

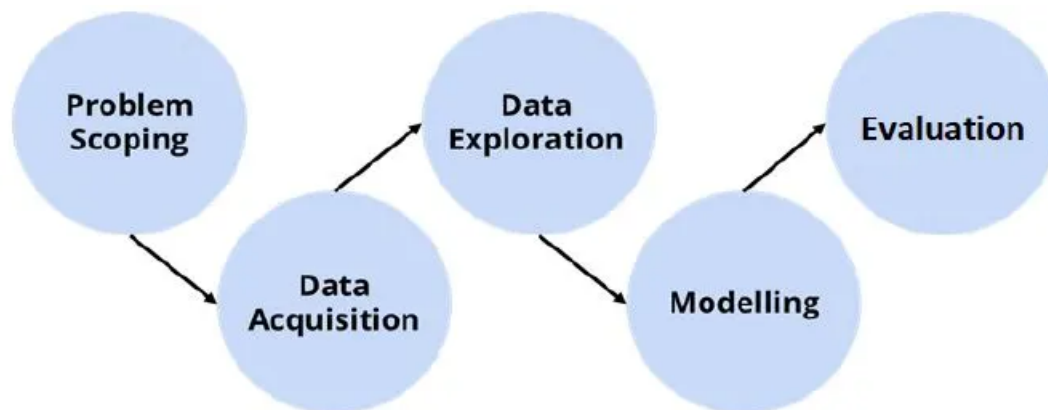
AI Project Cycle

The AI Project Cycle is a Step-by-step process to solve problems using proven scientific methods and drawing inferences about them. Starting with Problem Scoping, you set the goal for your AI project by stating the problem which you wish to solve with it. Under problem scoping, we look at various parameters which affect the problem we wish to solve so that the picture becomes clearer.

Stages of AI Project Cycle

- **Problem Scoping**- Understanding the problem
- **Data Acquisition**- Collecting accurate and reliable data
- **Data Exploration**- Arranging the data uniformly
- **Modeling**- Creating Models from the data
- **Evaluation**- Evaluating the project

This cycle has five main steps:



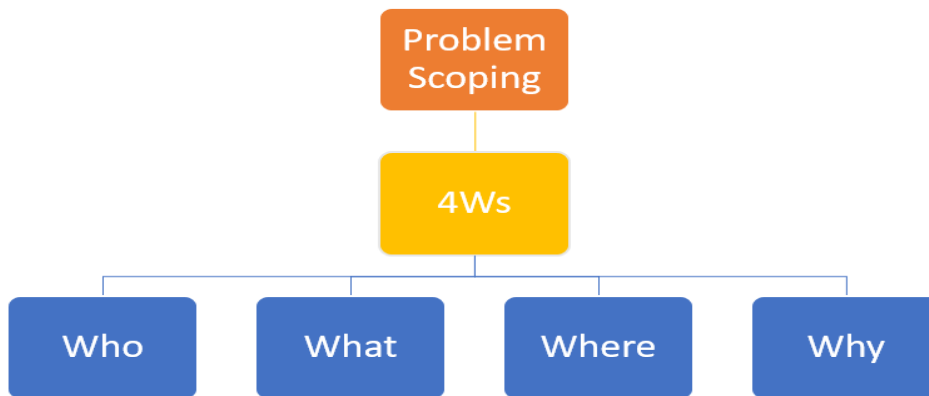
Stage 1 : Problem Scoping - Problem Scoping refers to understanding a problem, finding out various factors which affect the problem, define the goal or aim of the project. Some problems are small, and some are big. Sometimes we might not even notice them. But when we realize a problem and want to fix it, that's called Problem Scoping.

Sustainable Development Goals : The United Nations has created 17 important goals called Sustainable Development Goals. All the countries in the United Nations have agreed to work towards achieving these goals by the year 2030. These goals for sustainable development match the problems we may notice around us. It's important to find and fix these problems because it can make life better for many people and help our country achieve its goals.



4Ws Problem Canvas

Understanding a problem well can be challenging. To make it clearer for solving, we use the 4Ws Problem Canvas. The 4 W's of Problem Scoping stand for Who, What, Where, and Why. These 4 W's help us figure out and grasp the problem better.



- a. **Who:** This part helps us understand and sort out who is directly and indirectly impacted by the problem, including those known as Stakeholders.
- b. **What:** This section helps us examine and identify what the problem is all about. You can also gather proof to show that the problem really exists.
- c. **Where:** This part is about the situation and where the problem shows up.

d. **Why:** It's about why we should deal with the problem and how solving it will benefit the stakeholders.

