

PROGRAMMING FOR MACHINE LEARNING

Lecture 2

Mr. Doorgesh Sookarah

University of Technology, Mauritius



CONTENT

- Lecture 1: Revision
- Approaching a Machine Learning problem
- Steps involved Machine Learning

FROM LECTURE 1:



- **We provided the below definition for Artificial Intelligence.**
 - Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions.
- **There are five (5) components of Intelligence and they are as follows:**
 - Learning
 - Reasoning
 - Problem Solving
 - Perception
 - Language Understanding

FROM LECTURE 1:



- **We discussed intelligent systems.**
 - A system is artificially intelligent when it is equipped with at least one and at most of all forms of intelligence.
- **AI Systems can be classified either based on their capabilities or their functionalities.**
 - **Capability based AI** includes Narrow AI (ANI), General AI (AGI) and Super AI (ASI).
 - **Functionality based AI** includes Reactive AI, Limited Memory AI, Theory of Mind AI and Self Aware AI.

FROM LECTURE 1:

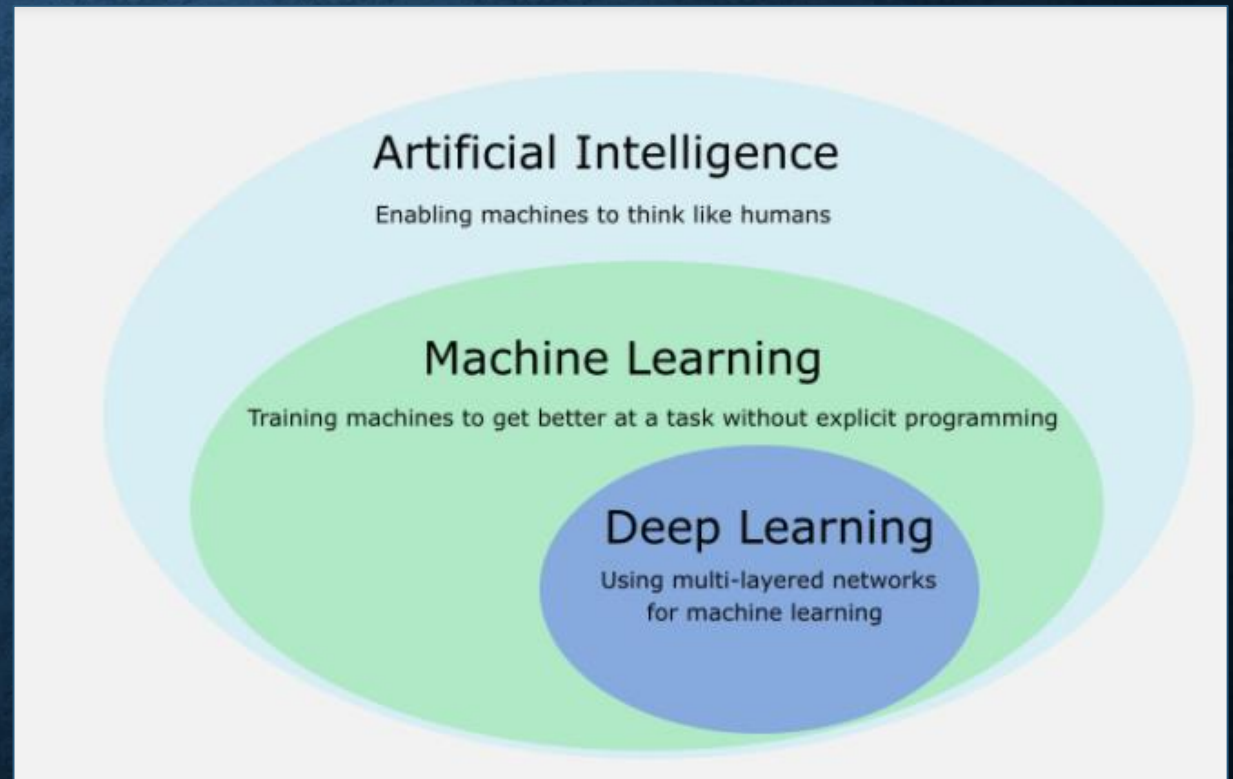


- **We provided the below definitions for Machine Learning.**
 - Machine learning is the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data.
 - Machine Learning enables a computer to automatically learn from data, improve its performance from past experiences and predict things without being programmed to do so.
- **We discussed the relationship between Artificial Intelligence, Machine Learning and Deep Learning.**

FROM LECTURE 1:



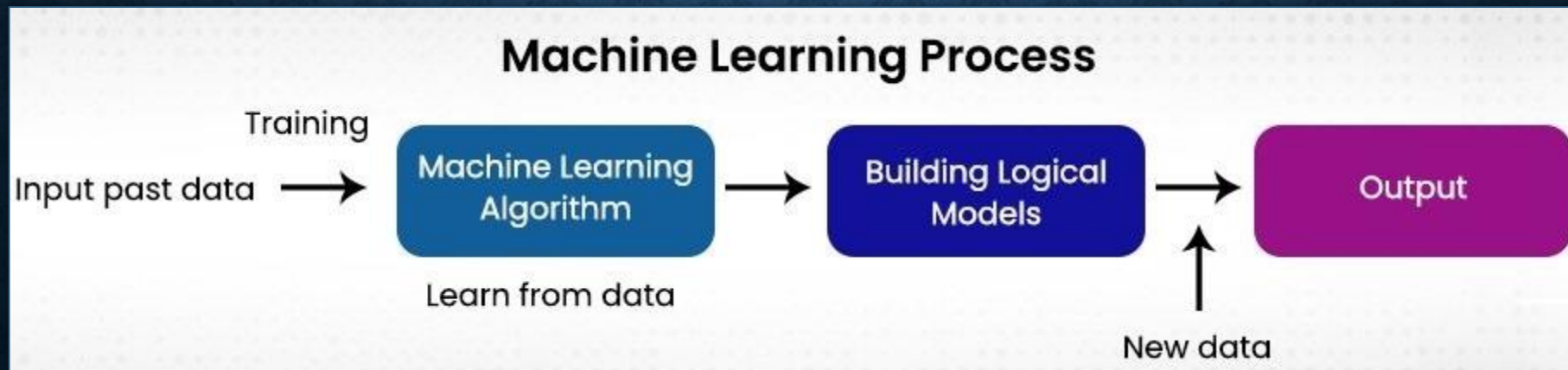
- AI is the ability of a machine to imitate human intelligence.
- ML uses algorithms to incorporate intelligence into machines by automatically learning from data.
- DL calls upon algorithms that mimic the human brain to incorporate intelligence into machines.



