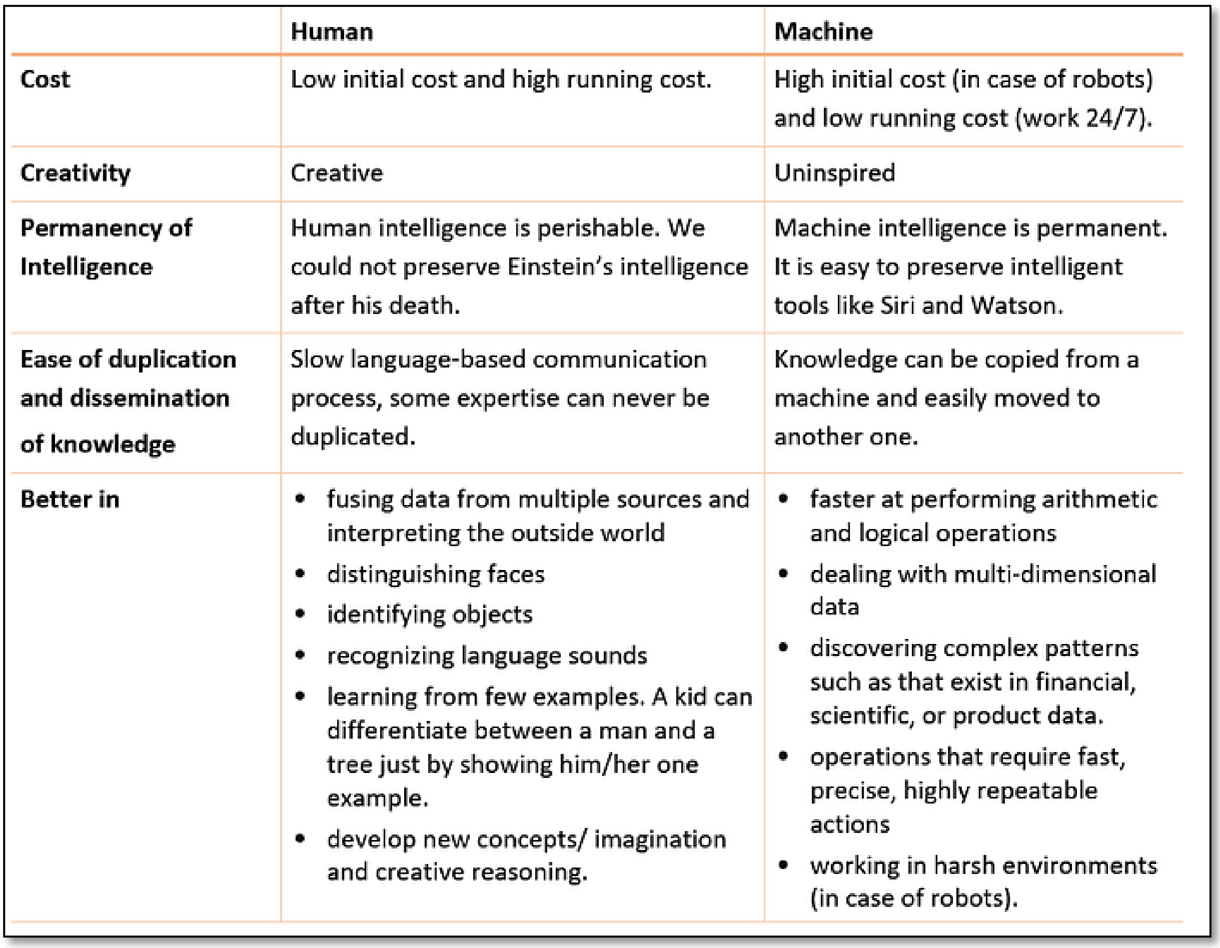
UNIT-1

INTRODUCTION TO MACHINE LEARNING

1.1. Overview of Machine Learning

- Machine Learning is a field of artificial intelligence that allows systems to learn and improve from experience without being explicitly programmed. It is predicated on the notion that computers can learn from data, spot patterns, and make judgments with little human assistance.
- It is the study of making machines more human-like in their behaviour and decisions by giving them the ability to learn and develop their programs. This is done with minimum human intervention, i.e., no explicit programming. The learning process is automated and improved based on the experiences of the machines throughout the process.
- Good quality data is fed to the machines, and different algorithms are used to build ML models to train the machines on this data. The choice of algorithm depends on the type of data at hand and the type of activity that needs to be automated.

