Introduction to Machine Learning

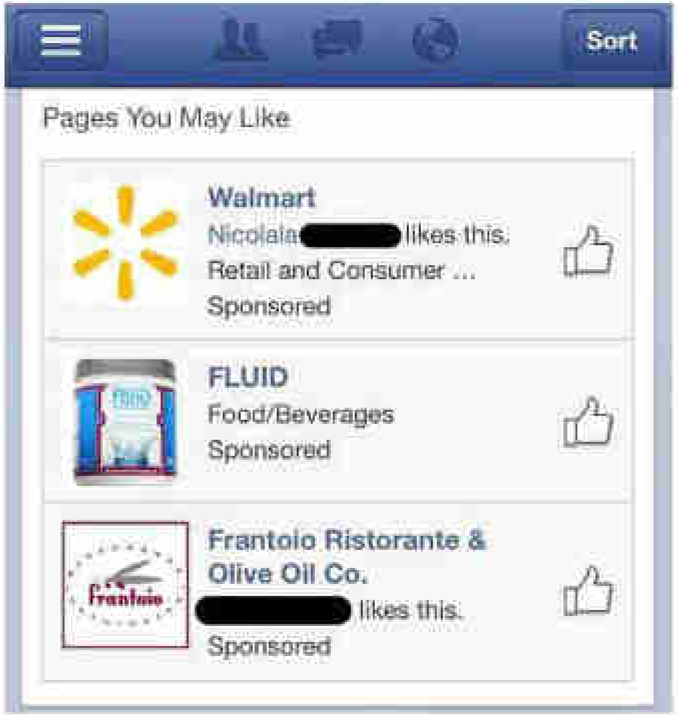
Unit 1

Introduction to Machine Learning: Definition, Terminology, Types of learning, Machine Learning Problem categories, Machine learning architecture, process, Lifecyle, Performance measures, tools and framework, data visualization

Application of Machine Learning

1. Social Media Features

- Social media platforms use machine learning algorithms and approaches to create some attractive and excellent features. For instance, Facebook notices and records your activities, chats, likes, and comments, and the time you spend on specific kinds of posts. Machine learning learns from your own experience and makes friends and page suggestions for your profile.



Pages You May Like

Walmart
Nicolaia [redacted] likes this.
Retail and Consumer ...
Sponsored

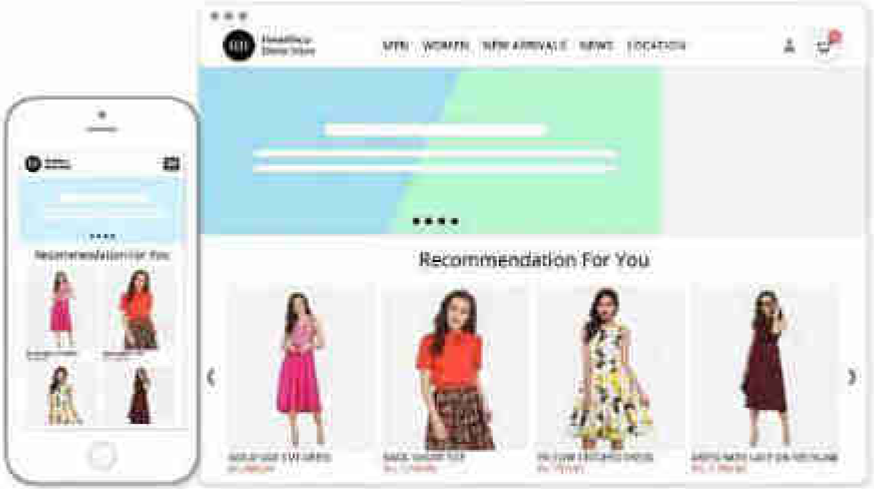
FLUID
Food/Beverages
Sponsored

Frantoio Ristorante & Olive Oil Co.
[redacted] likes this.
Sponsored

Application of Machine Learning

2. Product Recommendations

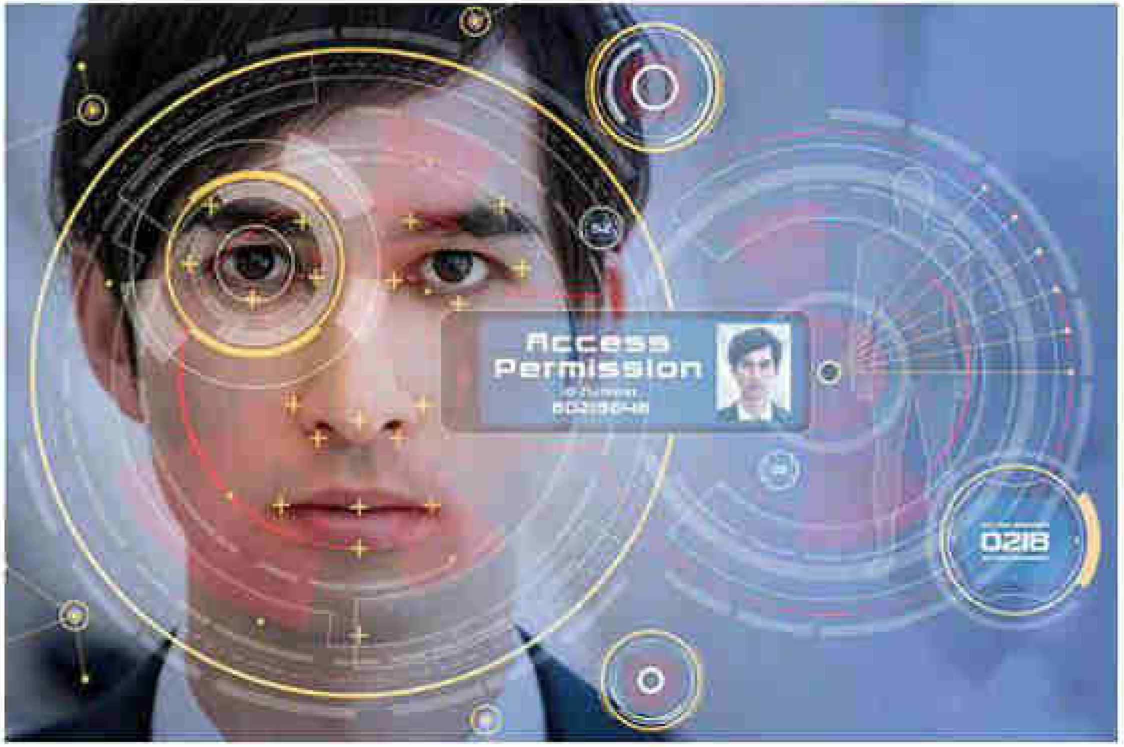
- Product recommendation is one of the most popular and known applications of machine learning. Product recommendation is one of the stark features of almost every e-commerce website today, which is an advanced application of machine learning techniques. Using machine learning and AI, websites track your behavior based on your previous purchases, searching patterns, and cart history, and then make product recommendations.



Application of Machine Learning

3. Image Recognition

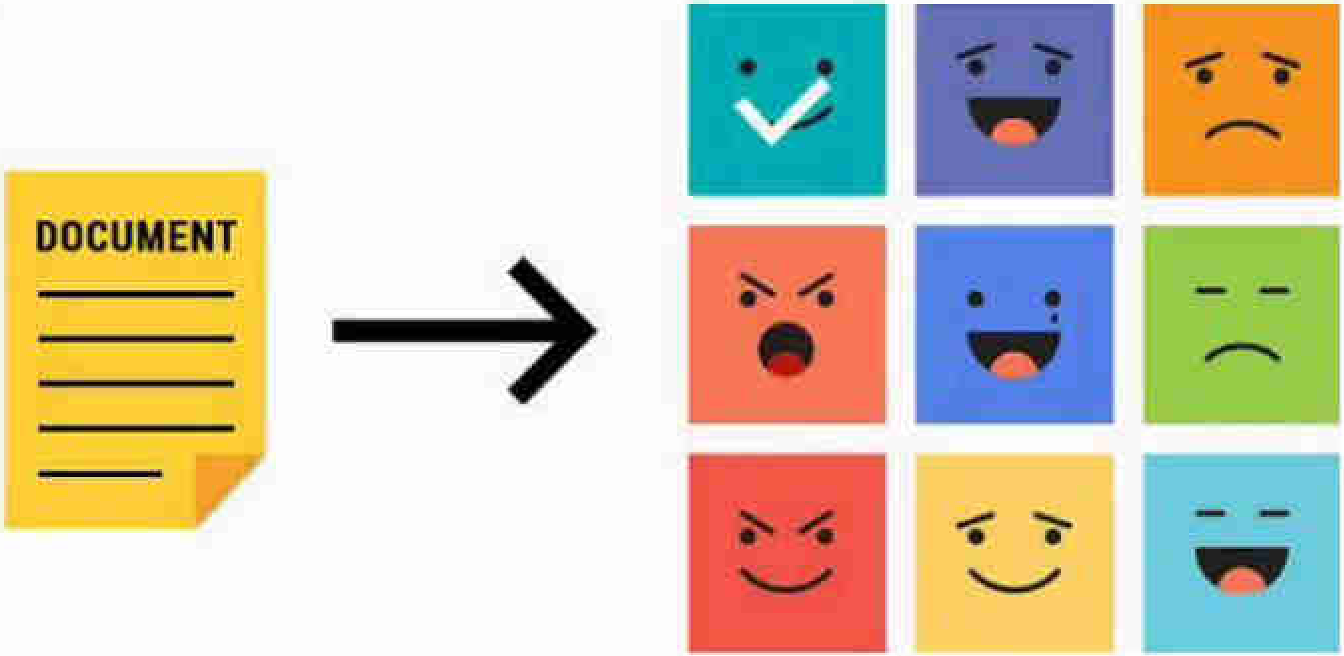
- Image recognition, which is an approach for cataloging and detecting a feature or an object in the digital image, is one of the most significant and notable machine learning and AI techniques. This technique is being adopted for further analysis, such as pattern recognition, face detection, and face recognition.



Application of Machine Learning

4. Sentiment Analysis

- Sentiment analysis is one of the most necessary applications of machine learning. Sentiment analysis is a real-time machine learning application that determines the emotion or opinion of the speaker or the writer. For instance, if someone has written a review or email (or any form of a document), a sentiment analyzer will instantly find out the actual thought and tone of the text. This sentiment analysis application can be used to analyze a review based website, decision-making applications, etc



Application of Machine Learning

5. Virtual Personal Assistants

As the name suggests, Virtual Personal Assistants assist in finding useful information, when asked via text or voice. Few of the major applications of Machine Learning here are:

Speech Recognition

Speech to Text Conversion

Natural Language Processing

Text to Speech Conversion

Application of Machine Learning

6. Self Driving Cars

- Well, here is one of the coolest application of Machine Learning. It's here and people are already using it. Machine Learning plays a very important role in Self Driving Cars and I'm sure you guys might have heard about **Tesla**. The leader in this business and their current ***Artificial Intelligence*** is driven by hardware manufacturer **NVIDIA**, which is based on Unsupervised Learning Algorithm.
- NVIDIA stated that they didn't train their model to detect people or any object as such. The model works on ***Deep Learning*** and it crowdsources data from all of its vehicles and its drivers. It uses internal and external sensors which are a part of **IOT**. According to the data gathered by McKinsey, the automotive data will hold a tremendous value of **$750 Billion.**

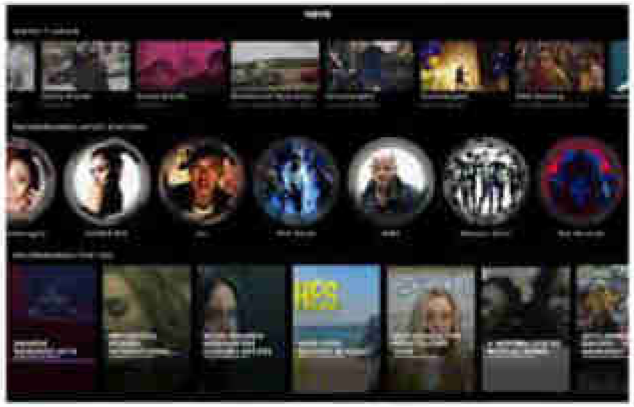
Application of Machine Learning

7. Entertainment

Companies such as Netflix, Amazon, YouTube, and Spotify give relevant movies, songs, and video recommendations to enhance their customer experience.

This is all thanks to Deep Learning. Based on a person's browsing history, interest, and behavior, online streaming companies give suggestions to help them make product and service choices.

Deep learning techniques are also used to add sound to silent movies and generate subtitles automatically.



What is Machine Learning

- **Definition of Machine Learning**

- Machine learning is an application of AI that enables systems to learn and improve from experience without being explicitly programmed. Machine learning focuses on developing computer programs that can access data and use it to learn for themselves.

- Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

- Machine learning (ML) is defined as a discipline of artificial intelligence (AI) that provides machines the ability to automatically learn from data and past experiences to identify patterns and make predictions with minimal human intervention.

What is Machine Learning

- **Definition of Machine Learning**

- A computer program is said to learn from experience **E** with respect to some class of tasks ***T*** and performance measure **P**, if its performance at tasks in T, as measured by P, improves with experience E.

