Introduction to Machine Learning

**What is Data Science**

Data is widely considered a crucial resource in different organizations across every industry. Data Science can be described in simple terms as a separate field of work that deals with the management and processing of data using statistical methods, artificial intelligence, and other tools in partnership with domain specialists. Pursuing Data Science encompasses concepts and epochs derived from different fields including Mathematics and Computer Science and Information Theory to interpret large data.

## Define Data Science.

The term “data science” combines two key elements: “data” and “science.”

1. **Data**: It refers to the raw information that is collected, stored, and processed. In today’s digital age, enormous amounts of data are generated from various sources such as sensors, social media, transactions, and more. This data can come in structured formats (e.g., databases) or unstructured formats (e.g., text, images, videos).
2. **Science**: It refers to the systematic study and investigation of phenomena using scientific methods and principles. Science involves forming hypotheses, conducting experiments, analyzing data, and drawing conclusions based on evidence.

When we put these two elements together, “**data+science**” refers to the scientific study of data. Data Science involves applying scientific methods, statistical techniques, computational tools, and domain expertise to explore, analyze, and extract insights from data. The term emphasizes the rigorous and systematic approach taken to understand and derive value from vast and complex datasets.

Essentially, data science is about using scientific methods to unlock the potential of data, uncover patterns, make predictions, and drive informed decision-making across various domains and industries.

## What is Data Science in Simple Words?

Imagine you’re scrolling through your favorite **social media** platform, and you notice that certain types of posts always seem to grab your attention. Maybe it’s cute animal videos, delicious food recipes, or inspiring travel photos.

Now, from the platform’s perspective, they want to keep you engaged and coming back for more. This is where data science comes into play. They collect a ton of information about what you like, share, and comment on. They use data science techniques to analyze all this information to understand your preferences better.

For instance, they might notice that you spend more time watching animal videos than looking at food recipes. Armed with this knowledge, they can then customize your feed to show you more of what you love – adorable pets! They might even predict what type of pet video you’re likely to enjoy next based on your past behavior.

In this scenario, data science is like the magic behind the scenes that helps social media platforms understand your interests and tailor your experience to keep you engaged. It’s all about using data to make your online experience more personalized and enjoyable.

## What is Data Science Course?

A data science course is a structured educational program designed to teach individuals the foundational concepts, tools, and techniques of data science. These data science courses typically cover a wide range of topics, including statistics, programming, machine learning, data visualization, and data analysis. They are suitable for beginners with little to no prior experience in data science, as well as professionals looking to expand their skills or transition into a data-related role.

*One such complete data science course which is trusted by students as well as professionals is*[***Complete Machine Learning & Data Science Program***](https://www.geeksforgeeks.org/courses/data-science-live)

Key components of a data science course may include:

1. **Foundational Concepts**: Introduction to basic concepts in data science, including data types, data manipulation, data cleaning, and exploratory data analysis.
2. **Programming Languages**: Instruction in programming languages commonly used in data science, such as Python or R. Students learn how to write code to analyze and manipulate data, create visualizations, and build machine learning models.
3. **Statistical Methods**: Coverage of statistical techniques and methods used in data analysis, hypothesis testing, regression analysis, and probability theory.
4. **Machine Learning**: Introduction to machine learning algorithms, including supervised learning, unsupervised learning, and deep learning. Students learn how to apply machine learning techniques to solve real-world problems and make predictions from data.
5. **Data Visualization**: Instruction in data visualization techniques and tools for effectively communicating insights from data. Students learn how to create plots, charts, and interactive visualizations to explore and present data.
6. **Practical Projects**: Hands-on experience working on data science projects and case studies, where students apply their knowledge and skills to solve real-world problems and analyze real datasets.
7. **Capstone Project**: A culminating project where students demonstrate their mastery of data science concepts and techniques by working on a comprehensive project from start to finish.

## What is Data Science Job?

A data science job involves using various techniques, algorithms, and tools to extract insights and knowledge from structured and unstructured data. Here are some of the key data science job roles:

1. **Data Scientist**:
   * Responsibilities: Analyzing large datasets, developing machine learning models, interpreting results, and providing insights to inform business decisions.
   * Skills: Proficiency in programming languages like Python or R, expertise in statistics and machine learning algorithms, data visualization skills, and domain knowledge in the relevant industry.
2. **Data Analyst**:
   * Responsibilities: Collecting, cleaning, and analyzing data to identify trends, patterns, and insights. Often involves creating reports and dashboards to communicate findings to stakeholders.
   * Skills: Strong proficiency in SQL for data querying, experience with data visualization tools like Tableau or Power BI, basic statistical knowledge, and familiarity with Excel or Google Sheets.
3. **Machine Learning Engineer**:
   * Responsibilities: Building and deploying machine learning models at scale, optimizing model performance, and integrating them into production systems.
   * Skills: Proficiency in programming languages like Python or Java, experience with machine learning frameworks like TensorFlow or PyTorch, knowledge of cloud platforms like AWS or Azure, and software engineering skills for developing scalable solutions.
4. **Data Engineer**:
   * Responsibilities: Designing and building data pipelines to collect, transform, and store large volumes of data. Ensuring data quality, reliability, and scalability.
   * Skills: Expertise in database systems like SQL and NoSQL, proficiency in programming languages like Python or Java, experience with big data technologies like Hadoop or Spark, and knowledge of data warehousing concepts.
5. **Business Intelligence (BI) Analyst**:
   * Responsibilities: Gathering requirements from business stakeholders, designing and developing BI reports and dashboards, and providing data-driven insights to support strategic decision-making.
   * Skills: Proficiency in BI tools like Tableau, Power BI, or Looker, strong SQL skills for data querying, understanding of data visualization principles, and ability to translate business needs into technical solutions.
6. **Data Architect**:
   * Responsibilities: Designing the overall structure of data systems, including databases, data lakes, and data warehouses. Defining data models, schemas, and data governance policies.
   * Skills: Deep understanding of database technologies and architectures, experience with data modeling tools like ERWin or Visio, knowledge of data integration techniques, and familiarity with data security and compliance regulations.

## What is Data Science Applications?

Data Science has a wide array of applications across various industries, significantly impacting the way businesses operate and how services are delivered. Here are some key applications of Data Science:

1. **Healthcare**:
   * **Predictive Analytics**: Predicting disease outbreaks, patient readmissions, and individual health risks.
   * **Medical Imaging**: Enhancing image recognition to diagnose conditions from X-rays, MRIs, and CT scans.
   * **Personalized Medicine**: Tailoring treatment plans based on genetic information and patient history.
2. **Finance**:
   * **Risk Management**: Identifying and mitigating financial risks through predictive modeling.
   * **Fraud Detection**: Analyzing transactions to detect fraudulent activities.
   * **Algorithmic Trading**: Using data-driven algorithms to execute high-frequency trading strategies.
3. **Marketing**:
   * **Customer Segmentation**: Grouping customers based on purchasing behavior and preferences for targeted marketing.
   * **Sentiment Analysis**: Analyzing customer feedback and social media interactions to gauge public sentiment.
   * **Predictive Analytics**: Forecasting sales trends and customer lifetime value.
4. **Retail**:
   * **Inventory Management**: Optimizing stock levels based on demand forecasting.
   * **Recommendation Systems**: Providing personalized product recommendations to customers.
   * **Price Optimization**: Adjusting prices dynamically based on market trends and consumer behavior.
5. **Transportation**:
   * **Route Optimization**: Enhancing logistics by determining the most efficient routes.
   * **Predictive Maintenance**: Forecasting equipment failures to schedule timely maintenance.
   * **Autonomous Vehicles**: Developing self-driving cars using machine learning algorithms.
6. **Education**:
   * **Personalized Learning**: Creating customized learning experiences based on student performance and preferences.
   * **Academic Analytics**: Analyzing data to improve student retention and graduation rates.
   * **Curriculum Development**: Using data to develop and refine educational programs.
7. **Entertainment**:
   * **Content Recommendation**: Suggesting movies, shows, and music based on user preferences.
   * **Audience Analytics**: Understanding audience behavior to improve content delivery.
   * **Production Analytics**: Optimizing production schedules and budgets through data analysis.
8. **Manufacturing**:
   * **Quality Control**: Using data to monitor and improve product quality.
   * **Supply Chain Optimization**: Streamlining supply chain processes through predictive analytics.
   * **Process Automation**: Implementing automated systems for efficient production workflows.
9. **Energy**:
   * **Smart Grids**: Enhancing the efficiency and reliability of energy distribution.
   * **Predictive Maintenance**: Forecasting and preventing equipment failures in power plants.
   * **Energy Consumption Analytics**: Analyzing patterns to optimize energy usage and reduce costs.
10. **Government**:
    * **Public Safety**: Analyzing crime data to improve law enforcement strategies.
    * **Urban Planning**: Using data to plan and develop smarter cities.
    * **Policy Making**: Leveraging data to make informed decisions and create effective policies.

**Artificial Intelligence**

**Artificial Intelligence (AI)** has become a discussed subject, in today’s fast-moving world. It has transitioned from being a concept in science fiction to a reality that impacts our daily lives. People all over the world are fascinated by AI and its ability to bring their imaginations to work in their daily lives.

## What is Artificial Intelligence?

Artificial Intelligence (AI) refers to the development of computer systems of performing tasks that require human intelligence. AI aids, in processing amounts of data identifying patterns and making decisions based on the collected information. This can be achieved through techniques like [Machine Learning](https://www.geeksforgeeks.org/machine-learning/), [Natural Language Processing](https://www.geeksforgeeks.org/natural-language-processing-nlp-tutorial/), [Computer Vision](https://www.geeksforgeeks.org/computer-vision/) and [Robotics](https://www.geeksforgeeks.org/robotics-introduction/). AI encompasses a range of abilities including learning, reasoning, perception, problem solving, data analysis and language comprehension. The ultimate***goal of AI*** is to create machines that can emulate capabilities and carry out diverse tasks, with enhanced efficiency and precision. The field of AI holds potential to revolutionize aspects of our daily lives.

## **Examples of Artificial Intelligence**

Artificial Intelligence (AI) has become increasingly integrated into various aspects of our lives, revolutionizing industries and impacting daily routines. Here are some examples illustrating the diverse applications of AI:

1. **Virtual Personal Assistants**: Popular examples like Siri, Google Assistant, and Amazon Alexa utilize AI to understand and respond to user commands. These assistants employ natural language processing (NLP) and machine learning algorithms to improve their accuracy and provide more personalized responses over time.
2. **Autonomous Vehicles**: AI powers the development of self-driving cars, trucks, and drones. Companies like Tesla, Waymo, and Uber are at the forefront of this technology, using AI algorithms to analyse sensory data from cameras, radar, and lidar to make real-time driving decisions.
3. **Healthcare Diagnosis and Treatment**: AI algorithms are used to analyse medical data, including patient records, imaging scans, and genetic information, to assist healthcare professionals in diagnosing diseases and planning treatments. IBM’s Watson for Health and Google’s DeepMind are examples of AI platforms employed in healthcare.
4. **Recommendation Systems**: Online platforms like Netflix, Amazon, and Spotify utilize AI to analyse user behaviour and preferences, providing personalized recommendations for movies, products, and music. These systems employ collaborative filtering and content-based filtering techniques to enhance user experience and increase engagement.
5. **Fraud Detection**: AI algorithms are employed by financial institutions to detect fraudulent activities in real-time. These systems analyse.

**AI** has the potential to revolutionize many industries and fields, such as healthcare, finance, transportation, and education. However, it also raises important ethical and societal questions, such as the impact on employment and privacy, and the responsible development and use of***AI technology***.

## Importance of Artificial Intelligence

Today, the amount of data in the world is so humongous that humans fall short of absorbing, interpreting, and making decisions of the entire data. This complex decision-making requires higher cognitive skills than human beings. This is why we’re trying to build machines better than us, in these task. Another major characteristic that AI machines possess but we don’t is repetitive learning. Let consider an example of how **Artificial Intelligence** is important to us. Data that is fed into the machines could be real-life incidents. **How people interact, behave and react ?**etc. So, in other words, machines learn to think like humans, by observing and learning from humans. That’s precisely what is called Machine Learning which is a subfield of AI.  Humans are observed to find repetitive tasks highly boring. Accuracy is another factor in which we humans lack. Machines have extremely high accuracy in the tasks that they perform. Machines can also take risks instead of human beings. AI is used in various fields like:

* **Health Care**
* **Retail**
* **Manufacturing**
* **Banking etc.**

## What are the types of Artificial intelligence?

AI can be broadly classified into two major categories:

### **Based on Capabilities:**

**1. Narrow AI:** Narrow AI, also known as [Weak AI](https://www.geeksforgeeks.org/what-is-artificial-narrow-intelligence-ani/), refers to artificial intelligence systems that are designed and trained to perform a specific task or a narrow range of tasks. These systems excel at their designated tasks but lack the broad cognitive abilities and understanding of human intelligence. Narrow AI is the most common form of AI currently in use and has found widespread application across various industries and domains.

Characteristics of Narrow AI include:

* **Specialized capabilities:**Narrow AI systems are highly specialized and focused on performing a specific function or solving a particular problem.
* **Limited scope**: These systems are not capable of generalizing their knowledge or skills to other domains outside of their designated tasks.
* **Lack of consciousness:** Narrow AI lacks self-awareness and consciousness. It operates based on predefined algorithms and data inputs without understanding the context or implications of its actions.
* **Examples**: Virtual personal assistants like Siri and Alexa, recommendation systems, image recognition software, chatbots, and autonomous vehicles are all examples of Narrow AI.