Human Learning vs. Machine Learning

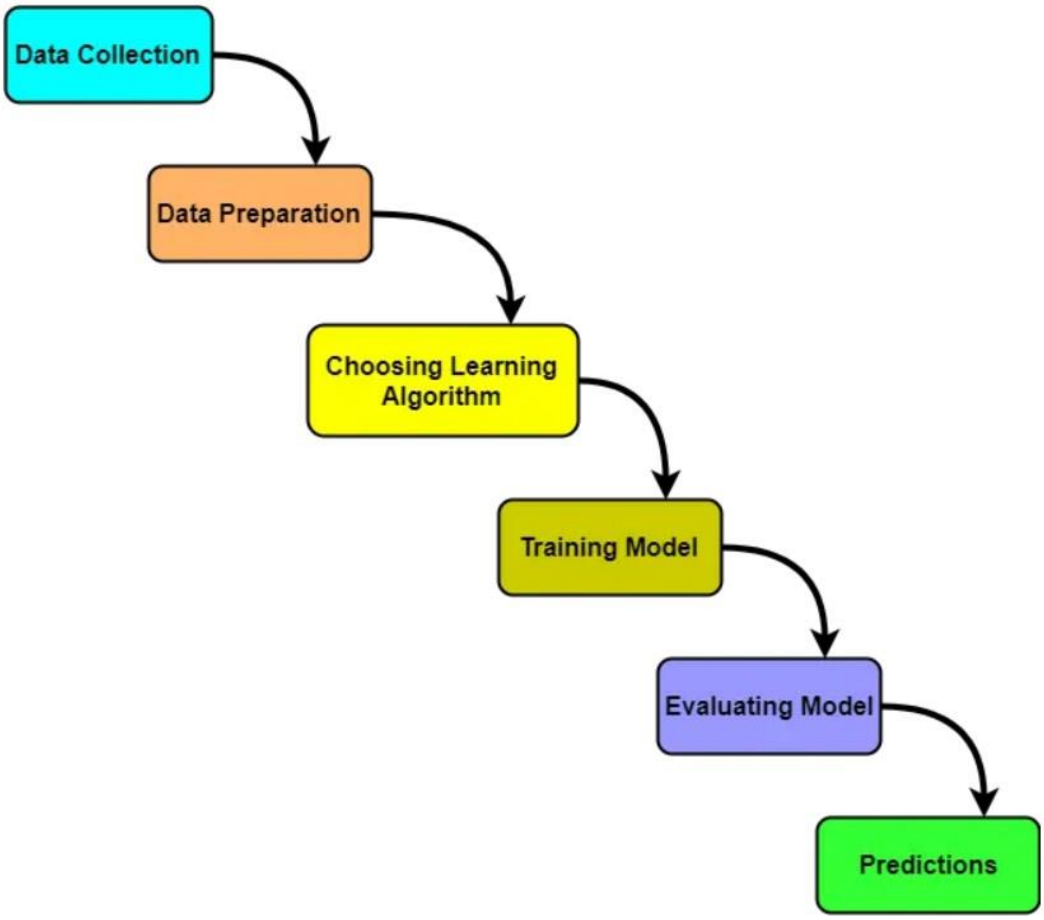


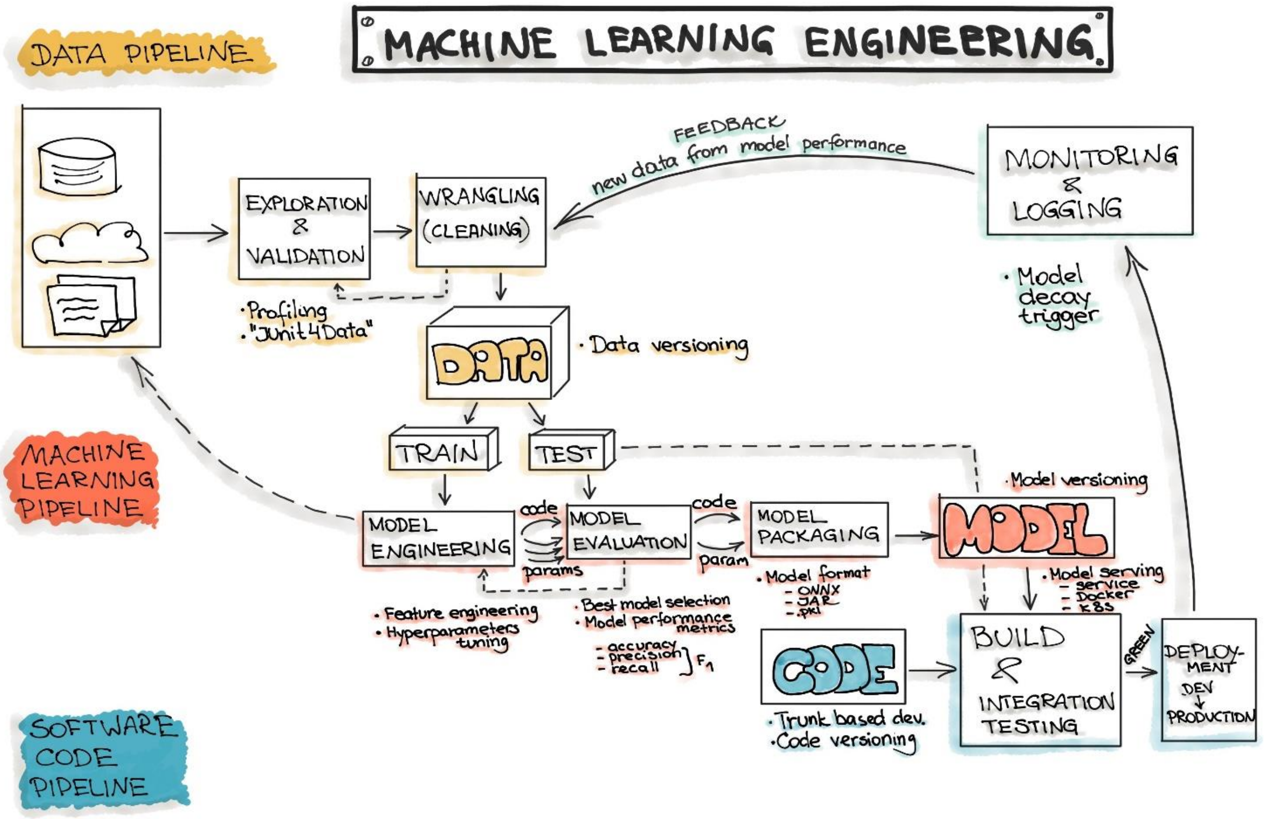
	Human	Machine
Cost	Low initial cost and high running cost.	High initial cost (in case of robots) and low running cost (work 24/7).
Creativity	Creative	Uninspired
Permanency of Intelligence	Human intelligence is perishable. We could not preserve Einstein's intelligence after his death.	Machine intelligence is permanent. It is easy to preserve intelligent tools like Siri and Watson.
Ease of duplication and dissemination of knowledge	Slow language-based communication process, some expertise can never be duplicated.	Knowledge can be copied from a machine and easily moved to another one.
Better in	• fusing data from multiple sources and interpreting the outside world • distinguishing faces • identifying objects • recognizing language sounds • learning from few examples. A kid can differentiate between a man and a tree just by showing him/her one example. • develop new concepts/ imagination and creative reasoning.	• faster at performing arithmetic and logical operations • dealing with multi-dimensional data • discovering complex patterns such as that exist in financial, scientific, or product data. • operations that require fast, precise, highly repeatable actions • working in harsh environments (in case of robots).

1.2. Machine Learning Terminology:

- **Model**: Also known as “hypothesis”, a Machine Learning model is the mathematical representation of a real-world process. A Machine Learning algorithm along with the training data builds a Machine Learning model.
- **Feature**: A feature is a measurable property or parameter of the dataset.
- **Feature Vector**: It is a set of multiple numeric features. We use it as an input to the Machine Learning model for training and prediction purposes.
- **Training**: An algorithm takes a set of data known as “training data” as input. The learning algorithm finds patterns in the input data and trains the model for expected results (target). The output of the training process is the Machine Learning model.
- **Prediction**: Once the Machine Learning model is ready, it can be fed with input data to provide a predicted output.
- **Target (Label)**: The value that the Machine Learning model has to predict is called the target or label.
- **Overfitting**: When a massive amount of data trains a Machine Learning model, it tends to learn from the noise and inaccurate data entries. Here the model fails to characterize the data correctly.
- **Underfitting**: It is the scenario when the model fails to decipher the underlying trend in the input data. It destroys the accuracy of the Machine Learning model. In simple terms, the model or the algorithm does not fit the data well enough.

