What Machine Learning does?

*Do you want to****predict a category****?*

* Machine Learning has **Classification**

Example

Predict if the stock price will increase or decrease.

*Do you want to****predict a quantity****?*

* Machine Learning has **Regression**

Example

Predict the age of a person based on their height, weight, and health factors.

What Machine Learning does?

*Do you want to****detect an anomaly****?*

* Machine Learning has **Anomaly Detection**

Example

Money withdrawal anomalies can be discovered.

*Do you want to****discover structure in unexplored data****?*

* Machine Learning has **Clustering**

Example

Finding a group of customers with similar behavior based on their buying data history.

*While a great deal of engrossment has been towards model building, model tuning, and model evaluation, many individuals still find themselves asking basic inquisitive questions like*

***What is the life cycle of Machine Learning?***

*This section of the course will aid in answering this question. Keep reading to know more!*

Big Picture

The big picture of Machine Learning process lies in the following 9 steps namely

1. Defining Project Objectives
2. Gathering Data
3. Exploratory Data Analysis(EDA) and Data Cleaning
4. Choosing a Model
5. Training
6. Evaluation
7. Hyper parameter Tuning
8. Interpret and Communicate
9. Deployment and Documentation

Defining Project Objectives

* The first step of the life cycle is to recognize the opportunity for tangible improvement of activities, to enhance customer satisfaction, or to create value otherwise.
* It is critical that you understand the problem you are trying to solve. In this stage, you should also be identifying the central objectives of your project by identifying the variables that need to be predicted.

Gathering Data

* Considered to be the primary step of Machine Learning process.
* The quality and quantity of data you gather in this step will determine how efficient your model will be.

Some important things to remember while gathering data:

* Data can be collected from anywhere in any format.
* More training examples will aid the model to be more efficient.
* Make sure the number of samples for every class or topic is not overly imbalanced.
* Make sure that your samples adequately cover the space of possible inputs, not only the common cases.

EDA and Data Cleaning

**EDA**

* Analyzing datasets to summarize their notable characteristics is called Exploratory Data Analysis.
* Helps in performing investigations on data so as to discover hidden patterns, anomalies etc.
* Aids in checking assumptions and hypothesis with the help of summary statistics.

**Data Cleaning**

* Data can have several shortcomings. To list a few are
  1. *Missing values*
  2. *Duplicate data*
  3. *Invalid data*
* The process of detecting, correcting and ensuring that the given dataset is error free, consistent enough to use is called Data Cleaning.

Choosing a Model

* There are numerous models that researchers and Data scientists have created over the years.
* Some are very well-suited for image data, while others are suited for sequences, text-based data and many more.
* Choosing the right model for the problem will impact the efficiency of the model.

*Explore this video to know the different constraints for choosing different models.*

<https://youtube.videoken.com/embed/_3eaVy8c-xk>

Training

* The next step of the Machine Learning process, often known as the the bulk of ML is **Training the model.**
* This step is very similar to a person who is learning to drive for the first time. Though at first they dont know any of pedals, switches, breaks but eventually after lots of practice and feedbacks a licensed driver emerges.
* The data is split into Training Data and Testing Data.
* Model is trained with the training data using different ML algorithms by adjusting the parameters in multiple iterations.
* Testing Data are put aside as unseen data to evaluate your models.

Evaluation

* Once training is complete, it’s time to see if the model is any good, using Evaluation.
* This is where that dataset that we set aside earlier comes into play(i.e) Testing Data.
* Evaluation allows us to test our model against data that has never been used for training.
* This metric allows us to see how the model might perform against data that it has not yet seen.
* This is meant to be representative of how the model might perform in the real world.

Hyperparameter tuning

* After the evaluation step, it's time to see if we can improve our training furthermore by tuning different parameters that were implicitly assumed in the training process and this process is called Hyperparameter Tuning.
* The tuned model is once again evaluated for model performance, and this cycle continues until the final best performing model is chosen.

