Introduction to Machine Learning

*Present 1-2: The Title slide and the Learning Objectives slide. An overview of what we will achieve in this course.*

Lesson Objectives

By the end of this lesson you will be able to:

* Describe three types of learning
* Understand how machine learning works through real-life examples
* Take a decision on which method to use where

Introduction:

*Announce 3: Welcome the class. Introduce yourself and discuss what the course will cover. Talk about the topics that will be covered in this lesson*

Machine Learning is the methodical learning of Procedures and statistical representations that are used to accomplish computations on tasks which does not require human interventions. So, in other words, it is the process of teaching computers to perform tasks by itself without explicit instructions, relying only on patterns and inferences.

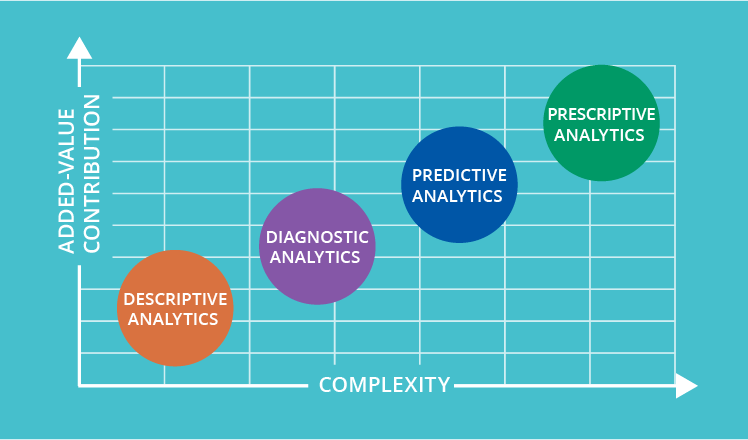
Machine Learning Algorithms are used in a variety of places, and recently has it found scopes beyond imagination. Some common use of machine learning algorithms is in *Email-filtering, Computer Vision, Computational Linguistics* and so on. Machine learning is called computational statistics because it focuses on making predictions using computers and algorithms and these algorithms are designed using the concepts of Statistic.

* + - Types of Analytics:

There are totally 4 types of analytics:

* Prescriptive Analytics
* Predictive Analytics
* Diagnostic Analysis
* Descriptive Analysis

Data Analytics is the process of Extraction, Transformation, Loading, Modelling and deriving conclusions from the data so that effective decisions can be made benefitting the business. Deriving effective conclusions from the data is called as business intelligence that would support a useful future forecast.



Here the Descriptive and Diagnostic analysis are mostly concerned with the past details of the data. These two acts as the base of the Analytics and they cannot be neglected.

Descriptive Analysis is subjected to answering questions like *“what happened?”* and *“how many, when and where?”.*

Diagnostic Analysis goes into reporting at an even deeper level like answering questions like “*why”* using queries and in order to query out the even more detailed report “*where”* we should be looking at.

Predictive Analysis deals with future and we will be working with “*Statistical Modelling”* to identify the “*Patterns”* in the data and do “*Predictive Modelling”* to see and predict what will happen next.

Prescriptive Analysis is where the Concept of Machine Learning, Artificial Intelligence and Big data will come into play. Statistical modelling is more related to finding the correlation and thereby proving the hypothesis right or wrong. Here we will go on to perform “testing” and *“optimize”* the based on the results from the testing.

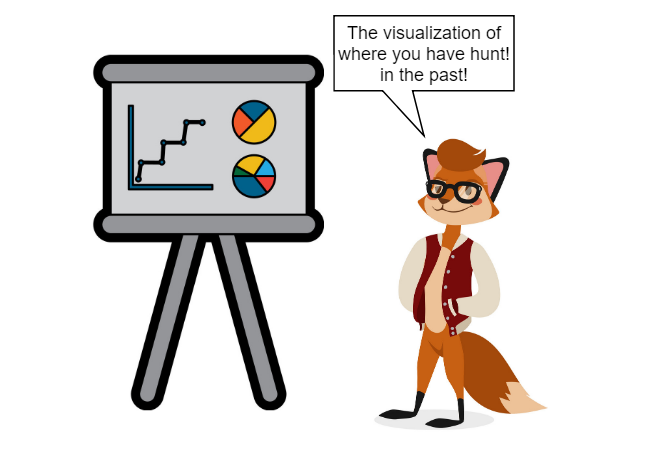
The as we move from Descriptive Analysis towards the right the complexity and the value increases.

*Real Life E.g: Restaurant chain expansion!*

A CEO of a famous restaurant chain in the city was looking to find out the most liked dish in each of his restaurant so that he can expand a small outlet outside the restaurants to serve it as take away. He hired a consultant to help him find out the most liked dishes of his restaurants.

CEO was quite confused as consultant never spent a day in kitchen to look at dishes, however he was doing a continuous research on the computer where the data of the daily sales, online orders and customer feedback was stored.

After a week, a consultant made a presentation to the CEO explaining the details of how he figured out the most liked dishes of each restaurant (would you like to guess what's the most liked dish in your area?) . This helped the profitable expansion of restaurant. This is an example of Descriptive Analytics.



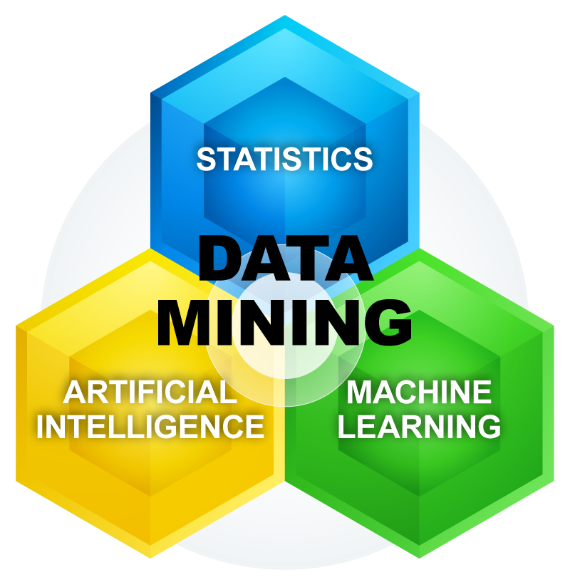
Following week, the consultant built a model to predict the probability of a most liked dish being preferred at a given place and time with the help of Machine Learning techniques. Additionally, the consultant predicted the demand for other dishes and suggested new places to install outlets/restaurants to maximize profit. This is an example for Predictive analysis and Optimization.

