Data science

1. Data Mining– study of data to extract information from data. Making the data communicate to you. This is done through descriptive statistics
   1. Data Visualizations
   2. Aggregating / pivoting data with the help of group by commands
   3. Measures of central tendency: mean / median / mode
      1. Mean – average
      2. Median – middle value
      3. Mode – value with the highest frequency
   4. Measures of dispersion
      1. Range
      2. Inter-Quartile Range
      3. Standard Deviation / Variance
2. Inferential Statistics / Analytics – Any decision / conclusion made on the sample data can be concluded / inferred on the population data, only if it passes the inferential statistical tests.

Representative sample – is the random sample that is large enough.

1. Predictive Analytics
   1. Regression – when our response variable is continuous.
      1. Eg. prediction of sales of a company.
      2. Prediction of cost
      3. Prediction of house prices in India
      4. Prediction of stock market prices
   2. Classification
      1. Binary Classification – Response variable is binary.
         1. Eg. predicting whether Amar will default on loan or not.
         2. Amar will survive or not if he travels from same source to destination as Titanic ship.
         3. Email is spam or ham
      2. Multi-class Classification – y is multi-class or mulinomial discrete (having a finite number of categories).
         1. Predicting the rating given to the latest movie by Pratyaksh 1/2/3/4/5.
         2. Image detection plane / truck / car
         3. Voice is of a male adult / a male child / female adult / female child
         4. Sentiment prediction - +ve, - ve, neutral

**Vocabulary**

Response Variable / Target Variable / Predicted variable / Dependent Variable / y variables – the variable that needs to be predicted

Independent variables / Dimensions / Attributes / Features / X variables

**Machine Learning –** The if-then / decision rules are not programmed by the programmer, rather they are learned from the data. The engineer is only give the algorithm - how to learn an if-then rule. But data will decide the if-then rules. This means, the data scientist is only giving the experience of the data and certain algorithm / baseline to decide the decision rules. The machine is learning the rules through data.

**Traditional programming approaches -**  Developer is defining the if-then rules

**Deep Learning –** based on neural networks. Our endeavour here is to simulate the human brain.

**Artificial Intelligence**

When we deploy multiple algorithms to create an application, which is achieving certain purpose

* Machine Learning algos
* Deep Learning algos
* Rule based engines

**Types of AI applications:**

* **Natural Language Processing –** when AI machine is showing the intelligence of understanding the human language
  + Alexa
  + Google home
  + Google Assistant
  + Chatbots
* **Robotics –** When the AI application is showing the intelligence similar to human motion. Eg. Sophia, Self-driving cars, autonomous helicopters, Humanoids
* **Computer Vision -**  When the AI application is showing the intelligence similar to human eyes. Eg. handwriting recognition , login with retina, finger-print scan, video analysis, google lens

**Structured Data –** When the independent variables are in the columns and the transactions in the rows.

**Semi-structured Data -**  When there is a structure to the data, but that structure is not in the rows and columns format. The structure can be easily converted to a structured format with a few lines of code. Eg. html, xml, json files

**Unstructured Data -**  when the data is having no structure eg. text data, voice data, images, videos

Machine learning algos can’t be applied to unstructured data. So, we have to find a way to convert this data to a structured format.

**Machine Learning is of 2 types:**

* **Supervised Learning**
  + When we have a response variable
  + We will use the training data (or past data) to train the models to predict the y variable. Once the model is ready based on the past data, we will deploy it on the unseen dataset
    - Regression
    - Time- Series problems
    - Classification
      * Binary
      * Multi-class classification
* **Unsupervised Learning**
  + When we don’t have a response variable and we are trying to group similar data to together. Eg. Clustering similar customers together of a grocery store to identify the different types of customers coming to the store.

**Big Data**

* Volume
* Variety
* Velocity