UNIT - 1

Unit 1 : INTRODUCTION: What is Data Science? - Big Data and Data Science hype – and getting past the hype - Why now? – Datafication - Current landscape of perspectives – the Skill needed to do data science. Statistical Inference - Populations and samples – Modeling - statistical modeling, probability distributions, fitting a model.

What Is Data Science?

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models. The data used for analysis can come from many different sources and presented in various formats.

Importance of Data Science

- ✓ Data is a valuable asset for various industries to help make careful and sound business-related decisions. Data science has the ability to churn raw data into meaningful insights.
- ✓ An expert data scientist has the capability of digging out meaningful information from whatever data is available to them. They lead organizations in the right direction through sound data-driven decisions and suggestions.

1. **Importance of Data Science in Various Industries**
   Large databases of structured and unstructured data must be mined using data science techniques to find hidden patterns and derive useful insights. Data science is crucial because of the numerous applications it may be used for, ranging from simple activities, such as asking Siri or Alexa for recommendations, to more sophisticated ones, such as operating a self-driving automobile.

2. Importance of Data Science in Business
   Data science's goal is to assist organizations in comprehending the patterns of variance in data, including client information, business growth rates, data volume, or any measurable quantity.

3. Importance of Data Science in Healthcare
   The healthcare sector produces enormous datasets of valuable data on patient demographics, treatment plans, outcomes of medical exams, insurance, etc. The vast amounts of dispersed, structured, and unstructured data generated by healthcare systems can be processed, analyzed, assimilated and managed with the help of data science.

4. **Importance of Data Science in the Future**
   Companies today have access to massive databases due to documenting every aspect of client engagement. Data science plays a crucial role in analyzing and developing these data-driven machine-learning models.

Real-Life Examples of Data Science Applications

1. Medical Image Analysis
Procedures like detecting neoplasia, artery stiffness, and organ delineation use methods and frameworks like MapReduce to find the best parameters for tasks like lung texture categorization.

2. Fraud and Risk Detection
Banking organizations have mastered the art of data science and control through the use of consumer profiling, historical purchases, and other crucial factors to calculate the likelihood of risk and default.

3. Airline Route Planning
The airline business has a reputation for overcoming challenges. A few airline service providers are trying to keep their working conditions and occupancy rates high.

History of Data Science
While the term data science is not new, the meanings and connotations have changed over time. The word first appeared in the ’60s as an alternative name for statistics. In the late ’90s, computer science professionals formalized the term. A proposed definition for data science saw it as a separate field with three aspects: data design, collection, and analysis. It still took another decade for the term to be used outside of academia.

1963: John W. Tukey, an American mathematician, initially expressed the data science dream in 1962. Nearly two decades before the first personal computers, he predicted the rise of a new field in his now-famous paper "The Future of Data Analysis."

1977: The International Association for Statistical Computing (IASC) was founded, with the mission of "linking traditional statistical methodology, the knowledge of domain experts and modern computer technology, to transform data into knowledge and information.

The 1980s & 1090s: With the inaugural Knowledge Discovery in Databases (KDD) workshop and the foundation of the International Federation of Classification Societies in the 1980s and 1990s, data science began to make considerable advancements (IFCS).

1994: The emerging phenomena of "Database Marketing" was covered in Business Week in 1994. It represented the process by which organizations gathered and analyzed massive volumes of data in order to understand more about their consumers, competitors, and advertising strategies.

The 1990s and early 2000s: Data science has definitely evolved as a recognized and specialized field in the 1990s and early 2000s.

2000s: Technology made huge strides by making internet connectivity, communication, and (of course) data collection practically widespread.