After finding all of these, the consultant starts to setup outlets/takeaway with liked dishes counters in various identified places of the city so that the profit of restaurant shoots up automatically. This is the example for Automation

Thus, the consultant goes from Descriptive Analytics -> Predictive Analytics and Optimization -> Automation which is exactly the process cycle of Analytics.

Data Mining:

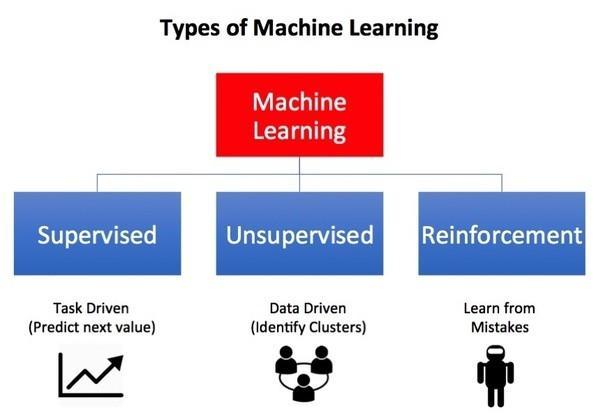
Data mining is a phenomenon wherein the concepts of S*tatistics, is intersected with both Machine Learning* and Artificial Intelligence.



The task of identifying the underlying patterns in the larger dataset is called Data Mining. Its main goal is to extract information from the dataset and transform the information into a complete structure. The variance between analysis of data and mining of data is that the former verifies the models and theory on datasets irrespective of the expanse of data. Whereas the latter employs machine learning and statistical models to discover concealed patterns in a huge size of data.

Types of learning in algorithms

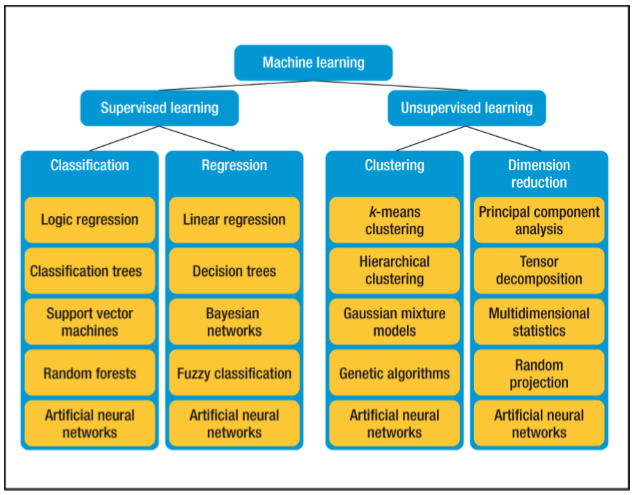
* Supervised and Semi-supervised learning
* Un-supervised learning
* Reinforcement learning



Supervised Learning is more of a Task Driven as we will try to guide the model to predict the next value of the target variable.

Unsupervised Learning is used when there is no specific target variable and the cluster must be formed with the help of the pattern in data so-called data-driven.

Reinforcement Learning is defined by Win-Lose situation and the model learns from its mistakes. Don’t worry we will see all the concepts in detail in the upcoming sessions.



* 1. Supervised Learning:

*Present 4: Slide introducing the topic*

Supervised learning as the name suggests will have a supervisor i.e., a trainer. Principally, it is a learning system which helps the machine in getting trained with the labelled data. Subsequently, the trained machine is fed with a new set of data on which the supervised learning algorithm performs its analysis, resulting in the correct outcome from the labelled data. This can also be called predictive modeling.

*Real Life E.g.1: SPAM MESSAGE FILTERING!*

Consider a company ABC that has a software-based support system to handle the customer queries sent via email or messages or website or support portal. The support agents of the ABC company face a very big problem of getting spam tickets sourced from various channels and reduces the working velocity of agents.

The team of experts (Data Scientist, Software Architect and Business Analyst) were asked to solve the above issue. After various discussions and brainstorming, the team converges to use supervised Machine Learning classifier to build an expert system to classify the valid tickets vs spam tickets. In order to do this, the Data scientist starts mining the historical data that has been already labelled as valid and spam ticket based on the various parameters like size of message, sender’s email id, ip addresses, words in the subject and content etc. The supervised machine learning classifier is trained with huge number of historical tickets that are labelled. The model learns as it sees the variation of values for each parameters and labels associated to it.

The following are the criterion based on which the E-mail is classified as **SPAM/VALID**

* If the size of the email is big, email addresses is too short or too big and body contains suspicious words like sale, offer, big amount etc., – **SPAM**.
* If the size of the email is moderate, email addresses follows email standards and body contains words like support, not working, trouble connecting etc., – **VALID**.

Once the training is done, evaluation of the model is done by presenting a new ticket to the machine to identify it based on the previous learning. As the machine has the intelligence to classify the e-mail as valid/spam, the model classifies the new ticket based on its size, content and e-mail address and confirms as valid/spam

within no time. Thus, the machine is trained using the train data set or already labeled data set (historical tickets with proper classification of valid or spam) and the knowledge learned is applied to the Test set (the single ticket). The objective of a machine learning model is to take a broad view from the training data to any data of the problem domain. This helps the model to make efficient predictions in the future, whenever a never seen data is given to it.

