data points:

1. Randomly chooses K centers within the data.
2. Assigns each data point to the closest of the randomly created centers.
3. Re-computes the center of each cluster.
4. If centers don’t change (or change very little), the process is finished. Otherwise, we return to step 2. (To prevent ending up in an infinite loop if the centers continue to change, set a maximum number of iterations in advance.)

The next plot applies K-Means to a data set of buildings. Each column in the plot indicates the efficiency for each building. The four measurements are related to air conditioning, plugged-in equipment (microwaves, refrigerators, etc…), domestic gas, and heating gas. We chose K=2 for clustering, which makes it easy to interpret one of the clusters as the group of efficient buildings and the other cluster as the group of inefficient buildings. To the left you see the location of the buildings and to right you see two of the four dimensions we used as inputs: plugged-in equipment and heating gas.



Clustering Buildings into Efficient (Green) and Inefficient (Red) Groups.

As you explore clustering, you’ll encounter very useful algorithms such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Mean Shift Clustering, Agglomerative Hierarchical Clustering, Expectation–Maximization Clustering using Gaussian Mixture Models, among others.

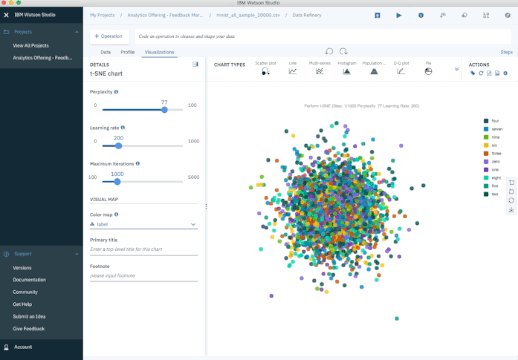
# Dimensionality Reduction

As the name suggests, we use dimensionality reduction to remove the least important information (sometime redundant columns) from a data set. In practice, I often see data sets with hundreds or even thousands of columns (also called features), so reducing the total number is vital. For instance, images can include thousands of pixels, not all of which matter to your analysis. Or when testing microchips within the manufacturing process, you might have thousands of measurements and tests applied to every chip, many of which provide redundant information. In these cases, you need dimensionality reduction algorithms to make the data set manageable.

The most popular dimensionality reduction method is Principal Component Analysis (PCA), which reduces the dimension of the feature space by finding new vectors that maximize the linear variation of the data. PCA can reduce the dimension of the data dramatically and without losing too much information when the linear correlations of the data are strong. (And in fact you can also measure the actual extent of the information loss and adjust accordingly.)

Another popular method is t-Stochastic Neighbor Embedding (t-SNE), which does non-linear dimensionality reduction. People typically use t-SNE for data visualization, but you can also use it for machine learning tasks like reducing the feature space and clustering, to mention just a few.

The next plot shows an analysis of the MNIST database of handwritten digits. MNIST contains thousands of images of digits from 0 to 9, which researchers use to test their clustering and classification algorithms. Each row of the data set is a vectorized version of the original image (size 28 x 28 = 784) and a label for each image (zero, one, two, three, …, nine). Note that we’re therefore reducing the dimensionality from 784 (pixels) to 2 (dimensions in our visualization). Projecting to two dimensions allows us to visualize the high-dimensional original data set.



t-SNE Iterations on MNIST Database of Handwritten Digits.

# Ensemble Methods

Imagine you’ve decided to build a bicycle because you are not feeling happy with the options available in stores and online. You might begin by finding the best of each part you need. Once you assemble all these great parts, the resulting bike will outshine all the other options.

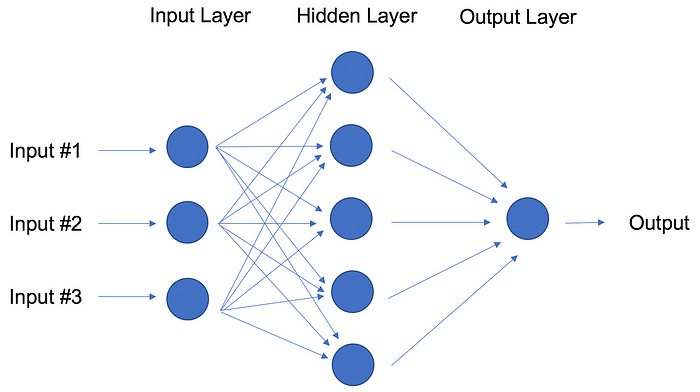
Ensemble methods use this same idea of combining several predictive models (supervised ML) to get higher quality predictions than each of the models could provide on its own. For example, the Random Forest algorithms is an ensemble method that combines many Decision Trees trained with different samples of the data sets. As a result, the quality of the predictions of a Random Forest is higher than the quality of the predictions estimated with a single Decision Tree.

Think of ensemble methods as a way to reduce the variance and bias of a single machine learning model. That’s important because any given model may be accurate under certain conditions but inaccurate under other conditions. With another model, the relative accuracy might be reversed. By combining the two models, the quality of the predictions is balanced out.

The great majority of top winners of [Kaggle](https://www.kaggle.com/" \t "_blank) competitions use ensemble methods of some kind. The most popular ensemble algorithms are Random Forest, [XGBoost](https://xgboost.readthedocs.io/" \t "_blank) and [LightGBM](https://github.com/Microsoft/LightGBM" \t "_blank).

# Neural Networks and Deep Learning

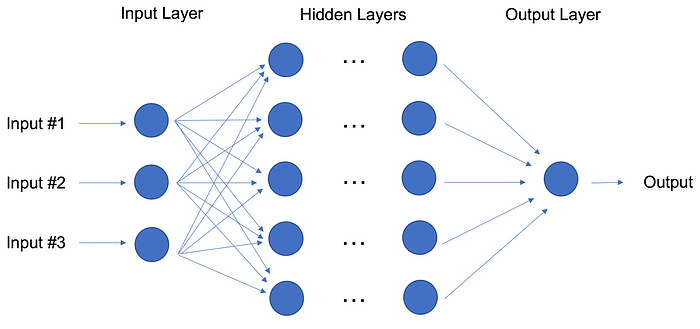
In contrast to linear and logistic regressions which are considered linear models, the objective of neural networks is to capture non-linear patterns in data by adding layers of parameters to the model. In the image below, the simple neural net has three inputs, a single hidden layer with five parameters, and an output layer.



Neural Network with One Hidden Layer.

In fact, the structure of neural networks is flexible enough to build our well-known linear and logistic regression. The term Deep learning comes from a neural net with many hidden layers (see next Figure) and encapsulates a wide variety of architectures.

It’s especially difficult to keep up with developments in deep learning, in part because the research and industry communities have doubled down on their deep learning efforts, spawning whole new methodologies every day.



Deep Learning: Neural Network with Many Hidden Layers.