FROM LECTURE 1:



- **High Level Machine Learning Process**



TYPES OF MACHINE LEARNING

- **Supervised Learning**

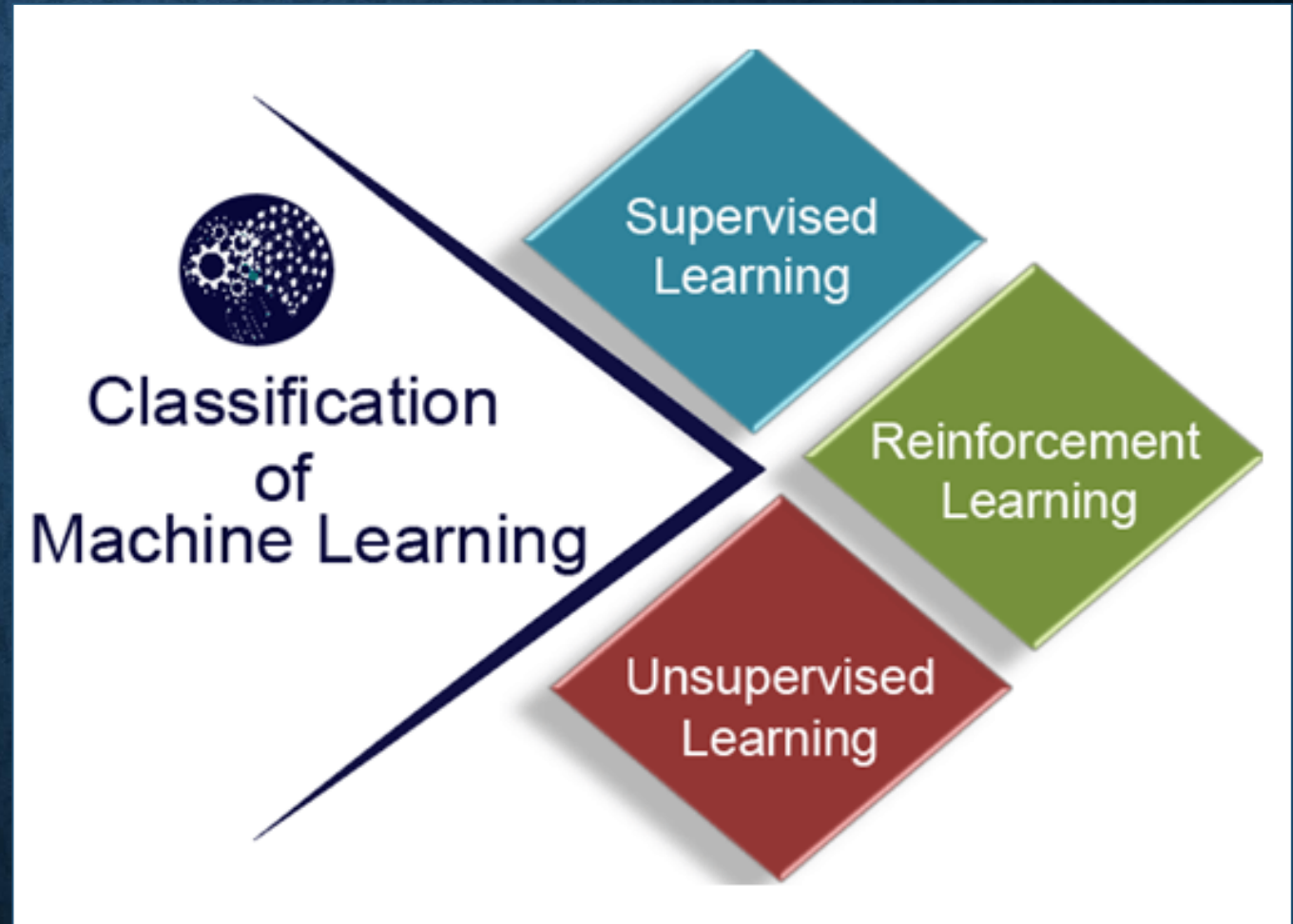
- Task driven
- Includes classification & Regression