Interpret and Communicate

* The most challenging task of the ML project is explaining the model's output.
* Earlier days, Machine learning is considered to be a BlackBox because it was hard to interpret their insights and values.
* The more interpretable your model is, then more it is easier to communicate your model's importance to the stakeholders

Deployment and Documentation

* Model deployment often poses a problem because of the coding and data science experience it requires and because the time-to-implementation of traditional data science methods from the start of the cycle is prohibitively long.
* The trained model has to be deployed in a real-world system for it to be efficient to humans.
* It can be deployed using any of the frameworks like FLASK, Cloud, Azure etc.
* Document your project well for your descendants to handle it.

Types of Machine Learning

The types of Machine Learning are as follows:

* Supervised Learning
* Unsupervised Learning
* Reinforcement Learning

Supervised Learning

* Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.
* It infers a function from labeled training data.
* Each training example is a pair consisting of an input object and a desired output value.
* A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples.

**Applications**

1. Spam Detection
2. Pattern Recognition
3. Speech Recognition

Unsupervised Learning

**Unsupervised Learning** helps in uncovering hidden patterns from unlabeled data.

**Applications**

1. Recommender Systems
2. Targetted Marketing
3. Customer Segmentation
4. Structure Discovery

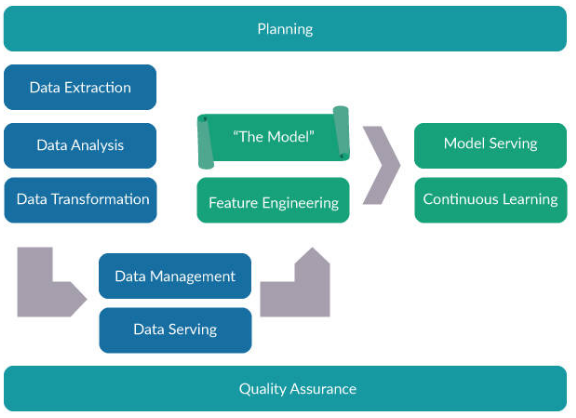
Reinforcement Learning

**Reinforcement Learning** is a type of machine learning in which software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.

**Applications**

1. Genetics
2. Economics
3. Robot Navigation

##### Machine Learning in SDLC

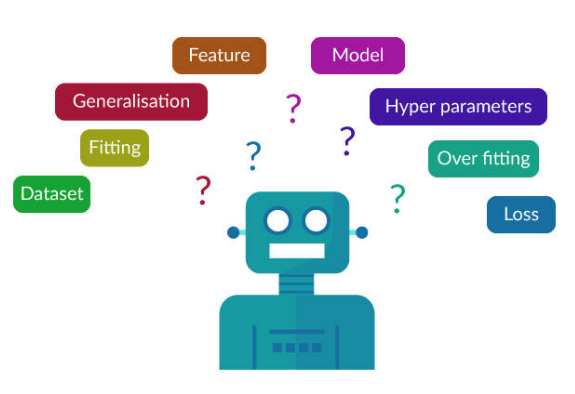


*The image depicted above illustrates how to integrate the process of Machine Learning into the traditional Software Development Life Cycle (SDLC).*

The three phases include:

1. Planning
2. Data Engineering
3. Modeling

##### Prelude



Confused about the jargons and terms in Machine Learning? This section is here to help you.

Few **key terminologies** to be known while using Machine Learning model are discussed in this section of the course.

##### Machine Learning Terminologies

**Accuracy**

Accuracy is the percentage of correct predictions made by the model.

**Algorithm**

* Machine learning algorithms are programs (math and logic) that adjust themselves to perform better as they are exposed to more data.
* The learning part of machine learning means that those programs change how they process data over time, much as humans change how they process data by learning.
* So a machine-learning algorithm is a program with a specific way to adjusting its own parameters, given feedback on its previous performance making predictions about a dataset.

Examples

* Linear regression
* Decision trees
* Support vector machines
* Neural networks

##### Machine Learning Terminologies

**Categorical Variables**

* Categorical variables are variables with a discrete set of possible values.
* They can be ordinal or nominal.

**Classification**

Classification aids in predicting the categorical output.

**Clustering**

Clustering is the unsupervised grouping of data into buckets.

Machine Learning Terminologies

**Dimension**

The dimension of data denotes the number of features in a dataset.

**Feature**

For a dataset, a feature represents the combination of attribute and value.