For the best performance, deep learning techniques require a lot of data — and a lot of compute power since the method is self-tuning many parameters within huge architectures. It quickly becomes clear why deep learning practitioners need very powerful computers enhanced with GPUs (graphical processing units).

In particular, deep learning techniques have been extremely successful in the areas of vision (image classification), text, audio and video. The most common software packages for deep learning are [Tensorflow](https://www.tensorflow.org/" \t "_blank) and [PyTorch](https://pytorch.org/" \t "_blank).

# Transfer Learning

Let’s pretend that you’re a data scientist working in the retail industry. You’ve spent months training a high-quality model to classify images as shirts, t-shirts and polos. Your new task is to build a similar model to classify images of dresses as jeans, cargo, casual, and dress pants. Can you transfer the knowledge built into the first model and apply it to the second model? Yes, you can, using Transfer Learning.

Transfer Learning refers to re-using part of a previously trained neural net and adapting it to a new but similar task. Specifically, once you train a neural net using data for a task, you can transfer a fraction of the trained layers and combine them with a few new layers that you can train using the data of the new task. By adding a few layers, the new neural net can learn and adapt quickly to the new task.

The main advantage of transfer learning is that you need less data to train the neural net, which is particularly important because training for deep learning algorithms is expensive in terms of both time and money (computational resources) — and of course it’s often very difficult to find enough labeled data for the training.

Let’s return to our example and assume that for the shirt model you use a neural net with 20 hidden layers. After running a few experiments, you realize that you can transfer 18 of the shirt model layers and combine them with one new layer of parameters to train on the images of pants. The pants model would therefore have 19 hidden layers. The inputs and outputs of the two tasks are different but the re-usable layers may be summarizing information that is relevant to both, for example aspects of cloth.

Transfer learning has become more and more popular and there are now many solid pre-trained models available for common deep learning tasks like image and text classification.

# Reinforcement Learning

Imagine a mouse in a maze trying to find hidden pieces of cheese. The more times we expose the mouse to the maze, the better it gets at finding the cheese. At first, the mouse might move randomly, but after some time, the mouse’s experience helps it realize which actions bring it closer to the cheese.

The process for the mouse mirrors what we do with Reinforcement Learning (RL) to train a system or a game. Generally speaking, RL is a machine learning method that helps an agent learn from experience. By recording actions and using a trial-and-error approach in a set environment, RL can maximize a cumulative reward. In our example, the mouse is the agent and the maze is the environment. The set of possible actions for the mouse are: move front, back, left or right. The reward is the cheese.

You can use RL when you have little to no historical data about a problem, because it doesn’t need information in advance (unlike traditional machine learning methods). In a RL framework, you learn from the data as you go. Not surprisingly, RL is especially successful with games, especially games of “[perfect information](https://en.wikipedia.org/wiki/Perfect_information)” like chess and Go. With games, feedback from the agent and the environment comes quickly, allowing the model to learn fast. The downside of RL is that it can take a very long time to train if the problem is complex.

Just as IBM’s Deep Blue beat the best human chess player in 1997, AlphaGo, a RL-based algorithm, beat the best Go player in 2016. The current pioneers of RL are the teams at DeepMind in the UK. More on AlphaGo and DeepMind [here](https://deepmind.com/research/alphago/).

On April, 2019, the OpenAI Five team was the first AI to beat a world champion team of e-sport Dota 2, a very complex video game that the OpenAI Five team chose because there were no RL algorithms that were able to win it at the time. The same AI team that beat Dota 2’s champion human team also developed a robotic hand that can reorient a block. Read more about the OpenAI Five team [here](https://openai.com/blog/how-to-train-your-openai-five/).

You can tell that Reinforcement Learning is an especially powerful form of AI, and we’re sure to see more progress from these teams, but it’s also worth remembering the method’s limitations.

# Natural Language Processing

A huge percentage of the world’s data and knowledge is in some form of human language. Can you imagine being able to read and comprehend thousands of books, articles and blogs in seconds? Obviously, computers can’t yet fully understand human text but we can train them to do certain tasks. For example, we can train our phones to autocomplete our text messages or to correct misspelled words. We can even teach a machine to have a simple conversation with a human.

Natural Language Processing (NLP) is not a machine learning method per se, but rather a widely used technique to prepare text for machine learning. Think of tons of text documents in a variety of formats (word, online blogs, ….). Most of these text documents will be full of typos, missing characters and other words that needed to be filtered out. At the moment, the most popular package for processing text is [NLTK](https://www.nltk.org/) (Natural Language ToolKit), created by researchers at Stanford.

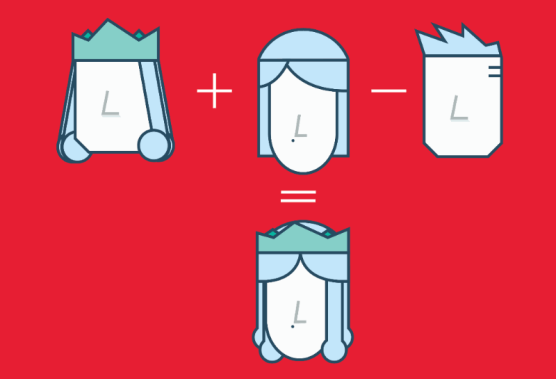
The simplest way to map text into a numerical representation is to compute the frequency of each word within each text document. Think of a matrix of integers where each row represents a text document and each column represents a word. This matrix representation of the word frequencies is commonly called Term Frequency Matrix (TFM). From there, we can create another popular matrix representation of a text document by dividing each entry on the matrix by a weight of how important each word is within the entire corpus of documents. We call this method Term Frequency Inverse Document Frequency (TFIDF) and it typically works better for machine learning tasks.

# Word Embeddings

TFM and TFIDF are numerical representations of text documents that only consider frequency and weighted frequencies to represent text documents. By contrast, word embeddings can capture the context of a word in a document. With the word context, embeddings can quantify the similarity between words, which in turn allows us to do arithmetic with words.

Word2Vec is a method based on neural nets that maps words in a corpus to a numerical vector. We can then use these vectors to find synonyms, perform arithmetic operations with words, or to represent text documents (by taking the mean of all the word vectors in a document). For example, let’s assume that we use a sufficiently big corpus of text documents to estimate word embeddings. Let’s also assume that the words king, queen, man and woman are part of the corpus. Let say that vector(‘word’) is the numerical vector that represents the word ‘word’. To estimate vector(‘woman’), we can perform the arithmetic operation with vectors:

*vector(‘king’) + vector(‘woman’)* — *vector(‘man’) ~ vector(‘queen’)*



Arithmetic with Word (Vectors) Embeddings.

Word representations allow finding similarities between words by computing the cosine similarity between the vector representation of two words. The cosine similarity measures the angle between two vectors.

We compute word embeddings using machine learning methods, but that’s often a pre-step to applying a machine learning algorithm on top. For instance, suppose we have access to the tweets of several thousand Twitter users. Also suppose that we know which of these Twitter users bought a house. To predict the probability of a new Twitter user buying a house, we can combine Word2Vec with a logistic regression.