- **Unsupervised Learning**

- Data Driven
- Includes Clustering

- **Reinforcement Learning**

- Focuses on learning from mistakes
- Commonly used in game simulation



SUPERVISED LEARNING

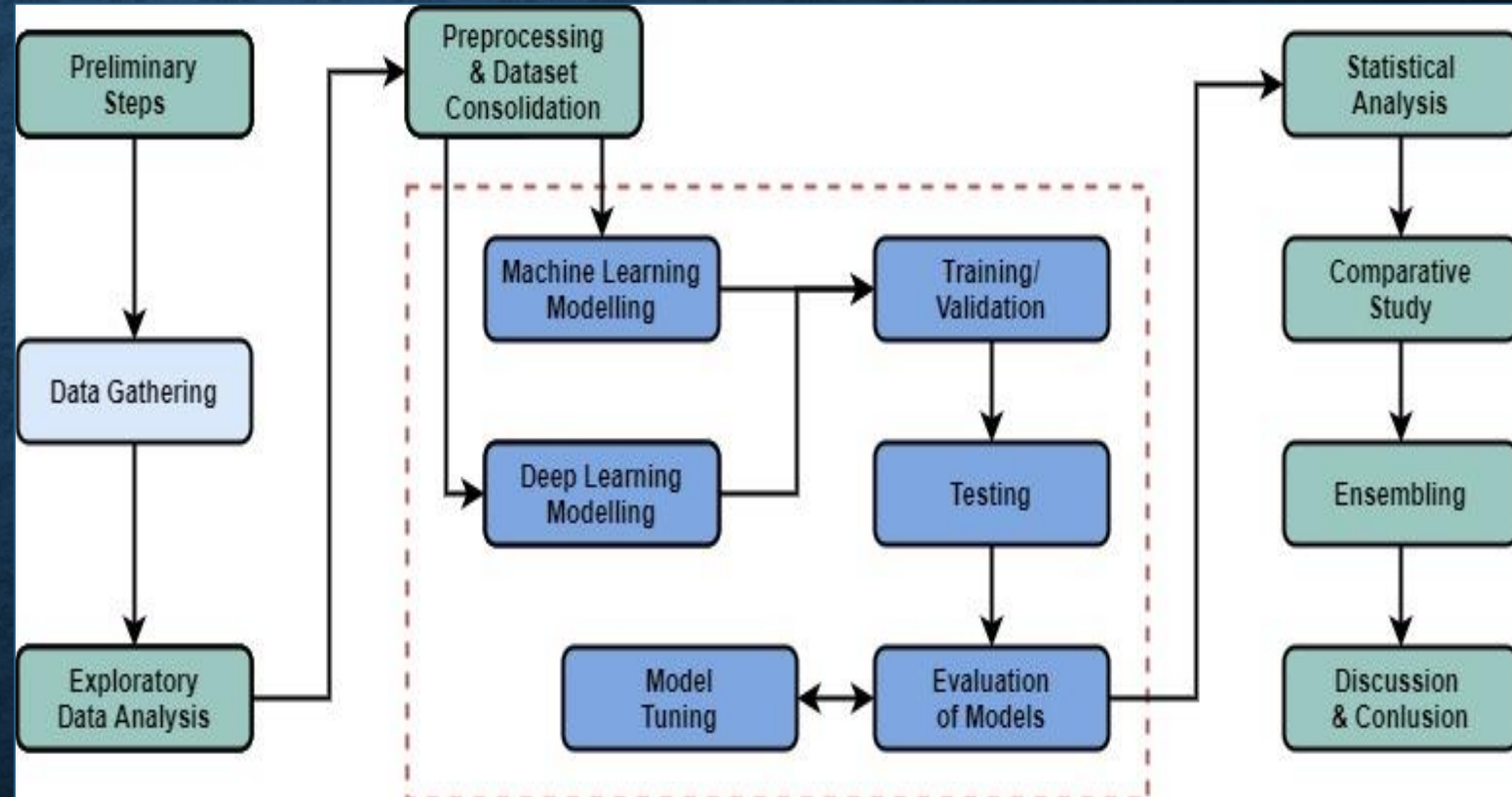
- In this module we will focus on the **Supervised Learning** method.
- When using supervised learning, we provide sample labelled data to the machine learning algorithm in order to train it.
- In turn, the machine learning algorithm will be basing itself on the results of its training to make predictions.
- The algorithm then creates a logical model using labelled data to understand the datasets and learn about each data.
- Once the training is completed, the model is tested to check if it is predicting the correct output or not.
- The goal of supervised learning is to map input data with the output data.

SUPERVISED LEARNING SUMMARISED

- Supervised Learning uses a training set to teach models to yield the desired output.
- The training dataset includes inputs and correct outputs, which allow the model to learn over time.
- The algorithm measures its accuracy through a performance metric such as the loss function.
- It will then be making adjustments until the prediction error has been sufficiently minimised.
- Note: Loss Functions measure how far an estimated value is from its true value.

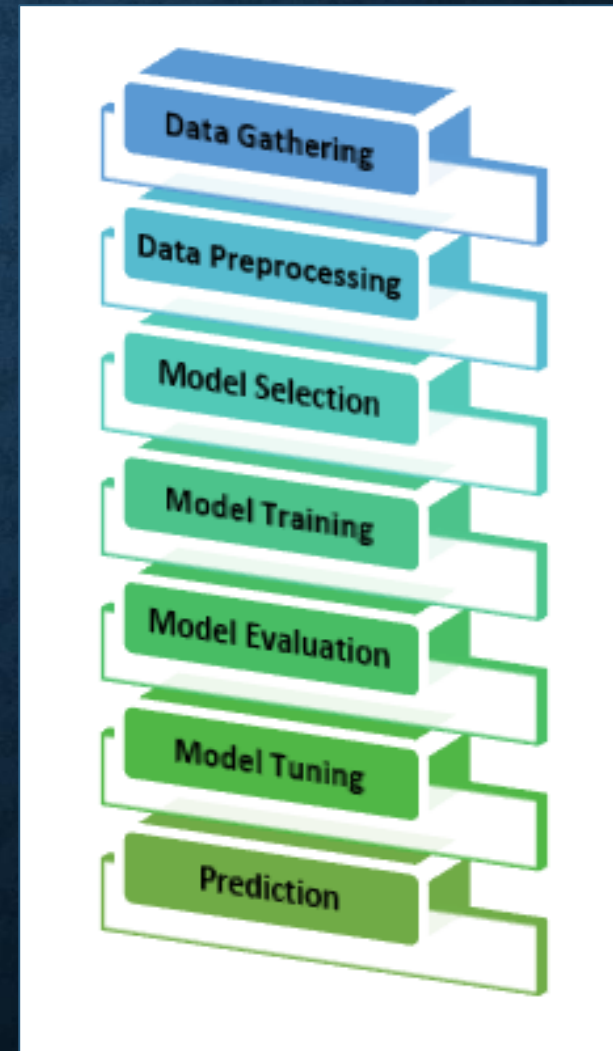
APPROACHING A MACHINE LEARNING PROBLEM?