1.3. Machine Learning Workflow





1. Data Collection-

- Data is collected from different sources.
- The type of data collected depends upon the type of desired project.
- Data may be collected from various sources such as files, databases, etc.
- The quality and quantity of gathered data directly affect the accuracy of the desired system.

2. Data Preparation-

In this stage,

- Data preparation is done to clean the raw data.
- Data collected from the real world is transformed into a clean dataset.
- Raw data may contain missing values, inconsistent values, duplicate instances, etc.
- So, raw data cannot be directly used for building a model.

Different methods of cleaning the dataset are-

- Ignoring the missing values
- Removing instances having missing values from the dataset.
- Estimating the missing values of instances using mean, median, or mode.
- Removing duplicate instances from the dataset.
- Normalizing the data in the dataset.

3. Choosing Learning Algorithm-

In this stage,

- The best-performing learning algorithm is researched.
- It depends upon the type of problem that needs to be solved and the type of data we have.
- If the problem is to classify and the data is labeled, classification algorithms are used.
- If the problem is to perform a regression task and the data is labeled, regression algorithms are used.

- If the problem is to create clusters and the data is unlabeled, clustering algorithms are used.

4. Training Model-

In this stage,

- The model is trained to improve its ability.
- The dataset is divided into a training dataset and a testing dataset.
- The training and testing split in order of 80/20 or 70/30.
- It also depends upon the size of the dataset.
- Training dataset is used for training purposes.
- Testing dataset is used for testing purposes.
- Training dataset is fed to the learning algorithm.
- The learning algorithm finds a mapping between the input and the output and generates the model.

5. Evaluating Model-

In this stage,

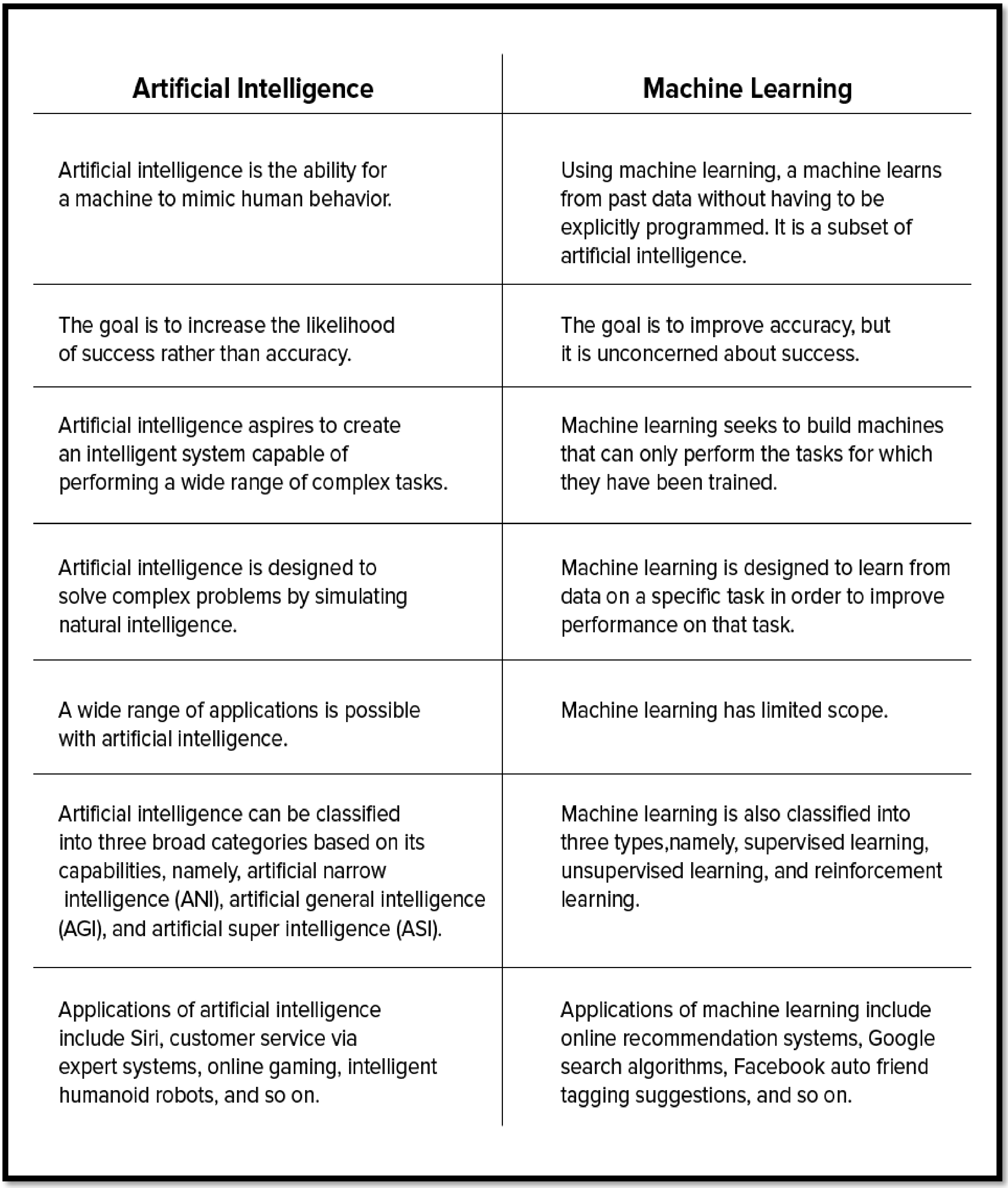
- The model is evaluated to test if the model is any good.
- The model is evaluated using the kept-aside testing dataset.
- It allows to test of the model against data that has never been used before for training.
- Metrics such as accuracy, precision, recall, etc are used to test the performance.
- If the model does not perform well, the model is re-built using different hyperparameters.
- The accuracy may be further improved by tuning the hyperparameters.