**Feature Selection**

Feature selection is the process of selecting relevant features from a dataset for creating a Machine Learning model.

**Hyperparameters**

Hyperparameters are higher-level properties of a model, such as how fast it can learn or the complexity of a model.

**Instance**

An instance is a data point, row, or sample in a dataset.

**Label**

The label is the answer part of the observation in supervised learning.

**Outlier**

An outlier is an observation that deviates significantly from other observations in the dataset.

**Regression**

Regression predicts the continuous form of output (For example, price, sales, and so on).

**Validation Set**

The validation set is a set of observations used during model training to provide feedback on how well the current parameters generalize beyond the training set.

*Let us now explore the following****popular Machine Learning techniques****:*

* Classification
* Clustering
* Association Rule Mining
* Outlier Detection
* Regression

##### Classification

#### Definition

Classification is the process of identifying a category to which a new observation belongs, based on a training set of data containing observations *whose categories are already known*.

* It follows a two-step process, namely:
  + **Learning Step** - Training phase where a model is constructed.
  + **Classification Step** - Predicting the class labels and testing the same for accuracy.
* Classification predicts the value of the categorical variables.

##### Clustering

Clustering is the task of *grouping a set of objects*, such that objects in the same cluster are similar to each other when compared to the objects in the other clusters.

* Distance measure plays a significant role in clustering.
* Clustering is an unsupervised learning method.
* The common distance measures used in various datasets are as follows.

Numeric Dataset

- Manhattan distance

- Minkowski distance

- Hamming distance

Non-Numeric Dataset

- Jaccard index

- Cosine Similarity

- Dice Coefficient

##### Association Rule Mining

##### 

Association Rule Mining aids in identifying the associations, correlations, and frequent patterns in data.

The derived relationships are represented in the form of Association Rules.

##### Association Rule Mining with Apriori

[**https://youtube.videoken.com/embed/guVvtZ7ZClw**](https://youtube.videoken.com/embed/guVvtZ7ZClw)

##### Outlier Detection

***Jiawei Han*** defines ***Outlier*** as

A data object that *deviates significantly from the normal objects*as if it were generated by a different mechanism.

The **types** of Outlier are as follows:

* **Global Outlier**

Global Outlier significantly deviates from the entire dataset.

* **Contextual Outlier**

Contextual Outlier significantly deviates based on the context selected.

* **Collective Outlier**

Collective Outlier is a subset of data objects that collectively deviates from the entire dataset.

##### Regression

* Regression analysis is a statistical method that aids in examining the relationship between two or more variables of interest.
* Examines the influence of **one or more independent variables** on a **dependent variable**.

##### Decision Tree

* A Decision Tree (DT) is a tree-like model of decisions and possible consequences, chance event outcomes, resource costs, and utility.
* Decision Trees are a non-parametric supervised learning method used for classification and regression.

Watch this video to know more.

<https://youtube.videoken.com/embed/7VeUPuFGJHk>

##### Naive Bayes

A Naive Bayes classifier is a probabilistic Machine Learning model that is used for classification tasks. The crux of the classifier is based on the following Bayes theorem formula.

P(A|B)=\dfrac{P(B|A)P(A)}{P(B)}*P*(*A*∣*B*)=​*P*(*B*)​​*P*(*B*∣*A*)*P*(*A*)​​

<https://youtube.videoken.com/embed/CPqOCI0ahss>

##### Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm. It is used for classification or regression type of problems.

<https://youtube.videoken.com/embed/Y6RRHw9uN9o>

##### K-means Clustering

Delve into this video to know about a type of clustering algorithm called K-means Clustering.

<https://youtube.videoken.com/embed/aiJ8II94qck>

##### Random Forest

Know more about the Random Forest algorithm through this video.

<https://youtube.videoken.com/embed/D_2LkhMJcfY>

##### Linear Regression

[**https://youtube.videoken.com/embed/zPG4NjIkCjc**](https://youtube.videoken.com/embed/zPG4NjIkCjc)

##### Logistic Regression

[**https://youtube.videoken.com/embed/yIYKR4sgzI8**](https://youtube.videoken.com/embed/yIYKR4sgzI8)