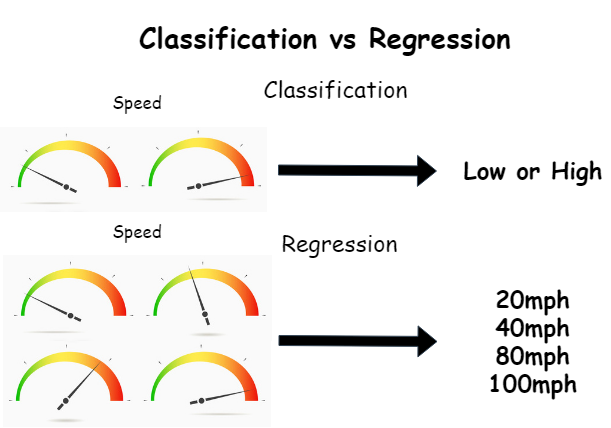
Supervised learning is broadly categorized into two:

**Classification:** A classification algorithm helps in identifying to which of a group of classes (sub-populations) a new class belongs, on the basis of a training data set which contains instances, whose class association is known. Examples are assigning a given loan account as "default" or "non-default" class, and conveying a investigation to a given patient built on experiential features of the patient (sex, blood pressure, presence or absence of certain symptoms, etc.).

**Regression:** It is a set of statistical processes for approximating the associations among variables. It comprises many practices for demonstrating and studying numerous variables, when the focus is on the relationship between a [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) and one or more [independent variables](https://en.wikipedia.org/wiki/Independent_variable) (or 'predictors'). Additionally, regression helps one understand how the characteristic value of the dependent variable (or 'criterion variable') varies when any one of the independent variables is varied, while the other independent variables are held fixed.

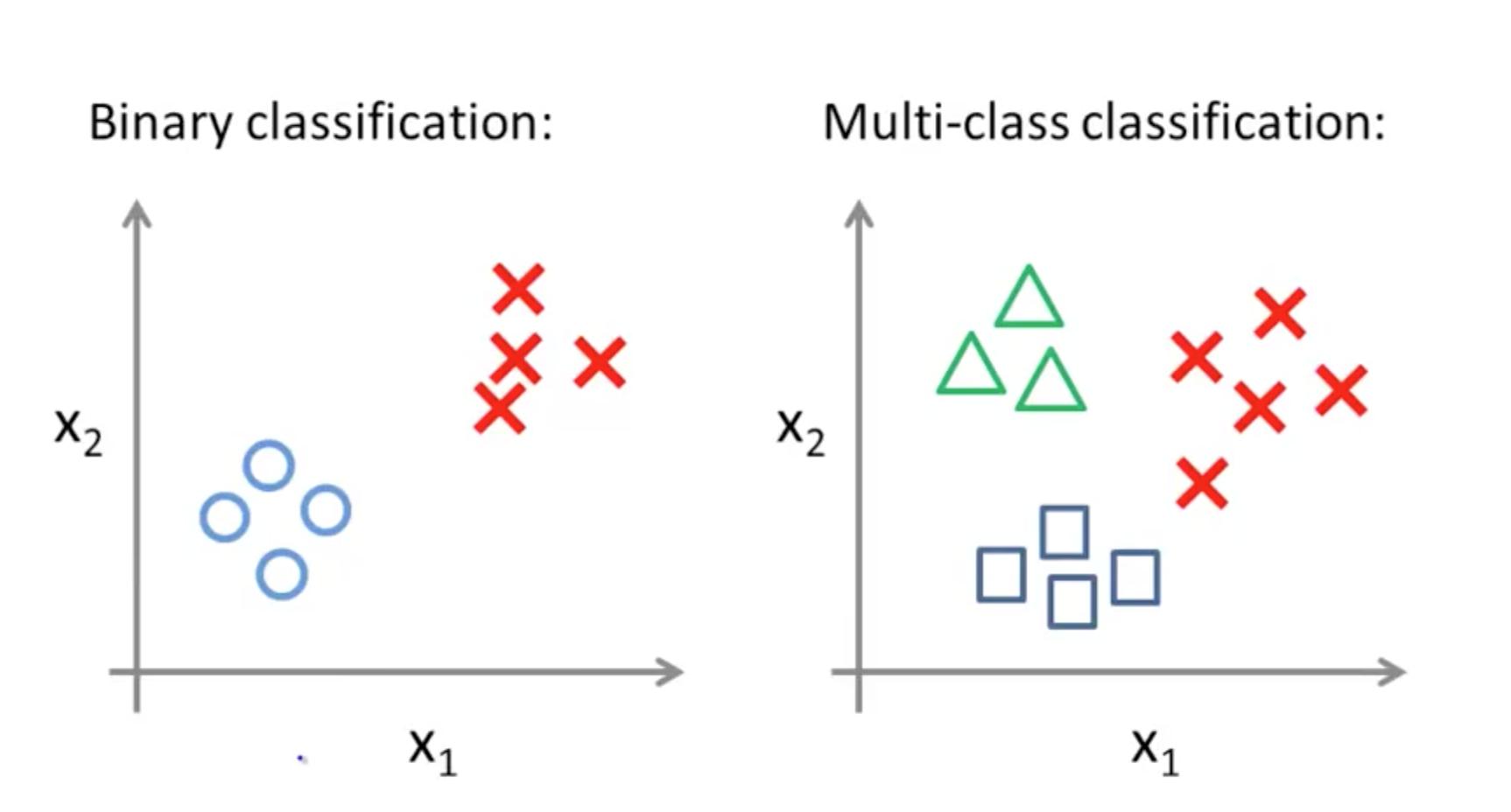
**Classification and Regression**

*Real Life E.g.2: Speed up!*



**Classification:**

The classification algorithm is used when we have to predict the target class (Yes/ No). When the prediction is between two classes, it is known as binary classification. With Speed as the target variable, predicting whether the driver is over speeding or not, and also to predict whether a customer will buy a new product or not having in mind the transaction details are examples of binary classification. If the prediction involves more than two target classes, it is known as multi-classification. Considering the speed of the car and the speed limit of the road we can predict if the driver is over speeding or not. Finding the objects in the image. These problems are called a multi-classification problem.