8. Can you explain the various forms of supervised learning? Explain each one with an example application.

# Supervised Machine Learning

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y)**.

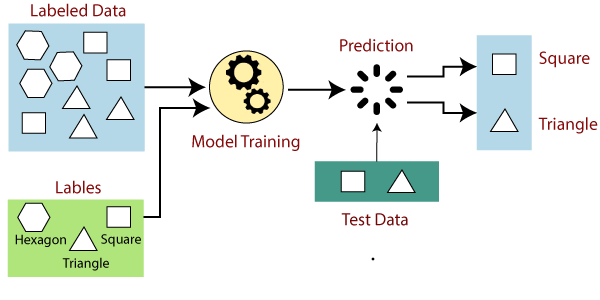
In the real-world, supervised learning can be used for **Risk Assessment, Image classification, Fraud Detection, spam filtering**, etc.

ADVERTISEMENT

## How Supervised Learning Works?

In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.

The working of Supervised learning can be easily understood by the below example and diagram:



Suppose we have a dataset of different types of shapes which includes square, rectangle, triangle, and Polygon. Now the first step is that we need to train the model for each shape.

ADVERTISEMENT

* If the given shape has four sides, and all the sides are equal, then it will be labelled as a **Square**.
* If the given shape has three sides, then it will be labelled as a **triangle**.
* If the given shape has six equal sides then it will be labelled as **hexagon**.

Now, after training, we test our model using the test set, and the task of the model is to identify the shape.

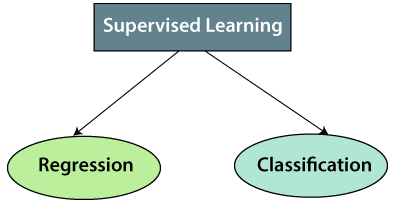
The machine is already trained on all types of shapes, and when it finds a new shape, it classifies the shape on the bases of a number of sides, and predicts the output.

## Steps Involved in Supervised Learning:

* First Determine the type of training dataset
* Collect/Gather the labelled training data.
* Split the training dataset into training **dataset, test dataset, and validation dataset**.
* Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
* Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.
* Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.
* Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

## Types of supervised Machine learning Algorithms:

Supervised learning can be further divided into two types of problems:



**1. Regression**

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc. Below are some popular Regression algorithms which come under supervised learning:

* Linear Regression
* Regression Trees
* Non-Linear Regression
* Bayesian Linear Regression
* Polynomial Regression

**2. Classification**

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

Spam Filtering,

* Random Forest
* Decision Trees
* Logistic Regression
* Support vector Machines

#### Note: We will discuss these algorithms in detail in later chapters.

## Advantages of Supervised learning:

* With the help of supervised learning, the model can predict the output on the basis of prior experiences.
* In supervised learning, we can have an exact idea about the classes of objects.
* Supervised learning model helps us to solve various real-world problems such as **fraud detection, spam filtering**, etc.

## Disadvantages of supervised learning:

* Supervised learning models are not suitable for handling the complex tasks.
* Supervised learning cannot predict the correct output if the test data is different from the training dataset.
* Training required lots of computation times.
* In supervised learning, we need enough knowledge about the classes of object.

9. What is the difference between supervised and unsupervised learning? With a sample application in each region, explain the differences.

**Supervised Learning**

1. Supervised learning algorithms are trained using labeled data.
2. Supervised learning model takes direct feedback to check if it is predicting correct output or not.
3. Supervised learning model predicts the output.
4. In supervised learning, input data is provided to the model along with the output.
5. The goal of supervised learning is to train the model so that it can predict the output when it is given new data.
6. Supervised learning needs supervision to train the model.
7. Supervised learning can be categorized in **Classification** and **Regression** problems.
8. Supervised learning can be used for those cases where we know the input as well as corresponding outputs.
9. Supervised learning model produces an accurate result.
10. Supervised learning is not close to true Artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output.
11. It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, Bayesian Logic, etc.

**Unsupervised Learning**

1. Unsupervised learning algorithms are trained using unlabeled data.
2. Unsupervised learning model does not take any feedback.
3. Unsupervised learning model finds the hidden patterns in data.
4. In unsupervised learning, only input data is provided to the model.
5. The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.
6. Unsupervised learning does not need any supervision to train the model.
7. Unsupervised Learning can be classified in **Clustering** and **Associations** problems.
8. Unsupervised learning can be used for those cases where we have only input data and no corresponding output data.
9. Unsupervised learning model may give less accurate result as compared to supervised learning.
10. Unsupervised learning is more close to the true Artificial Intelligence as it learns similarly as a child learns daily routine things by his experiences.
11. It includes various algorithms such as Clustering, KNN, and Apriori algorithm.

10. Describe the machine learning process in depth.

a. Make brief notes on any two of the following:

MATLAB is one of the most widely used programming languages.

ii. Deep learning applications in healthcare

Deep learning models can analyze electronic health records (EHR) that contain structured and unstructured data, including clinical notes, laboratory test results, diagnosis, and medications at exceptional speeds with the most possible accuracy.

Also, smartphones and[wearable devices provide useful information about lifestyle. They have the](https://www.businessinsider.com/wearable-technology-healthcare-medical-devices?r=US&IR=T) potential to transform data by using mobile apps to monitor medical risk factors for deep learning models. In 2019, Current Health’s AI wearable device [became](https://www.docwirenews.com/docwire-pick/future-of-medicine-picks/first-ai-wearable-approved-by-fda-for-home-use-monitoring-vitals/) one of the first AI medical monitoring wearables approved by Food and Drug Administration (FDA) for use at home. This device can measure the pulse, respiration, oxygen saturation, temperature, and mobility of patients.

Feel free to read our [examples on Healthcare Analytics](https://research.aimultiple.com/healthcare-analytics-examples/) for more.

iii. Study of the market basket

A data mining technique that is used to uncover purchase patterns in any retail setting is known as **Market Basket Analysis**. In simple terms Basically, Market basket analysis in data mining is to analyze the combination of products which been bought together

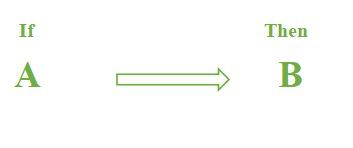
.

This is a technique that gives the careful study of purchases done by a customer in a supermarket. This concept identifies the pattern of frequent purchase items by customers. This analysis can help to promote deals, offers, sale by the companies, and data mining techniques helps to achieve this analysis task. Example:

* [Data mining](https://www.geeksforgeeks.org/data-mining/) concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.
* Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.
* Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

Market basket analysis mainly works with the ASSOCIATION RULE {IF} -> {THEN}.

* **IF** means **Antecedent:**An antecedent is an item found within the data
* **THEN** means **Consequent:**A consequent is an item found in combination with the antecedent.