**2. General AI:** Also referred to as **“General AI”**. Here is where there is no difference between a machine and a human being. This is the kind of AI we see in the movies, the robots. A close example (not the perfect example) would be the world’s first citizen robot, Sophia. She was introduced to the world on October 11, 2017. Sophia talks like she has emotions.

General AI, also known as Strong AI or Artificial General Intelligence (AGI), refers to artificial intelligence systems that replicate human-like cognitive abilities and understanding across a wide range of tasks and domains. Unlike Narrow AI, which is task-specific, General AI aims to simulate human intelligence comprehensively, including reasoning, problem-solving, learning, and adaptation to new situations.

Characteristics of General AI include:

* **Human-like cognitive abilities:** General AI systems can understand, learn, and apply knowledge across various tasks and domains, similar to human intelligence.
* **Adaptability**: These systems have the ability to generalize their knowledge and skills to new situations, tasks, and environments.
* **Consciousness**: General AI is theorized to possess self-awareness, consciousness, and subjective experiences, although achieving this level of intelligence remains a theoretical challenge.
* **Examples**: General AI remains largely theoretical and speculative, as researchers have not yet achieved human-level artificial intelligence. Development in this area continues to be a subject of ongoing research and exploration.

**3. Super Intelligent AI:** Super intelligent AI refers to artificial intelligence systems that surpass human intelligence in virtually every aspect. This type of AI, also known as [Artificial Superintelligence](https://www.geeksforgeeks.org/what-is-artificial-super-intelligence-asi/) (ASI), represents the highest level of AI capabilities and poses significant implications for society and the future of humanity.

Characteristics of Super intelligent AI include:

* **Cognitive superiority**: Super intelligent AI outperforms humans in terms of cognitive abilities, including problem-solving, creativity, and strategic planning.
* **Rapid learning and adaptation**: These systems can acquire and process information at an unprecedented speed and scale, leading to rapid advancements in various fields.
* **Ethical and existential risks**: The development of Super intelligent AI raises concerns about its potential impact on society, including risks related to control, alignment with human values, and existential threats to humanity.
* **Examples**: Super intelligent AI remains theoretical, and no concrete examples exist as of now. However, researchers are actively exploring the implications and challenges associated with its development.

### **Based on Functionalities :**

**1. Reactive machines:** These are the most basic type of AI and are purely reactive as the name suggests. They neither can form memories nor can use past experiences to form decisions. An example would be ***IBM’s Deep Blue chess-playing supercomputer*** which is mentioned above. Deep Blue beat the international grandmaster **Garry Kasparov in 1997**. It can choose the most optimal of the chess moves and beat the opponent. Apart from a rarely used chess-specific rule against repeating the same move three times, ***Deep Blue***ignores everything before the present moment, thus not storing any memories. This type of AI just perceives the world, the chess game in the case of **Deep Blue,** and acts on it.

**2. Limited memory:** These machines can look into the past. Not the ability to predict what happened in the past, but the usage of memories to form decisions. A common example could include self-driving cars. For example, they observe other cars’ speed and directions and act accordingly. This requires monitoring of how a car is driven for a specific amount of time. Just like how humans observe and learn the specifics. These pieces of information are not stored in the library of experiences of the machines, unlike humans. We humans automatically save everything in the library of our experiences and can learn from it, but limited memory machines can’t.

**3. Theory of mind:** These are types of machines that can understand that people have beliefs, emotions, expectations, etc., and have some of their own. A **“theory of mind”** machine can think emotionally and can respond with emotions. Even though there are close examples of this kind of AI like Sophia, the research is not complete yet. In other words, these machines have a notion of not just the world, but also the existing entities of the world, like human beings, animals, etc. These machines will be capable of answering simple **“what if”** questions. They’ll have a sense of empathy.

**4. Self-Awareness:** These types of machines can be called human equivalents. Of course, no such machines exist and the invention of them would be a milestone in the field of AI. These basically will have a sense of consciousness of who they are. The sense of**“I”** or**“me”.** Here’s a basic example of the difference between “**theory of mind**” and **“self-awareness”**AI. The feeling of I want to play is different from the feeling of I know I want to play. In the latter, if you notice, there is a sense of consciousness and is a characteristic of a self-aware machine, while the former feeling is a characteristic of a theory-of-mind machine. Self-aware machines will have the ability to predict others’ feelings. Let’s hope the invention is not so far away.

## How Does AI Work ?

Artificial Intelligence (AI) uses a wide range of techniques and approaches that enable machines to simulate human-like intelligence and perform tasks that traditionally require human assistance. AI systems work through a combination of algorithms, data, and computational power. Here’s an overview of how AI works:

1. **Data Collection**: AI systems rely on vast amounts of data to learn and make decisions. Data can be collected from various sources, including sensors, digital devices, databases, the internet, and user interactions. The quality and quantity of data are crucial for training accurate and reliable AI models.
2. **Data Pre-processing**: Once data is collected, it needs to be pre-processed to ensure it’s clean, structured, and suitable for analysis. This pre-processing stage may involve tasks such as cleaning noisy data, handling missing values, standardizing formats, and encoding categorical variables.
3. **Algorithm Selection**: AI algorithms are chosen based on the specific task or problem the AI system aims to solve. Different algorithms are suited for different types of tasks, such as classification, regression, clustering, and pattern recognition. Common AI algorithms include neural networks, decision trees, support vector machines, and k-nearest neighbours.
4. **Model Training**: In the training phase, AI models are fed with labelled data (supervised learning) or unlabelled data (unsupervised learning) to learn patterns and relationships. During training, the model adjusts its parameters iteratively to minimize errors and improve its performance on the given task. This process involves optimization techniques like gradient descent and backpropagation in neural networks.
5. **Model Evaluation**: After training, the AI model is evaluated using separate validation data to assess its performance and generalization ability. Performance metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC) are used to quantify the model’s effectiveness in making predictions or decisions.
6. **Model Deployment**: Once the AI model meets the desired performance criteria, it can be deployed into production environments to perform real-world tasks. Deployment involves integrating the model into existing systems, such as mobile apps, web services, or embedded devices, to provide AI-driven functionalities.
7. **Continuous Learning and Improvement**: AI systems can adapt and improve over time through continuous learning. They can be updated with new data and retrained periodically to stay relevant and accurate in dynamic environments. Techniques like online learning, transfer learning, and reinforcement learning enable AI models to learn from new experiences and feedback.
8. **Inference and Decision-Making**: During inference, the trained AI model applies its learned knowledge to make predictions or decisions on new, unseen data. Inference involves feeding input data into the model and obtaining output predictions or classifications based on the model’s learned patterns and representations.

Overall, AI systems work by leveraging data, algorithms, and computational power to learn from experience, make decisions, and perform tasks autonomously. The specific workings of an AI system depend on its architecture, algorithms, and the nature of the tasks it’s designed to accomplish.

## What are the Applications of AI?

**Artificial Intelligence (AI)** has a wide range of applications and has been adopted in many industries to improve efficiency, accuracy, and productivity. Some of the most common uses of AI are:

* **Healthcare:** AI is used in healthcare for various purposes such as diagnosing diseases, predicting patient outcomes, drug discovery, and personalized treatment plans.
* **Finance:** AI is used in the finance industry for tasks such as credit scoring, fraud detection, portfolio management, and financial forecasting.
* **Retail:**AI is used in the retail industry for applications such as customer service, demand forecasting, and personalized marketing.
* **Manufacturing:** AI is used in manufacturing for tasks such as quality control, predictive maintenance, and supply chain optimization.
* **Transportation:**AI is used in transportation for optimizing routes, improving traffic flow, and reducing fuel consumption.
* **Education:** AI is used in education for personalizing learning experiences, improving student engagement, and providing educational resources.
* **Marketing:** AI is used in marketing for tasks such as customer segmentation, personalized recommendations, and real-time audience analysis.
* **Gaming:**AI is used in gaming for developing intelligent game characters and providing personalized gaming experiences.
* **Security:** AI is used in security for tasks such as facial recognition, intrusion detection, and cyber threat analysis.
* **Natural Language Processing (NLP):** AI is used in [NLP](https://www.geeksforgeeks.org/natural-language-processing-nlp-tutorial/) for tasks such as speech recognition, machine translation, and sentiment analysis.

These are some of the most common uses of AI, but the[Applications of AI](https://www.geeksforgeeks.org/artificial-intelligence-applications/) are constantly expanding, evolving, and it is likely that new uses will emerge in the future.

## What will be the future of AI?

The future of AI is likely to involve continued advancements in machine learning, natural language processing, and computer vision, which will enable AI systems to become increasingly capable and integrated into a wide range of applications and industries. Some potential areas of growth for AI include healthcare, finance, transportation, and customer service. Additionally, there may be increasing use of AI in more sensitive areas such as decision making in criminal justice, hiring and education, which will raise ethical and societal implications that need to be addressed. It is also expected that there will be more research and development in areas such as [explainable A](https://www.geeksforgeeks.org/introduction-to-explainable-aixai-using-lime/)I, trustworthy AI and AI safety to ensure that AI systems are transparent, reliable and safe to use.

# Machine Learning

Machine Learning tutorial covers basic and advanced concepts, specially designed to cater to both students and experienced working professionals.

This machine learning tutorial helps you gain a solid introduction to the fundamentals of machine learning and explore a wide range of techniques, including supervised, unsupervised, and reinforcement learning.

Machine learning (ML) is a subdomain of artificial intelligence (AI) that focuses on developing systems that learn—or improve performance—based on the data they ingest. Artificial intelligence is a broad word that refers to systems or machines that resemble human intelligence. Machine learning and AI are frequently discussed together, and the terms are occasionally used interchangeably, although they do not signify the same thing. A crucial distinction is that, while all machine learning is AI, not all AI is machine learning.

## What is Machine Learning?

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that 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.

## Features of Machine learning

* Machine learning is data driven technology. Large amount of data generated by organizations on daily bases. So, by notable relationships in data, organizations makes better decisions.
* Machine can learn itself from past data and automatically improve.
* From the given dataset it detects various patterns on data.
* For the big organizations branding is important and it will become more easy to target relatable customer base.
* It is similar to data mining because it is also deals with the huge amount of data.

# Introduction to Deep Learning

In the fast-evolving era of artificial intelligence, Deep Learning stands as a cornerstone technology, revolutionizing how machines understand, learn, and interact with complex data. At its essence, Deep Learning AI mimics the intricate neural networks of the human brain, enabling computers to autonomously discover patterns and make decisions from vast amounts of unstructured data. This transformative field has propelled breakthroughs across various domains, from computer vision and natural language processing to healthcare diagnostics and autonomous driving.

As we dive into this introductory exploration of Deep Learning, we uncover its foundational principles, applications, and the underlying mechanisms that empower machines to achieve human-like cognitive abilities. This article serves as a gateway into understanding how Deep Learning is reshaping industries, pushing the boundaries of what’s possible in AI, and paving the way for a future where intelligent systems can perceive, comprehend, and innovate autonomously.

## **What is Deep Learning?**

The definition of Deep learning is that it is the branch of [machine learning](https://www.geeksforgeeks.org/machine-learning/) that is based on artificial neural network architecture. An artificial neural network or [ANN](https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/)uses layers of interconnected nodes called neurons that work together to process and learn from the input data.

In a fully connected Deep neural network, there is an input layer and one or more hidden layers connected one after the other. Each neuron receives input from the previous layer neurons or the input layer. The output of one neuron becomes the input to other neurons in the next layer of the network, and this process continues until the final layer produces the output of the network. The layers of the neural network transform the input data through a series of nonlinear transformations, allowing the network to learn complex representations of the input data.

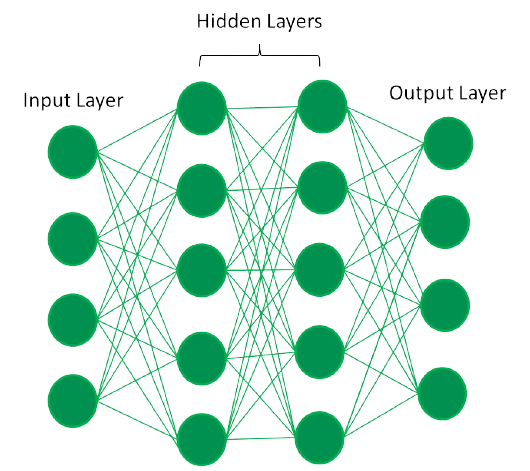
Today Deep learning AI has become one of the most popular and visible areas of machine learning, due to its success in a variety of applications, such as computer vision, natural language processing, and Reinforcement learning.

Deep learning AI can be used for supervised, unsupervised as well as reinforcement machine learning. it uses a variety of ways to process these.

* **Supervised Machine Learning:** [Supervised machine learning](https://www.geeksforgeeks.org/supervised-unsupervised-learning/) is the [machine learning](https://www.geeksforgeeks.org/machine-learning/) technique in which the neural network learns to make predictions or classify data based on the labeled datasets. Here we input both input features along with the target variables. the neural network learns to make predictions based on the cost or error that comes from the difference between the predicted and the actual target, this process is known as backpropagation.  Deep learning algorithms like Convolutional neural networks, Recurrent neural networks are used for many supervised tasks like image classifications and recognization, sentiment analysis, language translations, etc.
* **Unsupervised Machine Learning:** [Unsupervised machine learning](https://www.geeksforgeeks.org/supervised-unsupervised-learning/) is the [machine learning](https://www.geeksforgeeks.org/machine-learning/) technique in which the neural network learns to discover the patterns or to cluster the dataset based on unlabeled datasets. Here there are no target variables. while the machine has to self-determined the hidden patterns or relationships within the datasets. Deep learning algorithms like autoencoders and generative models are used for unsupervised tasks like clustering, dimensionality reduction, and anomaly detection.
* **Reinforcement  Machine Learning**: [Reinforcement  Machine Learning](https://www.geeksforgeeks.org/what-is-reinforcement-learning/) is the [machine learning](https://www.geeksforgeeks.org/machine-learning/) technique in which an agent learns to make decisions in an environment to maximize a reward signal. The agent interacts with the environment by taking action and observing the resulting rewards. Deep learning can be used to learn policies, or a set of actions, that maximizes the cumulative reward over time. Deep reinforcement learning algorithms like Deep Q networks and Deep Deterministic Policy Gradient (DDPG) are used to reinforce tasks like robotics and game playing etc.