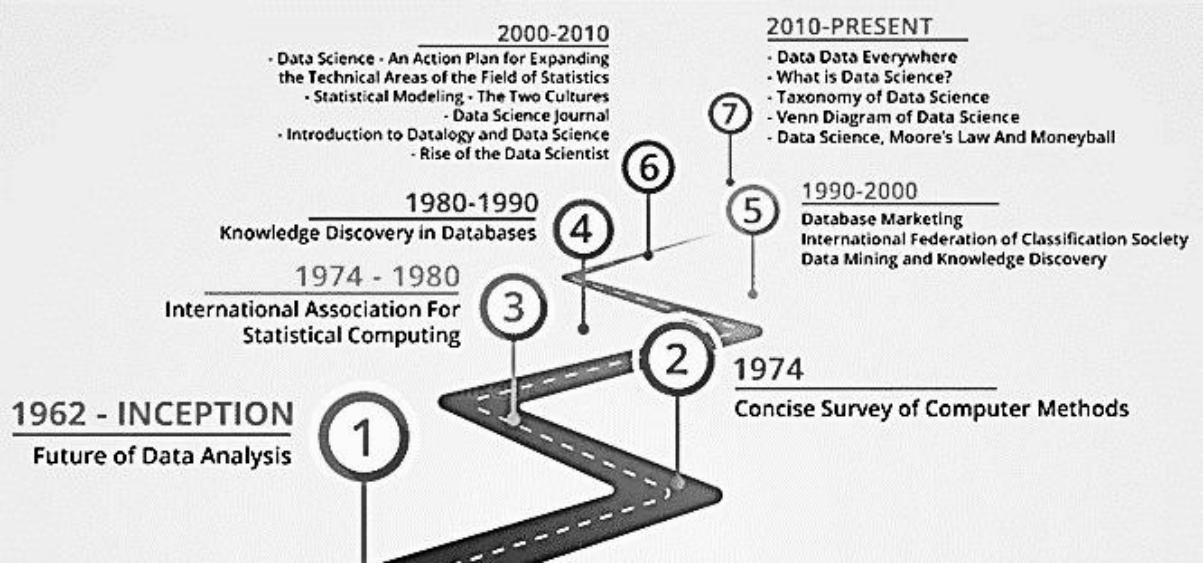
2005: Big data makes its debut in 2005. With tech behemoths like Google and Facebook amassing massive volumes of data.

2014: As data became more important and organizations became more interested in detecting patterns.

2015: Artificial Intelligence (AI), Machine Learning, and Deep learning all make their debut in the field of data science in 2015.

2018: One of the most significant aspects of the evolution of data science is the introduction of new regulations in the field.

2020s: We're seeing further advancements in AI and machine learning, as well as an ever-increasing demand for qualified Big Data specialists.



The Data Science Lifecycle

The major steps in the life cycle of a Data Science

1. *Problem identification & Business Understanding*

This is the most important stage of any Data Science endeavor; The first step is to understand how Data Science is useful in the domain under consideration and to identify appropriate tasks that are useful for the same. Domain specialists and data scientists play critical roles in problem identification. The business goals are formed by the customer's need to make predictions, boost sales, minimize losses, or optimize any given process, among other things.

- Clearly state the problem that requires solutions and why it should be resolved at once
- Define the potential value of the business project
- Find risks, including ethical aspects involved in the project
- Build and communicate a highly integrated, flexible project plan

2. **Data collection**

Data collection is the next stage in the data science lifecycle to gather raw data from relevant sources. The data captured can be either in structured or unstructured form.

3. **Data processing**

In this phase, data scientists analyse the data collected for biases, patterns, ranges, and distribution of values. It is done to determine the sustainability of the databases and predicts their usage in regression, machine learning and deep learning algorithms.

4. **Data analysis**

Data Analysis or Exploratory Data Analysis is another critical step in gaining some ideas about the solution and factors affecting the data science lifecycle. There are no set guidelines for this methodology, and it has no shortcuts.

5. **Data modelling**

Modelling Data is one of the major phases of data processes and is often mentioned as the heart of data analysis. A model should use prepared and analysed data to provide the desired output. The environment needed for executing the data model will be decided and created before meeting the specific requirements.

6. **Model deployment**

Now, we are at the final stage of the lifecycle of data science. After a rigorous evaluation process, the model is finally prepared to be deployed in the desired format and preferred channel.

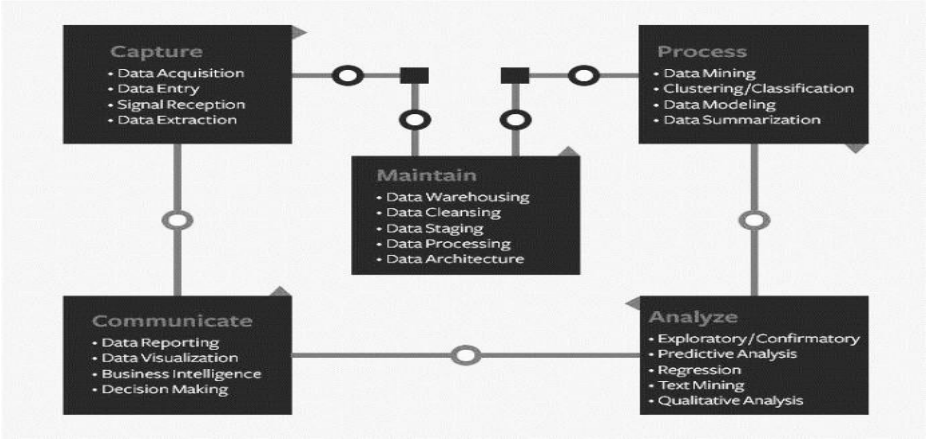
Process of Data Life Cycle:

Data science's lifecycle consists of five distinct stages, each with its own tasks:

1. Capture: Data Acquisition, Data Entry, Signal Reception, Data Extraction. This stage involves gathering raw structured and unstructured data.
2. Maintain: Data Warehousing, Data Cleansing, Data Staging, Data Processing, Data Architecture. This stage covers taking the raw data and putting it in a form that can be used.