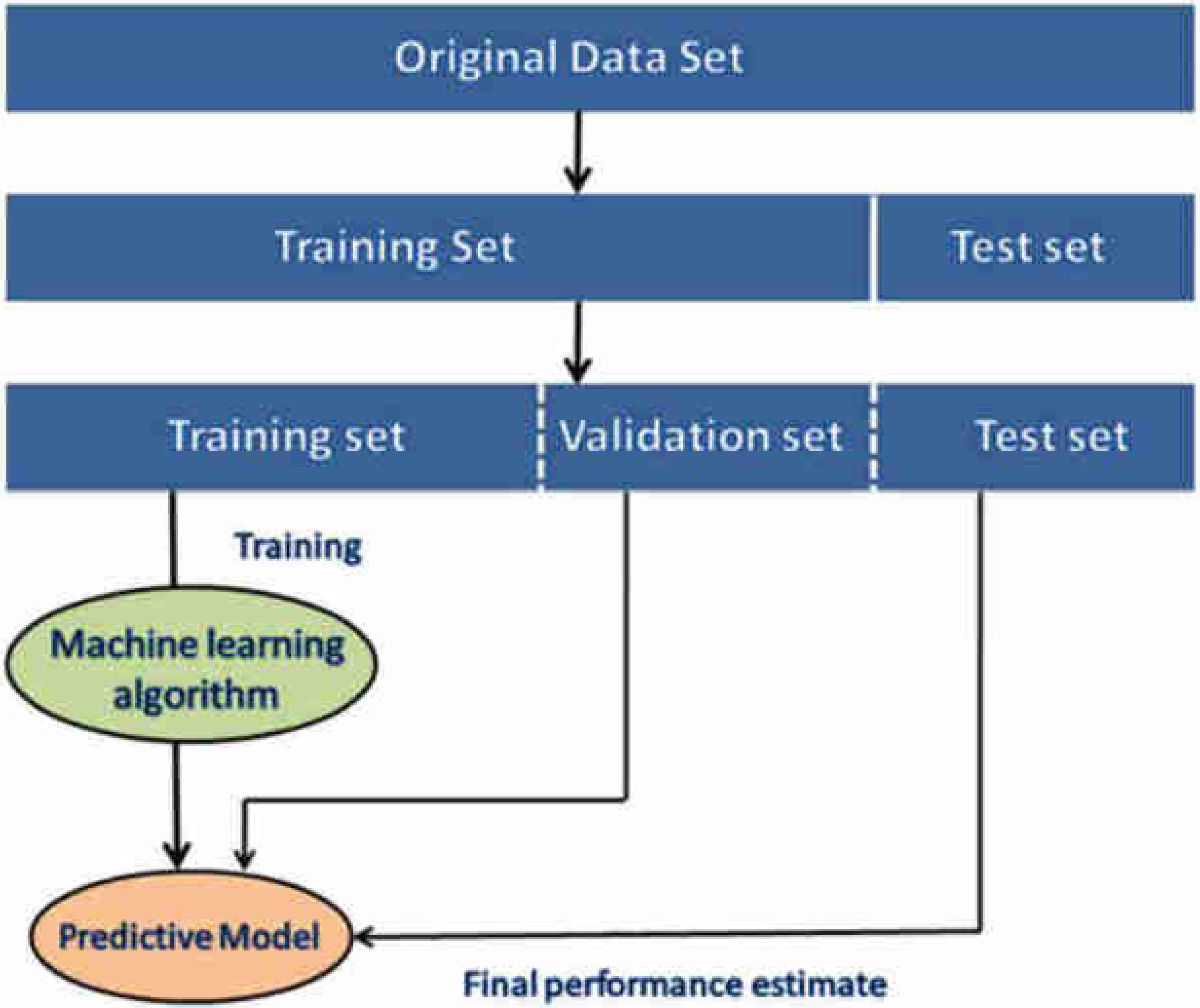
- **Task**: A task is defined as the main problem in which we are interested. This task/problem can be related to the predictions and recommendations and estimations, etc.
- **Experience**: It is defined as learning from historical or past data and used to estimate and resolve future tasks.
- **Performance**: It is defined as the capacity of any machine to resolve any machine learning task or problem and provide the best outcome for the same. However, performance is dependent on the type of machine learning problems.

Terminology in Machine Learning

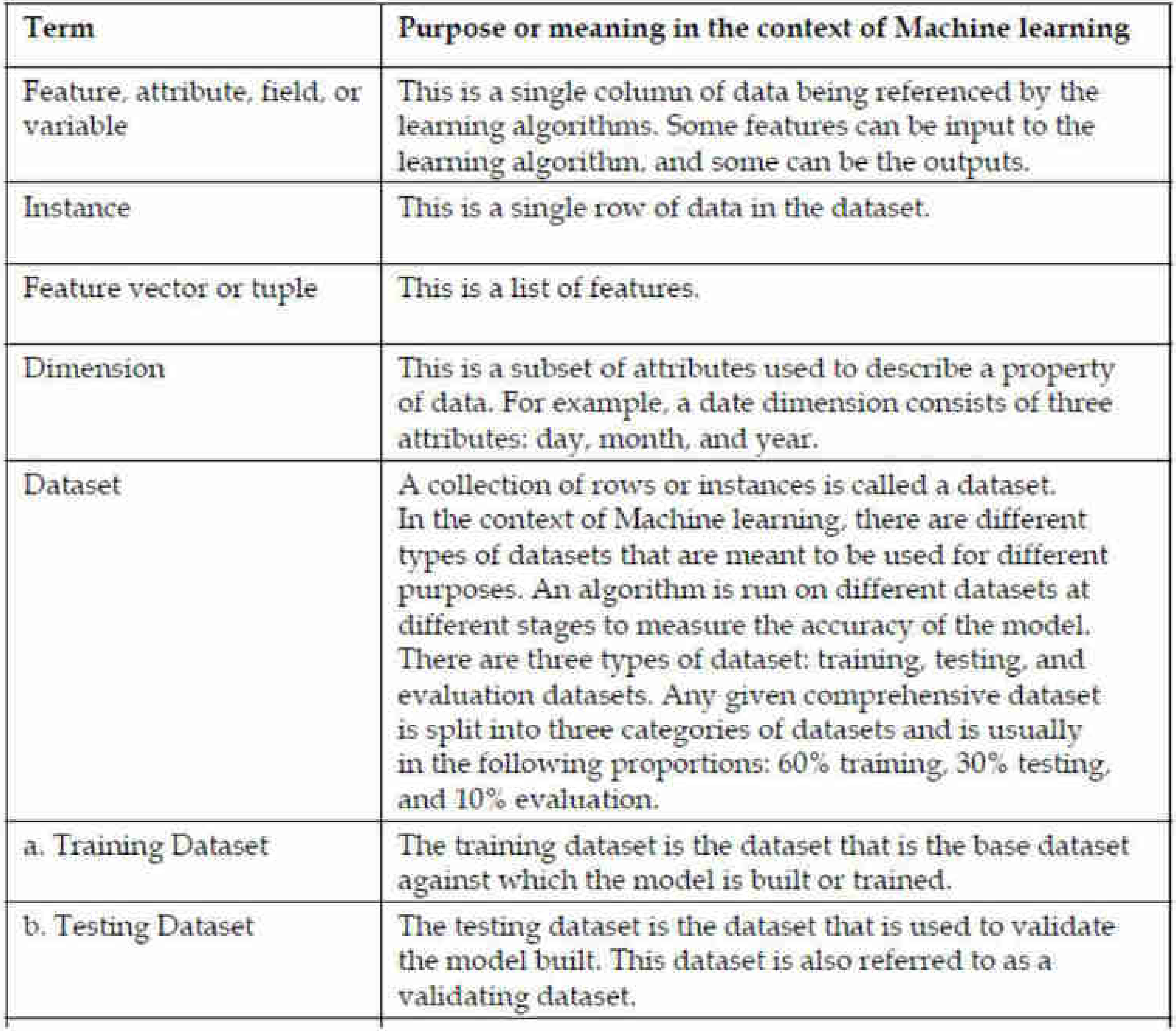
- An **Algorithm** is a set of rules that a machine follows to achieve a particular goal.
- **Machine Learning** is a set of methods that allow computers to learn from data to make and improve predictions (for example cancer, weekly sales, credit default).
- A **Learner** or **Machine Learning Algorithm** is the program used to learn a machine learning model from data. Another name is "inducer" (e.g. "tree inducer").
- A **Machine Learning Model** is the learned program that maps inputs to predictions. A trained machine is also called as a **Model**.
- A **Dataset** is a table with the data from which the machine learns. An **Instance** is a row in the dataset. A **feature** is a column in the dataset.
- The **Target** is the information the machine learns to predict. In mathematical formulas, the target is usually called y.
- The **Prediction** is what the machine learning model "guesses" what the target value should be based on the given features.
- A **Training set** is used to train a machine.
- A **Test Set** is used to test a already trained machine.
- A **validation set** is used to verify a trained machine.

https://christophm.github.io/interpretable-ml-book/terminology.html

Terminology in Machine Learning

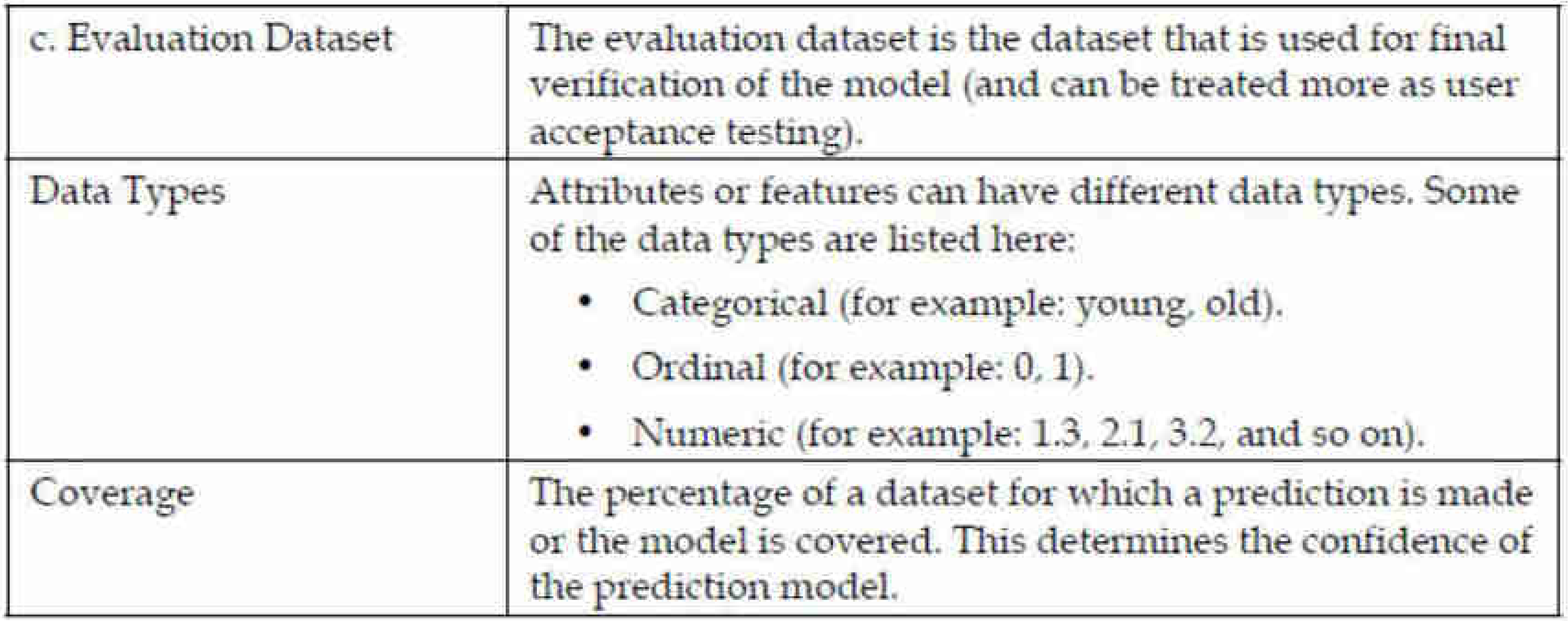


Terminology in Machine Learning



Term	Purpose or meaning in the context of Machine learning
Feature, attribute, field, or variable	This is a single column of data being referenced by the learning algorithms. Some features can be input to the learning algorithm, and some can be the outputs.
Instance	This is a single row of data in the dataset.
Feature vector or tuple	This is a list of features.
Dimension	This is a subset of attributes used to describe a property of data. For example, a date dimension consists of three attributes: day, month, and year.
Dataset	A collection of rows or instances is called a dataset. In the context of Machine learning, there are different types of datasets that are meant to be used for different purposes. An algorithm is run on different datasets at different stages to measure the accuracy of the model. There are three types of dataset: training, testing, and evaluation datasets. Any given comprehensive dataset is split into three categories of datasets and is usually in the following proportions: 60% training, 30% testing, and 10% evaluation.
a. Training Dataset	The training dataset is the dataset that is the base dataset against which the model is built or trained.
b. Testing Dataset	The testing dataset is the dataset that is used to validate the model built. This dataset is also referred to as a validating dataset.