### Artificial neural networks

[Artificial neural networks](https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/) are built on the principles of the structure and operation of human neurons. It is also known as neural networks or neural nets. An artificial neural network’s input layer, which is the first layer, receives input from external sources and passes it on to the hidden layer, which is the second layer. Each neuron in the hidden layer gets information from the neurons in the previous layer, computes the weighted total, and then transfers it to the neurons in the next layer. These connections are weighted, which means that the impacts of the inputs from the preceding layer are more or less optimized by giving each input a distinct weight. These weights are then adjusted during the training process to enhance the performance of the model.



*Fully Connected Artificial Neural Network*

Artificial neurons, also known as units, are found in artificial neural networks. The whole Artificial Neural Network is composed of these artificial neurons, which are arranged in a series of layers. The complexities of neural networks will depend on the complexities of the underlying patterns in the dataset whether a layer has a dozen units or millions of units.  Commonly, Artificial Neural Network has an input layer, an output layer as well as hidden layers. The input layer receives data from the outside world which the neural network needs to analyze or learn about.

In a fully connected artificial neural network, there is an input layer and one or more hidden layers connected one after the other. Each neuron receives input from the previous layer neurons or the input layer. The output of one neuron becomes the input to other neurons in the next layer of the network, and this process continues until the final layer produces the output of the network. Then, after passing through one or more hidden layers, this data is transformed into valuable data for the output layer. Finally, the output layer provides an output in the form of an artificial neural network’s response to the data that comes in.

Units are linked to one another from one layer to another in the bulk of neural networks. Each of these links has weights that control how much one unit influences another. The neural network learns more and more about the data as it moves from one unit to another, ultimately producing an output from the output layer.

## **Difference between Machine Learning and Deep Learning :**

[machine learning](https://www.geeksforgeeks.org/machine-learning/) and deep learning AI both are subsets of artificial intelligence but there are many similarities and differences between them.

| **Machine Learning** | **Deep Learning** |
| --- | --- |
| Apply statistical algorithms to learn the hidden patterns and relationships in the dataset. | Uses artificial neural network architecture to learn the hidden patterns and relationships in the dataset. |
| Can work on the smaller amount of dataset | Requires the larger volume of dataset compared to machine learning |
| Better for the low-label task. | Better for complex task like image processing, natural language processing, etc. |
| Takes less time to train the model. | Takes more time to train the model. |
| A model is created by relevant features which are manually extracted from images to detect an object in the image. | Relevant features are automatically extracted from images. It is an end-to-end learning process. |
| Less complex and easy to interpret the result. | More complex, it works like the black box interpretations of the result are not easy. |
| It can work on the CPU or requires less computing power as compared to deep learning. | It requires a high-performance computer with GPU. |

## Types of neural networks

Deep Learning models are able to automatically learn features from the data, which makes them well-suited for tasks such as image recognition, speech recognition, and natural language processing. The most widely used architectures in deep learning are feedforward neural networks, convolutional neural networks (CNNs), and recurrent neural networks (RNNs).

1. [Feedforward neural networks (FNNs)](https://www.geeksforgeeks.org/understanding-multi-layer-feed-forward-networks/) are the simplest type of ANN, with a linear flow of information through the network. FNNs have been widely used for tasks such as image classification, speech recognition, and natural language processing.
2. [Convolutional Neural Networks (CNNs)](https://www.geeksforgeeks.org/introduction-convolution-neural-network/) are specifically for image and video recognition tasks. CNNs are able to automatically learn features from the images, which makes them well-suited for tasks such as image classification, object detection, and image segmentation.
3. [Recurrent Neural Networks (RNNs)](https://www.geeksforgeeks.org/recurrent-neural-networks-explanation/)are a type of neural network that is able to process sequential data, such as time series and natural language. RNNs are able to maintain an internal state that captures information about the previous inputs, which makes them well-suited for tasks such as speech recognition, natural language processing, and language translation.

## **Deep Learning Applications:**

The main applications of deep learning AI can be divided into computer vision, natural language processing (NLP), and reinforcement learning.

### 1. [**Computer vision**](https://www.geeksforgeeks.org/applications-of-computer-vision/)

The first Deep Learning applications is Computer vision. In [computer vision](https://www.geeksforgeeks.org/applications-of-computer-vision/), Deep learning AI models can enable machines to identify and understand visual data. Some of the main applications of deep learning in computer vision include:

* **Object detection and recognition:**Deep learning model can be used to identify and locate objects within images and videos, making it possible for machines to perform tasks such as self-driving cars, surveillance, and robotics.
* **Image classification:**Deep learning models can be used to classify images into categories such as animals, plants, and buildings. This is used in applications such as medical imaging, quality control, and image retrieval.
* **Image segmentation:**Deep learning models can be used for image segmentation into different regions, making it possible to identify specific features within images.

### 2. [**Natural language processing (NLP)**](https://www.geeksforgeeks.org/natural-language-processing-nlp-tutorial/):

In Deep learning applications, second application is NLP. [NLP](https://www.geeksforgeeks.org/natural-language-processing-nlp-tutorial/), the Deep learning model can enable machines to understand and generate human language. Some of the main applications of deep learning in [NLP](https://www.geeksforgeeks.org/natural-language-processing-nlp-tutorial/) include:

* **Automatic Text Generation** – Deep learning model can learn the corpus of text and new text like summaries, essays can be automatically generated using these trained models.
* **Language translation:** Deep learning models can translate text from one language to another, making it possible to communicate with people from different linguistic backgrounds.
* **Sentiment analysis:**Deep learning models can analyze the sentiment of a piece of text, making it possible to determine whether the text is positive, negative, or neutral. This is used in applications such as customer service, social media monitoring, and political analysis.
* **Speech recognition:** Deep learning models can recognize and transcribe spoken words, making it possible to perform tasks such as speech-to-text conversion, voice search, and voice-controlled devices.

### 3. [**Reinforcement learning:**](https://www.geeksforgeeks.org/what-is-reinforcement-learning/)

In [reinforcement learning](https://www.geeksforgeeks.org/what-is-reinforcement-learning/), deep learning works as training agents to take action in an environment to maximize a reward. Some of the main applications of deep learning in reinforcement learning include:

* **Game playing:**Deep reinforcement learning models have been able to beat human experts at games such as Go, Chess, and Atari.
* **Robotics:**Deep reinforcement learning models can be used to train robots to perform complex tasks such as grasping objects, navigation, and manipulation.
* **Control systems:**Deep reinforcement learning models can be used to control complex systems such as power grids, traffic management, and supply chain optimization.

## **Challenges in Deep Learning**

Deep learning has made significant advancements in various fields, but there are still some challenges that need to be addressed. Here are some of the main challenges in deep learning:

1. **Data availability**: It requires large amounts of data to learn from. For using deep learning it’s a big concern to gather as much data for training.
2. **Computational Resources**: For training the deep learning model, it is computationally expensive because it requires specialized hardware like GPUs and TPUs.
3. **Time-consuming:** While working on sequential data depending on the computational resource it can take very large even in days or months.
4. I**nterpretability:**Deep learning models are complex, it works like a black box. it is very difficult to interpret the result.
5. **Overfitting:** when the model is trained again and again, it becomes too specialized for the training data, leading to overfitting and poor performance on new data.

## Advantages of Deep Learning:

1. **High accuracy:** Deep Learning algorithms can achieve state-of-the-art performance in various tasks, such as image recognition and natural language processing.
2. **Automated feature engineering:**Deep Learning algorithms can automatically discover and learn relevant features from data without the need for manual feature engineering.
3. **Scalability:** Deep Learning models can scale to handle large and complex datasets, and can learn from massive amounts of data.
4. **Flexibility:** Deep Learning models can be applied to a wide range of tasks and can handle various types of data, such as images, text, and speech.
5. **Continual improvement:** Deep Learning models can continually improve their performance as more data becomes available.

## Disadvantages of Deep Learning:

1. **High computational requirements:**Deep Learning AI models require large amounts of data and computational resources to train and optimize.
2. **Requires large amounts of labeled data**: Deep Learning models often require a large amount of labeled data for training, which can be expensive and time- consuming to acquire.
3. **Interpretability:** Deep Learning models can be challenging to interpret, making it difficult to understand how they make decisions.  
   **Overfitting:** Deep Learning models can sometimes overfit to the training data, resulting in poor performance on new and unseen data.
4. **Black-box nature**: Deep Learning models are often treated as black boxes, making it difficult to understand how they work and how they arrived at their predictions.

## Key Differences between Traditional Programming and Machine Learning

While both traditional programming and machine learning aim to solve problems and process data, they differ significantly in their approaches and methodologies. Here are some key differences between traditional programming and machine learning:

1. **Programming Paradigm:** Traditional programming follows a rule-based approach, where programmers explicitly define the rules and logic for the computer to follow. In contrast, machine learning follows a data-driven approach and focuses on learning patterns from data to make predictions or decisions.
2. **Flexibility:** Traditional programming is rigid, as it requires programmers to anticipate and define all possible scenarios and inputs. Any changes or updates to the program require manual modification of the code. Machine learning, on the other hand, is flexible and adaptable. The models can learn from new data and adjust their behavior accordingly without the need for explicit programming.
3. **Automation of Feature Extraction:** Traditional programming requires programmers to manually extract relevant features from raw data to train the models. In machine learning, the models can automate the feature extraction process, learning to extract relevant features and patterns from the data during the training phase.
4. **Handling Complexity:** Traditional programming is suitable for well-defined problems with fixed rules and inputs. Machine learning excels in handling complex and unstructured data, making it suitable for tasks such as natural language processing, image recognition, and sentiment analysis.
5. **Generalization:** Traditional programs are designed to perform specific tasks with explicit instructions, making them less capable of generalizing beyond the defined task. Machine learning models, once trained, can generalize their learning to make predictions on new, unseen data. This ability to generalize is a key strength of machine learning.
6. **Dependencies on Data:** Traditional programming relies on the programmer’s expertise and the predefined rules encoded in the program. Machine learning heavily relies on the quality and representativeness of the training data. Biased or incomplete data can lead to inaccurate predictions or biased models.
7. **Performance and Scalability:** Traditional programs can provide faster and deterministic results when the problem domain is well-defined. Machine learning algorithms, especially complex ones, might require more computational resources and time for training. However, once trained, machine learning models can provide efficient and scalable solutions to complex problems

# Understanding The Machine Learning Process: Key Steps and More

## **Machine Learning Process**

There are five main steps in the machine learning process:

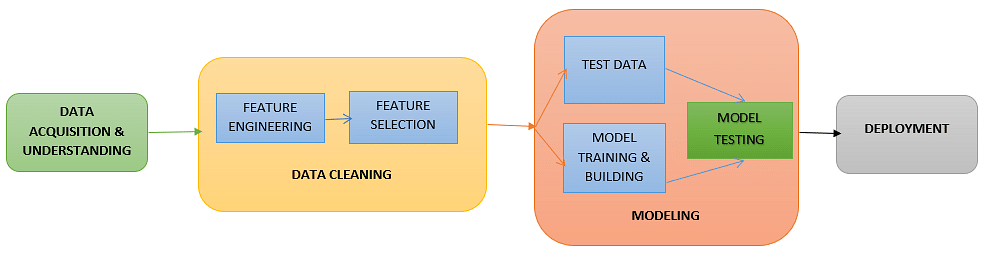


Fig: Machine learning process ([source](https://medium.com/analytics-vidhya/data-science-process-flow-and-machine-learning-fundamentals-6da8c591e8af))

### **Step 1: Data Acquisition**

The first step in the machine learning process is to get the [data](https://www.simplilearn.com/what-is-data-article). This will depend on the type of data you are gathering and the source of data. This can be either static data from an existing database or real-time data from an IoT system or data from other repositories.

### **Step 2: Data Cleaning**

All real-world data is often unorganized, redundant, or has missing elements. In order to feed data into the machine learning model, we need to first clean, prepare and manipulate the data. This is the most crucial step in the machine learning workflow and takes up the most time as well. Having clean data means that you can get a more accurate model down the road.

Data can be in any format - CSV, XML, JSON, etc. After cleaning the data, you need to then convert these data into valid formats that can be fed onto the [machine learning](https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-is-machine-learning) platform. Finally, these datasets are further divided into training and testing datasets. The training dataset is used to train the model. The testing dataset is used to validate the model.

Here are some things to keep in mind while splitting the dataset into training and testing sets:

* The split range is usually 20% to 80% between the testing and training stages
* You cannot mix or reuse the same data for the testing and training dataset
* Using the same data for both datasets can result in a faulty model

### **Step 3: Model Training**

The next step in the machine learning workflow is to train the model. A [machine learning algorithm](https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article) is used on the training dataset to train the model. This algorithm leverages mathematical modeling to learn and predict behaviors. These algorithms can fall into three broad categories - binary, classification, and regression.

### **Step 4: Model Testing**

After the model is trained, we need to test and validate it for further processing. By using the testing dataset obtained from Step 3, we can check the accuracy of the model. If the results are not satisfactory, the model should be further improved. The model is trained and improved over and over again until the results are satisfactory.