Problem Statement Template

The Problem Statement Template helps us to summarize all the key points into one single. Template so that in the future, whenever there is a need to look back at the basis of the problem, we can take a look at the Problem Statement Template and understand the key elements of it.

Our	[stakeholder(s)]	Who
	<hr/> <hr/> <hr/>	
has /have a problem that	[issue, problem, need]	What
	<hr/> <hr/> <hr/>	
when / while	[context, situation]	Where
	<hr/> <hr/> <hr/>	
An ideal solution would	[benefit of solution for them]	Why
	<hr/> <hr/> <hr/> <hr/> <hr/>	

Stage 2: Data Acquisition

The method of collecting accurate and trustworthy data to work with is referred to as data acquisition. Data can be acquired from a variety of sources, including websites, journals, newspapers, and other media, such as text, video, photographs, and audio.

Data can be a piece of information or facts and statistics collected together for reference or analysis. Whenever we want an AI project to be able to predict an output, we need to train it first using data.

For example : If you want to make an Artificially Intelligent system which can predict the weather, you would feed the data of previous days for that location into the machine. This is the data with which the machine can be trained. Now, once it is ready, it will predict the weather conditions efficiently.

Dataset

Dataset is a collection of data in tabular format. Dataset contains numbers or values that are related to a specific subject. For example, students' test scores in a class is a dataset.

The dataset is divided into two parts

a. Training dataset – Training dataset is a large dataset that teaches a machine learning model. Machine learning algorithms are trained to make judgments or perform a task through training datasets. Maximum part of the dataset comes under training data (Usually 80%)

b. Test dataset – Data that has been clearly identified for use in tests, usually of a computer program, is known as test data. 20% of data used in test data.

For any AI project to be efficient, the training data should be authentic and relevant to the problem statement scoped.

Data Features

Data Features refer to the type of data you want to collect. Data features for predicting used car prices include car age, mileage, brand, model, fuel type, transmission, number of owners, condition, accident history, service history, location, features, and market demand etc.

VARIOUS DATA SOURCES USED TO ACQUIRE RELIABLE DATA

There can be various ways in which you can collect data. Some of them are surveys, Web Scraping, Sensors, cameras, Observations, APIs (Application Program Interface) etc.

Surveys - Customer's feedback and reviews

Web Scraping Data extracted from various web pages

Sensors Data collected from various sensors to track the conditions of physical things can be monitored in real time.

Cameras - Live data from surveillance cameras, web cameras etc.

Observations - Reading and analysing trends

API - API stands for Application Programming interface. It is a program that generates data of their own while working, like data on their servers.

Open source Govt. Portals : data.gov.in , , india.gov.in .

Stage 3 : Data Exploration

Data Exploration is the process of visualization of collected data in a graphical format for better understanding to build the aimed project.

For Example, if you go to the library and choose a random book, you try to rapidly go through its content by turning pages and reading the description before you decide to borrow it for yourself. This helps you determine whether or not the book is suitable for your requirements and interests. This is data exploration.

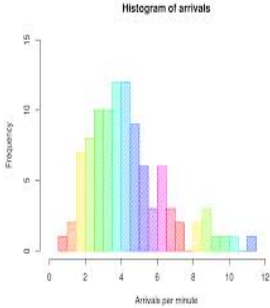
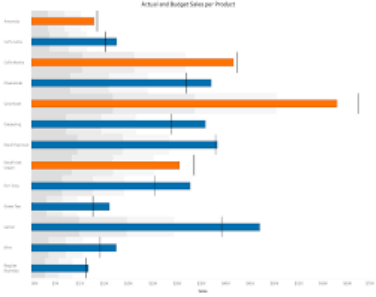
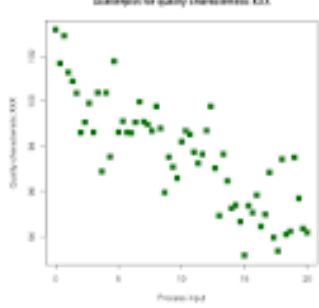
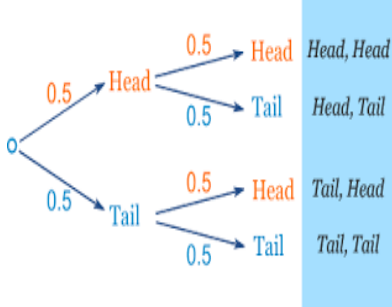
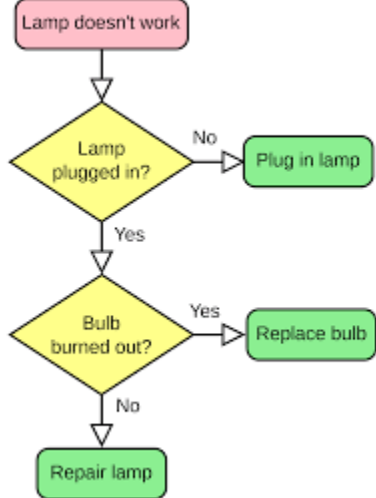
To analyze the data, you need to visualize it in some user-friendly format so that you can:

- Quickly get a sense of the trends, relationships and patterns contained within the data.
- Define strategy for which model to use at a later stage.
- Communicate the same to others effectively.

Different ways to visualize data are:

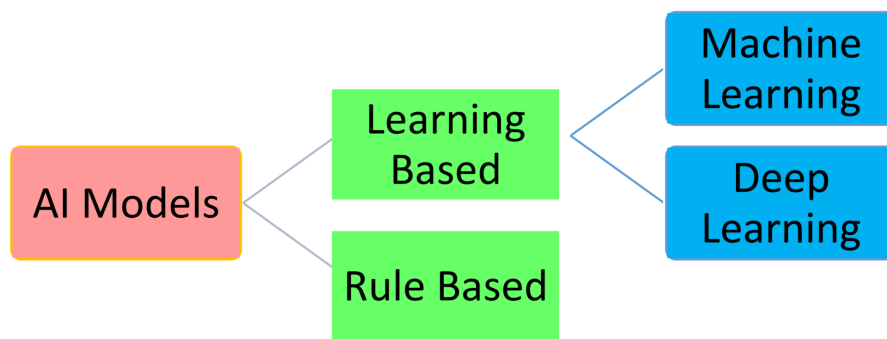
1. Bullet graphs - a Bullet Graph functions like a Bar Chart, but is accompanied by extra visual elements to pack in more context.

2. Histogram - A histogram is a graphical representation of data points organized into user-specified ranges.
3. Scatterplot
4. Tree Diagram
5. Flow Chart

 <p>A histogram showing the frequency of arrivals per minute. The x-axis is labeled 'Arrivals per minute' and ranges from 0 to 12. The y-axis is labeled 'Frequency' and ranges from 0 to 15. The bars are colored in a rainbow spectrum.</p> <p>Histogram</p>	 <p>A horizontal bullet graph titled 'Actual and Budget Sales per Product'. It shows multiple products on the y-axis. For each product, there are two bars: a blue bar representing budget sales and an orange bar representing actual sales. A vertical line indicates the target or maximum sales for each product.</p> <p>Bullet graphs</p>	 <p>A scatter plot titled 'Scatterplot for quality characteristics KEX'. The x-axis is labeled 'Process input' and ranges from 0 to 20. The y-axis is labeled 'Quality characteristic KEX' and ranges from 0 to 100. The plot shows a negative correlation between the process input and the quality characteristic.</p> <p>Scatter Plot</p>
 <p>A probability tree diagram starting from a root node '0'. It branches into 'Head' (0.5) and 'Tail' (0.5). From 'Head', it branches into 'Head' (0.5) and 'Tail' (0.5). From 'Tail', it branches into 'Head' (0.5) and 'Tail' (0.5). The final outcomes are listed in a blue box: 'Head, Head', 'Head, Tail', 'Tail, Head', and 'Tail, Tail'.</p> <p>Tree Diagram</p>	 <pre> graph TD Start([Lamp doesn't work]) --> D1{Lamp plugged in?} D1 -- No --> A1[Plug in lamp] D1 -- Yes --> D2{Bulb burned out?} D2 -- Yes --> A2[Replace bulb] D2 -- No --> A3[Repair lamp] </pre> <p>A flowchart for troubleshooting a lamp. It starts with 'Lamp doesn't work'. A decision diamond asks 'Lamp plugged in?'. If 'No', the action is 'Plug in lamp'. If 'Yes', another decision diamond asks 'Bulb burned out?'. If 'Yes', the action is 'Replace bulb'. If 'No', the final action is 'Repair lamp'.</p> <p>Flow Chart</p>	

Stage 4 : Modelling is choosing a suitable model type based on research and testing, building your AI project around it.