Terminology in Machine Learning



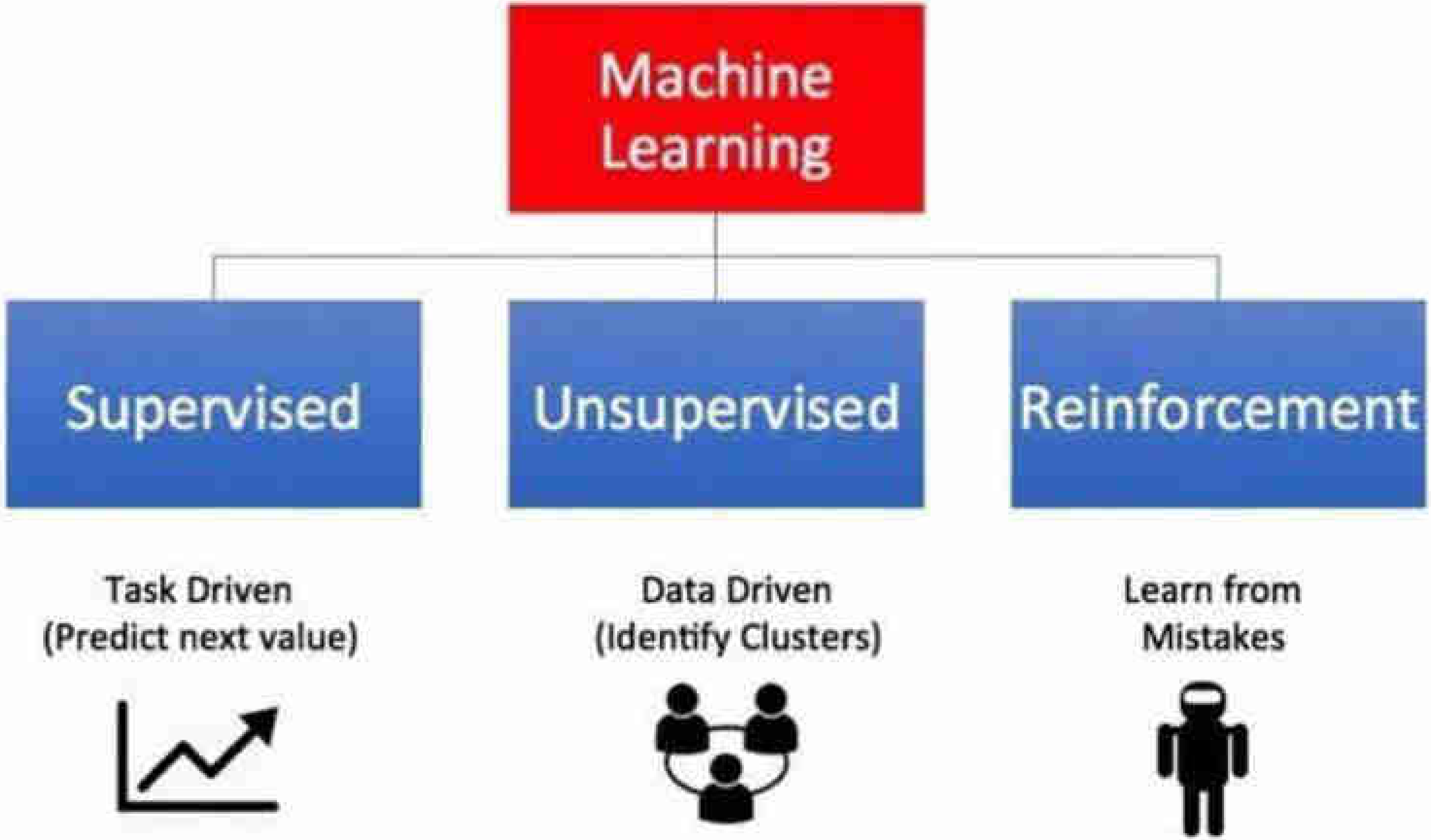
c. Evaluation Dataset	The evaluation dataset is the dataset that is used for final verification of the model (and can be treated more as user acceptance testing).
Data Types	Attributes or features can have different data types. Some of the data types are listed here: • Categorical (for example: young, old). • Ordinal (for example: 0, 1). • Numeric (for example: 1.3, 2.1, 3.2, and so on).
Coverage	The percentage of a dataset for which a prediction is made or the model is covered. This determines the confidence of the prediction model.

Terminology in Machine Learning

- **Bias** is the gap between predicted value by the model and the actual or target value.
- **Variance** tells how scattered the predicted values are.

https://christophm.github.io/interpretable-ml-book/terminology.html

Types of Machine Learning



Machine Learning

Supervised

Unsupervised

Reinforcement

Task Driven
(Predict next value)

Data Driven
(Identify Clusters)

Learn from
Mistakes

Types of Machine Learning

1. Supervised learning:

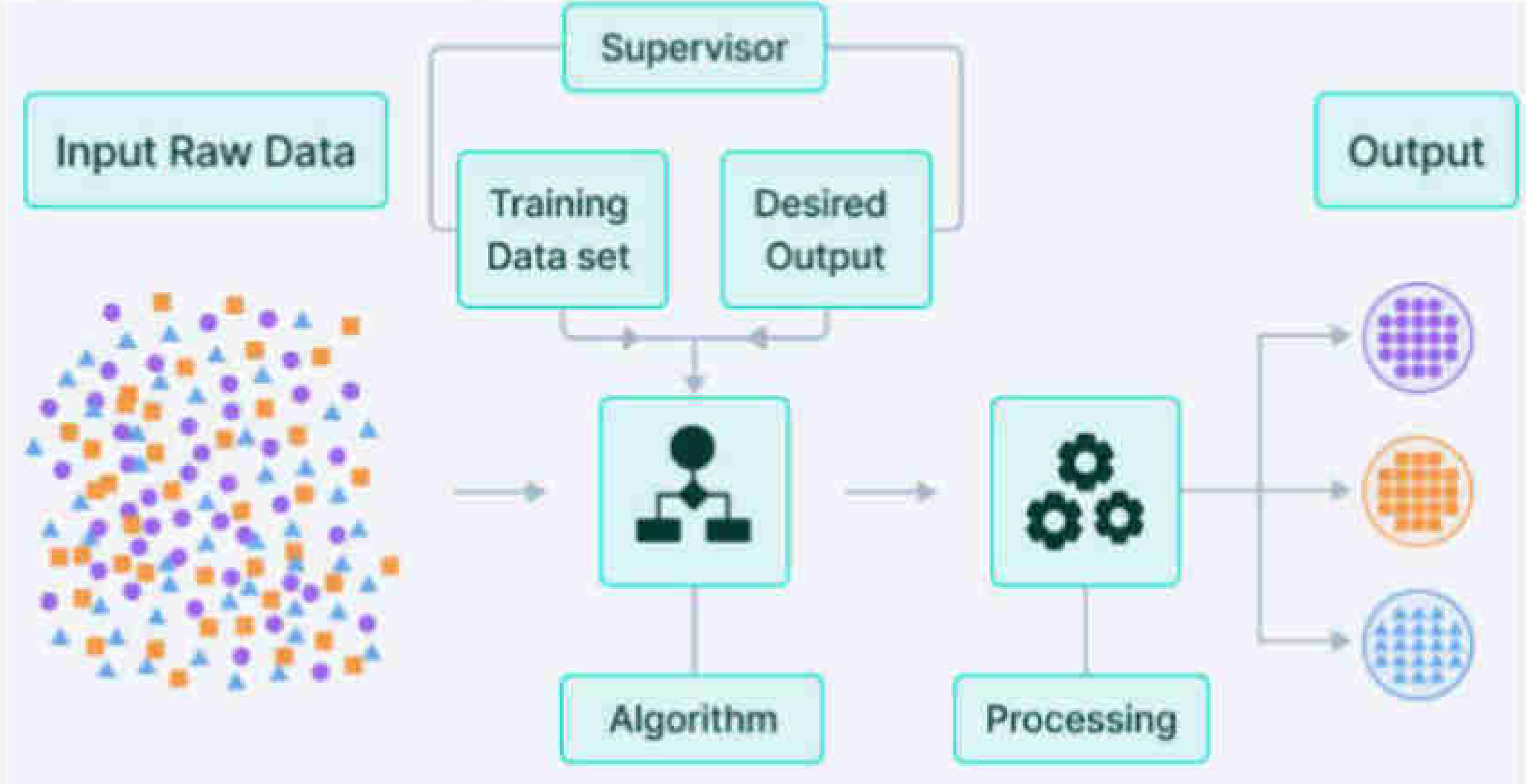
It is applicable when a machine has sample data, i.e., **input as well as output data with correct labels.** Correct labels are used to check the correctness of the model using some labels and tags.

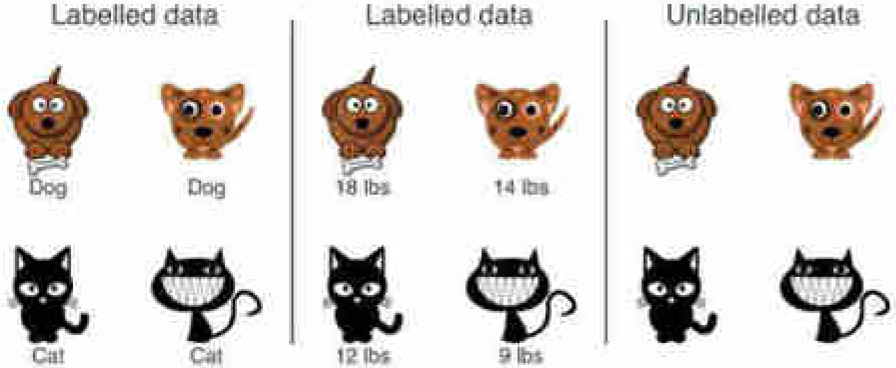
Supervised learning technique helps us to predict future events with the help of past experience and labeled examples. Initially, it analyses the known training dataset, and later it introduces an inferred function that makes predictions about output values. Further, it also predicts errors during this entire learning process and also corrects those errors through algorithms.

Example: Let's assume we have a set of images tagged as ''dog''. A machine learning algorithm is trained with these dog images so it can easily distinguish whether an image is a dog or not.

Types of Machine Learning

1. Supervised learning:





Types of Machine Learning

2. Unsupervised Learning:

As the name suggests, unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things.

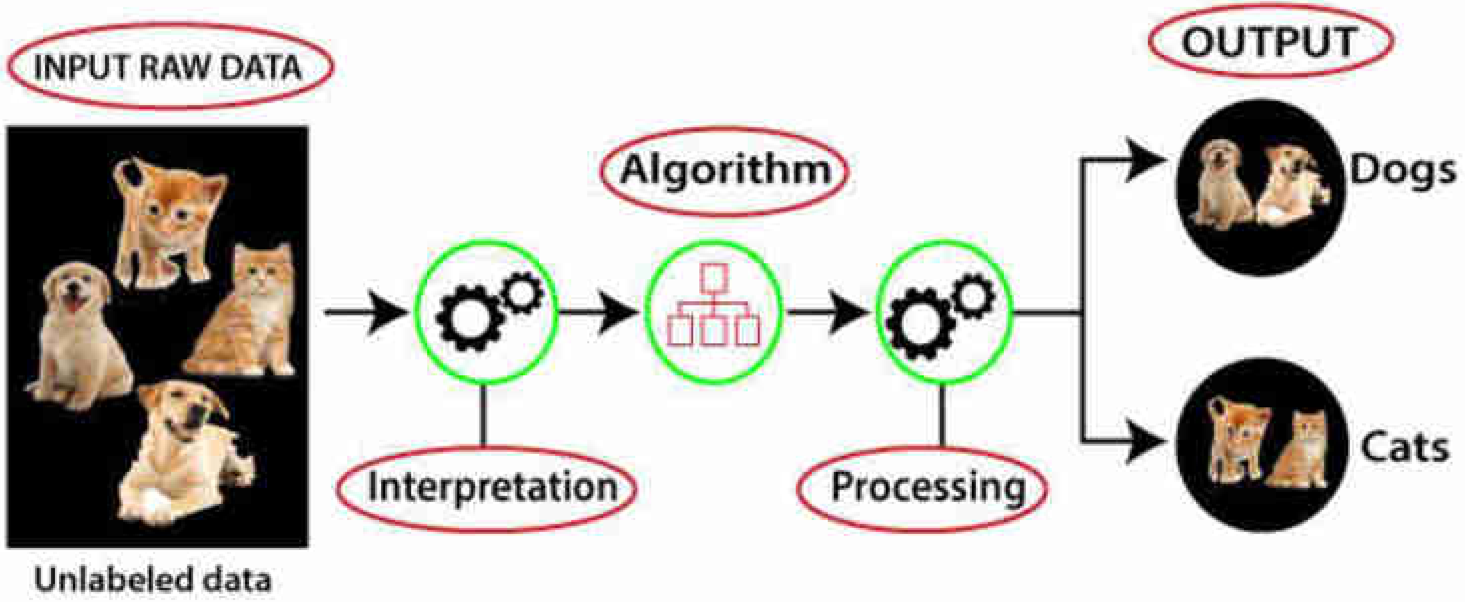
Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data.

Example: Suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cats and dogs. The algorithm is never trained upon the given dataset, which means it does not have any idea about the features of the dataset. The task of the unsupervised learning algorithm is to identify the image features on their own. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups according to similarities between images.

Types of Machine Learning

2. Unsupervised Learning:

Here, we have taken an unlabeled input data, which means it is not categorized and corresponding outputs are also not given. Now, this unlabeled input data is fed to the machine learning model in order to train it. Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc.



Types of Machine Learning

3. Reinforcement Learning:

Reinforcement Learning is a feedback-based machine learning technique.

In such type of learning, agents (computer programs) need to explore the environment, perform actions, and on the basis of their actions, they get rewards as feedback.

For each good action, they get a positive reward, and for each bad action, they get a negative reward.