Here are some things you can do to refine and improve the model:

* Review the model with the business stakeholders and take in their inputs
* Reconsider the algorithm you have chosen to train the model
* Adjust the parameters of the algorithm you have chosen (even small adjustments can have significant impacts)

### **Step 5: Deployment**

Once the model is trained, deploy and pipeline it to production for application consumption.

The machine learning process that we have outlined here is a fairly standard process. As you go through this process on your own with your own problems, you will start to discover a few more [machine learning steps](https://www.simplilearn.com/tutorials/machine-learning-tutorial/machine-learning-steps) that might work for you. For example, as you clean your data, you may find better questions to ask or feed the model. As you tune your model, you may realize you need more data, and so on. The important part is to keep iterating until you find a model that fits your project the most

# What is a Dataset: Types

## What is a Dataset?

A **Dataset is a set of data grouped into a collection** with which developers can work to meet their goals. In a dataset, the rows represent the number of data points and the columns represent the features of the Dataset. They are mostly used in fields like machine learning, business, and government to gain insights, make informed decisions, or train algorithms. Datasets may vary in size and complexity and they mostly require cleaning and preprocessing to ensure data quality and suitability for analysis or modeling.

## Types of Datasets

There are various types of datasets available out there. They are:

* **Numerical Dataset:** They include numerical data points that can be solved with equations. These include temperature, humidity, marks and so on.
* **Categorical Dataset:** These include categories such as colour, gender, occupation, games, sports and so on.
* **Web Dataset:** These include datasets created by calling APIs using HTTP requests and populating them with values for data analysis. These are mostly stored in JSON (JavaScript Object Notation) formats.
* **Time series Dataset:**These include datasets between a period, for example, changes in geographical terrain over time.
* **Image Dataset:** It includes a dataset consisting of images. This is mostly used to differentiate the types of diseases, heart conditions and so on.
* **Ordered Dataset:**These datasets contain data that are ordered in ranks, for example, customer reviews, movie ratings and so on.
* **Partitioned Dataset:** These datasets have data points segregated into different members or different partitions.
* **File-Based Datasets:** These datasets are stored in files, in Excel as .csv, or .xlsx files.
* **Bivariate Dataset:** In this dataset, 2 classes or features are directly correlated to each other. For example, height and weight in a dataset are directly related to each other.
* **Multivariate Dataset:** In these types of datasets, as the name suggests 2 or more classes are directly correlated to each other. For example, attendance, and assignment grades are directly correlated to a student’s overall grade.

## Properties of Dataset

* **Center of data:** This refers to the “middle” value of the data, often measured by mean, median, or mode. It helps understand where most of the data points are concentrated.
* **Skewness of data:** This indicates how symmetrical the data distribution is. A perfectly symmetrical distribution (like a normal distribution) has a skewness of 0. Positive skewness means the data is clustered towards the left, while negative skewness means it’s clustered towards the right.
* **Spread among data members:** This describes how much the data points vary from the center. Common measures include standard deviation or variance, which quantify how far individual points deviate from the average.
* **Presence of outliers:** These are data points that fall significantly outside the overall pattern. Identifying outliers can be important as they might influence analysis results and require further investigation.
* **Correlation among the data:** This refers to the strength and direction of relationships between different variables in the dataset. A positive correlation indicates values in one variable tend to increase as the other does, while a negative correlation suggests they move in opposite directions. No correlation means there’s no linear relationship between the variables.
* **Type of probability distribution that the data follows:** Understanding the distribution (e.g., normal, uniform, binomial) helps us predict how likely it is to find certain values within the data and choose appropriate statistical methods for analysis

# Feature Selection Techniques in Machine Learning

## Feature selection:

Feature selection is a process that chooses a subset of features from the original features so that the feature space is optimally reduced according to a certain criterion.

 Feature selection is a critical step in the feature construction process. In text categorization problems, some words simply do not appear very often. Perhaps the word “groovy” appears in exactly one training document, which is positive. Is it really worth keeping this word around as a feature ? It’s a dangerous endeavor because it’s hard to tell with just one training example if it is really correlated with the positive class or is it just noise. You could hope that your learning algorithm is smart enough to figure it out. Or you could just remove it.

There are three general classes of feature selection algorithms: **Filter methods, wrapper methods and embedded methods**.

The role of feature selection in machine learning is,

1. To reduce the dimensionality of feature space.

2. To speed up a learning algorithm.

3. To improve the predictive accuracy of a classification algorithm.

4. To improve the comprehensibility of the learning results.

**Features Selection Algorithms are as follows:**

**1**. **Instance based approaches:** There is no explicit procedure for feature subset generation. Many small data samples are sampled from the data. Features are weighted according to their roles in differentiating instances of different classes for a data sample. Features with higher weights can be selected.

**2. Nondeterministic approaches:**Genetic algorithms and simulated annealing are also used in feature selection.

**3. Exhaustive complete approaches:**Branch and Bound evaluates estimated accuracy and ABB checks an inconsistency measure that is monotonic. Both start with a full feature set until the preset bound cannot be maintained.

While building a machine learning model for real-life dataset, we come across a lot of features in the dataset and not all these features are important every time. Adding unnecessary features while training the model leads us to reduce the overall accuracy of the model, increase the complexity of the model and decrease the generalization capability of the model and makes the model biased. Even the saying “Sometimes less is better” goes as well for the machine learning model. Hence, **feature selection** is one of the important steps while building a machine learning model. Its goal is to find the best possible set of features for building a machine learning model.

Some popular techniques of feature selection in machine learning are:

* Filter methods
* Wrapper methods
* Embedded methods

**Filter Methods**

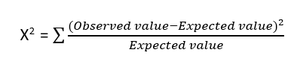
These methods are generally used while doing the pre-processing step. These methods select features from the dataset irrespective of the use of any machine learning algorithm. In terms of computation, they are very fast and inexpensive and are very good for removing duplicated, correlated, redundant features but these methods do not remove multicollinearity. Selection of feature is evaluated individually which can sometimes help when features are in isolation (don’t have a dependency on other features) but will lag when a combination of features can lead to increase in the overall performance of the model.



*Filter Methods Implementation*

Some techniques used are:

* **Information Gain –** It is defined as the amount of information provided by the feature for identifying the target value and measures reduction in the entropy values. Information gain of each attribute is calculated considering the target values for feature selection.
* **Chi-square test —** Chi-square method (X2) is generally used to test the relationship between categorical variables. It compares the observed values from different attributes of the dataset to its expected value.

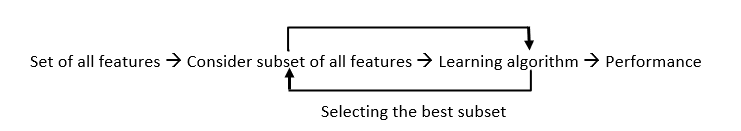


*Chi-square Formula*

* **Fisher’s Score –** Fisher’s Score selects each feature independently according to their scores under Fisher criterion leading to a suboptimal set of features. The larger the Fisher’s score is, the better is the selected feature.
* **Correlation Coefficient –** Pearson’s Correlation Coefficient is a measure of quantifying the association between the two continuous variables and the direction of the relationship with its values ranging from *-1 to 1*.
* **Variance Threshold –** It is an approach where all features are removed whose variance doesn’t meet the specific threshold. By default, this method removes features having zero variance. The assumption made using this method is higher variance features are likely to contain more information.
* **Mean Absolute Difference (MAD) –** This method is similar to variance threshold method but the difference is there is no square in MAD. This method calculates the mean absolute difference from the mean value.
* **Dispersion Ratio –** Dispersion ratio is defined as the ratio of the Arithmetic mean (AM) to that of Geometric mean (GM) for a given feature. Its value ranges from *+1 to ∞ as AM ≥ GM* for a given feature. Higher dispersion ratio implies a more relevant feature.
* **Mutual Dependence –**This method measures if two variables are mutually dependent, and thus provides the amount of information obtained for one variable on observing the other variable. Depending on the presence/absence of a feature, it measures the amount of information that feature contributes to making the target prediction.
* **Relief –** This method measures the quality of attributes by randomly sampling an instance from the dataset and updating each feature and distinguishing between instances that are near to each other based on the difference between the selected instance and two nearest instances of same and opposite classes.

**Wrapper methods:**

Wrapper methods, also referred to as greedy algorithms train the algorithm by using a subset of features in an iterative manner. Based on the conclusions made from training in prior to the model, addition and removal of features takes place. Stopping criteria for selecting the best subset are usually pre-defined by the person training the model such as when the performance of the model decreases or a specific number of features has been achieved. The main advantage of wrapper methods over the filter methods is that they provide an optimal set of features for training the model, thus resulting in better accuracy than the filter methods but are computationally more expensive.



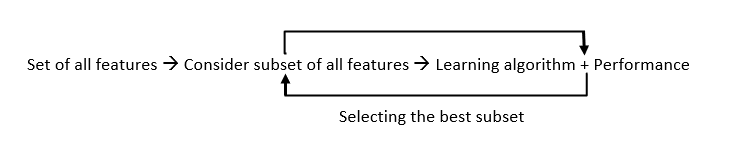
*Wrapper Methods Implementation*

Some techniques used are:

* **Forward selection –**This method is an iterative approach where we initially start with an empty set of features and keep adding a feature which best improves our model after each iteration. The stopping criterion is till the addition of a new variable does not improve the performance of the model.
* **Backward elimination –** This method is also an iterative approach where we initially start with all features and after each iteration, we remove the least significant feature. The stopping criterion is till no improvement in the performance of the model is observed after the feature is removed.
* **Bi-directional elimination –** This method uses both forward selection and backward elimination technique simultaneously to reach one unique solution.
* **Exhaustive selection –** This technique is considered as the brute force approach for the evaluation of feature subsets. It creates all possible subsets and builds a learning algorithm for each subset and selects the subset whose model’s performance is best.
* **Recursive elimination –** This greedy optimization method selects features by recursively considering the smaller and smaller set of features. The estimator is trained on an initial set of features and their importance is obtained using feature\_importance\_attribute. The least important features are then removed from the current set of features till we are left with the required number of features.

**Embedded methods:**

In embedded methods, the feature selection algorithm is blended as part of the learning algorithm, thus having its own built-in feature selection methods. Embedded methods encounter the drawbacks of filter and wrapper methods and merge their advantages. These methods are faster like those of filter methods and more accurate than the filter methods and take into consideration a combination of features as well.



*Embedded Methods Implementation*

Some techniques used are:

* **Regularization –** This method adds a penalty to different parameters of the machine learning model to avoid over-fitting of the model. This approach of feature selection uses Lasso (L1 regularization) and Elastic nets (L1 and L2 regularization). The penalty is applied over the coefficients, thus bringing down some coefficients to zero. The features having zero coefficient can be removed from the dataset.
* **Tree-based methods –**These methods such as Random Forest, Gradient Boosting provides us feature importance as a way to select features as well. Feature importance tells us which features are more important in making an impact on the target feature.

# Normal Distribution

## What is Normal Distribution?

We define Normal Distribution as the probability density function of any continuous random variable for any given system. Now for defining Normal Distribution suppose we take f(x) as the probability density function for any random variable X.

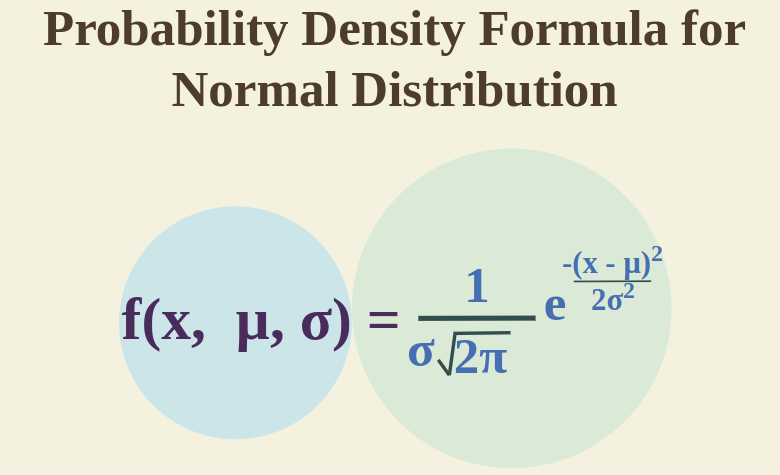
## Normal Distribution Examples

We can draw Normal Distribution for various types of data that include,

* Distribution of Height of People
* Distribution of Errors in any Measurement
* Distribution of Blood Pressure of any Patient, etc.

## Normal Distribution Formula

The formula for the probability density function of Normal Distribution (Gaussian Distribution) is added below,



*where,*

* *x is*[***Random Variable***](https://www.geeksforgeeks.org/random-variable/)
* *μ is*[***Mean***](https://www.geeksforgeeks.org/what-is-mean/)
* *σ is*[***Standard Deviation***](https://www.geeksforgeeks.org/standard-deviation-formula/)

## Normal Distribution Curve

In any **Normal Distribution, random variables are those variables that take unknown values related to the distribution and are generally bound by a range.**

An example of the random variable is, suppose take a **distribution of the height of students in a class then the random variable can take any value in this** **case but is bound by a boundary of 2 ft to 6 ft, as it is generally forced physically.**

* Range of any **normal distribution can be infinite in this case we say that normal distribution is not bothered by its range. In this case, range is extended from –∞ to + ∞.**
* Bell Curve still exist, in that case, **all the variables in that range are called Continuous variable** and their distribution is called Normal Distribution as all the values are generally closed aligned to the mean value.
* The **graph or the curve for the same is called the Normal Distribution Curve Or Normal Distribution Graph.**

## Normal Distribution Standard Deviation

We know that mean of any data spread out as a graph helps us to find the line of the symmetry of the graph whereas, Standard Deviation tells us how far the data is spread out from the mean value on either side. For smaller values of the standard deviation, the values in the graph come closer and the graph becomes narrower. While for higher values of the standard deviation the values in the graph are dispersed more and the graph becomes wider.