**Regression:**

Linear regressionmeasures the link between one or more predictor variablesand one outcome variable*.*Linear regression is used to predict the outcome using analysis and modeling. For example, linear regression helps to enumerate the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable).  Linear regression is also known as *multiple regression*,*multivariate regression*,*ordinary least squares (OLS)*, and *regression*.

**Time Series Analysis**

As name suggests time series data is the data that is distributed with respect to time i.e., data available in a chronological order. Stock market prediction, Customer churn prediction are some of the examples of Time series data. Depending on the requirement or the necessities time series can be either *Regression* or *Classification.*  Now let us look at the components of a time series data and their general definitions.

**Definitions:**

**Time Series:**

A time series is a sequence of recording of data points measured over a period of time. Most commonly, time series is defined as consecutive data points collected in correspondingly spread out points in time.

Components of time series:

The aspects that are accountable for carrying about changes in a time series, termed as components of the same are explained below.

**Trend:**

This is the key element of TS algorithm, due to the socio-economic and political issues over a long period of time. This trend may show the crest or troughs in a time series over a historical time. Variation of stock market prices depending on the political factors of a country is an example of Trend.

**Seasonal variations:**

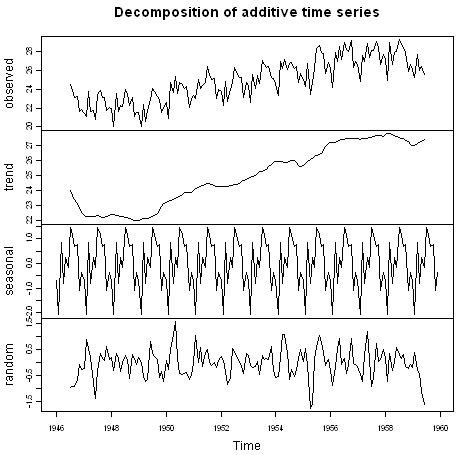
These are changes occurring in the data for a shorter duration of time. The short-term can be defined as very small period of time such as Rainy Season or festivities. During rainy season sale of raincoats will reach a new high. Similarly, the sale of garments and jewellery will be high during festive time.

**Cyclical variations:**

This long-term change is seen mostly in economic data and may be retained for more than a decade. These variations are mostly associated with the business cycles.

**Irregular Component:**

This random or residual variations of the data happen suddenly due to the unfavorable environmental condition. This component is unlikely to be repeated.



*Real Life E.g.3: Hold on to your Hat!*

Consider you are now in the windiest place in the world, Wellington, New Zealand. And you want to train the machine that will make a useful prediction in that place. What can be done? Assume you have all the data you will need for training the machine.

Since it is said you have all the data. The data will be in the form of *Time series*. Time Series as in the day to day information about wind speed and how frequently are the days being windy. So, with this data we can do two things.



1. Build a classification model:

The machine can be trained to forecast whether the next day will be windy or not. What is the frequency (if any) it is windy? And many more depending on the features available in the data.

1. Build a Regression model:

Okay! We now know for sure next day is going to be windy. Now we need to know if it’s going to be very worse that we cannot come out from home or is it going to be usual, just a normal day in Wellington? For this we can train the machine with the day to day wind speed data to forecast the wind speed.

***Note***

*Fb prophet is a forecasting procedure by facebook. It is fast and provides completely automated forecasts that can be tuned by hand by data scientists*

*Find more information about it here: d*[*https://facebook.github.io/prophet/docs/quick\_start.html*](https://facebook.github.io/prophet/docs/quick_start.html)

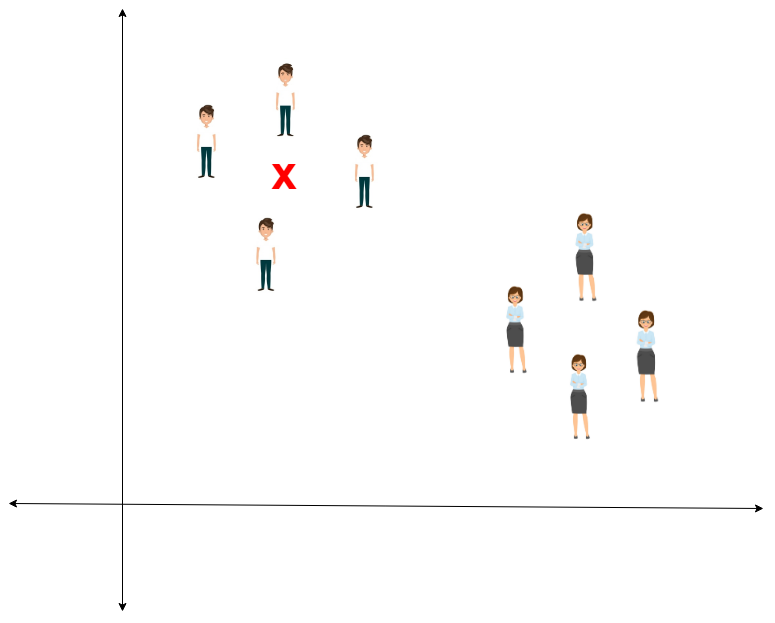
* + 1. **KNN Algorithms**

KNN is the abbreviation of K-Nearest Neighbors. Both regression and classification problems can be addressed with the help of KNN. But it is more widely used in classification problems in the industry. We generally look at 3 important things:

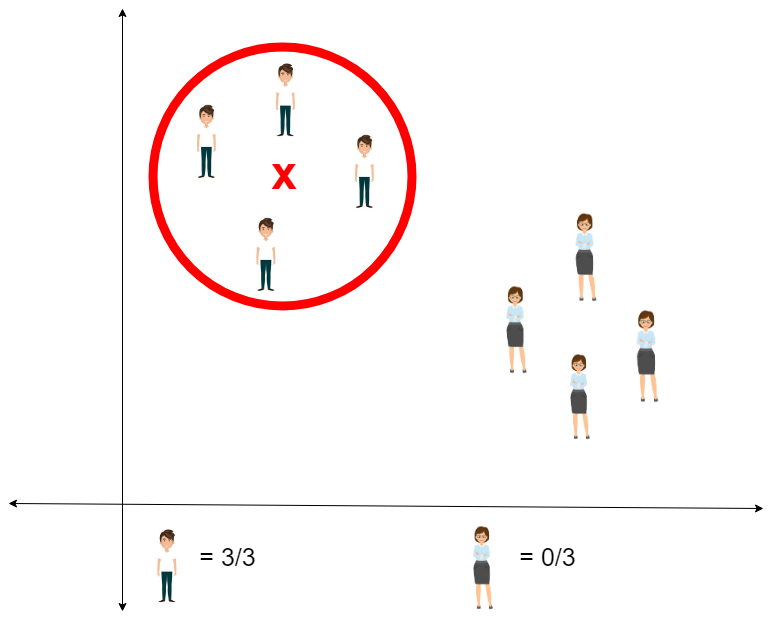
* Ease to interpret output
* Calculation time
* Predictive Power

*Real Life E.g.4: It’s a boy!*

Consider the example to understand this algorithm. Following is the data about Boys(B) and Girls(G).



You intend to find out the class of the Red **X**(X) . X can either be B or G and nothing else. The “K” is KNN algorithm is the nearest neighbors we wish to take vote from. Let’s say K = 3. Hence, we will now make a circle with X as center just as big as to enclose only three data points on the plane. Refer to following diagram for more details:



The three closest points (USING DISTANCE ESTIMATION) to X are all B. Hence, with good confidence level we can say that the X should belong to the class B. Here, the choice became obvious as all three votes from the closest neighbor went to B. The choice of the parameter K is very crucial in this algorithm.

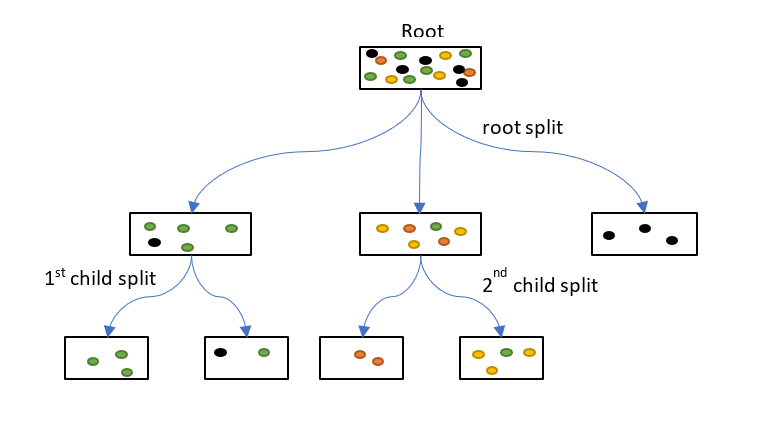
* + 1. **Decision Tree Algorithm**

Decision tree algorithm belongs to the family of supervised learning algorithms. This can also be used for both regression and classification.

The general motive of decision tree is to teach the machine which can be used to predict class or value of the target with the help of the training dataset.

The decision tree algorithm uses tree representation to solve the problem. The top most node is called the root node and each inner node of the tree corresponds to an attribute, and each leaf node corresponds to a class label. The leaf is the last portion of the decision tree just like a tree.

*Real Life E.g.5: Classify the marbles!*



As shown in the image above the decision tree begins the classification in the same way. The leaf nodes correspond to the class label. The first split is called the root split as it is the split of the root node and the remaining splits from the child nodes are the child splits.

We will look at the detailed workings in the upcoming chapters.

Unsupervised learning

*Present 4: Slide introducing the topic*

Unlike Supervised learning, the Unsupervised learning process involves the drilling of machine with data that is neither classified nor labeled and permitting the algorithm to perform based on data without guidance. Here the job of machine is to group un-clustered information with respect to likenesses, patterns, and variances without any erstwhile training of data.

The Unsupervised learning algorithm does not need a mentor to train it. The machine needs data to be fed and the machine takes care of the rest of the operation.

*Real Life E.g.4: Cars and Bikes!*

If the machine is fed with an image of Cars and bikes which the same is not aware of.



Nonetheless the machine can classify them according to their similarities, patterns, and differences i.e., the machine classifies the above picture into two. The former picture may have images of cars in it and the latter may contain all pictures with bikes in it. The inference from the above example is that the machine has the ability to classify among the object which it has not been trained with.

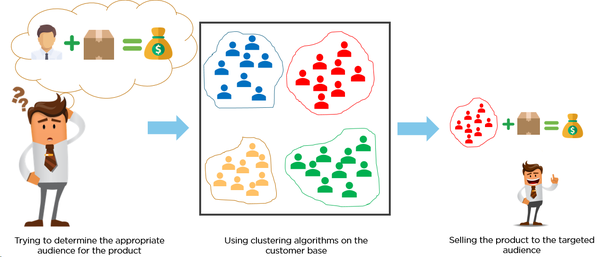
Unsupervised learning is broadly categorized into two:

* **Clustering:** A clustering procedure helps to discover the inherent patterns in the data, such as grouping customers by purchasing behavior and so on.

Visualize you have been given the job of selling a new product and bring huge profit. Now you need to inform your company about the target customer base, to whom if the product is marketed will bring the profit as expected.

Here Clustering lends a helping hand to find the target customers.

The data about the customers are given to you and now the customers are divided into small clusters and the algorithm selects the cluster with the target customers.



* **Association:** An association rule is a unique way to find patterns associated with a large amount of data, such as when someone buys product1 he also tends to buy product2.



**K-means Clustering**

K-Means Clustering is kind of unsupervised learning algorithm, which is used when we don’t have labelled data (i.e., data without defined categories or groups). The goal is to find the number of groups specified by ‘K’. The algorithm works in an iterative manner.

The K-means algorithm is composed of 3 steps:

* Initialization
* Cluster Assignment
* Reiterating the Centroid

Let us look at an example to understand more about the K-Means clustering algorithm.