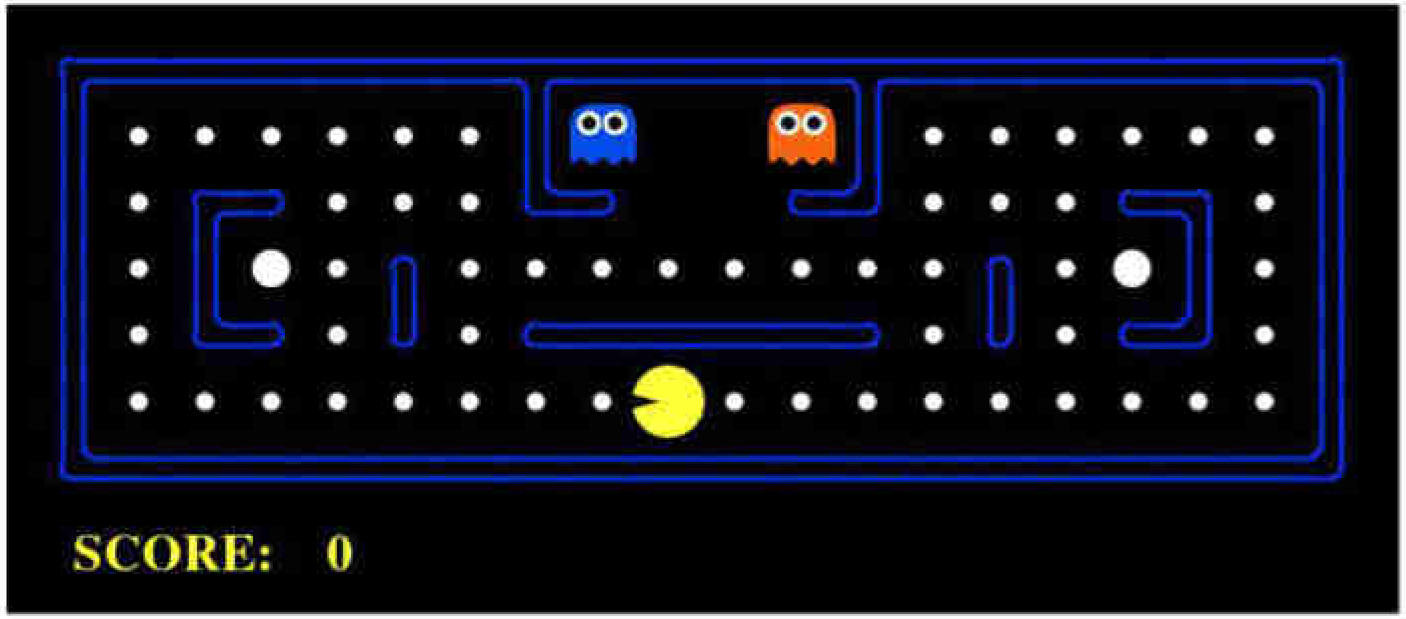
The goal of a Reinforcement learning agent is to maximize the positive rewards.

Since there is no labeled data, the agent is bound to learn by its experience only.

Types of Machine Learning

3. Reinforcement Learning:

An RL problem can be best explained through games. Let's take the game of PacMan where the goal of the **agent**(PacMan) is to eat the food in the grid while avoiding the ghosts on its way. In this case, the grid world is the interactive **environment** for the agent where it acts. Agent receives a **reward** for eating food and **punishment** if it gets killed by the ghost (loses the game). The states are the location of the agent in the grid world and the total cumulative reward is the agent winning the game.



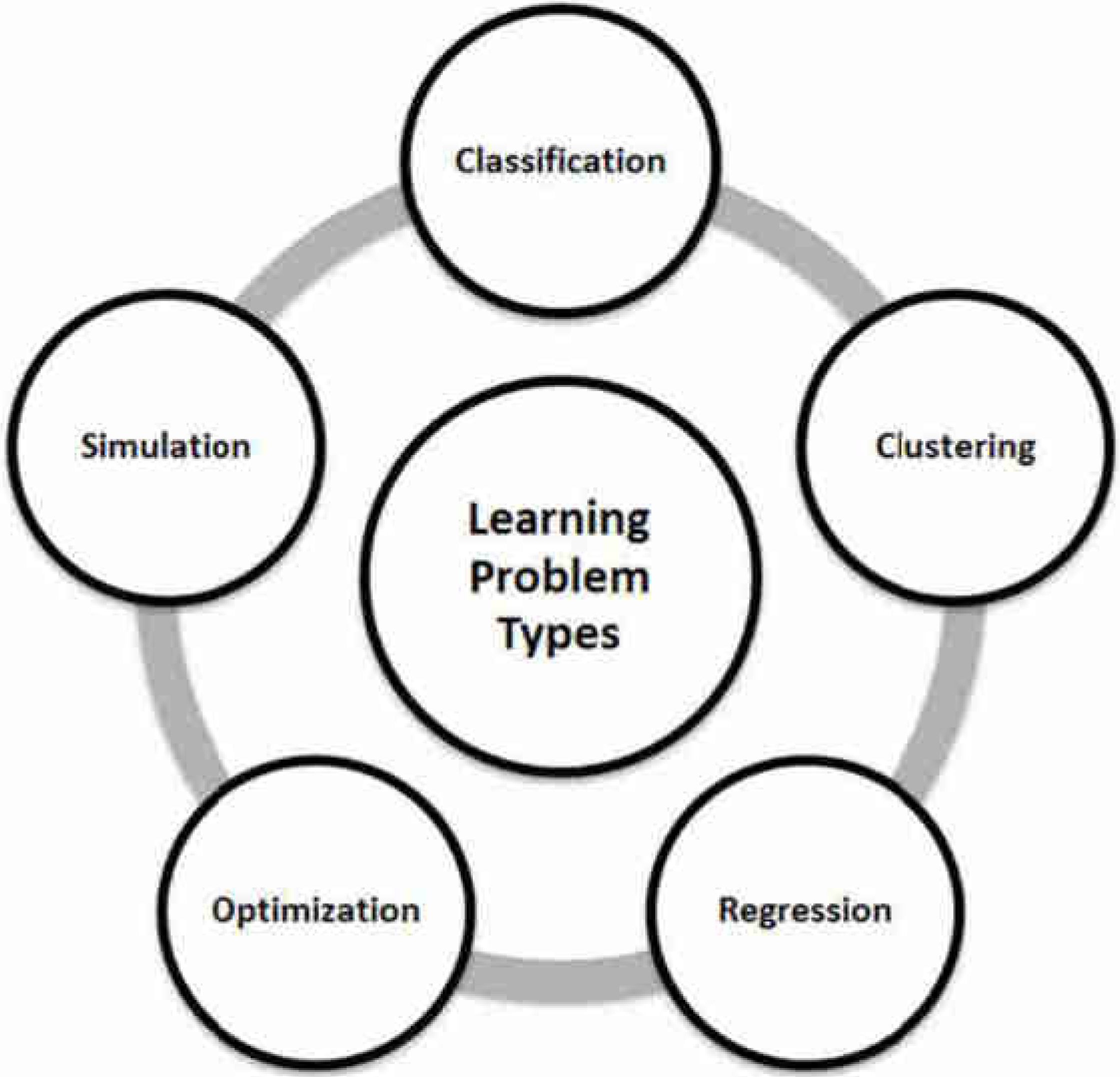
Types of Machine Learning

4. Semi-supervised Learning:

Semi-supervised Learning is an intermediate technique of **both supervised and unsupervised learning**. It performs actions on datasets having few labels as well as unlabeled data. However, it generally contains unlabeled data. Hence, it also reduces the cost of the machine learning model as labels are costly, but for corporate purposes, it may have few labels. Further, it also increases the accuracy and performance of the machine learning model.

Semi-supervised Learning is an intermediate technique of both supervised and unsupervised learning. It performs actions on datasets having few labels as well as unlabeled data. However, it generally contains unlabeled data. Hence, it also reduces the cost of the machine learning model as labels are costly, but for corporate purposes, it may have few labels. Further, it also increases the accuracy and performance of the machine learning model.

Machine Learning Problem Categories



Machine Learning Problem Categories

1. Classification:

Classification is a task that requires the use of machine learning algorithms that learn how to assign a class label to examples from the problem domain. An easy to understand example is classifying emails as “spam” or “not spam.”

Following are the types of classifications:-

1. Classification Predictive Modelling
2. Binary Classification
3. Multi-Class Classification
4. Multi-Label Classification
5. Imbalanced Classification

Machine Learning Problem Categories

1. Classification:

1. **Classification Predictive Modelling:** In machine learning, classification refers to a predictive modeling problem where a class label is predicted for a given example of input data.

Examples of classification problems include:

i. Given an example, classify if it is spam or not.

ii. Given a handwritten character, classify it as one of the known characters.

iii. Given recent user behavior, classify as churn or not.

Machine Learning Problem Categories

1. Classification:

2. **Binary Classification:** It refers to those classification tasks that have two class labels.

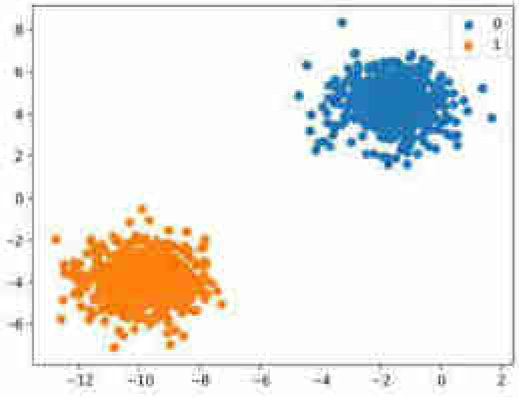
Examples include:

i. Email spam detection (spam or not).

ii. Churn prediction (churn or not).

iii. Conversion prediction (buy or not).

- Typically, binary classification tasks involve one class that is the **normal state** and another class that is the **abnormal state**.
- Bernoulli probability distribution is used to classify such a problem.
- Popular algorithms that can be used for binary classification include:
    - Logistic Regression
    - k-Nearest Neighbors
    - Decision Trees
    - Support Vector Machine
    - Naive Bayes



Scatter Plot of Binary Classification Dataset

Machine Learning Problem Categories

1. Classification:

3. **Multi-Class Classification:** It refers to those classification tasks that have more than two class labels.

Examples include:

i. Face classification.

ii. Plant species classification.

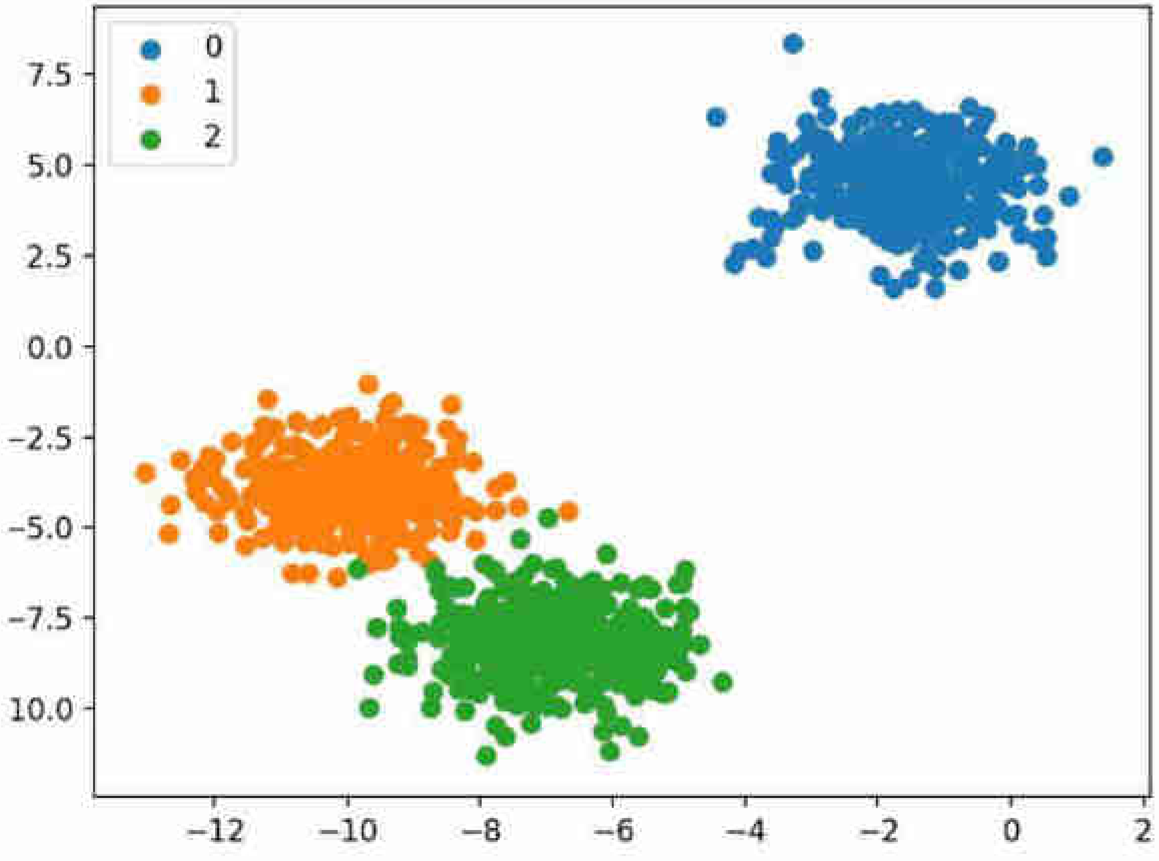
iii. Optical character recognition.

- Unlike binary classification, multi-class classification does not have the notion of normal and abnormal outcomes. Instead, examples are classified as belonging to one among a **range of known classes**.

- The number of class labels may be very large on some problems. For example, a model may predict a photo as belonging to one among thousands or tens of thousands of faces in a face recognition system.

Machine Learning Problem Categories

- Multinoulli probability distribution is used to classify such a problem.
- Popular algorithms that can be used for multi-class classification include:
  - i. k-Nearest Neighbors.
  - ii. Decision Trees.
  - iii. Naive Bayes.
  - iv. Random Forest.
  - v. Gradient Boosting.



Scatter Plot of Multi-Class Classification Dataset

Machine Learning Problem Categories

1. Classification:

4. **Multi-Label Classification:** It refers to those classification tasks that have two or more class labels, where one or more class labels may be predicted for each example.

- Consider the example of photo classification, where a given photo may have multiple objects in the scene and a model may predict the presence of multiple known objects in the photo, such as “bicycle,” “apple,” “person,” etc.

- Special multi-label versions of the algorithms given below are used for classification:-
  - i. Multi-label Decision Trees
  - ii. Multi-label Random Forests
  - iii. Multi-label Gradient Boosting

Machine Learning Problem Categories

1. **Classification:**

5. **Imbalanced Classification**: It refers to classification tasks where the number of examples in each class is unequally distributed.

- Typically, imbalanced classification tasks are binary classification tasks where the majority of examples in the training dataset belong to the normal class and a minority of examples belong to the abnormal class.

  Examples include:

  i. Fraud detection.

  ii. Outlier detection.

  iii. Medical diagnostic tests

- Specialized techniques may be used to change the composition of samples in the training dataset by under-sampling the majority class or oversampling the minority class.