- ML projects can only be implemented once a structured and logical plan has been elaborated.
- Defining methodology is therefore an essential step towards understanding and tackling the ML problem which is being investigated.
- The flowchart shown provides an example of an acceptable methodology for an ML project.



THE SEVEN STEPS IN MACHINE LEARNING

- Machine learning can be broadly decomposed into seven fundamental steps.
- Each of these steps need to be completed in a sequential manner to ensure the best outcome when deploying the machine learning model in live settings.



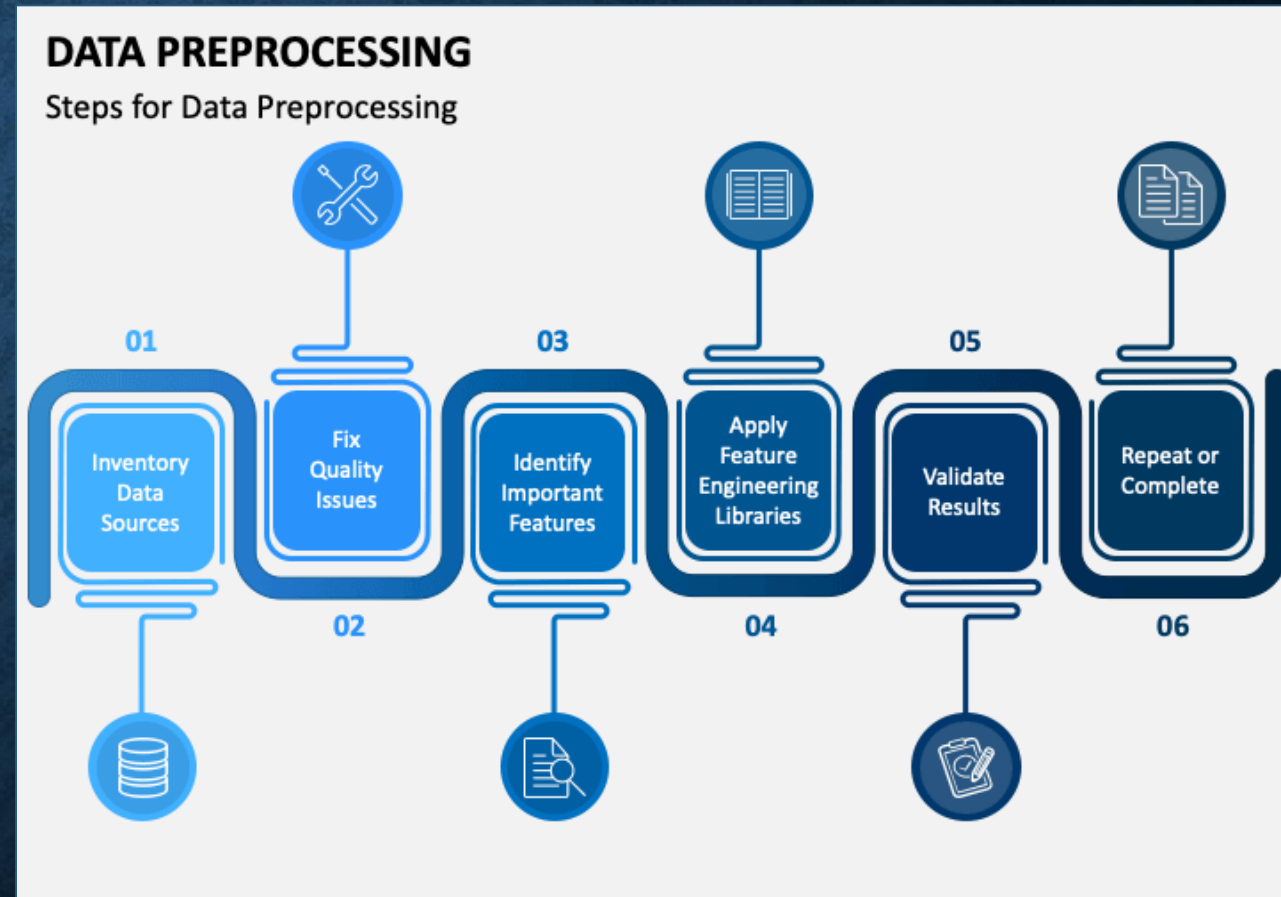
STEP 1: DATA GATHERING

- There is no way for a data science or machine learning project to move forward without first acquiring relevant data.
- The size of the dataset and the number of features it holds can directly affect the performance of the model
- In some instances, data availability may be limited for legal reasons or because they have not been released by the owners