3. Process: Data Mining, Clustering/Classification, Data Modeling, Data Summarization. Data scientists take the prepared data and examine its patterns, ranges, and biases to determine how useful it will be in predictive analysis.
4. Analyze: Exploratory/Confirmatory, Predictive Analysis, Regression, Text Mining, Qualitative Analysis. Here is the real meat of the lifecycle. This stage involves performing the various analyses on the data.
5. Communicate: Data Reporting, Data Visualization, Business Intelligence, Decision Making. In this final step, analysts prepare the analyses in easily readable forms such as charts, graphs, and reports.

The Data Science Life Cycle



How does Data Science Work?

The working of data science can be explained as follows:

1. Raw data is gathered from various sources that explain the business problem.
2. Using various statistical analysis, and machine learning approaches, data modeling is performed to get the optimum solutions that best explain the business problem.
3. Actionable insights that will serve as a solution for the business problems gathered through data science.

What are the data science techniques?

Data science professionals use computing systems to follow the data science process. The top techniques used by data scientists are:

Classification

Classification is the sorting of data into specific groups or categories. Computers are trained to identify and sort data. Known data sets are used to build decision algorithms in a computer that quickly processes and categorizes the data.

example:·

- Sort products as popular or not popular·
- Sort insurance applications as high risk or low risk·
- Sort social media comments into positive, negative, or neutral.

Data science professionals use computing systems to follow the data science process.

Regression

Regression is the method of finding a relationship between two seemingly unrelated data points. The connection is usually modeled around a mathematical formula and represented as a graph or curves. When the value of one data point is known, regression is used to predict the other data point.

example: ·

- The rate of spread of air-borne diseases.·
- The relationship between customer satisfaction and the number of employees.·
- The relationship between the number of fire stations and the number of injuries due to fire in a particular location.

Clustering

Clustering is the method of grouping closely related data together to look for patterns and anomalies. Clustering is different from sorting because the data cannot be accurately classified into fixed categories. Hence the data is grouped into most likely relationships. New patterns and relationships can be discovered with clustering.

example: ·

- Group customers with similar purchase behavior for improved customer service.·
- Group network traffic to identify daily usage patterns and identify a network attack faster.
- Cluster articles into multiple different news categories and use this information to find fake news content.

Data science practitioners work with complex technologies such as:

1. Artificial intelligence: Machine learning models and related software are used for predictive and prescriptive analysis.
2. Cloud computing: Cloud technologies have given data scientists the flexibility and processing power required for advanced data analytics.
3. Internet of things: IoT refers to various devices that can automatically connect to the internet. These devices collect data for data science initiatives. They generate massive data which can be used for data mining and data extraction.
4. Quantum computing: Quantum computers can perform complex calculations at high speed. Skilled data scientists use them for building complex quantitative algorithms.

What is Big Data?

Big Data is the extraction, analysis and management of processing a large volume of data. It revolves around the datatype – Big Data which is a collection of a colossal amount of data. 5 Vs that define big data are velocity, volume, value, variety and veracity.

Some data sets that we can consider truly big data include:

- Stock market data
- Social media
- Sporting events and games
- Scientific and research data

Characteristics of big data

- **Volume.** Big data is enormous, far surpassing the capabilities of normal data storage and processing methods. The volume of data determines if it can be categorized as big data.
- **Variety.** Large data sets are not limited to a single kind of data—instead, they consist of various kinds of data. Big data consists of different kinds of data, from tabular databases to images and audio data regardless of data structure.
- **Velocity.** The speed at which data is generated. In Big Data, new data is constantly generated and added to the data sets frequently. This is highly prevalent when dealing with continuously evolving data such as social media, IoT devices, and monitoring services.
- **Veracity or variability.** There will inevitably be some inconsistencies in the data sets due to the enormity and complexity of big data. Therefore, you must account for variability to properly manage and process big data.
- **Value.** The usefulness of Big Data assets. The worthiness of the output of big data analysis can be subjective and is evaluated based on unique business objectives.