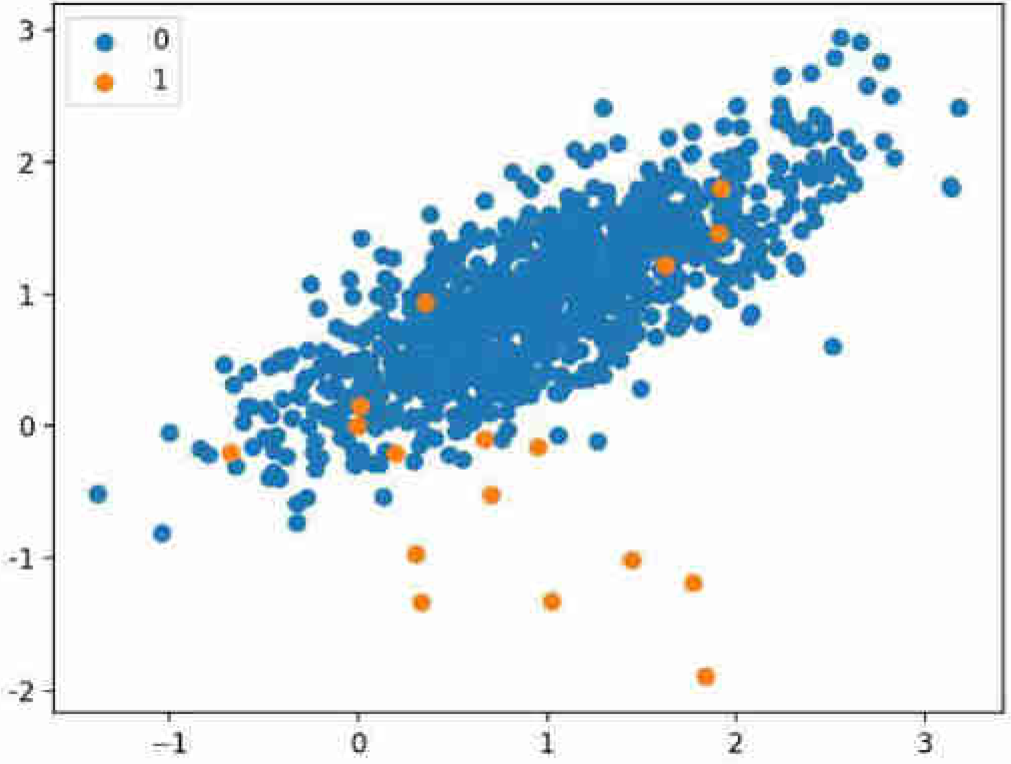
  Examples include:

  i. Random Under-sampling

  ii. SMOTE Oversampling

Machine Learning Problem Categories

- Specialized modeling algorithms may be used that pay more attention to the minority class when fitting the model on the training dataset, such as cost-sensitive machine learning algorithms.
  i. Examples include:
  ii. Cost-sensitive Logistic Regression
  iii. Cost-sensitive Decision Trees.



Scatter Plot of Imbalanced Binary Classification Dataset

Machine Learning Problem Categories

2. Clustering:

- Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as ***"A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."***
- It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.
- It is an **unsupervised learning method**, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset.
- After applying this clustering technique, each cluster or group is provided with a **cluster-ID**. ML system can use this id to simplify the processing of large and complex datasets.
- The clustering technique is commonly used for **statistical data analysis.**

Machine Learning Problem Categories

2. Clustering:

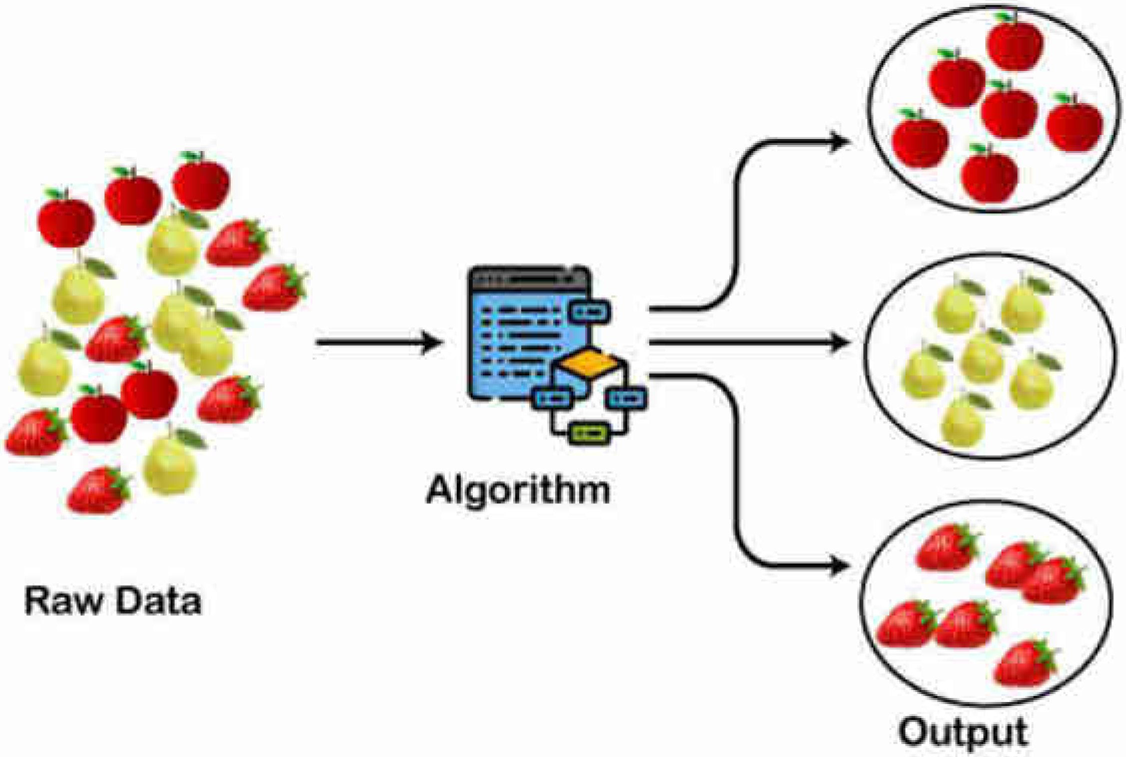
Example: Let's understand the clustering technique with the real-world example of Mall:

- When we visit any shopping mall, we can observe that the things with similar usage are grouped together. Such as the t-shirts are grouped in one section, and trousers are at other sections, similarly, at vegetable sections, apples, bananas, Mangoes, etc., are grouped in separate sections, so that we can easily find out the things. The clustering technique also works in the same way. Other examples of clustering are grouping documents according to the topic.
- The clustering technique can be widely used in various tasks. Some most common uses of this technique are:
  - Market Segmentation
  - Statistical data analysis
  - Social network analysis
  - Image segmentation
  - Anomaly detection, etc.
- Apart from these general usages, it is used by the **Amazon** in its recommendation system to provide the recommendations as per the past search of products. **Netflix** also uses this technique to recommend the movies and web-series to its users as per the watch history.

Machine Learning Problem Categories

2. Clustering:

The below diagram explains the working of the clustering algorithm. We can see the different fruits are divided into several groups with similar properties.



Machine Learning Problem Categories

2. Clustering:

- **Types of Clustering Methods:** The clustering methods are broadly divided into **Hard clustering** (data points belongs to only one group) and **Soft Clustering** (data points can belong to another group also). But there are also other various approaches of Clustering exist.

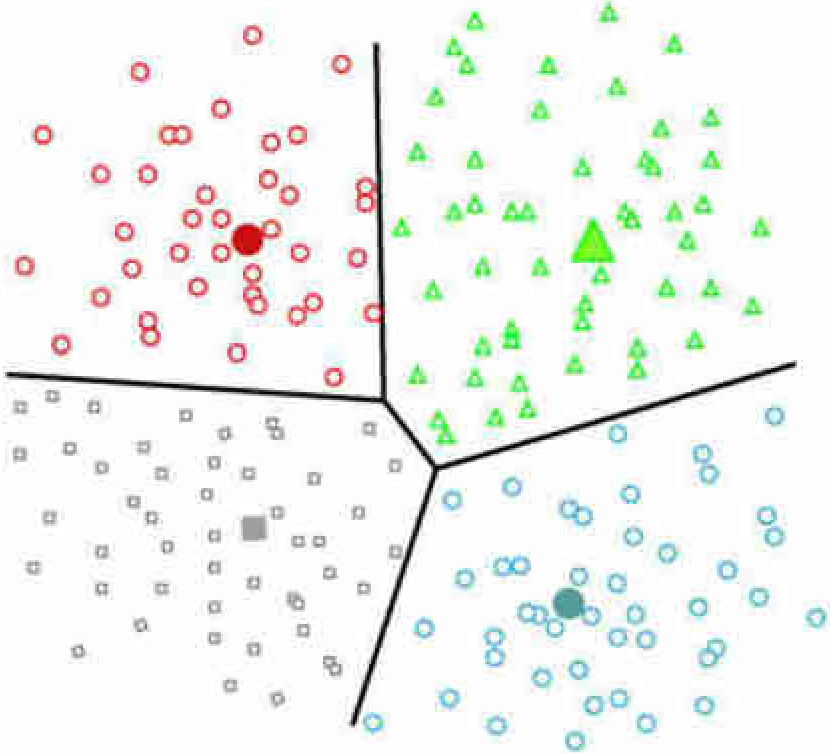
- Below are the main clustering methods used in Machine learning:
  a. Partitioning Clustering
  b. Density-Based Clustering
  c. Distribution Model-Based Clustering
  d. Hierarchical Clustering
  e. Fuzzy Clustering

Machine Learning Problem Categories

2. Clustering:

a. Partitioning Clustering:-

- It is a type of clustering that divides the data into **non-hierarchical groups**. It is also known as the **centroid-based method**. The most common example of partitioning clustering is the K-Means Clustering algorithm.
- In this type, the dataset is divided into a set of k groups, where K is used to define the number of pre-defined groups. The cluster center is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid.

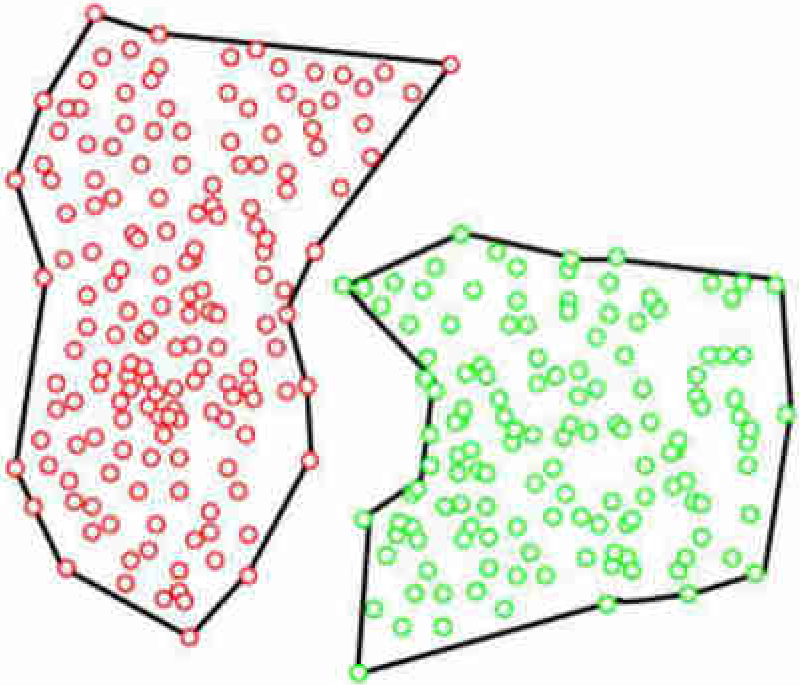


Machine Learning Problem Categories

2. Clustering:

b. Density-Based Clustering:-

- The density-based clustering method connects the highly-dense areas into clusters, and the **arbitrarily shaped distributions** are formed as long as the dense region can be connected. This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters. The dense areas in data space are divided from each other by sparser areas.
- These algorithms can face difficulty in clustering the data points if the dataset has varying densities and high dimensions.

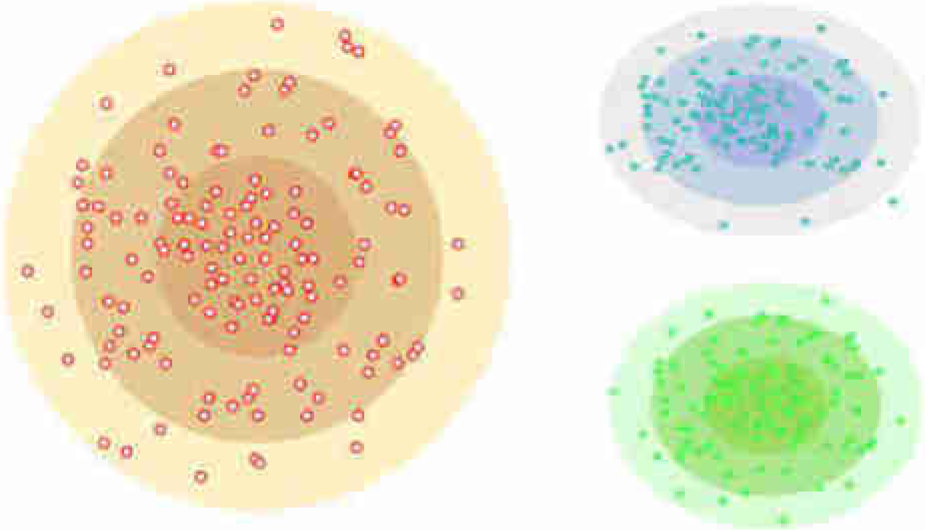


Machine Learning Problem Categories

2. Clustering:

c. Distribution Model-Based Clustering:-

- In the distribution model-based clustering method, the **data is divided based on the probability of how a dataset belongs to a particular distribution**. The grouping is done by assuming some distributions commonly **Gaussian Distribution**.
- The example of this type is the **Expectation-Maximization Clustering algorithm** that uses Gaussian Mixture Models (GMM).