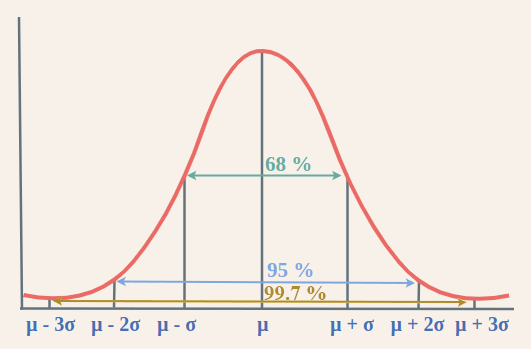
### **Empirical Rule of Standard Deviation**

Generally, the normal distribution has a positive standard deviation and the standard deviation divides the area of the normal curve into smaller parts and each part defines the percentage of data that falls into a specific region This is called the Empirical Rule of Standard Deviation in Normal Distribution.

***Empirical Rule states that,***

* *68% of the data approximately fall within one standard deviation of the mean, i.e. it falls between {****Mean – One Standard Deviation, and Mean + One Standard Deviation****}*
* *95% of the data approximately fall within two standard deviations of the mean, i.e. it falls between {****Mean – Two Standard Deviation, and Mean + Two Standard Deviation****}*
* *99.7% of the data approximately fall within a third standard deviation of the mean, i.e. it falls between {****Mean – Third Standard Deviation, and Mean + Third Standard Deviation****}*

## Normal Distribution Graph



Studying **the graph it is clear that using Empirical Rule we distribute data broadly in three parts. And thus, empirical rule is also called “68 – 95 – 99.7” rule.**

***Check:***[*Mathematics | Probability Distribution s Set 3 (Normal Distribution)*](https://www.geeksforgeeks.org/mathematics-probability-distributions-set-3-normal-distribution/)

## Normal Distribution Table

Normal Distribution Table which is also called, Normal Distribution Z Table is the table of z-value for normal distribution. This Normal Distribution Z Table is given as follows:

| **Z-Value** | **0** | **0.01** | **0.02** | **0.03** | **0.04** | **0.05** | **0.06** | **0.07** | **0.08** | **0.09** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0 | 0.004 | 0.008 | 0.012 | 0.016 | 0.0199 | 0.0239 | 0.0279 | 0.0319 | 0.0359 |
| **0.1** | 0.0398 | 0.0438 | 0.0478 | 0.0517 | 0.0557 | 0.0596 | 0.0636 | 0.0675 | 0.0714 | 0.0753 |
| **0.2** | 0.0793 | 0.0832 | 0.0871 | 0.091 | 0.0948 | 0.0987 | 0.1026 | 0.1064 | 0.1103 | 0.1141 |
| **0.3** | 0.1179 | 0.1217 | 0.1255 | 0.1293 | 0.1331 | 0.1368 | 0.1406 | 0.1443 | 0.148 | 0.1517 |
| **0.4** | 0.1554 | 0.1591 | 0.1628 | 0.1664 | 0.17 | 0.1736 | 0.1772 | 0.1808 | 0.1844 | 0.1879 |
| **0.5** | 0.1915 | 0.195 | 0.1985 | 0.2019 | 0.2054 | 0.2088 | 0.2123 | 0.2157 | 0.219 | 0.2224 |
| **0.6** | 0.2257 | 0.2291 | 0.2324 | 0.2357 | 0.2389 | 0.2422 | 0.2454 | 0.2486 | 0.2517 | 0.2549 |
| **0.7** | 0.258 | 0.2611 | 0.2642 | 0.2673 | 0.2704 | 0.2734 | 0.2764 | 0.2794 | 0.2823 | 0.2852 |
| **0.8** | 0.2881 | 0.291 | 0.2939 | 0.2967 | 0.2995 | 0.3023 | 0.3051 | 0.3078 | 0.3106 | 0.3133 |
| **0.9** | 0.3159 | 0.3186 | 0.3212 | 0.3238 | 0.3264 | 0.3289 | 0.3315 | 0.334 | 0.3365 | 0.3389 |
| **1** | 0.3413 | 0.3438 | 0.3461 | 0.3485 | 0.3508 | 0.3531 | 0.3554 | 0.3577 | 0.3599 | 0.3621 |
| **1.1** | 0.3643 | 0.3665 | 0.3686 | 0.3708 | 0.3729 | 0.3749 | 0.377 | 0.379 | 0.381 | 0.383 |
| **1.2** | 0.3849 | 0.3869 | 0.3888 | 0.3907 | 0.3925 | 0.3944 | 0.3962 | 0.398 | 0.3997 | 0.4015 |
| **1.3** | 0.4032 | 0.4049 | 0.4066 | 0.4082 | 0.4099 | 0.4115 | 0.4131 | 0.4147 | 0.4162 | 0.4177 |
| **1.4** | 0.4192 | 0.4207 | 0.4222 | 0.4236 | 0.4251 | 0.4265 | 0.4279 | 0.4292 | 0.4306 | 0.4319 |
| **1.5** | 0.4332 | 0.4345 | 0.4357 | 0.437 | 0.4382 | 0.4394 | 0.4406 | 0.4418 | 0.4429 | 0.4441 |
| **1.6** | 0.4452 | 0.4463 | 0.4474 | 0.4484 | 0.4495 | 0.4505 | 0.4515 | 0.4525 | 0.4535 | 0.4545 |
| **1.7** | 0.4554 | 0.4564 | 0.4573 | 0.4582 | 0.4591 | 0.4599 | 0.4608 | 0.4616 | 0.4625 | 0.4633 |
| **1.8** | 0.4641 | 0.4649 | 0.4656 | 0.4664 | 0.4671 | 0.4678 | 0.4686 | 0.4693 | 0.4699 | 0.4706 |
| **1.9** | 0.4713 | 0.4719 | 0.4726 | 0.4732 | 0.4738 | 0.4744 | 0.475 | 0.4756 | 0.4761 | 0.4767 |
| **2** | 0.4772 | 0.4778 | 0.4783 | 0.4788 | 0.4793 | 0.4798 | 0.4803 | 0.4808 | 0.4812 | 0.4817 |

## Properties of Normal Distribution

***Some important properties of normal distribution are,***

* *For normal distribution of data, mean, median, and mode are equal, (i.e.,****Mean = Median = Mode).***
* *Total area under the normal distribution curve is equal to 1.*
* *Normally distributed curve is symmetric at the center along the mean.*
* *In a normally distributed curve, there is exactly half value to the right of the central and exactly half value to the right side of the central value.*
* *Normal distribution is defined using the values of the mean and standard deviation.*
* *Normal distribution curve is a Unimodal Curve, i.e. a curve with only one peak*

### **People Also View:**

* [**Poisson Distribution**](https://www.geeksforgeeks.org/poisson-distribution-formula/)
* [**Binomial Distribution**](https://www.geeksforgeeks.org/what-is-binomial-probability-distribution-with-example/)
* [**Probability Distribution**](https://www.geeksforgeeks.org/probability-distribution/)

## Normal Distribution in Statistics

* **Normal distribution, also known as Gaussian distribution**, is a **bell-shaped curve that describes a large number of real-world phenomena**. It’s one of the most important concepts in statistics because it pops up in many areas of study.
* **Bell-Shaped Curve**: **Imagine a symmetrical bell where the middle is the highest point and the tails taper off on either side.**That’s the basic shape of a normal distribution. Most data points cluster around the center, and as you move further away from the center, the data points become less frequent.
* **Central Tendency: The center of the bell curve represents the central tendency of the data, which means it shows where most of the values are concentrated.** This could be the mean, median, or mode, depending on the specific data set.
* **Spread of Data:**The width of the bell curve indicates how spread out the data is a wider curve means the data points are more dispersed, while a narrower curve signifies the data points are closer together.
* **Random Variables: Normal distribution is typically used with continuous random variables, which can take on any value within a specific range.**Examples include heights, weights, IQ scores, or exam grades.

***Check****:*[*Normal Distribution in Business Statistics*](https://www.geeksforgeeks.org/normal-distribution-in-business-statistics/)

## Normal Distribution Problems and Solutions

Let’s solve some problems on Normal Distribution

**Example 1: Find the probability density function of the normal distribution of the following data. x = 2, μ = 3 and σ = 4.**

**Solution:**

*Given,*

* *Variable (x) = 2*
* *Mean = 3*
* *Standard Deviation = 4*

*Using formula of probability density of normal distribution*

*𝑓(𝑥,𝜇,𝜎)=1𝜎2𝜋𝑒−(𝑥−𝜇)22𝜎2f(x,μ,σ)=σ2π​1​e2σ2−(x−μ)2​*

*Simplifying,*

*f(2, 3, 4) = 0.09666703*

**Example 2: If the value of the random variable is 4, the mean is 4 and the standard deviation is 3, then find the probability density function of the Gaussian distribution.**

**Solution:**

*Given,*

* *Variable (x) = 4*
* *Mean = 4*
* *Standard Deviation = 3*

*Using formula of probability density of normal distribution*

*𝑓(𝑥,𝜇,𝜎)=1𝜎2𝜋𝑒−(𝑥−𝜇)22𝜎2f(x,μ,σ)=σ2π​1​e2σ2−(x−μ)2​*

*Simplifying,*

*f(4, 4, 3) = 1/(3√2π)e0*

*f(4, 4, 3) = 0.13301*

**Distance Measures (Euclidean and Manhattan)**

## What is Euclidean Distance?

The measure which gives the distance between any two points in an n-dimensional plane is known as Euclidean Distance. Euclidean distance between two points in the Euclidean space is defined as the length of the line segment between two points.

Euclidean distance is like measuring the straightest and shortest path between two points. Imagine you have a string and you stretch it tight between two points on a map; the length of that string is the Euclidean distance. It tells you how far apart the two points are without any turns or bends, just like a bird would fly directly from one spot to another.

## Euclidean Distance Formula

Consider two points (x1, y1) and (x2, y2) in a 2-dimensional space; the Euclidean Distance between them is given by using the formula:

***d = √[(x2 – x1)2 + (y2 – y1)2]***

Where,

* d is Euclidean Distance
* (x1, y1) is Coordinate of the first point
* (x2, y2) is Coordinate of the second point

### Euclidean Distance in 3D

If the two points (x1, y1, z1) and (x2, y2, z2) are in a 3-dimensional space, the Euclidean Distance between them is given by using the formula:

***d = √[(x2 – x1)2 + (y2 – y1)2+ (z2 – z1)2]***

## **Euclidean Distance and Manhattan Distance**

Differences between the Euclidean and Manhattan methods of measuring distance are listed in the following table:

| **Aspect** | **Euclidean Distance** | **Manhattan Distance** |
| --- | --- | --- |
| **Definition** | Measures the shortest straight-line distance between two points. | Measures the distance between two points along axes at right angles. |
| **Formula (2D)** | d = √[(x2 – x1)2 + (y2 – y1)2] | d = [|x2 – x1| + |y2 – y1|] |
| **Path** | Direct straight line. | Path that resembles city blocks or a grid pattern. |
| **Metric Name** | L2 norm or Euclidean norm. | L1 norm or Manhattan norm. |
| **Use Cases** | Used in scenarios where direct distances are needed (e.g., physics). | Commonly used in planning, urban design, and certain optimization algorithms. |
| **Sensitivity to Scaling** | Less sensitive to scaling of axes. | More sensitive to scaling of axes since it adds absolute differences. |

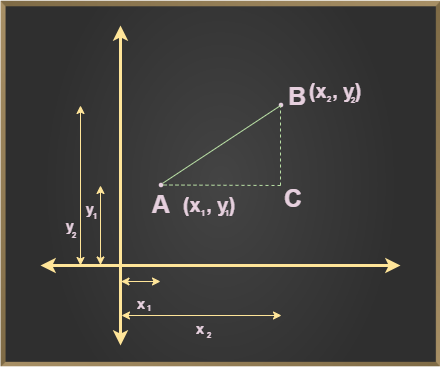
## Euclidean Distance Formula Derivation

Euclidean Distance Formula is derived by following the steps added below:

**Step 1:** Let us consider two points, A (x1, y1) and B (x2, y2), and d is the distance between the two points.

**Step 2:** Join the points using a straight line (AB).

**Step 3:** Now, let us construct a right-angled triangle whose hypotenuse is AB, as shown in the figure below.



**Step4:** Now, using [**Pythagoras theorem**](https://www.geeksforgeeks.org/pythagoras-theorem)we know that,

***(Hypotenuse)2 = (Base)2 + (Perpendicular)2***

⇒ **d2 = (x2 – x1)2 + (y2 – y1)2**

Now, take the square root on both sides of the equation, we get

***d = √(x2 – x1)2 + (y2 – y1)2***

## **Correlation Analysis**

Correlation analysis is a statistical technique for determining the strength of a link between two variables. It is used to detect patterns and trends in data and to forecast future occurrences.

* Consider a problem with different factors to be considered for making optimal conclusions
* Correlation explains how these variables are dependent on each other.
* Correlation quantifies how strong the relationship between two variables is. A higher value of the correlation coefficient implies a stronger association.
* The sign of the correlation coefficient indicates the direction of the relationship between variables. It can be either positive, negative, or zero.