Let’s see ASSOCIATION RULE {IF} -> {THEN} rules used in Market Basket Analysis in Data Mining. For example, customers buying a domain means they definitely need extra plugins/extensions to make it easier for the users.

Like we said above Antecedent is the item sets that are available in data. By formulating from the rules means**{if}** component and from the example is the domain.

Same as Consequent is the item that is found with the combination of Antecedents. By formulating from the rules means **{THEN}** component and from the example is extra plugins/extensions.

With the help of these, we are able to predict customer behavioral patterns. From this, we are able to make certain combinations with offers that customers will probably buy those products. That will automatically increase the sales and revenue of the company.

With the help of the [Apriori Algorithm,](https://www.geeksforgeeks.org/apriori-algorithm/) we can further classify and simplify the item sets which are frequently bought by the consumer.

There are three components in APRIORI ALGORITHM:

* SUPPORT
* CONFIDENCE
* LIFT

Now take an example, suppose 5000 transactions have been made through a popular eCommerce website. Now they want to calculate the support, confidence, and lift for the two products, let’s say pen and notebook for example out of 5000 transactions, 500 transactions for pen, 700 transactions for notebook, and  1000 transactions for both.

**SUPPORT:**It is been calculated with the number of transactions divided by the total number of transactions made,

support(pen) = transactions related to pen/total transactions

i.e support -> 500/5000=10 percent

**CONFIDENCE:**It is been calculated for whether the product sales are popular on individual sales or through combined sales. That is calculated with combined transactions/individual transactions.

Confidence =   combine transactions/individual transactions

i.e confidence-> 1000/500=20 percent

**LIFT:**Lift is calculated for knowing the ratio for the sales.

Lift-> 20/10=2

When the Lift value is below 1 means the combination is not so frequently bought by consumers. But in this case, it shows that the probability of buying both the things together is high when compared to the transaction for the individual items sold.

With this, we come to an overall view of the Market Basket Analysis in Data Mining and how to calculate the sales for combination products.

### Types of Market Basket Analysis

There are three types of Market Basket Analysis. They are as follow:

1. **Descriptive market basket analysis**: This sort of analysis looks for patterns and connections in the data that exist between the components of a market basket. This kind of study is mostly used to understand consumer behavior, including what products are purchased in combination and what the most typical item combinations. Retailers can place products in their stores more profitably by understanding which products are frequently bought together with the aid of descriptive market basket analysis.
2. **Predictive Market Basket Analysis**: Market basket analysis that predicts future purchases based on past purchasing patterns is known as predictive market basket analysis. Large volumes of data are analyzed using machine learning algorithms in this sort of analysis in order to create predictions about which products are most likely to be bought together in the future. Retailers may make data-driven decisions about which products to carry, how to price them, and how to optimize shop layouts with the use of predictive market basket research.
3. **Differential Market Basket Analysis**: Differential market basket analysis analyses two sets of market basket data to identify variations between them. Comparing the behavior of various client segments or the behavior of customers over time is a common usage for this kind of study. Retailers can respond to shifting consumer behavior by modifying their marketing and sales tactics with the help of differential market basket analysis.

### Benefits of Market Basket Analysis

1. **Enhanced Customer Understanding**: Market basket research offers insights into customer behavior, including what products they buy together and which products they buy the most frequently. Retailers can use this information to better understand their customers and make informed decisions.
2. **Improved Inventory Management**: By examining market basket data, retailers can determine which products are sluggish sellers and which ones are commonly bought together. Retailers can use this information to make well-informed choices about what products to stock and how to manage their inventory most effectively.
3. **Better Pricing Strategies**: A better understanding of the connection between product prices and consumer behavior might help merchants develop better pricing strategies. Using this knowledge, pricing plans that boost sales and profitability can be created.
4. **Sales Growth**: Market basket analysis can assist businesses in determining which products are most frequently bought together and where they should be positioned in the store to grow sales. Retailers may boost revenue and enhance customer shopping experiences by improving store layouts and product positioning.

### Applications of Market Basket Analysis

1. **Retail**: Market basket research is frequently used in the retail sector to examine consumer buying patterns and inform decisions about product placement, inventory management, and pricing tactics. Retailers can utilize market basket research to identify which items are sluggish sellers and which ones are commonly bought together, and then modify their inventory management strategy accordingly.
2. **E-commerce**: Market basket analysis can help online merchants better understand the customer buying habits and make data-driven decisions about product recommendations and targeted advertising campaigns. The behaviour of visitors to a website can be examined using market basket analysis to pinpoint problem areas.
3. **Finance**: Market basket analysis can be used to evaluate investor behaviour and forecast the types of investment items that investors will likely buy in the future. The performance of investment portfolios can be enhanced by using this information to create tailored investment strategies.
4. **Telecommunications**: To evaluate consumer behaviour and make data-driven decisions about which goods and services to provide, the telecommunications business might employ market basket analysis. The usage of this data can enhance client happiness and the shopping experience.
5. **Manufacturing**: To evaluate consumer behaviour and make data-driven decisions about which products to produce and which materials to employ in the production process, the manufacturing sector might use market basket analysis. Utilizing this knowledge will increase effectiveness and cut costs.

iv. Linear regression (simple)

# Simple Linear Regression in Machine Learning

Simple Linear Regression is a type of Regression algorithms that models the relationship between a dependent variable and a single independent variable. The relationship shown by a Simple Linear Regression model is linear or a sloped straight line, hence it is called Simple Linear Regression.

The key point in Simple Linear Regression is that the **dependent variable must be a continuous/real value**. However, the independent variable can be measured on continuous or categorical values.

Simple Linear regression algorithm has mainly two objectives:

ADVERTISEMENT

* **Model the relationship between the two variables.** Such as the relationship between Income and expenditure, experience and Salary, etc.
* **Forecasting new observations.** Such as Weather forecasting according to temperature, Revenue of a company according to the investments in a year, etc.

## Simple Linear Regression Model:

The Simple Linear Regression model can be represented using the below equation:

Backward Skip 10sPlay VideoForward Skip 10s

y= a0+a1x+ ε

Where,

**a0= It is the intercept of the Regression line (can be obtained putting x=0)**  
**a1= It is the slope of the regression line, which tells whether the line is increasing or decreasing.**  
**ε = The error term. (For a good model it will be negligible)**

## Implementation of Simple Linear Regression Algorithm using Python

**Problem Statement example for Simple Linear Regression:**

Here we are taking a dataset that has two variables: salary (dependent variable) and experience (Independent variable). The goals of this problem is:

* **We want to find out if there is any correlation between these two variables**
* **We will find the best fit line for the dataset.**
* **How the dependent variable is changing by changing the independent variable.**

In this section, we will create a Simple Linear Regression model to find out the best fitting line for representing the relationship between these two variables.