An AI model is a program that has been trained to recognize patterns using a set of data. AI modeling is the process of creating algorithms, also known as models, that may be educated to produce intelligent results. This is the process of programming code to create a machine artificially. Generally, AI models can be classified as follows:



Rule Based Approach AI modelling in which the developer sets the rules. The machine executes its duty in accordance with the rules or instructions specified by the developer.

Learning Based Approach AI modelling in which the computer learns on its own. The AI model is trained on the data provided to it under the Learning Based technique, and after that, it is able to create a model that is flexible to the change in data.

The learning-based approach can further be divided into three parts:

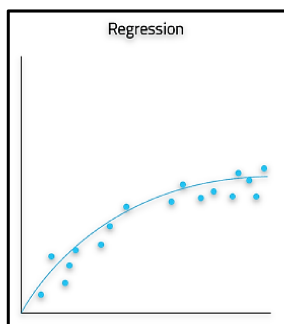
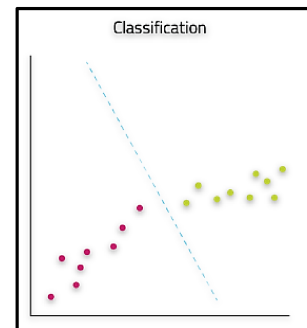
1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

Supervised Learning – In a supervised learning model, the dataset which is fed to the machine is labelled. In other words, we can say that the dataset is known to the person who is training the machine only then he/she is able to label the data. In supervised learning, the dataset used to train the machine learning model consists of input data (features) and corresponding output labels (target values). The goal is for the model to learn the mapping between the input features and their respective correct output labels. The term "supervised" comes from the fact that during the training process, the model is supervised by having

access to this labeled data, allowing it to adjust its parameters to make accurate predictions. Example: Let's consider a classic example of email classification. The task is to build a model that can distinguish between spam and non-spam (ham) emails. For this, we would need a dataset containing a large number of emails, each labeled as either "spam" or "ham." The input features might include the email's content, sender, subject, and other relevant attributes. The output labels would be "spam" or "ham."

There are two types of Supervised Learning models:

Classification: Where the data is classified according to the labels. For example, in the grading system, students are classified on the basis of the grades they obtain with respect to their marks in the examination. This model works on a discrete dataset which means the data need not be continuous.

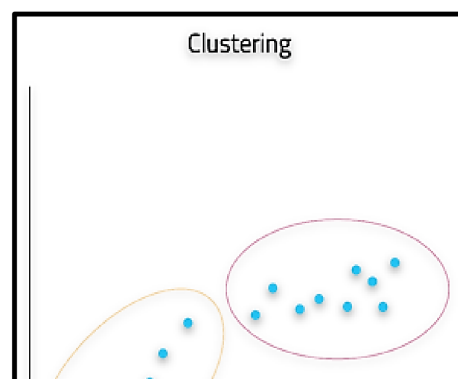


Regression: Such models work on continuous data. For example, if you wish to predict your next salary, then you would put in the data of your previous salary, any increments, etc., and would train the model. Here, the data which has been fed to the machine is continuous.

Unsupervised Learning – An unsupervised learning model works on an unlabelled dataset. This means that the data which is fed to the machine is random and there is a possibility that the person who is training the model does not have any information regarding it.

Unsupervised learning models can be further divided into two categories:

Clustering – Refers to the unsupervised learning algorithm which can cluster the unknown data according to the patterns or trends identified out of it. The patterns



observed might be the ones which are known to the developer or it might even come up with some unique patterns out of it.

Dimensionality Reduction – We humans are able to visualize upto 3-Dimensions only but according to a lot of theories and algorithms, there are various entities which exist beyond 3-Dimensions. For example, in Natural language Processing, the words are considered to be N-Dimensional entities. Which means that we cannot visualize them as they exist beyond our visualization ability. Hence, to make sense out of it, we need to reduce their dimensions. Here, dimensionality reduction algorithm is used.

As we reduce the dimension of an entity, the information which it contains starts getting distorted. For example, if we have a ball in our hand, it is 3-Dimensions right now. But if we click its picture, the data transforms to 2-D as an image is a 2-Dimensional entity. Now, as soon as we reduce one dimension, at least 50% of the information is lost as now we will not know about the back of the ball. Whether the ball was of the same color at the back or not? Or was it just a hemisphere? If we reduce the dimensions further, more and more information will get lost. Hence, to reduce the dimensions and still be able to make sense out of the data, we use Dimensionality Reduction.

Reinforcement Learning

Reinforcement learning is the closest machine learning type to how humans learn. The algorithm or agent used learns by interacting with its environment and getting a positive or negative reward.