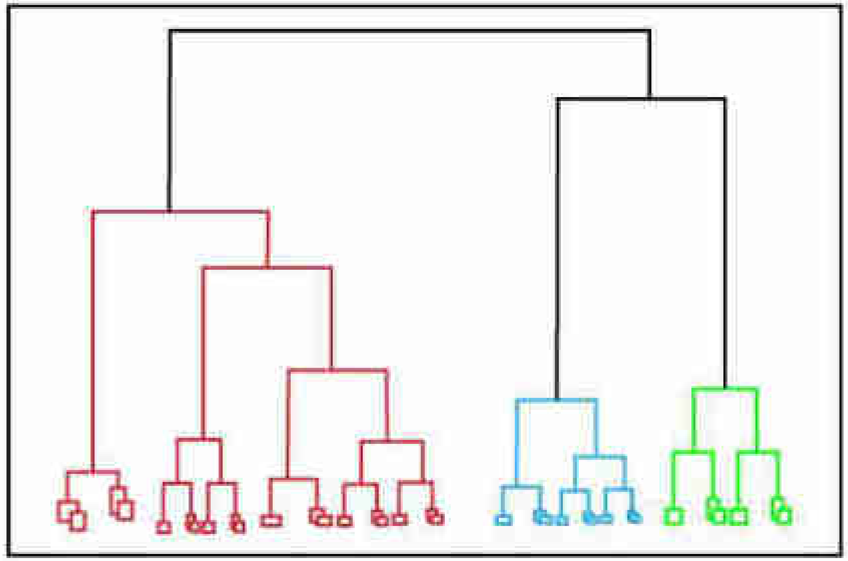


Machine Learning Problem Categories

2. Clustering:

d. Hierarchical Clustering:-

- Hierarchical clustering can be used as an alternative for the partitioned clustering as there is **no requirement of pre-specifying the number of clusters** to be created.
- In this technique, the dataset is divided into clusters to create a tree-like structure, which is also called a **dendrogram**.
- The observations or any number of clusters can be selected by cutting the tree at the correct level. The most common example of this method is the **Agglomerative Hierarchical algorithm**.



Machine Learning Problem Categories

2. Clustering:

e. Fuzzy Clustering:-

- Clustering is a type of soft method in which a data object may belong to more than one group or cluster.
- Each dataset has a set of membership coefficients, which depend on the **degree of membership** to be in a cluster.
- **Fuzzy C-means algorithm** is the example of this type of clustering; it is sometimes also known as the Fuzzy k-means algorithm.

Machine Learning Problem Categories

2. Clustering:

Clustering Algorithms:-

- **K-Means algorithm:** The k-means algorithm is one of the most popular clustering algorithms. It classifies the dataset by dividing the samples into different clusters of equal variances. The number of clusters must be specified in this algorithm. It is fast with fewer computations required, with the linear complexity of **O(n).**
- **Mean-shift algorithm:** Mean-shift algorithm tries to find the dense areas in the smooth density of data points. It is an example of a centroid-based model, that works on updating the candidates for centroid to be the center of the points within a given region.
- **DBSCAN Algorithm:** It stands **for Density-Based Spatial Clustering of Applications with Noise**. It is an example of a density-based model similar to the mean-shift, but with some remarkable advantages. In this algorithm, the areas of high density are separated by the areas of low density. Because of this, the clusters can be found in any arbitrary shape.
- **Expectation-Maximization Clustering using GMM:** This algorithm can be used as an alternative for the k-means algorithm or for those cases where K-means can be failed. In GMM, it is assumed that the data points are Gaussian distributed.
- **Agglomerative Hierarchical algorithm:** The Agglomerative hierarchical algorithm performs the bottom-up hierarchical clustering. In this, each data point is treated as a single cluster at the outset and then successively merged. The cluster hierarchy can be represented as a tree-structure.
- **Affinity Propagation:** It is different from other clustering algorithms as it does not require to specify the number of clusters. In this, each data point sends a message between the pair of data points until convergence. It has $O(N^2T)$ time complexity, which is the main drawback of this algorithm.

Machine Learning Problem Categories

3. Regression:

- Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as **temperature, age, salary, price,** etc.
- We can understand the concept of regression analysis using the below example:

Example: Suppose there is a marketing company A, who does various advertisement every year and get sales on that. The below list shows the advertisement made by the company in the last 5 years and the corresponding sales:

Now, the company wants to do the advertisement of $200 in the year 2019 and wants to know the prediction about the sales for this year. So to solve such type of prediction problems in machine learning, we need regression analysis.

Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

Machine Learning Problem Categories

3. Regression:

- Regression is a supervised learning technique which helps in finding the ***correlation between variables*** and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

- In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data. In simple words, *"Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum."* The distance between datapoints and line tells whether a model has captured a strong relationship or not.

- Some examples of regression can be as:
  - Prediction of rain using temperature and other factors
  - Determining Market trends
  - Prediction of road accidents due to rash driving.

Machine Learning Problem Categories

3. Regression:

Terminologies Related to the Regression Analysis:

- **Dependent Variable:** The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called **target variable**.
- **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a **predictor**.
- **Outliers:** Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.
- **Multicollinearity:** If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.
- **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is called **underfitting**.

Machine Learning Problem Categories

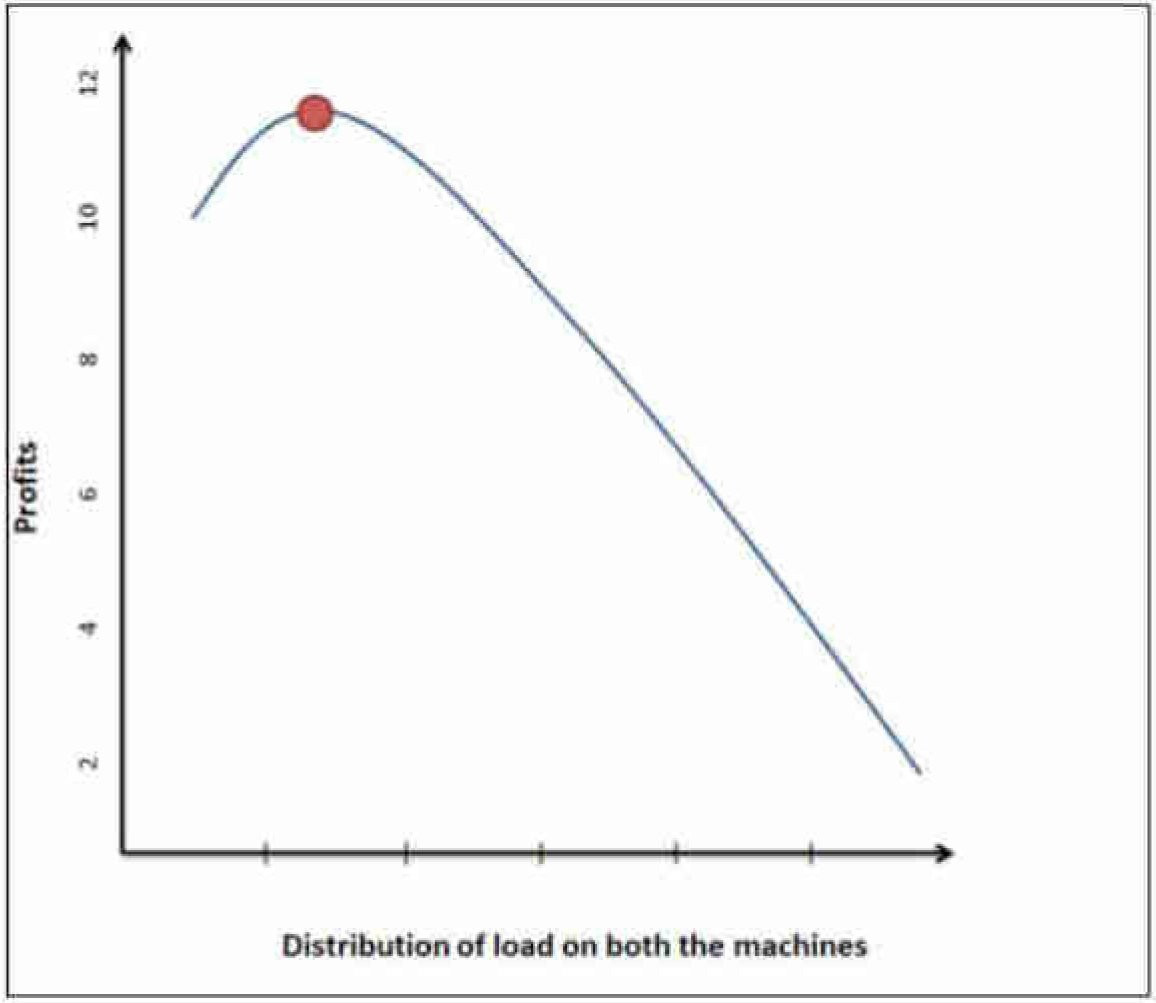
4. Optimization:

- **Optimization is the problem of finding a set of inputs to an objective function that results in a maximum or minimum function evaluation.**
- Optimization, in simple terms, is a mechanism to make something better or define a context for a solution that makes it the best.
- Consider a production scenario:- Let's assume there are two machines that produce the desired product-
  - one machine requires more energy for high speed in production and lower raw materials
  - other requires higher raw materials and less energy to produce the same output in the same time.
- It is important to understand the patterns in the output based on the variation in inputs; a combination that gives the highest profits would probably be the one the production manager would want to know.
- As an analyst, one needs to identify the best possible way to distribute the production between the machines that gives them the highest profit.

Machine Learning Problem Categories

4. Optimization:

- The following image shows the point of highest profit when a graph was plotted for various distribution options between the two machines. Identifying this point is the goal of this technique.