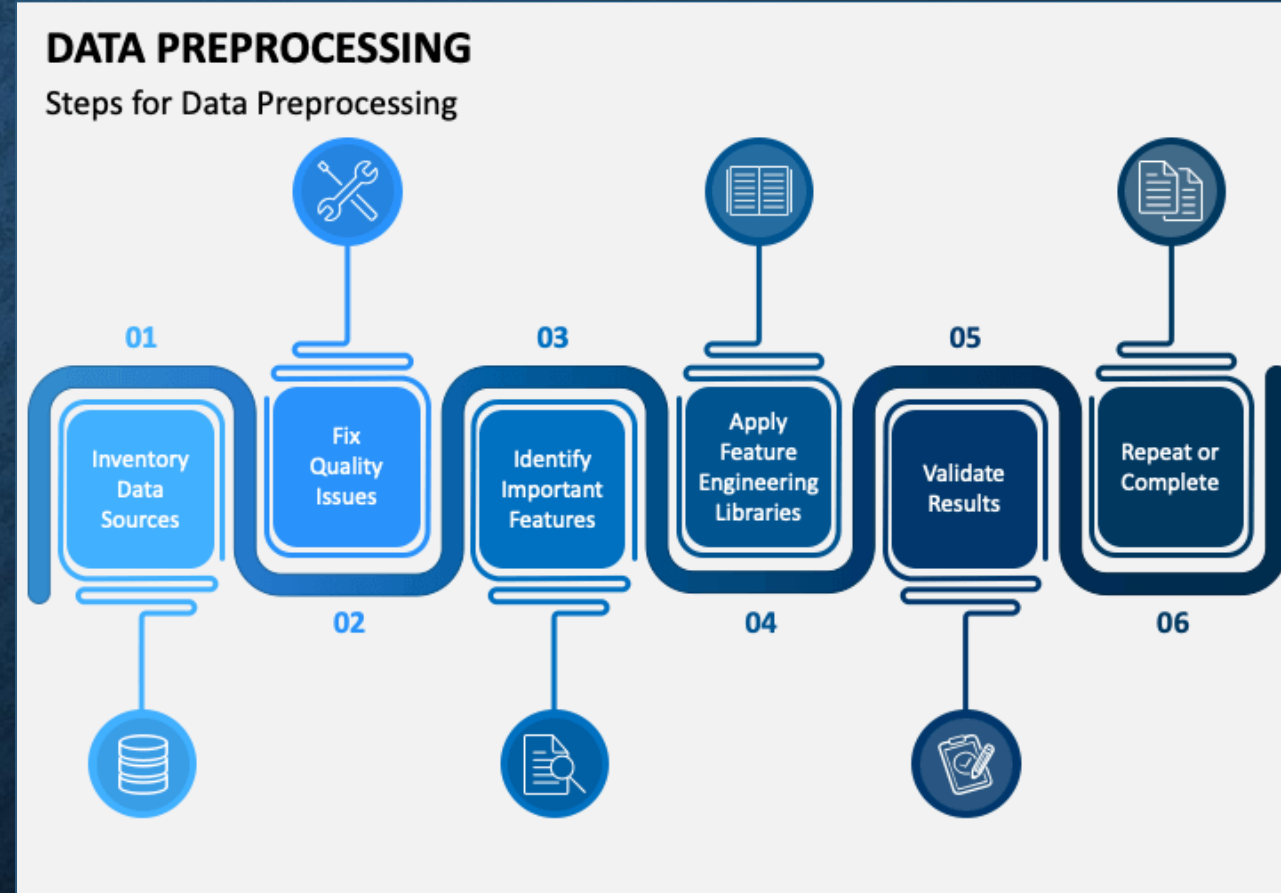
STEP 2: DATA PRE-PROCESSING

- A procedure called the exploratory data analysis (EDA) needs to be carried out once a usable dataset has been consolidated or identified.
- The EDA involves multiple statistical analysis such as measures of central tendency; visualisation of the data through multiple plots, charts and diagrams; and correlation analysis for every feature present in the dataset.
- This enables the identification of patterns in the dataset and can also provide insights on the importance of original or generated features