*Real Life E.g.6: Buyers!*

An online retailer has collected a data of all the customer those have purchased products online from their site for the period of last one year. The retail company wanted to send promotional and offer emails to the customer based on their interest and buying potential. Looking at the historical data, data experts couldn’t gauge any insights using supervised learning. Hence, they begin to explore more and eventually apply unsupervised learning to find out a pattern of similarities in the data by which the customer can be grouped into 2 or more clusters each representing a cluster of “Offer buyers”, “Regular buyers” and “Seasonal buyers”.

Later, when a new customer enrolls, retailer could very well position him into any one of the clusters (offer buyers or regular buyers or seasonal buyers) and target emails or messages accordingly. Because of this clustering technique, the retailer could effectively channelize their operational/advertising cost.

**Reinforcement Learning**

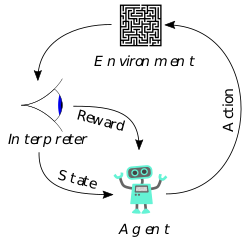
*Present 4: Slide introducing the topic*

Reinforcement learning is a broad area in Machine Learning where the machine learns to perform the next step in an environment by looking at the results to the ensuing actions performed.

The Reinforcement Learning does not have an answer and the reinforcement agent decides what should be done to perform the specified task. It learns from its knowledge.

This kind of learning involves both

Reward and Penalty. We will see more about how the reward acts as the central idea of the Reinforcement learning concept.



Let us consider a Robot as the Reinforcement learning agent and the prime job of the robot is to reach the end of the maze, which is our environment in this case. The robot must solve the maze without any prior knowledge about it.

When the robot travels in a progressive direction in a maze, a reward will be given and penalty will be nailed when the robot reaches the dead end. Depending on the knowledge of rewards and penalty the robot learns by itself the way to come out of the maze.

**Forms of Reinforcement:**

* **Positive:**

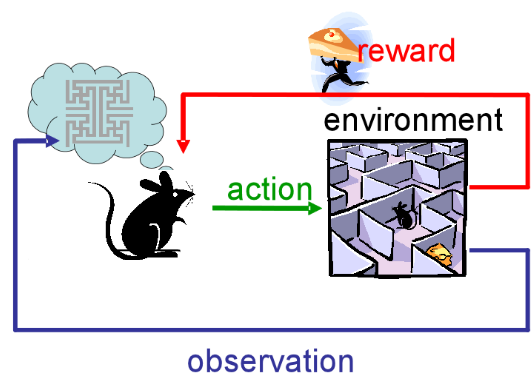
Positive Reinforcement is when an event, occurs due to a behavior, the strength increases along with the frequency of the behavior. In other words, it has a positive effect on the behavior which occurs due to a reward.

* **Negative:**

Negative Reinforcement is when an event, strengthens the behavior or actions when a negative condition is stopped or avoided. This occurs when a penalty is received.

*Real Life E.g.6: Mouse and the maze!*

Consider a mouse that is left in a maze. What will it do? I will stroll a stroll to each lane, and if it is a dead lane, it would try another lane. If it completes the maze it will get a cheese cube as a reward.

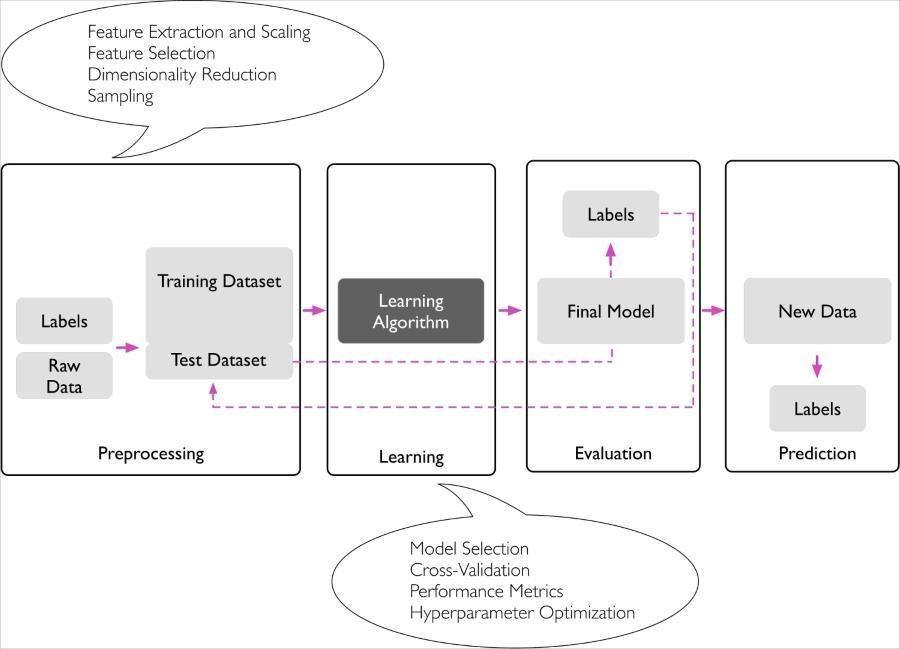


So, it takes action on the environment based on the observation of the maze and the reward it gets. If it is a positive reward it continues its actions and if it is a negative reward (penalty), it corrects its actions.

Until it achieves its rewards its actions are reiterative where each time it learns its way to achieve the rewards through experience.

**Roadmap for building ML models**

*Present 4: Slide introducing the topic*



Here the roadmap for building Machine Learning models are pretty straight forward and consists of 4 steps:

* Pre-processing
* Model Learning (will differ for each learning types)
* Model Evaluation
* Prediction

**Pre-processing**

This is the first and foremost step in building a Machine Learning model. Data preprocessing refers to the transformation of data before feeding it into the model. It deals with the techniques that are used to convert the raw – unusable data into clean reliable data.

The Raw data along with or without the labels cannot be fed into the model as such as that might compromise with the quality of the results. The different types of preprocessing steps are:

* Feature Extraction and Scaling
* Feature Selection
* Dimensionality Reduction
* Feature Engineering
* Feature Sampling
* Encoding

And so on. After we perform Pre-processing, we split the data into train, test and cross-validation sets. We then move on to the model building.