Architecture of Machine Learning

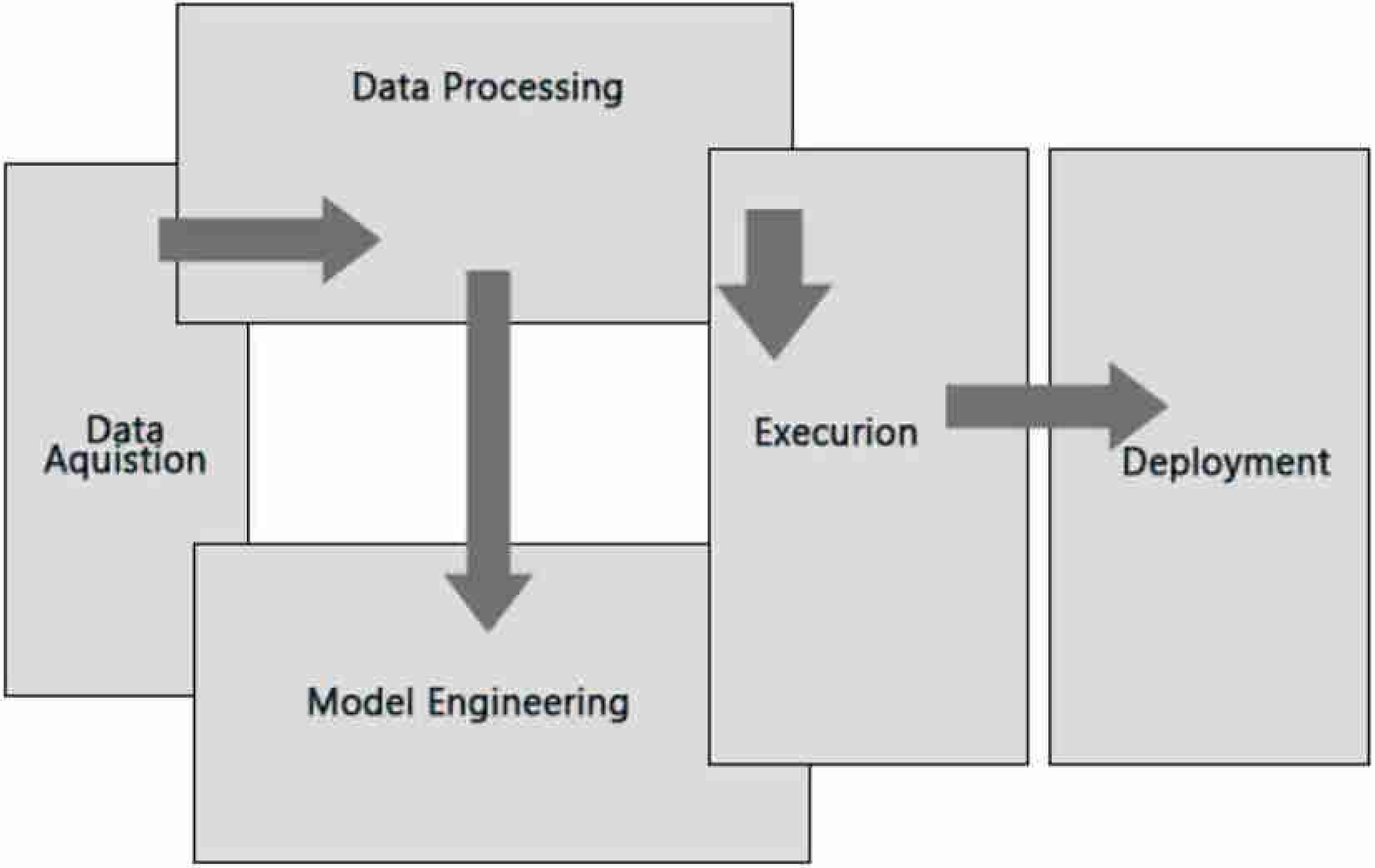


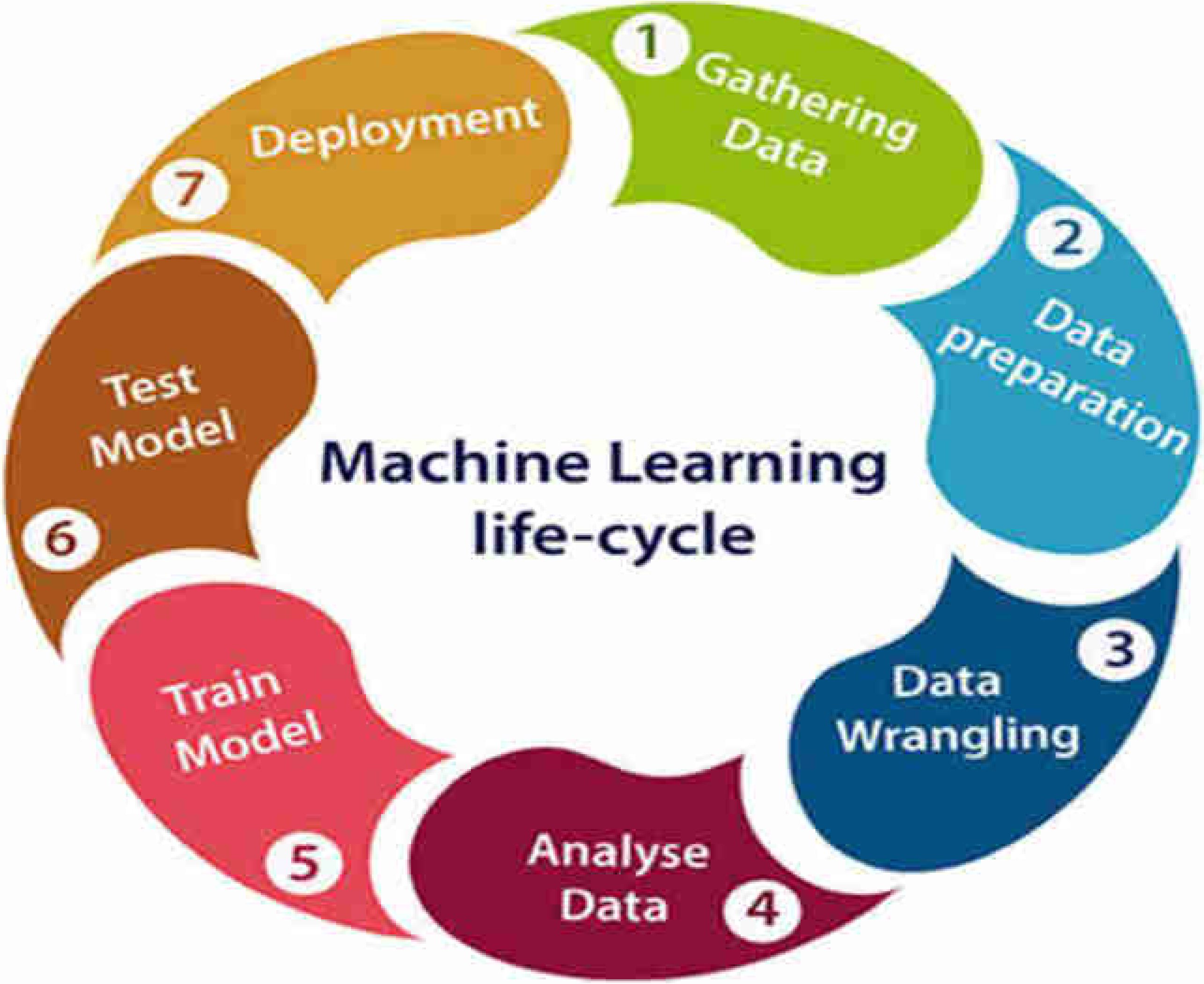
Fig:- Block diagram of decision flow architecture for Machine learning systems

Architecture of Machine Learning

1. **Data Acquisition:** This involves ***data collection, preparing and segregating*** the case scenarios based on certain features involved with the decision making cycle and forwarding the data to the processing unit for carrying out further categorization. This stage is sometimes called the **data preprocessing stage**.
2. **Data Processing:** The received data in the data acquisition layer is then sent forward to the data processing layer where it is subjected to advanced integration and processing and involves ***normalization of the data, data cleaning, transformation, and encoding***.
3. **Data Modelling:** This layer of the architecture involves the ***selection of different algorithms*** that might adapt the system to address the problem for which the learning is being devised.
4. **Execution:** This stage in machine learning is where the experimentation is done, ***testing is involved and tunings are performed***. The output of the step is a refined solution capable of providing the required data for the machine to make decisions.
5. **Deployment:** This is the system deployment phase. i.e. ***application phase***.

Process Life Cycle in Machine Learning

- Machine learning life cycle involves seven major steps, which are given below:
  1. Gathering Data
  2. Data preparation
  3. Data Wrangling
  4. Analyze Data
  5. Train the model
  6. Test the model
  7. Deployment



Process Life Cycle in Machine Learning

- The most important thing in the complete process is to ***understand the problem*** and to know the purpose of the problem. Therefore, before starting the life cycle, we need to understand the problem because the good result depends on the better understanding of the problem.
- In the complete life cycle process, to solve a problem, we create a machine learning system called "**model**", and this model is created by providing "**training**". But to train a model, we need data, hence, life cycle starts by collecting data.

1. **Gathering Data:** In this step, we need to identify the different data sources, as data can be collected from various sources such as **files**, **database**, **internet**, or **mobile devices**. The quantity and quality of the collected data will determine the efficiency of the output. The more will be the data, the more accurate will be the prediction.

- This step includes the below tasks:
  - Identify various data sources
  - Collect data
  - Integrate the data obtained from different sources
- By performing the above task, we get a coherent set of data, also called as a **dataset**. It will be used in further steps.

Process Life Cycle in Machine Learning

2. Data preparation: After collecting the data, we need to prepare it for further steps. In this step, first, we put all data together, and then randomize the ordering of data.

- This step can be further divided into two processes:
- **Data exploration:**
  It is used to understand the nature of data that we have to work with. We need to understand the characteristics, format, and quality of data.
  A better understanding of data leads to an effective outcome. In this, we find Correlations, general trends, and outliers.
- **Data pre-processing:**
  Now the next step is preprocessing of data for its analysis.

3. Data Wrangling: Data wrangling is the process of cleaning and converting raw data into a useable format. It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step. In real-world applications, collected data may have various issues, including- ***Missing Values, Duplicate data, Invalid data, Noise***. So, we use various filtering techniques to clean the data. It is mandatory to detect and remove the above issues because it can negatively affect the quality of the outcome.

Process Life Cycle in Machine Learning

4. Data Analysis: Now the cleaned and prepared data is passed on to the analysis step.

- This step involves:
  - Selection of analytical techniques
  - Building models
  - Review the result

- The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome.

- It starts with the determination of the type of the problems, where we select the machine learning techniques such as **Classification**, **Regression**, **Cluster analysis**, **Association**, etc. then build the model using prepared data, and evaluate the model.

- Hence, in this step, we take the data and use machine learning algorithms to build the model.

Process Life Cycle in Machine Learning

5. Train Model: Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem.

- We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various **patterns, rules, and, features.**

6. Test Model: Once our machine learning model has been trained on a given dataset, then we test the model. In this step, we check for the accuracy of our model by providing a test dataset to it.

- Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

7. Deployment: The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system.

Process Life Cycle in Machine Learning

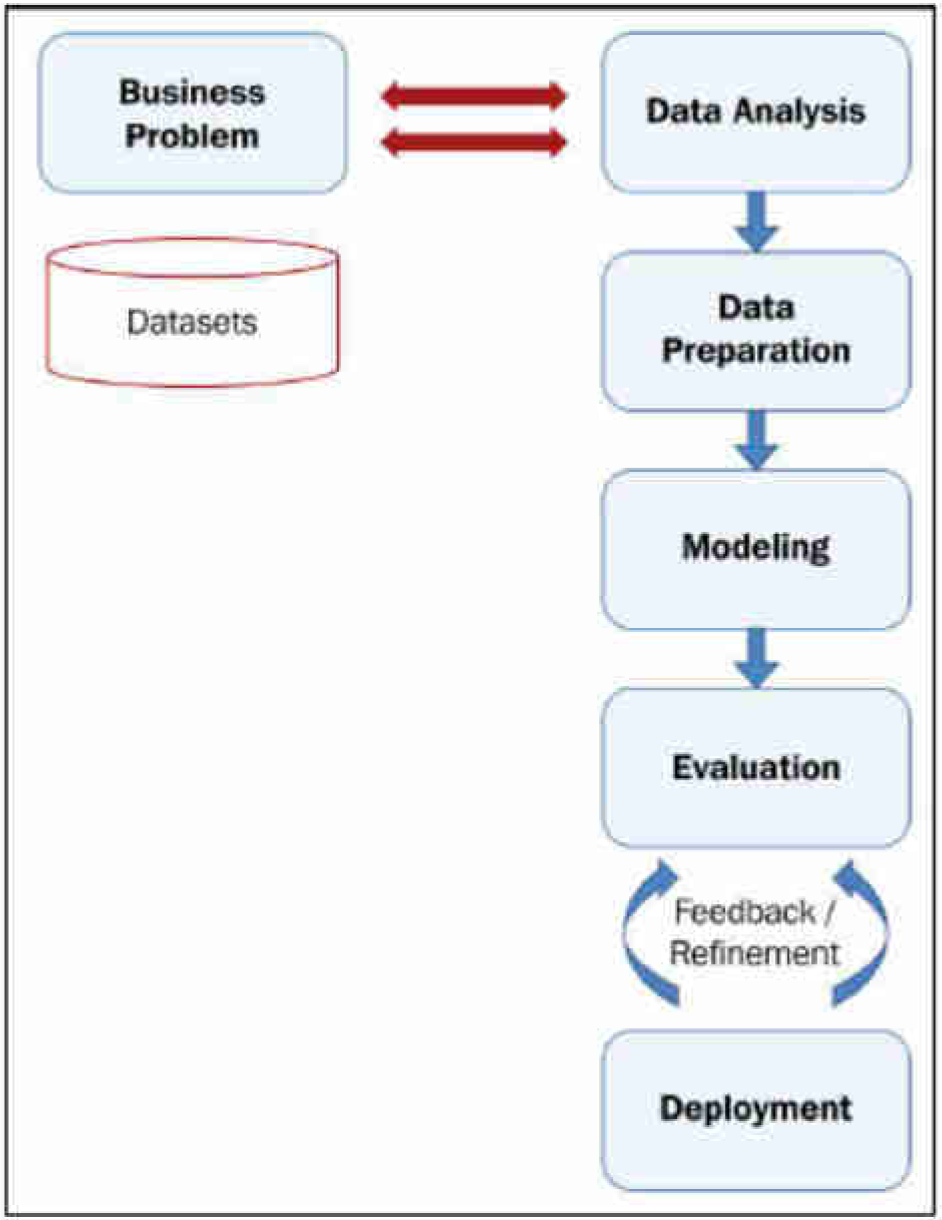
1. Defining the problem statement, which includes defining the goal, process, and assumptions.

2. Determine what problem type is this problem classified under? Whether it is a classification, regression, or optimization problem?

3. Choose a metric that will be used to measure the accuracy of the model.
4. In order to ensure the model works well with the unseen data:
   - Build the model using training data.
   - Tweak the model using test data.
   - Declare an accuracy based on the final version.

The following figure explains the flow and architecture of the underlying system:

Process Life Cycle in Machine Learning



Data and inconsistencies in Machine learning

- Before understanding the overfitting and underfitting, let's understand some basic term that will help to understand this topic well:

- **Signal:** It refers to the true underlying pattern of the data that helps the machine learning model to learn from the data.

- **Noise:** Noise is unnecessary and irrelevant data that reduces the performance of the model.

- **Bias:** Bias is a prediction error that is introduced in the model due to oversimplifying the machine learning algorithms. Or it is the difference between the predicted values and the actual values.

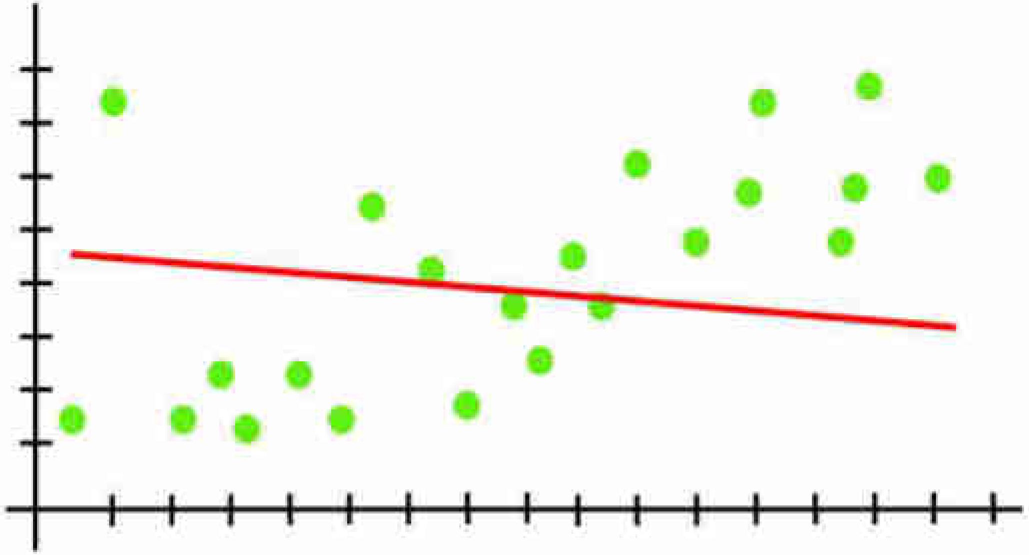
- **Variance:** If the machine learning model performs well with the training dataset, but does not perform well with the test dataset, then variance occurs.