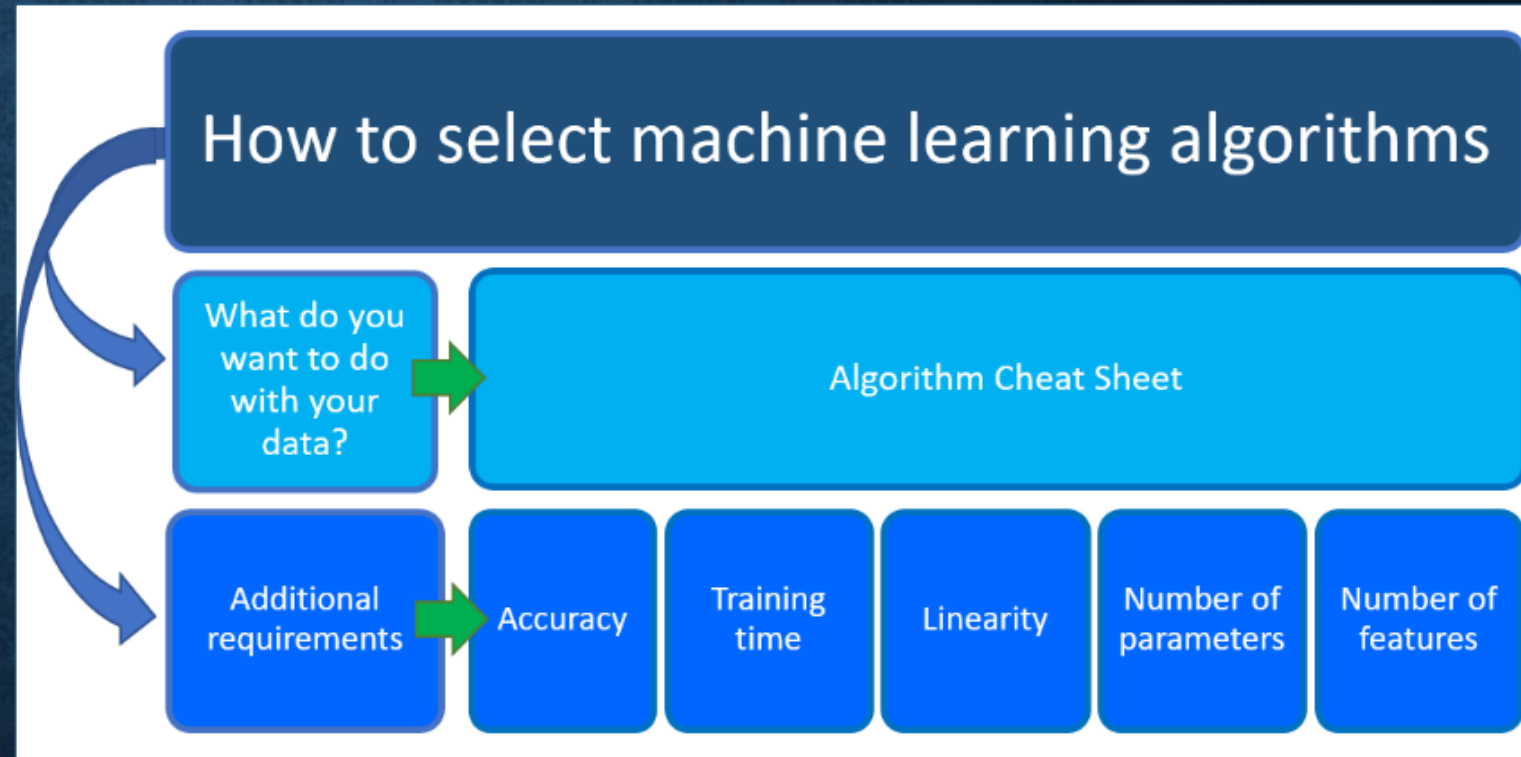
STEP 2: DATA PRE-PROCESSING

- The data is processed in such a way that it no more contains any class imbalances, missing attributes, duplicates and overall, this normally helps in reducing the execution time of code.
- The data pre-processing also includes normalization and splitting of the dataset into smaller segments for training, validation and testing.



STEP 3: MODEL SELECTION

- There should be a clear definition of the problem and the nature of the required predictions.
- Some Machine Learning problems are better investigated as a classification task while others are more suited as regression tasks.



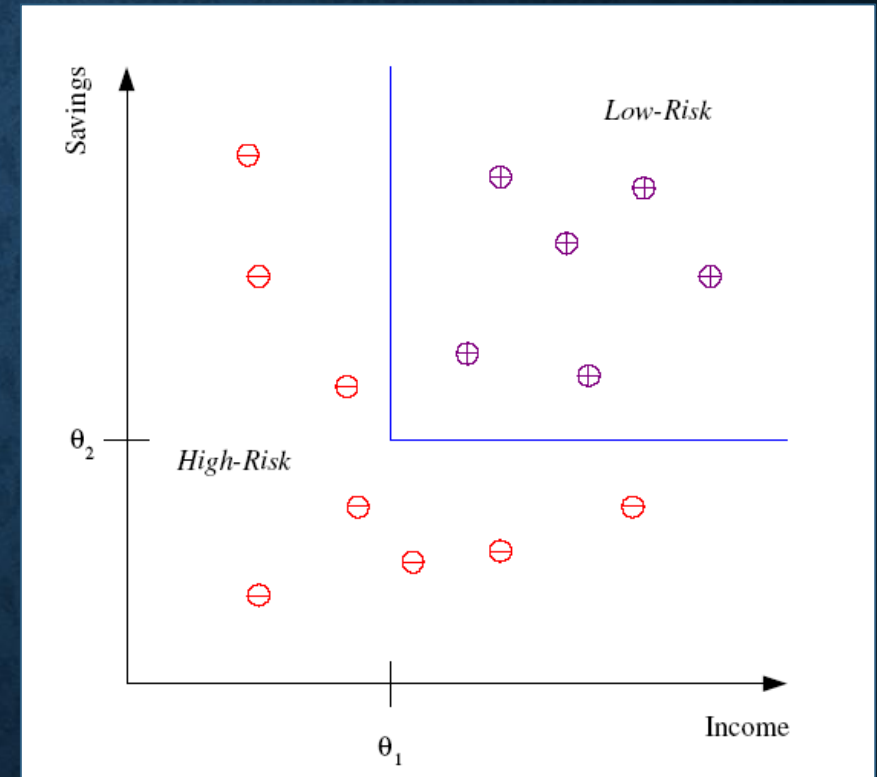
WHAT IS CLASSIFICATION?

- A classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data.
- In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups.
- Classes can be considered as targets/labels or categories. Typical examples of a class includes: **Yes or No, 0 or 1, Spam or Not Spam, cat or dog**, etc.

EXAMPLE OF CLASSIFICATION

- Example of a classification task: Credit Scoring by a bank
- Differentiating between **low-risk** and **high-risk** customers from their *income* and *savings*

Discriminant: IF $income > \theta_1$ AND $savings > \theta_2$
THEN **low-risk** ELSE **high-risk**



APPLICATIONS OF CLASSIFICATION

- It is also known as Pattern recognition.
- **Face recognition:** Pose, lighting, occlusion (glasses, beard), make-up, hair style
- **Character recognition:** Different handwriting styles.
- **Speech recognition:** Temporal dependency.
 - Use of a dictionary or the syntax of the language.
 - Sensor fusion: Combine multiple modalities; eg, visual (lip image) and acoustic for speech
- **Medical diagnosis:** From symptoms to illnesses

WHAT IS 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.
- It is the primary technique to solve the numerical problems in machine learning using data modelling.
- It involves determining the best fit line, which is a line that passes through all the data points in such a way that distance of the line from each data point is minimised.
- A regressor model predicts continuous or real values such as **temperature, age, salary, price**, etc.