**Learning**

This is the part where the models are trained with the train data set. The machine identifies the underlying patterns. The Learning algorithms will also have their parameters that needs to be initialized. These parameters are those that will determine the quality of the results (of course the quality of data comes first). These parameters are called as “Hyper Parameters”. Once we select the appropriate model and initialize it we will need to perform multiple cross validations with the validation/test sets.

**Evaluation**

These Cross-validations will be measured with the help of specific performance metrics. With the results of the cross-validations and measure of the Performance metrics we will go on to tune the Hyper-Parameters (this process is called as Optimization). We will do this step until we as satisfied with either the increased or decreased performance metric.

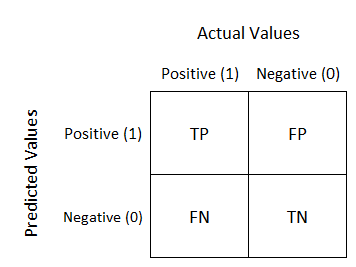
**Performance metrics**

There are different evaluation metrics in Machine Learning, And theses depend on the type of the data and the requirement. The different types of Metrics are:

* Confusion Matrix
* Precision
* Recall
* Accuracy
* F1 – Score
* Log loss
* AUC (Area Under the Curve)
* MAE (Mean Absolute Error)
* MSE (Mean Squared Error)

**Confusion Matrix**

A confusion Matrix is a table that is used to define the performance of the classification model on the test data for which the actual values are known.



Definition of Terms in the confusion matrix:

* Positive: The records are positive or true
* Negative: The records are Negative or False
* TP – True Positive: The record is positive, and it is also being predicted by the classifier as Positive
* FN – False Negative: The record is positive, but it is being predicted as Negative
* TN – True Negative: The record is Negative, and it is also predicted as negative by the classifier
* FP – False Positive: The record is Negative, but it is predicted as positive

**Precision**

It is the ratio of True positive to the total positive. The precision looks at how precise we have built our model. How much is really positive from the totally predicted positive.

**Precision** = =

**Recall**

It calculates how much of the total positive has our model captured correctly out of the total positive.

**Recall** = **=**

**Accuracy**

**Accuracy** =

Or in other words,

**Accuracy** = .

**F1-score**

Accuracy focuses or targets largely by True Negatives which will not be of much importance in the business case scenarios, here we need to focus more on False Negative and False Positive which might bear some business costs, thus F1 score is much better to use where we can seek a balance between Precision and Recall and when there is an uneven class(large number of Actual Negatives).

**F1-Score** =

**F1 Score is the measure of the test’s Accuracy**

**Log-Loss**

**Log-Loss** =

Here the Log – Loss function is mainly used for the classification models where we build the model and we change the estimates of the model to reduce the Log-Loss function.

Where,

– The output of the regression Equation

– The Probability threshold of calculation

**AUC (Area Under the Curve)**

This is also used for binary classification problem. Where it is a plot between the True positive rate and the False positive rate

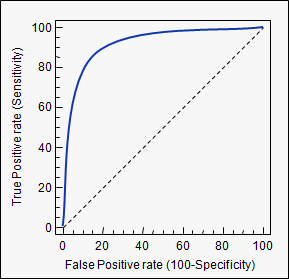
**True Positive Rate (Sensitivity):** True Positive rate refers to points correctly predicted as positive with respect to all positive. If you notice it right this is also called as Recall.

**True Positive Rate** = **=**

**False Positive Rate (Specificity):** False positive Rate is the set of negative points that are faultily predicted as positive, with respect to all negative points.

**False Positive Rate** = **=**

AUC curve is the area under the curve of the plot Sensitivity vs Specificity.



**Mean Absolute Error**

When we find the average of absolute differences between the predicted and the actual value over the set of test sample, then it is called as the Mean Absolute Error.

**MAE** =

**Mean Squared Error**

When we find the average of squared difference between the predicted value and actual value over a set of test sample, then it is called as the Mean Square Error.

**MSE** =

**Prediction**

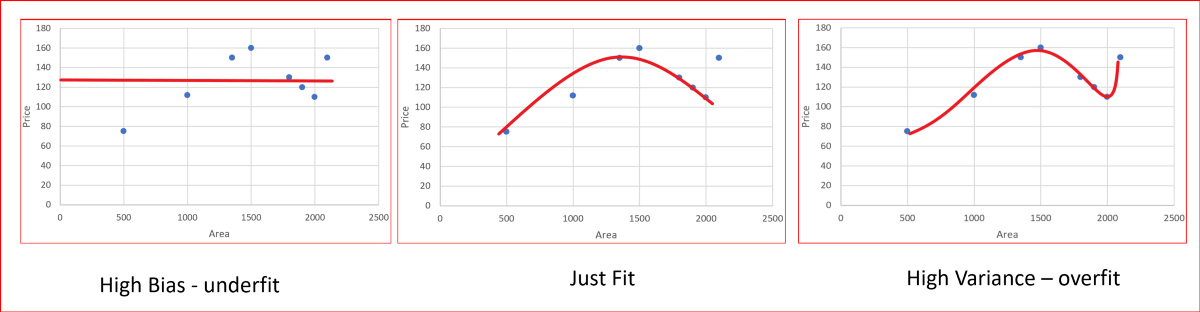
Once we get the near to ideal or satisfying results from the evaluation step, we will then move on to do the predictions. These predictions will be made in the labels by the trained model with the new data set. These predictions are final and will be taken over to the decision makers to make effective decisions. Here we will have to understand another important concept of prediction error (bias and variance).