To implement the Simple Linear regression model in machine learning using Python, we need to follow the below steps:

**Step-1: Data Pre-processing**

The first step for creating the Simple Linear Regression model is [data pre-processing](https://www.javatpoint.com/data-preprocessing-machine-learning). We have already done it earlier in this tutorial. But there will be some changes, which are given in the below steps:

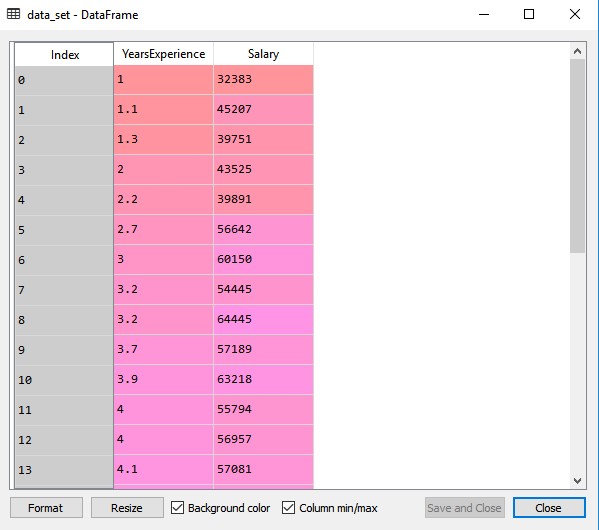
* First, we will import the three important libraries, which will help us for loading the dataset, plotting the graphs, and creating the Simple Linear Regression model.

1. **import** numpy as nm
2. **import** matplotlib.pyplot as mtp
3. **import** pandas as pd

* Next, we will load the dataset into our code:

1. data\_set= pd.read\_csv('Salary\_Data.csv')

By executing the above line of code (ctrl+ENTER), we can read the dataset on our Spyder IDE screen by clicking on the variable explorer option.



The above output shows the dataset, which has two variables: Salary and Experience.

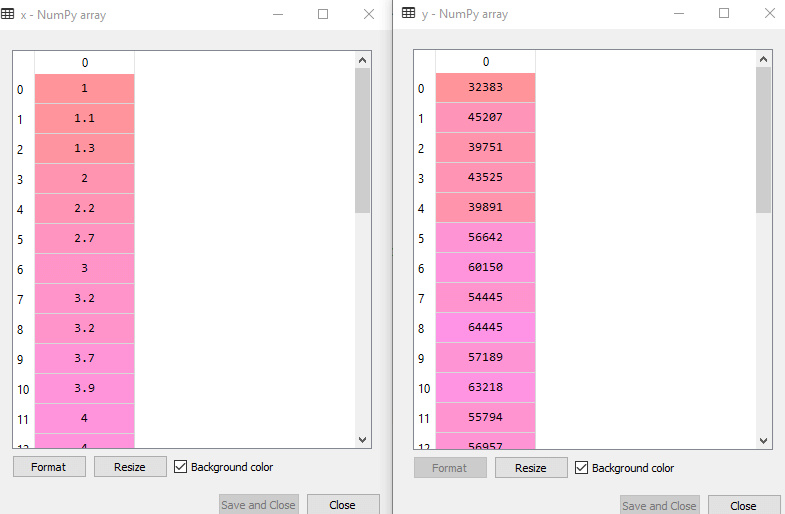
#### Note: In Spyder IDE, the folder containing the code file must be saved as a working directory, and the dataset or csv file should be in the same folder.

* After that, we need to extract the dependent and independent variables from the given dataset. The independent variable is years of experience, and the dependent variable is salary. Below is code for it:

1. x= data\_set.iloc[:, :-1].values
2. y= data\_set.iloc[:, 1].values

In the above lines of code, for x variable, we have taken -1 value since we want to remove the last column from the dataset. For y variable, we have taken 1 value as a parameter, since we want to extract the second column and indexing starts from the zero.

By executing the above line of code, we will get the output for X and Y variable as:



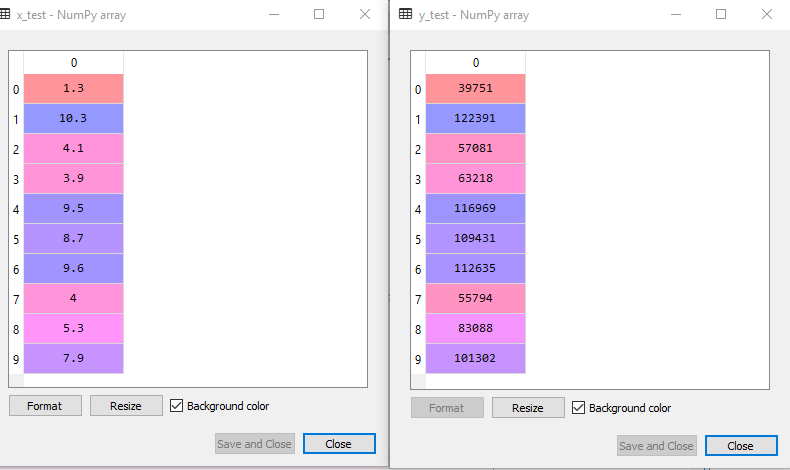
In the above output image, we can see the X (independent) variable and Y (dependent) variable has been extracted from the given dataset.

* Next, we will split both variables into the test set and training set. We have 30 observations, so we will take 20 observations for the training set and 10 observations for the test set. We are splitting our dataset so that we can train our model using a training dataset and then test the model using a test dataset. The code for this is given below:

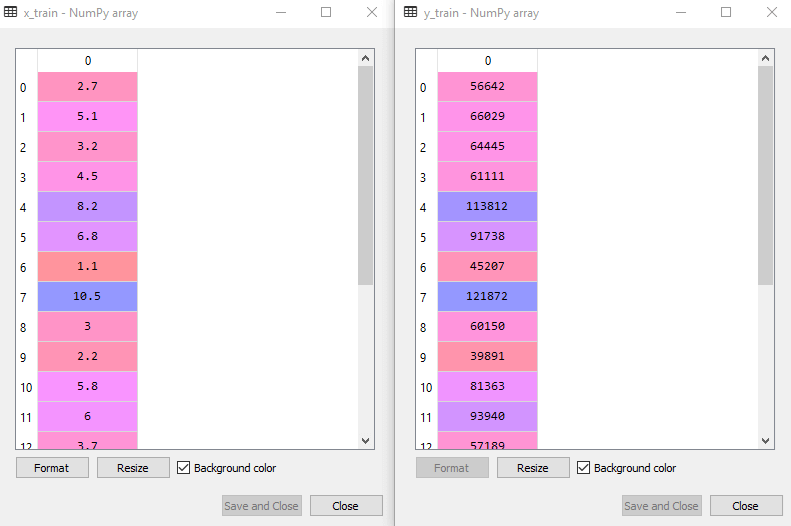
1. # Splitting the dataset into training and test set.
2. from sklearn.model\_selection **import** train\_test\_split
3. x\_train, x\_test, y\_train, y\_test= train\_test\_split(x, y, test\_size= 1/3, random\_state=0)

By executing the above code, we will get x-test, x-train and y-test, y-train dataset. Consider the below images:

**Test-dataset:**



**Training Dataset:**



* For simple linear Regression, we will not use Feature Scaling. Because Python libraries take care of it for some cases, so we don't need to perform it here. Now, our dataset is well prepared to work on it and we are going to start building a Simple Linear Regression model for the given problem.