Data and inconsistencies in Machine learning

- **Under-fitting:** Underfitting occurs when our machine learning model is not able to capture the underlying trend of the data.
- **Example:** We can understand the underfitting using below output of the linear regression model:
- As we can see from the above diagram, the model is unable to capture the data points present in the plot.

How to avoid underfitting:

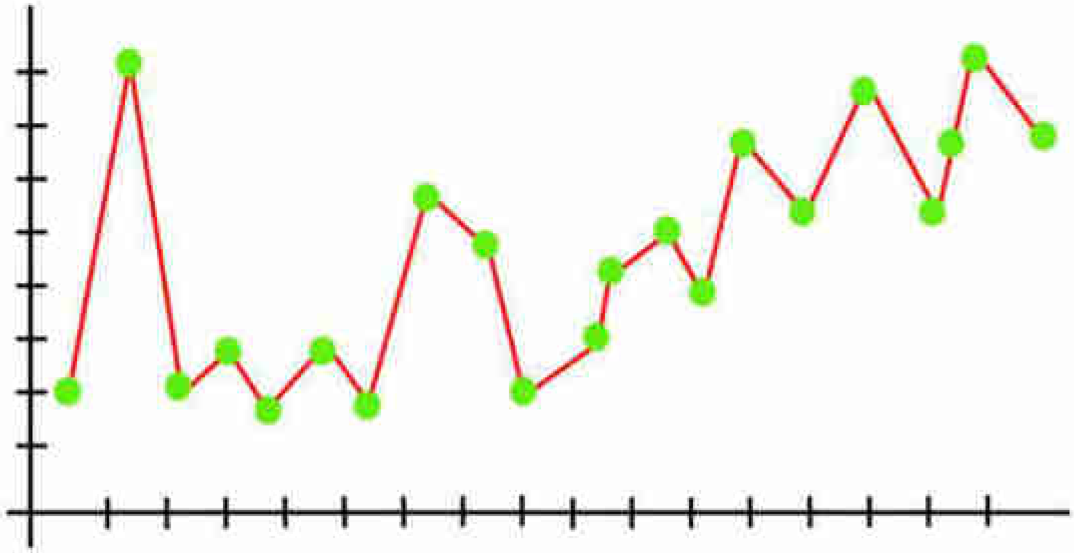
- By increasing the training time of the model.
- By increasing the number of features.



Data and inconsistencies in Machine learning

- **Overfitting:** Overfitting occurs when our machine learning model tries to cover all the data points or more than the required data points present in the given dataset.
- Because of this, the model starts caching noise and inaccurate values present in the dataset, and all these factors reduce the efficiency and accuracy of the model.
- The overfitted model has low bias and high variance.

- The chances of occurrence of overfitting increase as much we provide training to our model. It means the more we train our model, the more chances of occurring the overfitted model.
- Overfitting is the main problem that occurs in supervised learning.



Data and inconsistencies in Machine learning

- As we can see from the above graph, the model tries to cover all the data points present in the scatter plot. It may look efficient, but in reality, it is not so. Because the goal of the regression model to find the best fit line, but here we have not got any best fit, so, it will generate the prediction errors.

How to avoid the Overfitting in Model

- Both overfitting and underfitting cause the degraded performance of the machine learning model. But the main cause is overfitting, so there are some ways by which we can reduce the occurrence of overfitting in our model.
  - Cross-Validation
  - Training with more data
  - Removing features
  - Early stopping the training
  - Regularization
  - Ensembling

Data and inconsistencies in Machine learning

- **Data instability:** Machine learning algorithms are usually robust to noise within the data.
- A problem will occur if the ***outliers*** are due to manual error or misinterpretation of the relevant data. This will result in a ***skewing of the data***, which will ultimately end up in an incorrect model.
- Therefore, there is a strong need to have a process to correct or handle human errors that can result in building an incorrect model.

- **Unpredictable data formats:** Machine learning is meant to work with ***new data*** constantly coming into the system and learning from that data.
- Complexity will creep in when the new data entering the system comes in formats that are not supported by the machine learning system.
- It is now difficult to say if our models work well for the new data given the instability in the formats that we receive the data, unless there is a mechanism built to handle this.

Performance Measures in Machine Learning

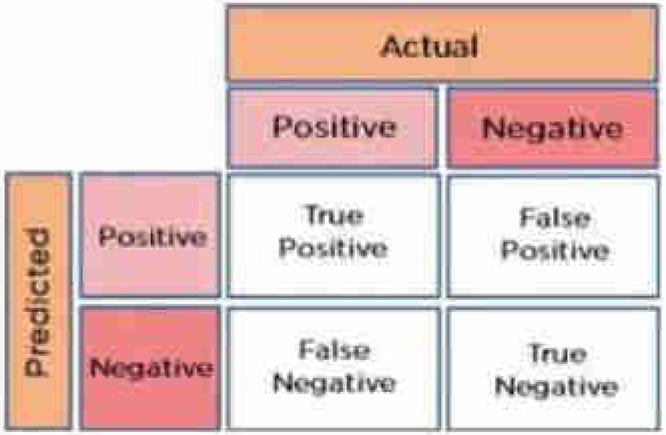
Probably Approximately Correct (PAC) Theory.

- There are two types of uncertainties as per the PAC theory:
- **Approximate**: This measures the extent to which an error is accepted for a hypothesis.
- **Probability**: This measure is the percentage certainty of the hypothesis being correct.
- The **confusion matrix** is a matrix used to determine the performance of the classification models.
- Since it shows the errors in the model performance in the form of a matrix, hence also known as an **error matrix**. Some features of Confusion matrix are given below:
  - For the 2 prediction classes of classifiers, the matrix is of 2*2 table, for 3 classes, it is 3*3 table, and so on.
  - The matrix is divided into two dimensions, that are **predicted values** and **actual values** along with the total number of predictions.
  - Predicted values are those values, which are predicted by the model, and actual values are the true values for the given observations.

Performance Measures in Machine Learning

- It looks like the below table:

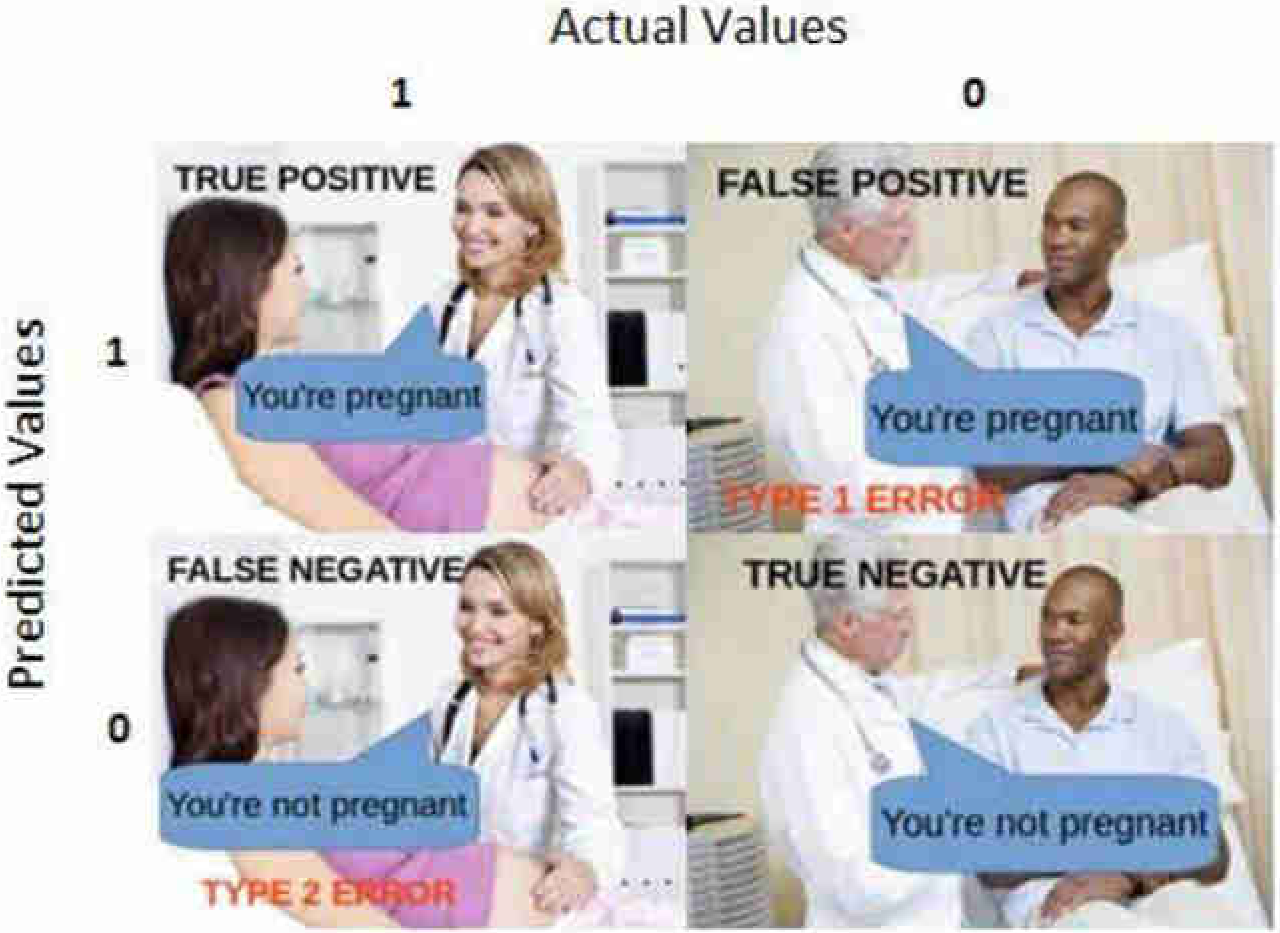
n = total predictions	Actual: No	Actual: Yes
Predicted: No	True Negative	False Positive
Predicted: Yes	False Negative	True Positive



		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

- **True Negative (TN):** Model has given prediction No, and the real or actual value was also No.
- **True Positive (TP):** The model has predicted yes, and the actual value was also true.
- **False Negative (FN):** The model has predicted no, but the actual value was Yes, it is also called as **Type-II error**.
- **False Positive (FP):** The model has predicted Yes, but the actual value was No. It is also called a **Type-I error.**

Performance Measures in Machine Learning



Confusion Matrix [Image 3] (Image courtesy: My Photoshopped Collection)

Performance Measures in Machine Learning

- **Need for Confusion Matrix in Machine learning:**
  - It evaluates the performance of the classification models, when they make predictions on test data, and tells how good our classification model is.
  - It not only tells the error made by the classifiers but also the ***type of errors*** such as it is either type-I or type-II error.
  - With the help of the confusion matrix, we can calculate the different parameters for the model, such as ***accuracy, precision***, etc.

https://www.simplilearn.com/tutorials/machine-learning-tutorial/confusion-matrix-machine-learning

https://www.w3schools.com/python/python_ml_confusion_matrix.asp

Performance Measures in Machine Learning

- **Calculations using Confusion Matrix:**
- We can perform various calculations for the model, such as the model's accuracy, using this matrix. These calculations are given below:

Classification Accuracy: It defines how often the model predicts the correct output. It can be calculated as the ratio of the ***number of correct predictions*** made by the classifier to ***all number of predictions*** made by the classifiers. The formula is given below:

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN}$$

- **Misclassification rate:** It is also termed as ***Error rate***, and it defines how often the model gives the wrong predictions. The value of error rate can be calculated as the ***number of incorrect predictions*** to ***all number of the predictions*** made by the classifier. The formula is given below:

$$\text{Error rate}=\frac{FP+FN}{TP+FP+FN+TN}$$

Performance Measures in Machine Learning

- **Calculations using Confusion Matrix:**
- **Precision:** It can be defined as the ***number of correct outputs*** provided by the model or ***out of all positive classes*** that have predicted correctly by the model, how many of them were actually true. It can be calculated using the below formula:

$$\text{Precision}= \frac{TP}{TP+FP}$$

- **Recall:** It is defined as the out of ***total positive classes***, how our model predicted correctly. The recall must be as high as possible.

$$\text{Recall} = \frac{TP}{TP+FN}$$

- **F-measure:** If two models have low precision and high recall or vice versa, it is difficult to compare these models. So, for this purpose, we can use F-score. This score helps us to evaluate the recall and precision at the same time. The F-score is maximum if the recall is equal to the precision. It can be calculated using the below formula:

$$\textbf{F-measure} = \frac{2*Recall*Precision}{Recall+Precision}$$

Performance Measures in Machine Learning

- **Calculations using Confusion Matrix:**

- Other important terms used in Confusion Matrix:

- **Null Error rate:** It defines how often our model would be incorrect if it always predicted the majority class. As per the accuracy paradox, it is said that "*the best classifier has a higher error rate than the null error rate.*"

- **ROC Curve:** The ROC is a graph displaying a classifier's performance for all possible thresholds. The graph is plotted between the true positive rate (on the Y-axis) and the false Positive rate (on the x-axis).

Performance Measures in Machine Learning

Mean squared error (MSE): To compute the MSE, we first take the ***square of the difference between the actual and predicted values of every record***.

We then take the ***average*** value of these squared errors. If the predicted value of the i^{th} record is *Pi* and the actual value is *Ai*, then the MSE is:

- It is also common to use the square root of this quantity called **root mean square error** (**RMSE**).

$$MSE = \frac{\sum_{i=1}^{n} (P_i - A_i)^2}{n}$$

- **Mean absolute error (MAE):** To compute the MAE, we take the ***absolute difference between the predicted and actual values of every record***. We then take the ***average*** of those absolute differences.

$$MAE = \frac{\sum_{i=1}^{n} |P_i - A_i|}{n}$$

Performance Measures in Machine Learning

Normalized MSE and MAE (NMSE and NMAE): Ratio of MSE of Developed model to native model.

$$NMSE = \frac{MSE\ of\ developed\ model}{MSE\ of\ naive\ model}$$