### **What is Correlation?**

The Pearson correlation coefficient is the most often used metric of correlation. It expresses the linear relationship between two variables in numerical terms. The Pearson correlation coefficient, written as “r,” is as follows:

𝑟=∑(𝑥𝑖−𝑥ˉ)(𝑦𝑖−𝑦ˉ)∑(𝑥𝑖−𝑥ˉ)2∑(𝑦𝑖−𝑦ˉ)2*r*=∑(*xi*​−*x*ˉ)2∑(*yi*​−*y*ˉ​)2​∑(*xi*​−*x*ˉ)(*yi*​−*y*ˉ​)​

where,

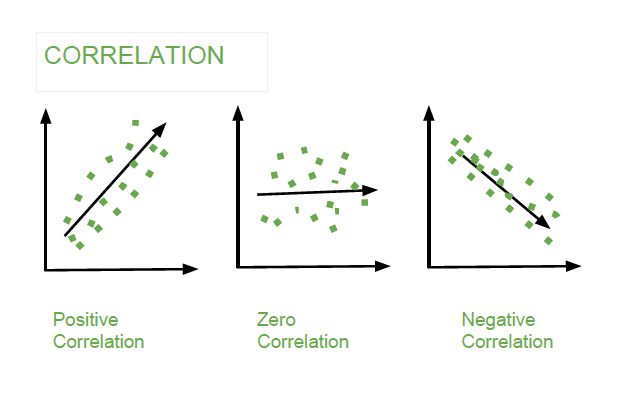
* r: Correlation coefficient
* 𝑥𝑖*xi*​ : i^th value first dataset X
* 𝑥ˉ*x*ˉ : Mean of first dataset X
* 𝑦𝑖*yi*​ : i^th value second dataset Y
* 𝑦ˉ*y*ˉ​ : Mean of second dataset Y

The [correlation coefficient](https://www.geeksforgeeks.org/correlation-coefficient-formula/), denoted by “r”, ranges between -1 and 1.

*r = -1 indicates a perfect negative correlation.  
r = 0 indicates no linear correlation between the variables.  
r = 1 indicates a perfect positive correlation.*

### **Types of Correlation**

There are three types of correlation:



*Correlation*

1. **Positive Correlation**: Positive correlation indicates that two variables have a direct relationship. As one variable increases, the other variable also increases. For example, there is a positive correlation between height and weight. As people get taller, they also tend to weigh more.
2. **Negative Correlation:**Negative correlation indicates that two variables have an inverse relationship. As one variable increases, the other variable decreases. For example, there is a negative correlation between price and demand. As the price of a product increases, the demand for that product decreases.
3. **Zero Correlation:** Zero correlation indicates that there is no relationship between two variables. The changes in one variable do not affect the other variable. For example, there is zero correlation between shoe size and intelligence

## **What is Cross-Validation?**

Cross validation is a technique used in machine learning to evaluate the performance of a model on unseen data. It involves dividing the available data into multiple folds or subsets, using one of these folds as a validation set, and training the model on the remaining folds. This process is repeated multiple times, each time using a different fold as the validation set. Finally, the results from each validation step are averaged to produce a more robust estimate of the model’s performance. Cross validation is an important step in the [machine learning](https://www.geeksforgeeks.org/machine-learning/) process and helps to ensure that the model selected for deployment is robust and generalizes well to new data.

## What is cross-validation used for?

The main purpose of cross validation is to prevent [overfitting](https://www.geeksforgeeks.org/overfitting-and-regularization-in-ml/), which occurs when a model is trained too well on the training data and performs poorly on new, unseen data. By evaluating the model on multiple validation sets, cross validation provides a more realistic estimate of the model’s generalization performance, i.e., its ability to perform well on new, unseen data.

## Types of Cross-Validation

There are several types of cross validation techniques, including **k-fold cross validation, leave-one-out cross validation, and Holdout validation, Stratified Cross-Validation.**The choice of technique depends on the size and nature of the data, as well as the specific requirements of the modeling problem.

### **1. Holdout Validation**

In[Holdout Validation](https://www.geeksforgeeks.org/introduction-of-holdout-method/), we perform training on the 50% of the given dataset and rest 50% is used for the testing purpose. It’s a simple and quick way to evaluate a model. The major drawback of this method is that we perform training on the 50% of the dataset, it may possible that the remaining 50% of the data contains some important information which we are leaving while training our model i.e. higher bias.

### **2. LOOCV (Leave One Out Cross Validation)**

In this method, we perform training on the whole dataset but leaves only one data-point of the available dataset and then iterates for each data-point. In [LOOCV](https://www.geeksforgeeks.org/loocvleave-one-out-cross-validation-in-r-programming/), the model is trained on  samples and tested on the one omitted sample, repeating this process for each data point in the dataset. It has some advantages as well as disadvantages also.

**An advantage** of using this method is that we make use of all data points and hence it is low bias.

The major**drawback**of this method is that it leads to **higher variation**in the testing model as we are testing against one data point. If the data point is an outlier it can lead to higher variation. Another drawback is it **takes a lot of execution time** as it iterates over ‘the number of data points’ times.

### **3. Stratified Cross-Validation**

It is a technique used in machine learning to ensure that each fold of the cross-validation process maintains the same class distribution as the entire dataset. This is particularly important when dealing with imbalanced datasets, where certain classes may be underrepresented. In this method,

1. The dataset is divided into k folds while maintaining the proportion of classes in each fold.
2. During each iteration, one-fold is used for testing, and the remaining folds are used for training.
3. The process is repeated k times, with each fold serving as the test set exactly once.

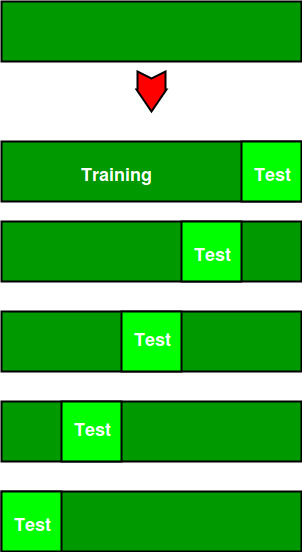
[Stratified Cross-Validation](https://www.geeksforgeeks.org/stratified-k-fold-cross-validation/)is essential when dealing with classification problems where maintaining the balance of class distribution is crucial for the model to generalize well to unseen data.

### **4. K-Fold Cross Validation**

In [K-Fold Cross Validation](https://www.geeksforgeeks.org/k-fold-cross-validation-in-r-programming/), we split the dataset into k number of subsets (known as folds) then we perform training on the all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purpose each time.

***Note:****It is always suggested that the value of k should be 10 as the lower value of k is takes towards validation and higher value of k leads to LOOCV method.*

#### **Example** of K Fold Cross Validation

The diagram below shows an example of the training subsets and evaluation subsets generated in k-fold cross-validation. Here, we have total 25 instances. In first iteration we use the first 20 percent of data for evaluation, and the remaining 80 percent for training ([1-5] testing and [5-25] training) while in the second iteration we use the second subset of 20 percent for evaluation, and the remaining three subsets of the data for training ([5-10] testing and [1-5 and 10-25] training), and so on.

Total instances: 25  
Value of k : 5   
No. Iteration Training set observations Testing set observations  
 1 [ 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24] [0 1 2 3 4]  
 2 [ 0 1 2 3 4 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24] [5 6 7 8 9]  
 3 [ 0 1 2 3 4 5 6 7 8 9 15 16 17 18 19 20 21 22 23 24] [10 11 12 13 14]  
 4 [ 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 20 21 22 23 24] [15 16 17 18 19]  
 5 [ 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19] [20 21 22 23 24]

## Comparison between cross-validation and hold out method

### **Advantages of train/test split:**

1. This runs K times faster than Leave One Out cross-validation because K-fold cross-validation repeats the train/test split K-times.
2. Simpler to examine the detailed results of the testing process.

**Advantages of cross-validation:**

1. More accurate estimate of out-of-sample accuracy.
2. More “efficient” use of data as every observation is used for both training and testing.

## Advantages and Disadvantages of Cross Validation

### Advantages:

1. Overcoming Overfitting: Cross validation helps to prevent overfitting by providing a more robust estimate of the model’s performance on unseen data.
2. Model Selection: Cross validation can be used to compare different models and select the one that performs the best on average.
3. Hyperparameter tuning: Cross validation can be used to optimize the hyperparameters of a model, such as the regularization parameter, by selecting the values that result in the best performance on the validation set.
4. Data Efficient: Cross validation allows the use of all the available data for both training and validation, making it a more data-efficient method compared to traditional validation techniques.

### Disadvantages:

1. Computationally Expensive: Cross validation can be computationally expensive, especially when the number of folds is large or when the model is complex and requires a long time to train.
2. Time-Consuming: Cross validation can be time-consuming, especially when there are many hyperparameters to tune or when multiple models need to be compared.
3. Bias-Variance Tradeoff: The choice of the number of folds in cross validation can impact the bias-variance tradeoff, i.e., too few folds may result in high variance, while too many folds may result in high bias.

# Feature Engineering: Scaling, Normalization, and Standardization

## What is Feature Scaling?

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If [feature scaling](https://www.geeksforgeeks.org/python-how-and-where-to-apply-feature-scaling/) is not done, then a [machine learning](https://www.geeksforgeeks.org/machine-learning/) algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

## Why use Feature Scaling?

In machine learning, feature scaling is employed for a number of purposes:

* Scaling guarantees that all features are on a comparable scale and have comparable ranges. This process is known as feature normalisation. This is significant because the magnitude of the features has an impact on many machine learning techniques. Larger scale features may dominate the learning process and have an excessive impact on the outcomes. You can avoid this problem and make sure that each feature contributes equally to the learning process by scaling the features.
* Algorithm performance improvement: When the features are scaled, several machine learning methods, including gradient descent-based algorithms, distance-based algorithms (such k-nearest neighbours), and support vector machines, perform better or converge more quickly. The algorithm’s performance can be enhanced by scaling the features, which can hasten the convergence of the algorithm to the ideal outcome.
* Preventing numerical instability: Numerical instability can be prevented by avoiding significant scale disparities between features. Examples include distance calculations or matrix operations, where having features with radically differing scales can result in numerical overflow or underflow problems. Stable computations are ensured and these issues are mitigated by scaling the features.
* Scaling features makes ensuring that each characteristic is given the same consideration during the learning process. Without scaling, bigger scale features could dominate the learning, producing skewed outcomes. This bias is removed through scaling, which also guarantees that each feature contributes fairly to model predictions.

### Absolute Maximum Scaling

This method of scaling requires two-step:

1. We should first select the maximum absolute value out of all the entries of a particular measure.
2. Then after this, we divide each entry of the column by this maximum value.

After performing the above-mentioned two steps we will observe that each entry of the column lies in the range of -1 to 1. But this method is not used that often the reason behind this is that it is too sensitive to the outliers. And while dealing with the real-world data presence of outliers is a very common thing.

For the demonstration purpose, we will use the dataset which you can download from [here](https://drive.google.com/file/d/1J7dPhnj2yBuzPwYraFU6cpCsa8Va3fiM/view?usp=share_link). This dataset is a simpler version of the original house price prediction dataset having only two columns from the original dataset. The first five rows of the original data are shown below:

* Python3

|  |
| --- |
| **import** pandas as pd  df **=** pd.read\_csv('SampleFile.csv')  print(df.head()) |

**Output:**

LotArea MSSubClass  
0 8450 60  
1 9600 20  
2 11250 60  
3 9550 70  
4 14260 60

Now let’s apply the first method which is of the absolute maximum scaling. For this first, we are supposed to evaluate the absolute maximum values of the columns.

* Python3

|  |
| --- |
| **import** numpy as np  max\_vals **=** np.max(np.abs(df))  max\_vals |

**Output:**

LotArea 215245  
MSSubClass 190  
dtype: int64

Now we are supposed to subtract these values from the data and then divide the results from the maximum values as well.

* Python3

|  |
| --- |
| print((df **-** max\_vals) **/** max\_vals) |

**Output:**

LotArea MSSubClass  
0 -0.960742 -0.684211  
1 -0.955400 -0.894737  
2 -0.947734 -0.684211  
3 -0.955632 -0.631579  
4 -0.933750 -0.684211  
... ... ...  
1455 -0.963219 -0.684211  
1456 -0.938791 -0.894737  
1457 -0.957992 -0.631579  
1458 -0.954856 -0.894737  
1459 -0.953834 -0.894737  
  
[1460 rows x 2 columns]

## Min-Max Scaling

This method of scaling requires below two-step:

1. First, we are supposed to find the minimum and the maximum value of the column.
2. Then we will subtract the minimum value from the entry and divide the result by the difference between the maximum and the minimum value.

As we are using the maximum and the minimum value this method is also prone to [outliers](https://www.geeksforgeeks.org/machine-learning-outlier/) but the range in which the data will range after performing the above two steps is between 0 to 1.

* Python3

|  |
| --- |
| **from** sklearn.preprocessing **import** MinMaxScaler    scaler **=** MinMaxScaler()  scaled\_data **=** scaler.fit\_transform(df)  scaled\_df **=** pd.DataFrame(scaled\_data,                           columns**=**df.columns)  scaled\_df.head() |

**Output:**

LotArea MSSubClass  
0 0.033420 0.235294  
1 0.038795 0.000000  
2 0.046507 0.235294  
3 0.038561 0.294118  
4 0.060576 0.235294

## Normalization

This method is more or less the same as the previous method but here instead of the minimum value, we subtract each entry by the mean value of the whole data and then divide the results by the difference between the minimum and the maximum value.