6. Predictions-

In this stage,

- The built system is finally used to do something useful in the real world.
- Here, the true value of machine learning is realized.

1.4. Artificial Intelligence vs. Machine Learning



Artificial Intelligence	Machine Learning
Artificial intelligence is the ability for a machine to mimic human behavior.	Using machine learning, a machine learns from past data without having to be explicitly programmed. It is a subset of artificial intelligence.
The goal is to increase the likelihood of success rather than accuracy.	The goal is to improve accuracy, but it is unconcerned about success.
Artificial intelligence aspires to create an intelligent system capable of performing a wide range of complex tasks.	Machine learning seeks to build machines that can only perform the tasks for which they have been trained.
Artificial intelligence is designed to solve complex problems by simulating natural intelligence.	Machine learning is designed to learn from data on a specific task in order to improve performance on that task.
A wide range of applications is possible with artificial intelligence.	Machine learning has limited scope.
Artificial intelligence can be classified into three broad categories based on its capabilities, namely, artificial narrow intelligence (ANI), artificial general intelligence (AGI), and artificial super intelligence (ASI).	Machine learning is also classified into three types,namely, supervised learning, unsupervised learning, and reinforcement learning.
Applications of artificial intelligence include Siri, customer service via expert systems, online gaming, intelligent humanoid robots, and so on.	Applications of machine learning include online recommendation systems, Google search algorithms, Facebook auto friend tagging suggestions, and so on.