**Step-2: Fitting the Simple Linear Regression to the Training Set:**

Now the second step is to fit our model to the training dataset. To do so, we will import the **LinearRegression** class of the **linear\_model** library from the **scikit learn**. After importing the class, we are going to create an object of the class named as a **regressor**. The code for this is given below:

1. #Fitting the Simple Linear Regression model to the training dataset
2. from sklearn.linear\_model **import** LinearRegression
3. regressor= LinearRegression()
4. regressor.fit(x\_train, y\_train)

In the above code, we have used a **fit()** method to fit our Simple Linear Regression object to the training set. In the fit() function, we have passed the x\_train and y\_train, which is our training dataset for the dependent and an independent variable. We have fitted our regressor object to the training set so that the model can easily learn the correlations between the predictor and target variables. After executing the above lines of code, we will get the below output.

**Output:**

Out[7]: LinearRegression(copy\_X=True, fit\_intercept=True, n\_jobs=None, normalize=False)

**Step: 3. Prediction of test set result:**

dependent (salary) and an independent variable (Experience). So, now, our model is ready to predict the output for the new observations. In this step, we will provide the test dataset (new observations) to the model to check whether it can predict the correct output or not.

We will create a prediction vector **y\_pred**, and **x\_pred**, which will contain predictions of test dataset, and prediction of training set respectively.

1. #Prediction of Test and Training set result
2. y\_pred= regressor.predict(x\_test)
3. x\_pred= regressor.predict(x\_train)

On executing the above lines of code, two variables named y\_pred and x\_pred will generate in the variable explorer options that contain salary predictions for the training set and test set.

**Output:**

You can check the variable by clicking on the variable explorer option in the IDE, and also compare the result by comparing values from y\_pred and y\_test. By comparing these values, we can check how good our model is performing.

**Step: 4. visualizing the Training set results:**

Now in this step, we will visualize the training set result. To do so, we will use the scatter() function of the pyplot library, which we have already imported in the pre-processing step. The **scatter () function** will create a scatter plot of observations.

In the x-axis, we will plot the Years of Experience of employees and on the y-axis, salary of employees. In the function, we will pass the real values of training set, which means a year of experience x\_train, training set of Salaries y\_train, and color of the observations. Here we are taking a green color for the observation, but it can be any color as per the choice.

Now, we need to plot the regression line, so for this, we will use the **plot() function** of the pyplot library. In this function, we will pass the years of experience for training set, predicted salary for training set x\_pred, and color of the line.

Next, we will give the title for the plot. So here, we will use the **title()** function of the **pyplot** library and pass the name ("Salary vs Experience (Training Dataset)".

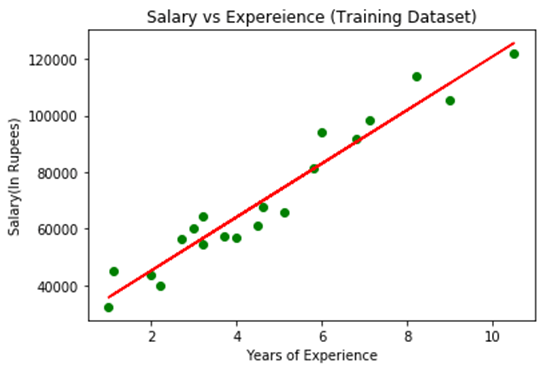
After that, we will assign labels for x-axis and y-axis using **xlabel() and ylabel() function**.

Finally, we will represent all above things in a graph using show(). The code is given below:

1. mtp.scatter(x\_train, y\_train, color="green")
2. mtp.plot(x\_train, x\_pred, color="red")
3. mtp.title("Salary vs Experience (Training Dataset)")
4. mtp.xlabel("Years of Experience")
5. mtp.ylabel("Salary(In Rupees)")
6. mtp.show()

**Output:**

By executing the above lines of code, we will get the below graph plot as an output.



In the above plot, we can see the real values observations in green dots and predicted values are covered by the red regression line. The regression line shows a correlation between the dependent and independent variable.

The good fit of the line can be observed by calculating the difference between actual values and predicted values. But as we can see in the above **plot, most of the observations are close to the regression line, hence our model is good for the training set**.

**Step: 5. visualizing the Test set results:**

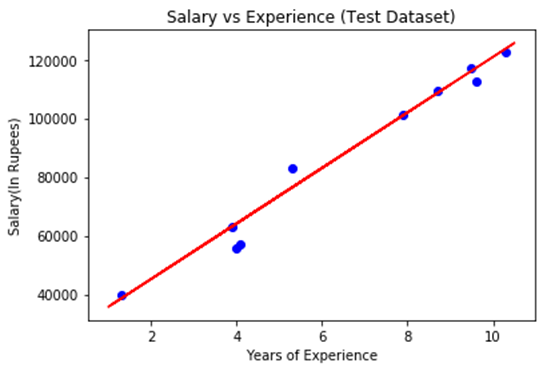
In the previous step, we have visualized the performance of our model on the training set. Now, we will do the same for the Test set. The complete code will remain the same as the above code, except in this, we will use x\_test, and y\_test instead of x\_train and y\_train.

Here we are also changing the color of observations and regression line to differentiate between the two plots, but it is optional.

1. #visualizing the Test set results
2. mtp.scatter(x\_test, y\_test, color="blue")
3. mtp.plot(x\_train, x\_pred, color="red")
4. mtp.title("Salary vs Experience (Test Dataset)")
5. mtp.xlabel("Years of Experience")
6. mtp.ylabel("Salary(In Rupees)")
7. mtp.show()

**Output:**

By executing the above line of code, we will get the output as:



In the above plot, there are observations given by the blue color, and prediction is given by the red regression line. As we can see, most of the observations are close to the regression line, hence we can say our Simple Linear Regression is a good model and able to make good predictions.

11. Make a comparison between:-

1. Generalization and abstraction

# Generalization, Specialization and Aggregation in ER Model

Using the ER model for bigger data creates a lot of complexity while designing a database model, So in order to minimizethe complexityGeneralization, Specialization, and Aggregation were introduced in the ER model and these were used for data abstraction in which an abstraction mechanism is used to hide details of a set of objects. Some of the terms were added to the Enhanced ER Model, where some new concepts were added. These new concepts are:

* Generalization
* Specialization
* Aggregation

## ****Generalization****

Generalization is the process of extracting common properties from a set of entities and creating a generalized entity from it. It is a bottom-up approach in which two or more entities can be generalized to a higher-level entity if they have some attributes in common. For Example, STUDENT and FACULTY can be generalized to a higher-level entity called PERSON as shown in Figure 1. In this case, common attributes like P\_NAME, and P\_ADD become part of a higher [entity](https://www.geeksforgeeks.org/difference-between-entity-and-object/) (PERSON), and specialized [attributes](https://www.geeksforgeeks.org/types-of-attributes-in-er-model/) like S\_FEE become part of a specialized entity (STUDENT).