Reinforcement learning is applicable in areas capable of being fully simulated that are either stationary or have large volumes of relevant data.

Some examples of uses include:

- Teaching cars to park themselves and drive autonomously
- Dynamically controlling traffic lights to reduce traffic jams
- Training robots to learn policies using raw video images as input that they can use to replicate the actions they see

Stage 5 Evaluation: Test the selected model using new data to assess its performance, refining and improving it as necessary. Once a model has been created and trained, it must undergo appropriate testing in order to determine the

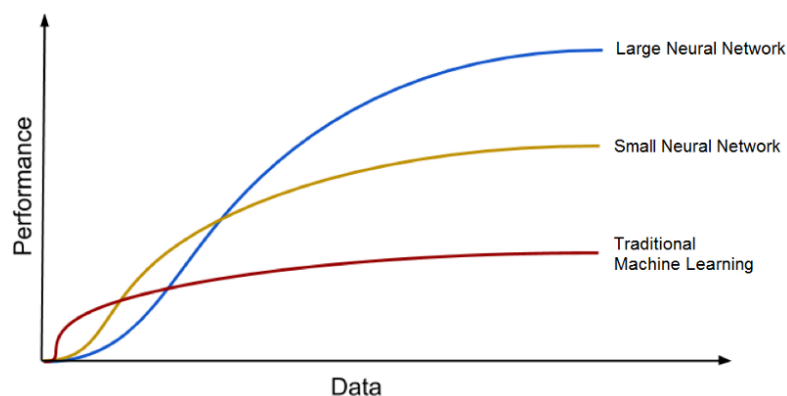
model's effectiveness and performance. Thus, the model is evaluated using Testing Data (which was taken from the dataset generated at the Data Acquisition stage), and the effectiveness of the model is determined using the parameters listed below –



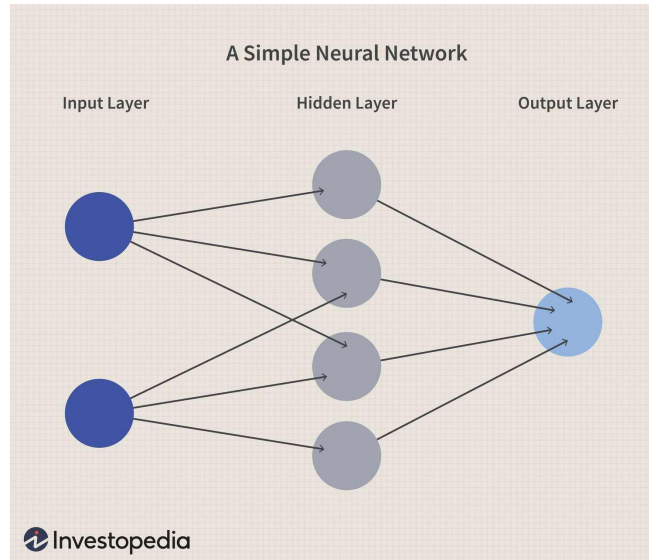
Neural Networks

A neural network is **a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain**. It is a type of machine learning process, called deep learning, which uses interconnected nodes or neurons in a layered structure that resembles the human brain.

Neural networks are loosely modelled after how neurons in the human brain behave. The key advantage of neural networks, are that they are able to extract data features automatically without needing the input of the programmer. A neural network is essentially a system of organizing machine-learning algorithms to perform certain tasks. It is a fast and efficient way to solve problems for which the dataset is very large, such as in images.



As seen in the figure given, the larger Neural Networks tend to perform better with larger amounts of data whereas the traditional machine learning algorithms stop improving after a certain saturation point.



This is a representation of how neural networks work. A Neural Network is divided into multiple layers and each layer is further divided into several blocks called nodes. Each node has its own task to accomplish which is then passed to the next layer.

- The first layer of a Neural Network is known as the **input layer**. The job of an input layer is to acquire data and feed it to the Neural Network. No processing occurs at the input layer.
- Next to it, are the **hidden layers**. Hidden layers are the layers in which the whole processing occurs. Their name essentially means that these layers are hidden and are not visible to the user. Each node of these hidden layers has its own machine learning algorithm which it executes on the data received from the input layer. The processed output is then fed to the subsequent hidden layer of the network.
- The last hidden layer passes the final processed data to the **output layer** which then gives it to the user as the final output. Similar to the input layer, the output layer too does not process the data which it acquires. It is meant for user-interface.