1.5. Types of Machine Learning

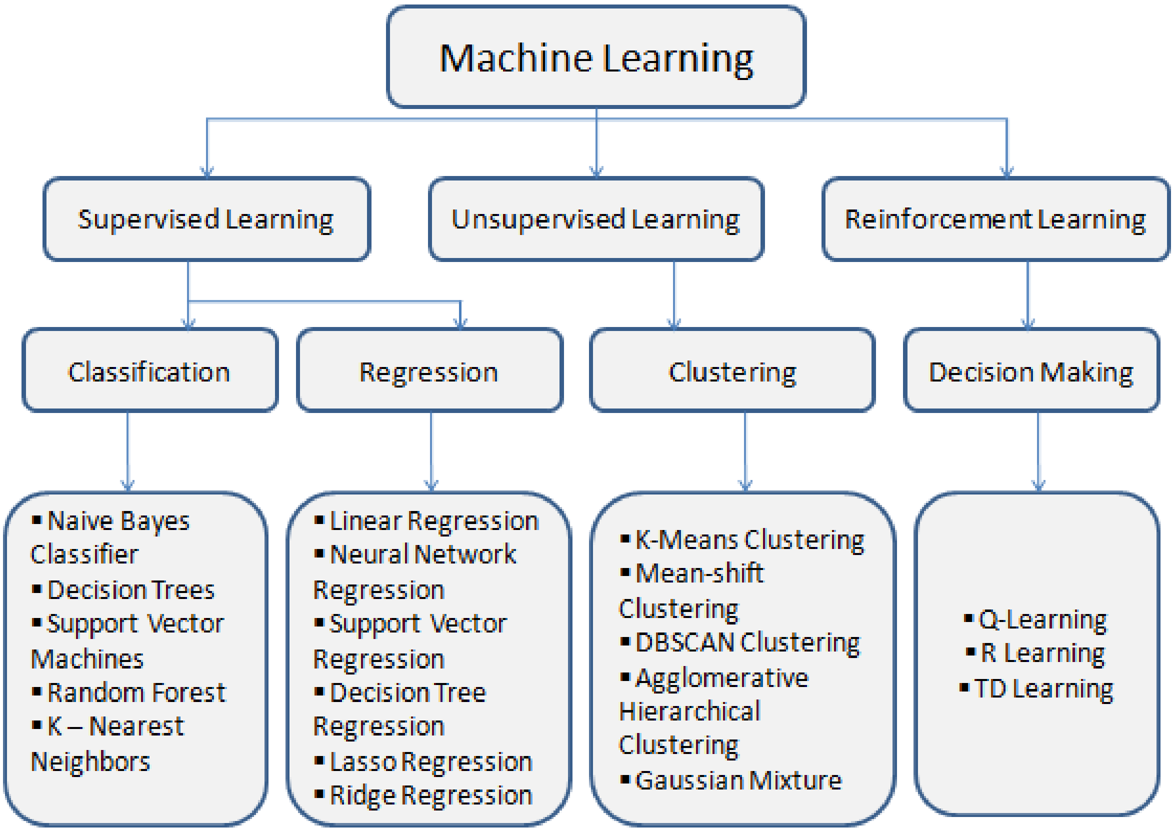
Machine Learning can be classified into three broad categories:

- **Supervised learning:**
    - Supervised learning is a class of problems that uses a model to learn the mapping between the input and target variables. Applications consisting of the training data describing the various input variables and the target variable are known as supervised learning tasks.
    - Let the set of input variable be (x) and the target variable be (y). A supervised learning algorithm tries to learn a hypothetical function which is a mapping given by the expression y=f(x), which is a function of x.
    - The learning process here is monitored or supervised. Since we already know the output the algorithm is corrected each time it makes a prediction, to optimize the results. Models are fit on training data which consists of both the input and the output variable and then it is used to make predictions on test data. Only the inputs are provided during the test phase and the outputs produced by the model are compared with the kept-back target variables and are used to estimate the performance of the model.
    - There are two types of supervised problems: Classification – which involves the prediction of a class label and Regression – which involves the prediction of a numerical value. 2. Unsupervised learning – Also called descriptive learning. A machine finds patterns in unknown objects by grouping similar objects.

- **Unsupervised learning:**
    - In an unsupervised learning problem, the model tries to learn by itself recognize patterns, and extract the relationships among the data. As in the case of supervised learning, there is no supervisor or teacher to drive the model. Unsupervised learning operates only on the input variables. There are no target variables to guide the learning process. The goal here is to interpret the underlying patterns in the data to obtain more proficiency in the underlying data.
    - There are two main categories in unsupervised learning: clustering – where the task is to find out the different groups in the data. And the next is Density Estimation – which tries to consolidate the distribution of data. These operations are performed to understand the patterns in the data. Visualization and Projection may also be considered unsupervised as they try to provide more insight into the data. Visualization involves creating plots and graphs on the data and Projection is involved with the dimensionality reduction of the data.

- **Reinforcement learning**
    - Reinforcement learning is type a of problem where there is an agent and the agent is operating in an environment based on the feedback or reward given to the agent by the environment in which it is operating. The rewards could be either positive or negative. The agent then proceeds in the environment based on the rewards gained.
    - The reinforcement agent determines the steps to perform a particular task. There is no fixed training dataset here and the machine learns on its own.

- Playing a game is a classic example of a reinforcement problem, where the agent's goal is to acquire a high score. It makes successive moves in the game based on the feedback given by the environment which may be in terms of rewards or penalization. Reinforcement learning has shown tremendous results in Google's AplhaGo of Google which defeated the world's number one Go player.