EXAMPLE OF REGRESSION ANALYSIS

- Example: Price of a used car

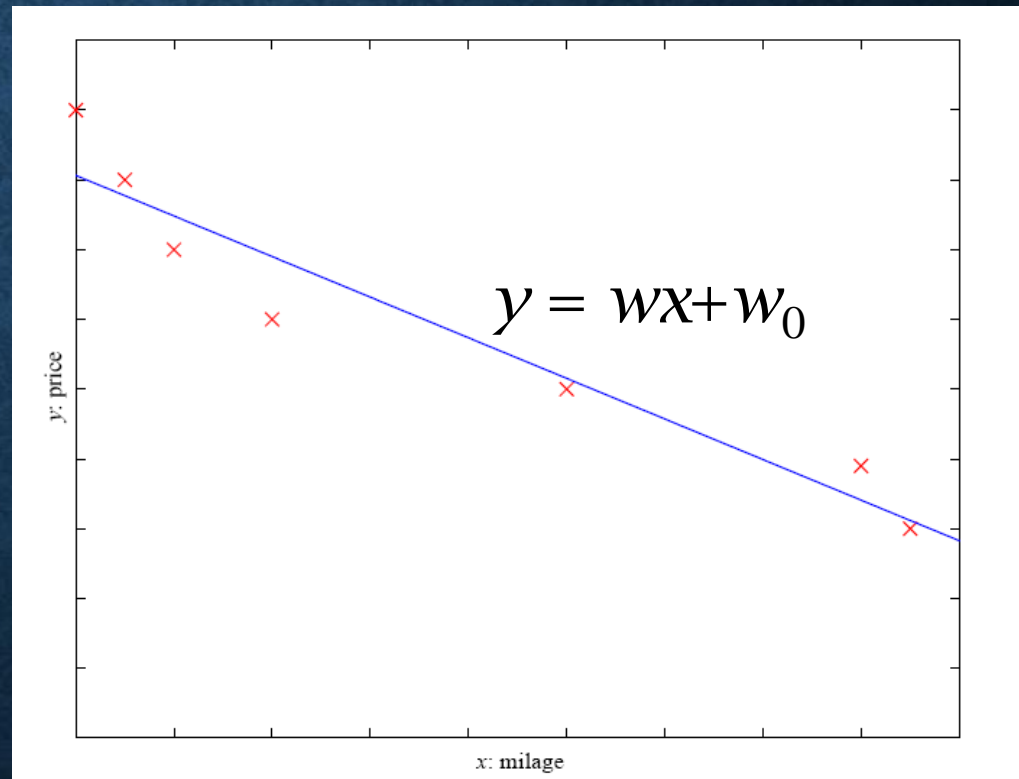
- x : car attributes

y : price

$$y = g(x | \theta)$$

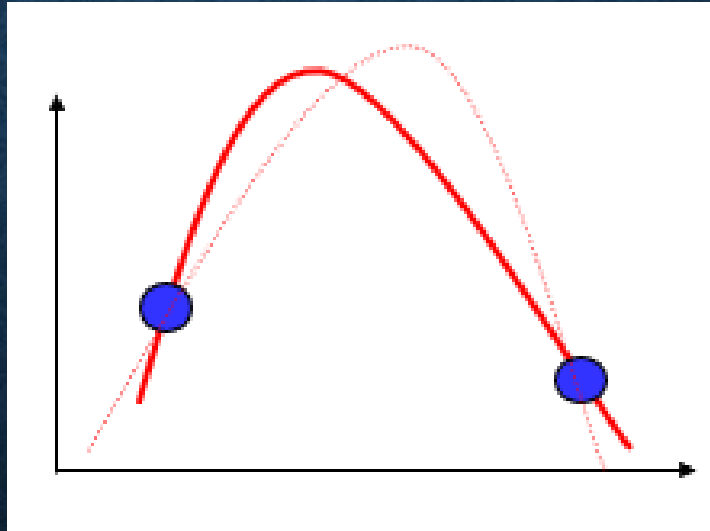
$g(\cdot)$ model,

θ parameters



APPLICATIONS OF REGRESSION

- Navigating a car: Angle of the steering wheel
- Kinematics of a robot arm
- Response surface design



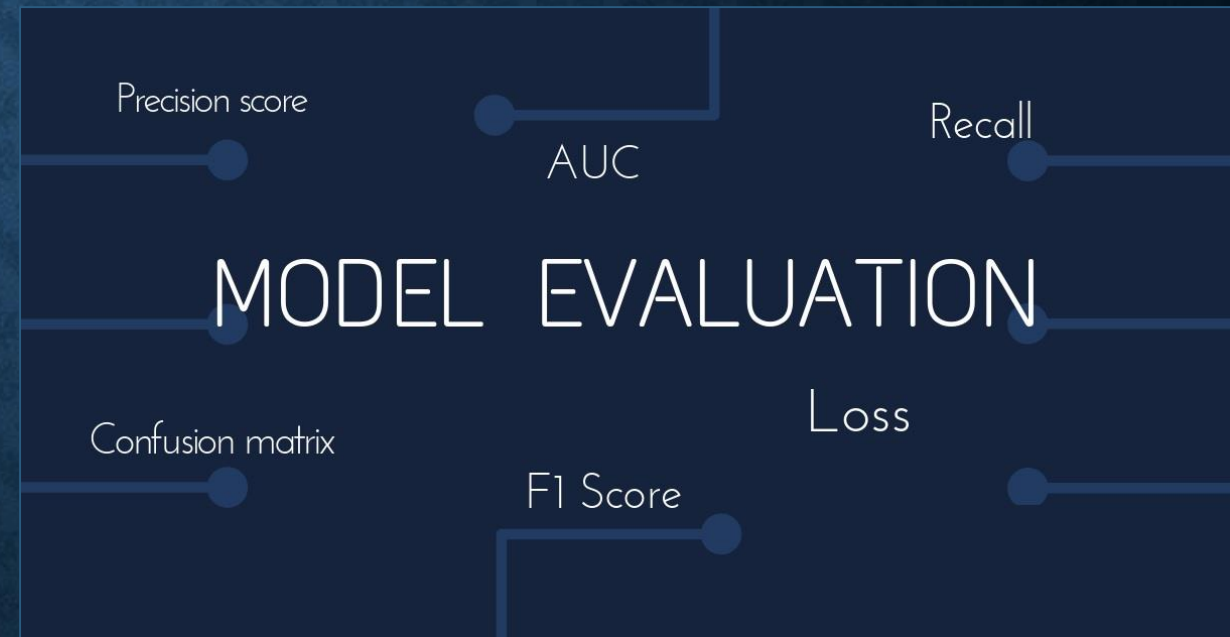
STEP 4: MODEL TRAINING

- Training of a machine learning model is crucial as it enables an incremental improvement in the generated predictions.
- The model will consider each example that is present in the dataset and will try to match the target variable as closely as possible.
- A complete run of the model on the whole dataset is called a training **epoch**.