* Python3

|  |
| --- |
| **from** sklearn.preprocessing **import** Normalizer    scaler **=** Normalizer()  scaled\_data **=** scaler.fit\_transform(df)  scaled\_df **=** pd.DataFrame(scaled\_data,                           columns**=**df.columns)  print(scaled\_df.head()) |

**Output:**

LotArea MSSubClass  
0 0.999975 0.007100  
1 0.999998 0.002083  
2 0.999986 0.005333  
3 0.999973 0.007330  
4 0.999991 0.004208

## Standardization

This method of scaling is basically based on the central tendencies and variance of the data.

1. First, we should calculate the [mean and standard deviation](https://www.geeksforgeeks.org/mathematics-mean-variance-and-standard-deviation/) of the data we would like to normalize.
2. Then we are supposed to subtract the mean value from each entry and then divide the result by the standard deviation.

This helps us achieve a [normal distribution](https://www.geeksforgeeks.org/mathematics-probability-distributions-set-3-normal-distribution/)(if it is already normal but skewed) of the data with a mean equal to zero and a standard deviation equal to 1.

* Python3

|  |
| --- |
| **from** sklearn.preprocessing **import** StandardScaler    scaler **=** StandardScaler()  scaled\_data **=** scaler.fit\_transform(df)  scaled\_df **=** pd.DataFrame(scaled\_data,                           columns**=**df.columns)  print(scaled\_df.head()) |

**Output:**

LotArea MSSubClass  
0 -0.207142 0.073375  
1 -0.091886 -0.872563  
2 0.073480 0.073375  
3 -0.096897 0.309859  
4 0.375148 0.073375

## Robust Scaling

In this method of scaling, we use two main statistical measures of the data.

* [Median](https://www.geeksforgeeks.org/median/)
* [Inter-Quartile Range](https://www.geeksforgeeks.org/interquartile-range-formula/)

After calculating these two values we are supposed to subtract the median from each entry and then divide the result by the interquartile range.

* Python3

|  |
| --- |
| **from** sklearn.preprocessing **import** RobustScaler    scaler **=** RobustScaler()  scaled\_data **=** scaler.fit\_transform(df)  scaled\_df **=** pd.DataFrame(scaled\_data,                           columns**=**df.columns)  print(scaled\_df.head()) |

**Output:**

LotArea MSSubClass  
0 -0.254076 0.2  
1 0.030015 -0.6  
2 0.437624 0.2  
3 0.017663 0.4  
4 1.181201 0.2

# Types of Machine Learning

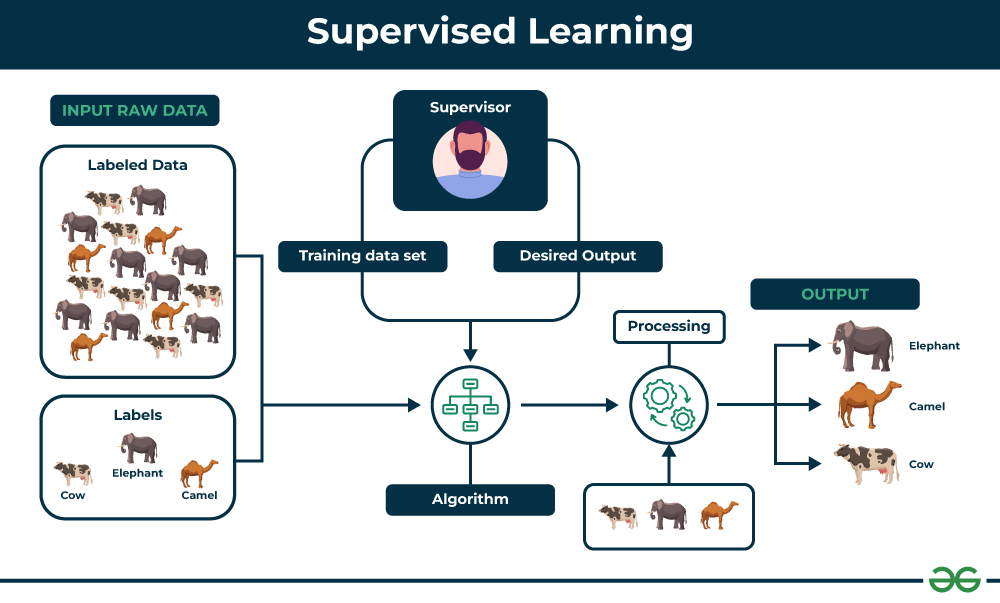
## Types of Machine Learning

There are several types of machine learning, each with special characteristics and applications. Some of the main types of machine learning algorithms are as follows:

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

Here are some regression algorithms:

* [**Linear Regression**](https://www.geeksforgeeks.org/ml-linear-regression/)
* [**Polynomial Regression**](https://www.geeksforgeeks.org/videos/polynomial-regression-algorithm-machine-learning/)
* [**Ridge Regression**](https://www.geeksforgeeks.org/videos/lasso-ridge-regression-algorithm-machine-learning/)
* [**Lasso Regression**](https://www.geeksforgeeks.org/videos/lasso-ridge-regression-algorithm-machine-learning/)
* [**Decision tree**](https://www.geeksforgeeks.org/decision-tree-introduction-example/)
* [**Random Forest**](https://www.geeksforgeeks.org/random-forest-regression-in-python/)

#### Advantages of Supervised Machine Learning

* **Supervised Learning** models can have high accuracy as they are trained on **labelled data**.
* The process of decision-making in supervised learning models is often interpretable.
* It can often be used in pre-trained models which saves time and resources when developing new models from scratch.

#### Disadvantages of Supervised Machine Learning

* It has limitations in knowing patterns and may struggle with unseen or unexpected patterns that are not present in the training data.
* It can be time-consuming and costly as it relies on**labeled**data only.
* It may lead to poor generalizations based on new data.

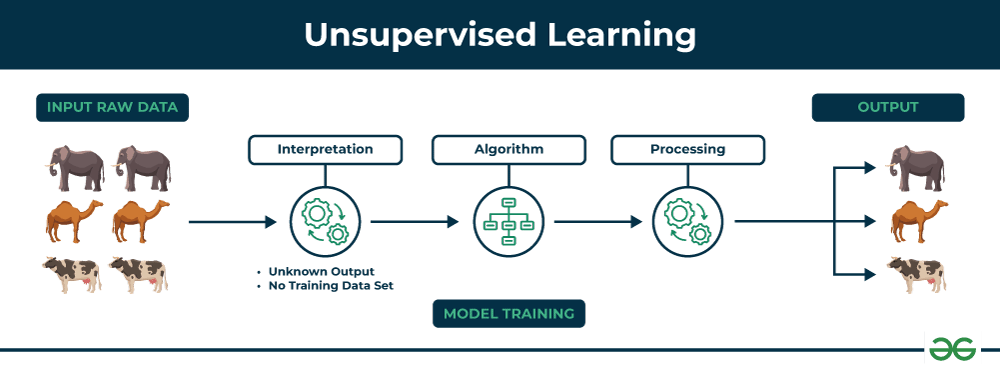
#### Applications of Supervised Learning

Supervised learning is used in a wide variety of applications, including:

* **Image classification**: Identify objects, faces, and other features in images.
* **Natural language processing:** Extract information from text, such as sentiment, entities, and relationships.
* **Speech recognition**: Convert spoken language into text.
* **Recommendation systems**: Make personalized recommendations to users.
* **Predictive analytics**: Predict outcomes, such as sales, customer churn, and stock prices.
* **Medical diagnosis**: Detect diseases and other medical conditions.
* **Fraud detection**: Identify fraudulent transactions.
* **Autonomous vehicles**: Recognize and respond to objects in the environment.
* **Email spam detection**: Classify emails as spam or not spam.
* **Quality control in manufacturing**: Inspect products for defects.
* **Credit scoring**: Assess the risk of a borrower defaulting on a loan.
* **Gaming**: Recognize characters, analyze player behavior, and create NPCs.
* **Customer support**: Automate customer support tasks.
* **Weather forecasting**: Make predictions for temperature, precipitation, and other meteorological parameters.
* **Sports analytics**: Analyze player performance, make game predictions, and optimize strategies.

### 2. Unsupervised Machine Learning

[Unsupervised Learning](https://www.geeksforgeeks.org/unsupervised-machine-learning-the-future-of-cybersecurity/) Unsupervised learning is a type of machine learning technique in which an algorithm discovers patterns and relationships using unlabeled data. Unlike supervised learning, unsupervised learning doesn’t involve providing the algorithm with labeled target outputs. The primary goal of Unsupervised learning is often to discover hidden patterns, similarities, or clusters within the data, which can then be used for various purposes, such as data exploration, visualization, dimensionality reduction, and more.



*Unsupervised Learning*

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

**Example:**Consider that you have a dataset that contains information about the purchases you made from the shop. Through clustering, the algorithm can group the same purchasing behavior among you and other customers, which reveals potential customers without predefined labels. This type of information can help businesses get target customers as well as identify outliers.

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

* [Clustering](https://www.geeksforgeeks.org/clustering-in-machine-learning/)
* [Association](https://www.geeksforgeeks.org/association-rule/)

#### Clustering

[Clustering](https://www.geeksforgeeks.org/clustering-in-machine-learning/) is the process of grouping data points into clusters based on their similarity. This technique is useful for identifying patterns and relationships in data without the need for labeled examples.

Here are some clustering algorithms:

* [**K-Means Clustering algorithm**](https://www.geeksforgeeks.org/k-means-clustering-introduction/)
* [**Mean-shift algorithm**](https://www.geeksforgeeks.org/ml-mean-shift-clustering/)
* [**DBSCAN Algorithm**](https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/)
* [**Principal Component Analysis**](https://www.geeksforgeeks.org/principal-component-analysis-pca/)
* [**Independent Component Analysis**](https://www.geeksforgeeks.org/ml-independent-component-analysis/)

#### Association

[Association rule learn](https://www.geeksforgeeks.org/association-rule/)ing is a technique for discovering relationships between items in a dataset. It identifies rules that indicate the presence of one item implies the presence of another item with a specific probability.

Here are some association rule learning algorithms:

* [**Apriori Algorithm**](https://www.geeksforgeeks.org/apriori-algorithm/)
* [**Eclat**](https://www.geeksforgeeks.org/ml-eclat-algorithm/)
* [**FP-growth Algorithm**](https://www.geeksforgeeks.org/frequent-pattern-growth-algorithm/)

#### Advantages of Unsupervised Machine Learning

* It helps to discover hidden patterns and various relationships between the data.
* Used for tasks such as**customer segmentation, anomaly detection,**and **data exploration**.
* It does not require labeled data and reduces the effort of data labeling.

#### Disadvantages of Unsupervised Machine Learning

* Without using labels, it may be difficult to predict the quality of the model’s output.
* Cluster Interpretability may not be clear and may not have meaningful interpretations.
* It has techniques such as[autoencoders](https://www.geeksforgeeks.org/auto-encoders/) and [dimensionality reduction](https://www.geeksforgeeks.org/dimensionality-reduction/) that can be used to extract meaningful features from raw data.

#### Applications of Unsupervised Learning

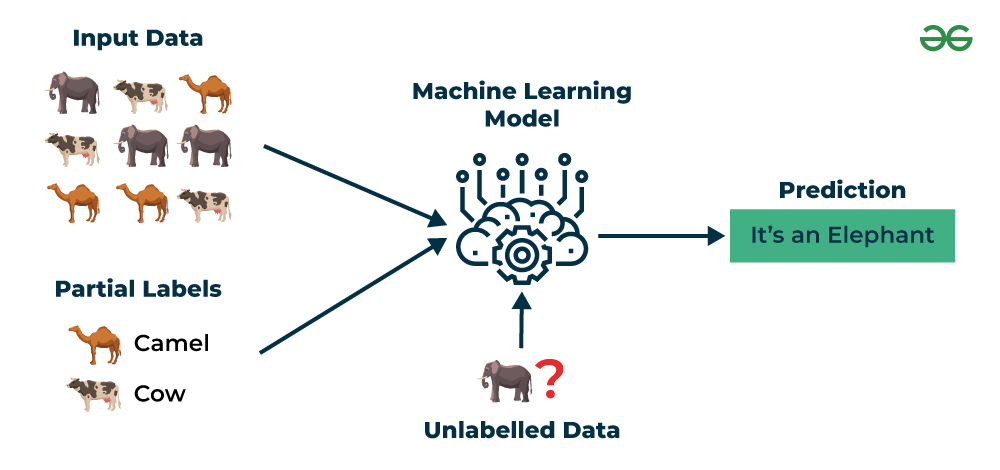
Here are some common applications of unsupervised learning:

* **Clustering**: Group similar data points into clusters.
* **Anomaly detection**: Identify outliers or anomalies in data.
* **Dimensionality reduction**: Reduce the dimensionality of data while preserving its essential information.
* **Recommendation systems**: Suggest products, movies, or content to users based on their historical behavior or preferences.
* **Topic modeling**: Discover latent topics within a collection of documents.
* **Density estimation**: Estimate the probability density function of data.
* **Image and video compression**: Reduce the amount of storage required for multimedia content.
* **Data preprocessing**: Help with data preprocessing tasks such as data cleaning, imputation of missing values, and data scaling.
* **Market basket analysis**: Discover associations between products.
* **Genomic data analysis**: Identify patterns or group genes with similar expression profiles.
* **Image segmentation**: Segment images into meaningful regions.
* **Community detection in social networks**: Identify communities or groups of individuals with similar interests or connections.
* **Customer behavior analysis**: Uncover patterns and insights for better marketing and product recommendations.
* **Content recommendation**: Classify and tag content to make it easier to recommend similar items to users.
* **Exploratory data analysis (EDA)**: Explore data and gain insights before defining specific tasks.