Generalization is also called as ‘ Bottom-up approach”.

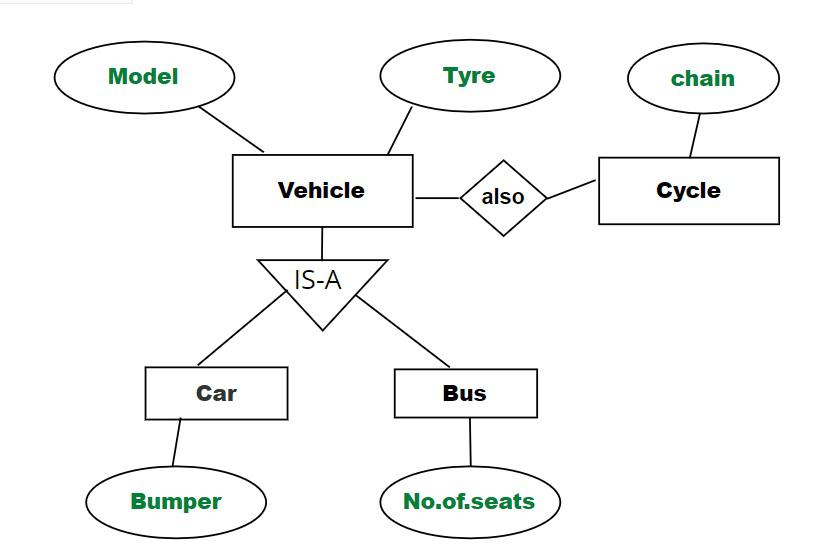
## ****Specialization****

In specialization, an entity is divided into sub-entities based on its characteristics. It is a top-down approach where the higher-level entity is specialized into two or more lower-level [entities](https://www.geeksforgeeks.org/difference-between-entity-entity-set-and-entity-type/). For Example, an EMPLOYEE entity in an Employee management system can be specialized into DEVELOPER, TESTER, etc. as shown in Figure 2. In this case, common attributes like E\_NAME, E\_SAL, etc. become part of a higher entity (EMPLOYEE), and specialized attributes like TES\_TYPE become part of a specialized entity (TESTER).

Specialization is also called as ” Top-Down approch”.

**Inheritance:** It is an important feature of generalization and specialization

* **Attribute inheritance**: allows lower level entities to inherit the attributes of higher level entities and vice versa.
* in diagram: **Car** entity is an inheritance of **Vehicle**entity ,So Car can acquire attributes of **Vehicle** example:car can acquire **Model** attribute of **Vehicle**.
* **Participation inheritance**: In participation inheritance, relationships involving higher level entity set also inherited by lower level entity and vice versa.
* in diagram: Vehicle entity has an relationship with Cycle entity ,So **Cycle entity** can acquire attributes of lower level entities i.e**Car** and **Bus** since it is inheritance of **Vehicle.**



*Example of Relation*

## ****Aggregation****

An ER diagram is not capable of representing the relationship between an entity and a relationship which may be required in some scenarios. In those cases, a relationship with its corresponding entities is aggregated into a higher-level entity. Aggregation is an abstraction through which we can represent relationships as higher-level entity sets.

For Example, an Employee working on a project may require some machinery. So, REQUIRE relationship is needed between the relationship WORKS\_FOR and entity MACHINERY. Using aggregation, WORKS\_FOR relationship with its entities EMPLOYEE and PROJECT is aggregated into a single entity and relationship REQUIRE is created between the aggregated entity and MACHINERY.

*Aggregation*

## ****Representing Aggregation Via Schema****

To represent aggregation, create a schema containing the following things.

* the [primary key](https://www.geeksforgeeks.org/primary-key-constraint-in-sql/) to the aggregated relationship
* the primary key to the associated entity set
* descriptive attribute, if exists

2. Learning that is guided and unsupervised

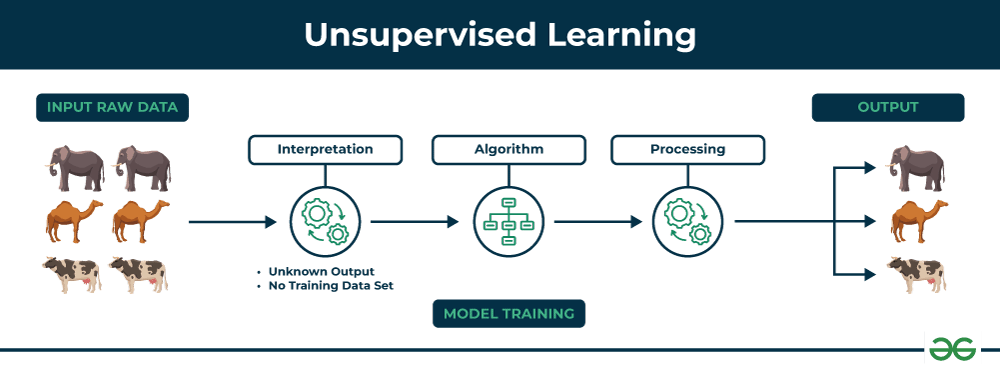
## ****What is Unsupervised learning?****

Unsupervised learning is a type of machine learning that learns from unlabeled data. This means that the data does not have any pre-existing labels or categories. The goal of unsupervised learning is to discover patterns and relationships in the data without any explicit guidance.

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.

You can use unsupervised learning to examine the animal data that has been gathered and distinguish between several groups according to the traits and actions of the animals. These groupings might correspond to various animal species, providing you to categorize the creatures without depending on labels that already exist.



**Key Points**

* Unsupervised learning allows the model to discover patterns and relationships in unlabeled data.
* Clustering algorithms group similar data points together based on their inherent characteristics.
* Feature extraction captures essential information from the data, enabling the model to make meaningful distinctions.
* Label association assigns categories to the clusters based on the extracted patterns and characteristics.

### Example

Imagine you have a machine learning model trained on a large dataset of unlabeled images, containing both dogs and cats. The model has never seen an image of a dog or cat before, and it has no pre-existing labels or categories for these animals. Your task is to use unsupervised learning to identify the dogs and cats in a new, unseen image.

**For instance**, suppose it is given an image having both dogs and cats which it has never seen.

Thus the machine has no idea about the features of dogs and cats so we can’t categorize it as ‘dogs and cats ‘. But it can categorize them according to their similarities, patterns, and differences, i.e., we can easily categorize the above picture into two parts. The first may contain all pics having **dogs** in them and the second part may contain all pics having **cats** in them. Here you didn’t learn anything before, which means no training data or examples.

It allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with unlabelled data.

## Types of Unsupervised Learning

Unsupervised learning is classified into two categories of algorithms:

* **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
* **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

### ****Clustering****

Clustering is a type of unsupervised learning that is used to group similar data points together. [Clustering algorithms](https://www.geeksforgeeks.org/clustering-in-machine-learning/) work by iteratively moving data points closer to their cluster centers and further away from data points in other clusters.

1. Exclusive (partitioning)
2. Agglomerative
3. Overlapping
4. Probabilistic

**Clustering Types:-**

1. Hierarchical clustering
2. K-means clustering
3. Principal Component Analysis
4. Singular Value Decomposition
5. Independent Component Analysis
6. Gaussian Mixture Models (GMMs)
7. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

### ****Association rule learning****

Association rule learning is a type of unsupervised learning that is used to identify patterns in a data. [Association rule](https://www.geeksforgeeks.org/association-rule/)learning algorithms work by finding relationships between different items in a dataset.

Some common association rule learning algorithms include:

* Apriori Algorithm
* Eclat Algorithm
* FP-Growth Algorithm

### Evaluating Non-Supervised Learning Models

Evaluating non-supervised learning models is an important step in ensuring that the model is effective and useful. However, it can be more challenging than evaluating supervised learning models, as there is no ground truth data to compare the model’s predictions to.

There are a number of different metrics that can be used to evaluate non-supervised learning models, but some of the most common ones include:

* **Silhouette score:** The silhouette score measures how well each data point is clustered with its own cluster members and separated from other clusters. It ranges from -1 to 1, with higher scores indicating better clustering.
* **Calinski-Harabasz score:** The Calinski-Harabasz score measures the ratio between the variance between clusters and the variance within clusters. It ranges from 0 to infinity, with higher scores indicating better clustering.
* **Adjusted Rand index:** The adjusted Rand index measures the similarity between two clusterings. It ranges from -1 to 1, with higher scores indicating more similar clusterings.
* **Davies-Bouldin index:** The Davies-Bouldin index measures the average similarity between clusters. It ranges from 0 to infinity, with lower scores indicating better clustering.
* **F1 score:** The F1 score is a weighted average of precision and recall, which are two metrics that are commonly used in supervised learning to evaluate classification models. However, the F1 score can also be used to evaluate non-supervised learning models, such as clustering models.

### Application ****of Unsupervised learning****

Non-supervised learning can be used to solve a wide variety of problems, including:

* Anomaly detection: Unsupervised learning can identify unusual patterns or deviations from normal behavior in data, enabling the detection of fraud, intrusion, or system failures.
* Scientific discovery: Unsupervised learning can uncover hidden relationships and patterns in scientific data, leading to new hypotheses and insights in various scientific fields.
* Recommendation systems: Unsupervised learning can identify patterns and similarities in user behavior and preferences to recommend products, movies, or music that align with their interests.
* Customer segmentation: Unsupervised learning can identify groups of customers with similar characteristics, allowing businesses to target marketing campaigns and improve customer service more effectively.
* Image analysis: Unsupervised learning can group images based on their content, facilitating tasks such as image classification, object detection, and image retrieval.

### Advantages ****of Unsupervised learning****

* It does not require training data to be labeled.
* Dimensionality reduction can be easily accomplished using unsupervised learning.
* Capable of finding previously unknown patterns in data.
* Unsupervised learning can help you gain insights from unlabeled data that you might not have been able to get otherwise.
* Unsupervised learning is good at finding patterns and relationships in data without being told what to look for. This can help you learn new things about your data.

### Disadvantages ****of Unsupervised learning****

* Difficult to measure accuracy or effectiveness due to lack of predefined answers during training.
* The results often have lesser accuracy.
* The user needs to spend time interpreting and label the classes which follow that classification.
* Unsupervised learning can be sensitive to data quality, including missing values, outliers, and noisy data.
* Without labeled data, it can be difficult to evaluate the performance of unsupervised learning models, making it challenging to assess their effectiveness.

1. Regression and classification

**Regression**

* In this problem statement, the target variables are continuous.
* Problems like [House Price Prediction](https://www.geeksforgeeks.org/house-price-prediction-using-machine-learning-in-python/), [Rainfall Prediction](https://www.geeksforgeeks.org/ml-rainfall-prediction-using-linear-regression/) like problems are solved using regression Algorithms.
* In this algorithm, we try to find the best-fit line which can represent the overall trend in the data.
* Evaluation metrics like [Mean Squared Error,](https://www.geeksforgeeks.org/python-mean-squared-error/) [R2-Score](https://www.geeksforgeeks.org/ml-r-squared-in-regression-analysis/), and  [MAPE](https://www.geeksforgeeks.org/how-to-calculate-mape-in-python/) are used here to evaluate the performance of the regression algorithms.
* Here we face the problems like [Linear Regression](https://www.geeksforgeeks.org/ml-linear-regression/) models as well as non-linear models
* Input Data are Independent variables and continuous dependent variable.
* The regression algorithm’s task is mapping input value (x) with continuous output variable (y).
* Output is Continuous numerical values
* Objective is to Predicting continuous numerical values.
* Example use cases are Stock price prediction, house price prediction, demand forecasting.
* **Examples of regression algorithms are:**

Linear Regression, Polynomial Regression, Ridge Regression, Lasso Regression, Support Vector Regression (SVR), Decision Trees for Regression, Random Forest Regression, K-Nearest Neighbors (K-NN) Regression, Neural Networks for Regression, etc.

**Classification**

* In this problem statement, the target variables are discrete.
* Problems like [Spam Email Classification](https://www.geeksforgeeks.org/detecting-spam-emails-using-tensorflow-in-python/), [Disease prediction](https://www.geeksforgeeks.org/disease-prediction-using-machine-learning/) like problems are solved using Classification Algorithms.
* In this algorithm, we try to find the best possible decision boundary which can separate the two classes with the maximum possible separation.
* [Evaluation metrics](https://www.geeksforgeeks.org/metrics-for-machine-learning-model/) like Precision, Recall, and F1-Score are used here to evaluate the performance of the classification algorithms.
* Here we face the problems like [binary Classification](https://www.geeksforgeeks.org/getting-started-with-classification/) or [Multi-Class Classification](https://www.geeksforgeeks.org/multiclass-classification-using-scikit-learn/) problems.
* Input Data are Independent variables and categorical dependent variable.
* The classification algorithm’s task mapping the input value of x with the discrete output variable of y.
* Output is Categorical labels.
* Objective is to  Predict categorical/class labels.
* Example use cases are Spam detection, image recognition, sentiment analysis
* **Examples of classification algorithms are:**

Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), Naive Bayes, Neural Networks, K-Means Clustering, Multi-layer Perceptron (MLP), etc.