STEP 5: MODEL EVALUATION

- A performance metric is used to assess how good or bad the model is doing.
- This parameter is computed by comparing the results obtained for an evaluation dataset.
- The evaluation dataset contains examples that were previously unseen by the model.
- Model evaluation can also be used to break the training cycle if the optimal solution has been reached.



PERFORMANCE OF A MODEL

- There are four different outcomes that can occur when a model performs classification predictions:
 - **True positives** occur when the model predicts that an observation belongs to a class and it does belong to that class.
 - **True negatives** occur when the model predicts that an observation does not belong to a class and it does not belong to that class.
 - **False positives** occur when it is predicted that an observation belongs to a class when in reality it does not. This also known as a ***Type 2 error***.
 - **False negatives** occur when it is predicted that an observation does not belong to a class when in fact it does. This also known as a ***Type 1 error***.

PERFORMANCE METRICS - ACCURACY

- From the outcomes listed previously, we can evaluate a model using various performance metrics.
- **Accuracy** measures the proportion of true results to total cases. We should always aim for a high accuracy rate.

$$\text{accuracy} = \text{number of correct predictions} / \text{total number of data points}$$

- Accuracy is expressed as a percentage and thus varies between 0% and 100%.

PERFORMANCE METRICS - PRECISION

- **Precision** is the proportion of true results over all positive results.

$$\text{Precision} = \text{True Positive} / \text{True Positive} + \text{False Positive}$$

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

- Precision values vary between 0 and 1. A high precision value is desirable.
- Hence, **precision helps us to visualise the reliability of the machine learning model in classifying the model as positive.**

PERFORMANCE METRICS - RECALL

- Recall is the fraction of all correct results returned by the model.
- It is calculated as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples.
- *Recall measures the model's ability to detect positive samples.*

$$\text{Recall} = \text{True Positive} / \text{True Positive} + \text{False Negative}$$

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

- Unlike Precision, Recall is independent of the number of negative sample classifications. Further, if the model classifies all positive samples as positive, then Recall will be 1.

PERFORMANCE METRICS – F1 SCORE

- **F1-Score** is the weighted average of precision and recall between 0 and 1.

$$\begin{aligned} F_1 &= \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \\ &= \frac{2 \times \text{tp}}{2 \times \text{tp} + \text{fp} + \text{fn}} \end{aligned}$$

- The formula for the standard F1-score is the harmonic mean of the precision and recall.
- A perfect model has an F1-Score of 1.

THE CONFUSION MATRIX

- **Confusion Matrix** illustrates the correlation between a label and the model's classification.
- One axis of a confusion matrix is the label that the model predicted, and the other axis is the actual label.
- N represents the number of classes. In a **binary classification** problem, $N=2$.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

LOSS FUNCTION: MEAN SQUARE ERROR (MSE)

- **Mean Squared Error** is the average of the squared differences between the actual and the predicted values where n is the total number of data points in the dataset.

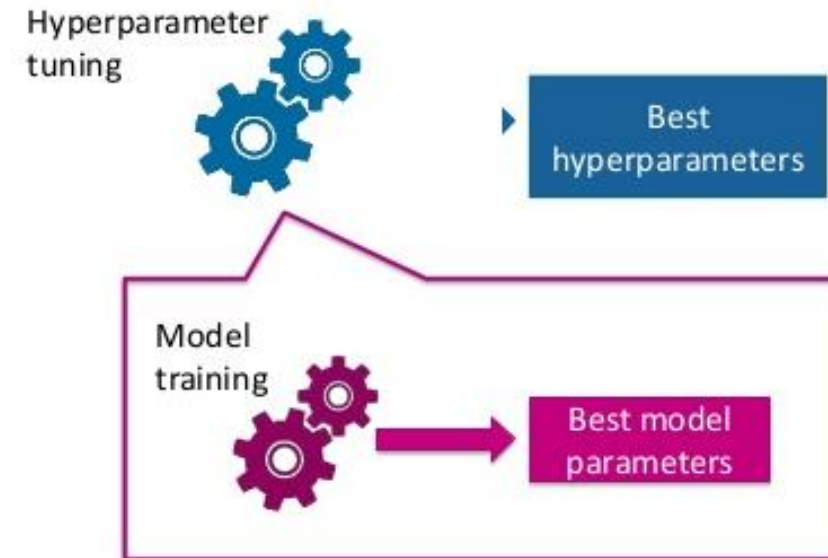
$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

- It is one of the most popular performance metrics when dealing with regression problems.

STEP 6: MODEL TUNING

- Model tuning is an essential aspect of any Machine Learning project as it allows the optimisation of the achieved performance.
- Additionally, each model has a set of hyperparameters that can be adjusted to alter the way the model learns from data and predict the results.

Hyperparameter tuning vs. model training



UNDERFITTING VS OVERFITTING

- Both conditions will usually be apparent during the evaluation phase as predictions may not meet the benchmark for acceptance.
- **Underfitting** occurs when the model does not have enough data to learn from and thus cannot find the function that correctly matches the inputs to the outputs.
- In contrast, **overfitting** occurs when the model has learned too much from the data presented and cannot generalize its predictions for unseen data.
- When facing either underfitting or overfitting, it is important to investigate both the dataset and the training approach. This ensures that the model will be able to perform generalisation.
- **Generalisation** refers to how well the model predicts outcomes for a new set of data.

MANAGING NOISE

- Noise management is very important in Machine Learning.
- Noise is defined as unwanted anomalies that disguise the underlying relationship in the data set and weakens the learning process.
- Reasons why noise occurs in training include:
 1. Large training data set
 2. Errors in input data
 3. Data labelling errors
 4. Unobservable attributes that might affect the algorithms/model.

STEP 7: PREDICTION

- This is the last phase of implementing a machine learning model. Predictions can now be made on the real-life application or using live data.



Machine Learning
Prediction

QUESTIONS TIME