1.6. Tools and Technology for Machine Learning

Various tools and technologies support different stages of the Machine Learning workflow, from data preprocessing to model deployment. Here's a brief overview of some key tools and technologies in the field of Machine Learning:

- **Programming Languages:**
  - Python: Widely used for ML due to its extensive libraries (NumPy, Pandas, Scikit-learn, TensorFlow, PyTorch).
  - R: Commonly used for statistical analysis and data visualization in ML.
- **Libraries and Frameworks:**
  - Scikit-learn: A simple and efficient tool for data analysis and modeling, built on NumPy, SciPy, and Matplotlib.
  - TensorFlow: Developed by Google, it's an open-source ML framework used for building and training deep learning models.
  - PyTorch: Developed by Facebook, it's another popular deep learning framework known for its dynamic computation graph.
  - Keras: High-level neural networks API running on top of TensorFlow or Theano, simplifying the process of building and training models.
- **Data Processing and Analysis:**
  - NumPy and Pandas: Fundamental libraries for numerical operations and data manipulation in Python
- **Visualization Tools:**
  - Matplotlib and Seaborn: Python libraries for creating static, animated, and interactive visualizations.
  - TensorBoard: A web-based tool provided with TensorFlow for visualizing Machine Learning experiments.

1.7. Application of Machine Learning

- **Facial recognition/Image recognition**
    - There are a lot of use cases of facial recognition, mostly for security purposes like identifying criminals, searching for missing individuals, aiding forensic investigations, etc. Intelligent marketing, diagnosing diseases, and tracking attendance in schools, are some other uses.

- **Automatic Speech Recognition**
        - Abbreviated as ASR, automatic speech recognition is used to convert speech into digital text. Its applications lie in authenticating users based on their voice and performing tasks based on human voice inputs. Speech patterns and vocabulary are fed into the system to train the model. Presently ASR systems find a wide variety of applications in the following domains:
            - Medical Assistance
            - Industrial Robotics
            - Forensic and Law enforcement
            - Defense & Aviation
            - Telecommunications Industry
            - Home Automation and Security Access Control
            - I.T. and Consumer Electronics
    - Speech recognition is a process of converting voice instructions into text, and it is also known as "**Speech to text**", or "**Computer speech recognition**." At present, Machine Learning algorithms are widely used in various applications of speech recognition. **Google Assistant**, **Siri**, **Cortana**, and **Alexa** are using speech recognition technology to follow voice instructions.

- **Financial Services**
  - Machine Learning has many use cases in Financial Services. Machine Learning algorithms prove to be excellent at detecting fraud by monitoring the activities of each user and assessing that if an attempted activity is typical of that user or not. Financial monitoring to detect money laundering activities is also a critical security use case.
  - It also helps in making better trading decisions with the help of algorithms that can analyze thousands of data sources simultaneously. Credit scoring and underwriting are some of the other applications. The most common application in our day-to-day activities is the virtual personal assistants like Siri and Alexa.

- **Traffic predictions**
  - When you use Google Maps to map your commute to work or a new restaurant in town, it provides an estimated time of arrival. Google uses Machine Learning to build models of how long trips will take based on historical traffic data (gleaned from satellites). It then takes that data based on your current trip and traffic levels to predict the best route according to these factors.

- **Healthcare**
  - A vital application is in the diagnosis of diseases and ailments, which are otherwise difficult to diagnose. Radiotherapy is also becoming better.
  - Early-stage drug discovery is another crucial application that involves technologies such as precision medicine and next-generation sequencing. Clinical trials cost a lot of time and money to complete

    and deliver results. Applying ML-based predictive analytics could improve on these factors and give better results.
  - These technologies are also critical to making outbreak predictions. Scientists around the world are using ML technologies to predict epidemic outbreaks.

- **Recommendation Systems**
  - Many businesses today use recommendation systems to effectively communicate with the users on their sites. It can recommend relevant products, movies, web series, songs, and much more. The most prominent use cases of recommendation systems are e-commerce sites like Amazon, Flipkart, and many others, along with Spotify, Netflix, and other web-streaming channels.

- **Credit card fraud detection**
  - Predictive analytics can help determine whether a credit card transaction is fraudulent or legitimate. Fraud examiners use AI and Machine Learning to monitor variables involved in past fraud events. They use these training examples to measure the likelihood that a specific event was fraudulent activity.