### 3. Semi-Supervised Learning

[Semi-Supervised learning](https://www.geeksforgeeks.org/ml-semi-supervised-learning/)is a machine learning algorithm that works between the [supervised and unsupervised](https://www.geeksforgeeks.org/supervised-unsupervised-learning/) learning so it uses both **labelled and unlabelled** data. It’s particularly useful when obtaining labeled data is costly, time-consuming, or resource-intensive. This approach is useful when the dataset is expensive and time-consuming. Semi-supervised learning is chosen when labeled data requires skills and relevant resources in order to train or learn from it.

We use these techniques when we are dealing with data that is a little bit labeled and the rest large portion of it is unlabeled. We can use the unsupervised techniques to predict labels and then feed these labels to supervised techniques. This technique is mostly applicable in the case of image data sets where usually all images are not labeled.



*Semi-Supervised Learning*

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

**Example**: Consider that we are building a language translation model, having labeled translations for every sentence pair can be resources intensive. It allows the models to learn from labeled and unlabeled sentence pairs, making them more accurate. This technique has led to significant improvements in the quality of machine translation services.

#### Types of Semi-Supervised Learning Methods

There are a number of different semi-supervised learning methods each with its own characteristics. Some of the most common ones include:

* **Graph-based semi-supervised learning:** This approach uses a graph to represent the relationships between the data points. The graph is then used to propagate labels from the labeled data points to the unlabeled data points.
* **Label propagation:** This approach iteratively propagates labels from the labeled data points to the unlabeled data points, based on the similarities between the data points.
* **Co-training:** This approach trains two different machine learning models on different subsets of the unlabeled data. The two models are then used to label each other’s predictions.
* **Self-training:** This approach trains a machine learning model on the labeled data and then uses the model to predict labels for the unlabeled data. The model is then retrained on the labeled data and the predicted labels for the unlabeled data.
* [**Generative adversarial networks (GANs)**](https://www.geeksforgeeks.org/generative-adversarial-network-gan/)**:** GANs are a type of deep learning algorithm that can be used to generate synthetic data. GANs can be used to generate unlabeled data for semi-supervised learning by training two neural networks, a generator and a discriminator.

#### Advantages of Semi- Supervised Machine Learning

* It leads to better generalization as compared to **supervised learning,** as it takes both labeled and unlabeled data.
* Can be applied to a wide range of data.

#### Disadvantages of Semi- Supervised Machine Learning

* **Semi-supervised**methods can be more complex to implement compared to other approaches.
* It still requires some **labeled data** that might not always be available or easy to obtain.
* The unlabeled data can impact the model performance accordingly.

#### Applications of Semi-Supervised Learning

Here are some common applications of semi-supervised learning:

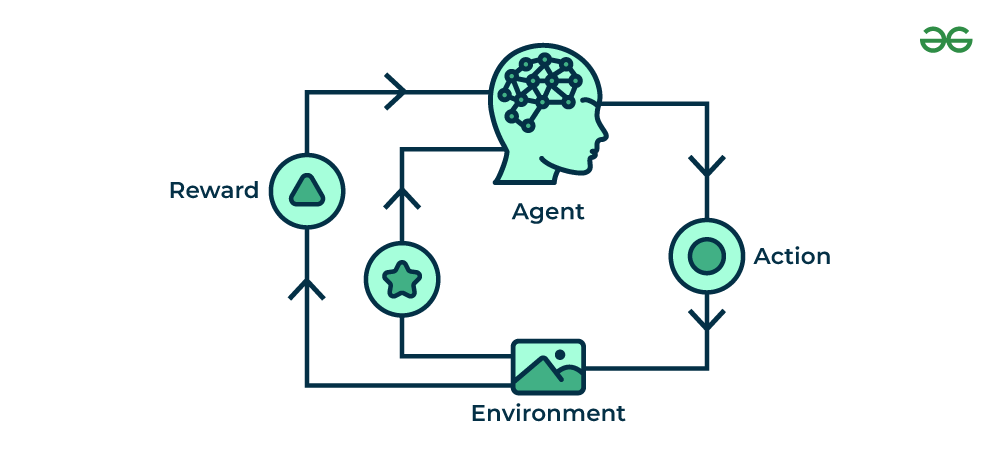
* **Image Classification and Object Recognition**: Improve the accuracy of models by combining a small set of labeled images with a larger set of unlabeled images.
* **Natural Language Processing (NLP)**: Enhance the performance of language models and classifiers by combining a small set of labeled text data with a vast amount of unlabeled text.
* **Speech Recognition:** Improve the accuracy of speech recognition by leveraging a limited amount of transcribed speech data and a more extensive set of unlabeled audio.
* **Recommendation Systems**: Improve the accuracy of personalized recommendations by supplementing a sparse set of user-item interactions (labeled data) with a wealth of unlabeled user behavior data.
* **Healthcare and Medical Imaging**: Enhance medical image analysis by utilizing a small set of labeled medical images alongside a larger set of unlabeled images.

### 4. Reinforcement Machine Learning

[Reinforcement machine learning](https://www.geeksforgeeks.org/what-is-reinforcement-learning/)algorithm is a learning method that interacts with the environment by producing actions and discovering errors. **Trial, error, and delay** are the most relevant characteristics of reinforcement learning. In this technique, the model keeps on increasing its performance using Reward Feedback to learn the behavior or pattern. These algorithms are specific to a particular problem e.g. Google Self Driving car, AlphaGo where a bot competes with humans and even itself to get better and better performers in Go Game. Each time we feed in data, they learn and add the data to their knowledge which is training data. So, the more it learns the better it gets trained and hence experienced.

Here are some of most common reinforcement learning algorithms:

* [**Q-learning:**](https://www.geeksforgeeks.org/q-learning-in-python/) Q-learning is a model-free RL algorithm that learns a Q-function, which maps states to actions. The Q-function estimates the expected reward of taking a particular action in a given state.
* [**SARSA (State-Action-Reward-State-Action):**](https://www.geeksforgeeks.org/sarsa-reinforcement-learning/) SARSA is another model-free RL algorithm that learns a Q-function. However, unlike Q-learning, SARSA updates the Q-function for the action that was actually taken, rather than the optimal action.
* [**Deep Q-learning**](https://www.geeksforgeeks.org/deep-q-learning/)**:** Deep Q-learning is a combination of Q-learning and deep learning. Deep Q-learning uses a neural network to represent the Q-function, which allows it to learn complex relationships between states and actions.



*Reinforcement Machine Learning*

Let’s understand it with the help of examples.

**Example:**Consider that you are training an [AI](https://www.geeksforgeeks.org/artificial-intelligence-an-introduction/) agent to play a game like chess. The agent explores different moves and receives positive or negative feedback based on the outcome. Reinforcement Learning also finds applications in which they learn to perform tasks by interacting with their surroundings.

#### Types of Reinforcement Machine Learning

There are two main types of reinforcement learning:

**Positive reinforcement**

* Rewards the agent for taking a desired action.
* Encourages the agent to repeat the behavior.
* Examples: Giving a treat to a dog for sitting, providing a point in a game for a correct answer.

**Negative reinforcement**

* Removes an undesirable stimulus to encourage a desired behavior.
* Discourages the agent from repeating the behavior.
* Examples: Turning off a loud buzzer when a lever is pressed, avoiding a penalty by completing a task.

#### Advantages of Reinforcement Machine Learning

* It has autonomous decision-making that is well-suited for tasks and that can learn to make a sequence of decisions, like robotics and game-playing.
* This technique is preferred to achieve long-term results that are very difficult to achieve.
* It is used to solve a complex problems that cannot be solved by conventional techniques.

#### Disadvantages of Reinforcement Machine Learning

* Training Reinforcement Learning agents can be computationally expensive and time-consuming.
* Reinforcement learning is not preferable to solving simple problems.
* It needs a lot of data and a lot of computation, which makes it impractical and costly.

#### Applications of Reinforcement Machine Learning

Here are some applications of reinforcement learning:

* **Game Playing**: RL can teach agents to play games, even complex ones.
* **Robotics**: RL can teach robots to perform tasks autonomously.
* **Autonomous Vehicles**: RL can help self-driving cars navigate and make decisions.
* **Recommendation Systems**: RL can enhance recommendation algorithms by learning user preferences.
* **Healthcare**: RL can be used to optimize treatment plans and drug discovery.
* **Natural Language Processing (NLP)**: RL can be used in dialogue systems and chatbots.
* **Finance and Trading**: RL can be used for algorithmic trading.
* **Supply Chain and Inventory Management**: RL can be used to optimize supply chain operations.
* **Energy Management**: RL can be used to optimize energy consumption.
* **Game AI**: RL can be used to create more intelligent and adaptive NPCs in video games.
* **Adaptive Personal Assistants**: RL can be used to improve personal assistants.
* **Virtual Reality (VR) and Augmented Reality (AR):** RL can be used to create immersive and interactive experiences.
* **Industrial Control**: RL can be used to optimize industrial processes.
* **Education**: RL can be used to create adaptive learning systems.
* **Agriculture**: RL can be used to optimize agricultural operations.

**Components of Generalization Error**

## Bias and Variance in Machine Learning

* [**Bias**](https://www.geeksforgeeks.org/bias-vs-variance-in-machine-learning/): Bias refers to the error due to overly simplistic assumptions in the learning algorithm. These assumptions make the model easier to comprehend and learn but might not capture the underlying complexities of the data. It is the error due to the model’s inability to represent the true relationship between input and output accurately. When a model has poor performance both on the training and testing data means high bias because of the simple model, indicating underfitting.
* [**Variance**](https://www.geeksforgeeks.org/bias-vs-variance-in-machine-learning/): Variance, on the other hand, is the error due to the model’s sensitivity to fluctuations in the training data. It’s the variability of the model’s predictions for different instances of training data. High variance occurs when a model learns the training data’s noise and random fluctuations rather than the underlying pattern. As a result, the model performs well on the training data but poorly on the testing data, indicating overfitting.

## **Underfitting in Machine Learning**

A [statistical model](https://www.geeksforgeeks.org/difference-between-statistical-model-and-machine-learning/) or a machine learning algorithm is said to have underfitting when a model is too simple to capture data complexities. It represents the inability of the model to learn the training data effectively result in poor performance both on the training and testing data. In simple terms, an underfit model’s are inaccurate, especially when applied to new, unseen examples. It mainly happens when we uses very simple model with overly simplified assumptions. To address underfitting problem of the model, we need to use more complex models, with enhanced feature representation, and less regularization.

**Note: The underfitting model has High bias and low variance.**

### **Reasons for** **Underfitting**

1. The model is too simple, So it may be not capable to represent the complexities in the data.
2. The input features which is used to train the model is not the adequate representations of underlying factors influencing the target variable.
3. The size of the training dataset used is not enough.
4. Excessive regularization are used to prevent the overfitting, which constraint the model to capture the data well.
5. Features are not scaled.

### **Techniques to Reduce Underfitting**

1. Increase model complexity.
2. Increase the number of features, performing [feature engineering](https://www.geeksforgeeks.org/what-is-feature-engineering/).
3. Remove noise from the data.
4. Increase the number of [epochs](https://www.geeksforgeeks.org/epoch-in-machine-learning/) or increase the duration of training to get better results.

## **Overfitting in Machine Learning**

A [statistical model](https://www.geeksforgeeks.org/difference-between-statistical-model-and-machine-learning/) is said to be overfitted when the model does not make accurate predictions on testing data. When a model gets trained with so much data, it starts learning from the noise and inaccurate data entries in our data set. And when testing with test data results in High variance. Then the model does not categorize the data correctly, because of too many details and noise. The causes of overfitting are the non-parametric and non-linear methods because these types of machine learning algorithms have more freedom in building the model based on the dataset and therefore they can really build unrealistic models. A solution to avoid overfitting is using a linear algorithm if we have linear data or using the parameters like the maximal depth if we are using decision trees.

In a nutshell, [Overfitting](https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/) is a problem where the evaluation of machine learning algorithms on training data is different from unseen data.

### Reasons for Overfitting:

1. High variance and low bias.
2. The model is too complex.
3. The size of the training data.

### **Techniques to Reduce Overfitting**

1. Improving the quality of training data reduces overfitting by focusing on meaningful patterns, mitigate the risk of fitting the noise or irrelevant features.
2. Increase the training data can improve the model’s ability to generalize to unseen data and reduce the likelihood of overfitting.
3. Reduce model complexity.
4. [Early stopping](https://www.geeksforgeeks.org/regularization-by-early-stopping/) during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
5. [Ridge Regularization](https://www.geeksforgeeks.org/lasso-vs-ridge-vs-elastic-net-ml/) and [Lasso Regularization](https://www.geeksforgeeks.org/implementation-of-lasso-regression-from-scratch-using-python/).
6. Use [dropout](https://www.geeksforgeeks.org/dropout-in-neural-networks/) for [neural networks](https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/) to tackle overfitting.



*Underfitting and Overfitting*

## **Good Fit in a Statistical Model**

Ideally, the case when the model makes the predictions with 0 error, is said to have a good fit on the data. This situation is achievable at a spot between overfitting and underfitting. In order to understand it, we will have to look at the performance of our model with the passage of time, while it is learning from the training dataset.

With the passage of time, our model will keep on learning, and thus the error for the model on the training and testing data will keep on decreasing. If it will learn for too long, the model will become more prone to overfitting due to the presence of noise and less useful details. Hence the performance of our model will decrease. In order to get a good fit, we will stop at a point just before where the error starts increasing. At this point, the model is said to have good skills in training datasets as well as our unseen testing dataset

s