**What is bias?**

Bias is how far are the predicted value from the actual value. High bias means the model is very simple and is not capable enough to capture the data complexity and is thus called **Underfitting.**

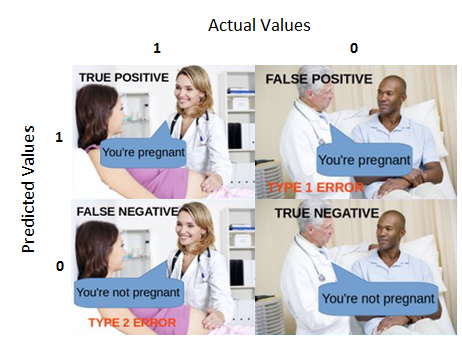
**What is variance?**

High variance is when the model performs very well on the trained dataset and this will tell us how the predicted values are scattered from the actual values. This causes **Overfitting** and makes the model so much specific to the train data and not good to perform on test data.



*Real Life E.g.7: Congratulations you are pregnant!*

Let us understand the concept of Confusion matrix and its metrics TP, TN, FP, FN in detail. Consider you are building a model that predicts pregnancy.



TP (True Positive):

Say Gender is Female and she is actually ‘Pregnant’ and your model also predicted ‘True’.

FP (False Positive):

Say Gender is Male and your model predicted ‘True’. Which cannot happen and this gives rise to a type of error called ‘Type 1’ error.

FN (False Negative):

Say Gender is Female and she is actually pregnant and the model predicts ‘False’. Which is also an error and this is called ‘Type 2’ error.

TN (True Negative):

Say Gender is Male and prediction is also ‘False’ then that is True Negative.

Here if you can see the Type 1 error is a blunder which is more dangerous than Type 2 error.

**Why Python for Machine Learning?**

**A close up of a map

Description automatically generated**

Image Credits: [Stack Overflow](https://stackoverflow.blog/2017/09/06/incredible-growth-python/)

A close up of a map

Description automatically generated

Image Credits: [Stack Overflow](https://insights.stackoverflow.com/survey/2017#technology)

Python is most wanted Language as per StackOverflow developer survey 2017

[19 principle of python from Wikipedia](https://en.wikipedia.org/wiki/Zen_of_Python)

This can be displayed on your own python terminal

**Exercise 1:**

**Displaying the principal of python from Command-Prompt (Windows)/Terminal(Linux or Windows)**

**Note**

Assuming python is installed. if not, I recommend to install Anaconda which installs python jupyter notebook from this [link](https://www.anaconda.com/distribution/#download-section)

1. Open a command prompt if Windows user or terminal if Mac/Linux
2. Type python and press Enter

A screenshot of a cell phone

Description automatically generated

1. Type import this and press Enter

**Note**

**Import** is used to access to the code from pre-written module.

A screenshot of a cell phone

Description automatically generated

Python is simple, easy to learn, better to read, portable, scripting tool, web development framework and has lots of library pre written for Data Science purpose.

Machine learning deals with data, to find pattern of it. Task involved in a daily life of Data Scientist is to extract the data from the source, apply the necessary processing, which in turns cleaning, making up the data for inputting it to the intelligent and sophisticated algorithms.

Data is fuel for current generation. It is available everywhere in a raw format with not structured, not good and not complete.

Now with this problem persisting. How python is going to help us?

Yes, Python has a ready-made solution in terms of pre written code called **Packages.** Still open-source, easy to use and lot of support in terms of forums.

For dealing with data – pandas, matplotlib, seaborn, …

For dealing with text – nltk, spacy, gensim

For dealing with numbers – numpy, scipy

For dealing with audio – librosa

For dealing with machine learning – scikit-learn

For dealing with deep learning – tensorflow, pytorch, fast.ai

For dealing with web development – flask, django

And many more…

Thus, with all those ready-made libraries with easy implementation python has got lot of interest and attention among developers especially with data science space.

**Why Pandas?**

A close up of a map

Description automatically generated

Image Credits: Stack Overflow

Pandas is most discussed topic in stack overflow

Pandas is an open-source package. It has lot functions to load and process the data ready for machine learning. It has lot of tools to analyze and manipulate the data. It will represent the data in a DataFrame data structures. Please visit the [link](https://pandas.pydata.org/pandas-docs/stable/) to learn more about pandas.

**Note**

Pandas is pre-installed in Anaconda python flavour. If not, to install pandas go to command-prompt/terminal and enter **pip install pandas**

Once install go to the python terminal and type command **import pandas** to check if it is installed.

Data can be read from many formats using pandas. We will be using csv for our data reads throughout our chapters.

A screenshot of a cell phone

Description automatically generated

Image Credits: [Pandas Documentation](https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html)

To read the data, use read\_csv function by passing the “filename.csv” as an argument.

Example:

A black sign with white text

Description automatically generated

pd is an alias name given to pandas and it is not mandatory to give alias.

To visualize the pandas dataframe. use a function head() to list down top 5 rows.

With this we will stop the discussion and move to the next chapter on making the hands dirty with data.

**Activity 1: Load the Customer Segments Dataset using pandas**

The [Customer Segments Dataset](https://archive.ics.uci.edu/ml/datasets/Wholesale+customers) of clients of a wholesale distributor. It includes the annual spending in monetary units on diverse product categories. The column details are given below.

**Attribute Information:**

* + - 1. FRESH: annual spending (m.u.) on fresh products (Continuous);
      2. MILK: annual spending (m.u.) on milk products (Continuous);
      3. GROCERY: annual spending (m.u.)on grocery products (Continuous);
      4. FROZEN: annual spending (m.u.)on frozen products (Continuous)
      5. DETERGENTS\_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)
      6. DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous);
      7. CHANNEL: customersâ€™ Channel - Horeca (Hotel/Restaurant/CafÃ©) or Retail channel (Nominal)
      8. REGION: customersâ€™ Region â€“ Lisnon, Oporto or Other (Nominal)

Now perform the following steps:

1. Import the pandas library with an alias name pd
2. Load the Data using pandas read\_csv() function
3. Display the first 5 rows of the dataframe

**Solution for Activity 1:**

1. Import the pandas with an alias name pd

import pandas as pd

1. Load the data using read\_csv()

df = pd.read\_csv("Wholesale customers data.csv")

1. Display the first 5 rows of the dataframe

df.head()

A screenshot of a cell phone

Description automatically generated