Types of big data

- **Structured data:** Any data set that adheres to a specific structure can be called structured data. A good example for structured data will be a distributed RDBMS which contains data in organized table structures.
- **Semi-structured data:** This type of data does not adhere to a specific structure yet retains some kind of observable structure such as a grouping or an organized hierarchy. Some examples of semi-structured data will be markup languages (XML), web pages, emails, etc.
- **Unstructured data:** This type of data consists of data that does not adhere to a schema or a preset structure. It is the most common type of data when dealing with big data—things like text, pictures, video, and audio all come up under this type.

Comparison of Big data Vs Data Science

Basis	Data Science	Big Data
Meaning	Skewed towards the scientific approach of interpreting the data and retrieves the information from a given data set	Revolves around the huge volumes of data which cannot be handled using the conventional data analysis method
Concept	Obtained with big data is heterogeneous that indicates a diversified data set which has to be per-cleaned and sorted before running analytics on them	Scientific techniques to process data, extract information and interpret results which help in the decision-making process
Formation	Internet users/ traffic, live feeds, and data generated from system logs	Data filtering, preparation, and analysis
Application areas	Internet search, digital advertisements, text-to-speech recognition, risk detection, and other activities	Telecommunication, financial service, health and sports, research and development, and security and law enforcement
Approach	Uses mathematics and statistics extensively along with programming skills to develop a model to test the hypothesis and make decisions in the business	Used by businesses to track their presence in the market which helps them develop agility and gain a competitive advantage over others

Big Data and Data Science Hype

Big Data is one of THE biggest buzzwords around at the moment and I believe big data will change the world.

Basically, big data refers to our ability to collect and analyse the vast amounts of data we are now generating in the world.

- Volume – the vast amounts of data generated every second
- Velocity – the speed at which new data is generated and moves around (credit card fraud detection is a good example where millions of transactions are checked for unusual patterns in almost real time)
- Variety – the increasingly different types of data (from financial data to social media feeds, from photos to sensor data, from video capture to voice recordings)
- Veracity – the messiness of the data (just think of Twitter posts with hash tags, abbreviations, typos and colloquial speech)

business example: Wal-Mart is able to take data from your past buying patterns, their internal stock information, your mobile phone location data, social media as well as external weather information and analyse all of this in seconds.

Why Now:

Technology makes this possible: infrastructure for large-scale data processing, increased memory, and bandwidth, as well as a cultural acceptance of technology in the fabric of our lives. This wasn't true a decade ago. Consideration should be to the ethical and technical responsibilities for the people responsible for the process.

In the past we had traditional database and analytics tools that couldn't deal with extremely large, messy, unstructured and fast moving data. Without going into too much detail, we now

have software like Hadoop and others which enable us to analyse large, messy and fast moving volumes of structured and unstructured data.

Let's look at some real examples of how big data is used today to make a difference:

- The FBI is combining data from social media, CCTV cameras, phone calls and texts to track down criminals and predict the next terrorist attack.
- Facebook is using face recognition tools to compare the photos you have up-loaded with those of others to find potential friends.
- Politicians are using social media analytics to determine where they have to campaign the hardest to win the next election.
- Video analytics and sensor data of Baseball or Football games is used to improve performance of players and teams.
- Google's self-driving car is analysing a gigantic amount of data from sensor and cameras in real time to stay on the road safely.
- The GPS information on where our phone is and how fast it is moving is now used to provide live traffic up-dates.
- Companies are using sentiment analysis of Facebook and Twitter posts to determine and predict sales volume and brand equity.
- Supermarkets are combining their loyalty card data with social media information to detect and leverage changing buying patterns. For example, it is easy for retailers to predict that a woman is pregnant simply based on the changing buying patterns. This allows them to target pregnant women with promotions for baby related goods.
- A hospital unit that looks after premature and sick babies is generating a live steam of every heartbeat. It then analyses the data to identify patterns. Based on the analysis the system can now detect infections 24hrs before the baby would show any visible symptoms, which allows early intervention and treatment.

Companies are barely starting to get to grips with the new world of big data.

Just think about it for a minute for why now Big Data and Data Science is needed:

- When you were reading a book in the past, no external data was generated. If you now use a Kindle or Nook device, they track what you are reading, when you are reading it, how often you read it, how quickly you read it, and so on.
- When you were listening to CDs in the past no data was generated. Now we listen to Music on your iPhone or digital music player.
- Today, most of us carry smart phones and they are constantly collecting and generating data by logging our location, tracking our speed, monitoring what apps we are using as well as who we are ringing or texting.
- Sensors are increasingly used to monitor and capture everything from temperature to power consumption, from ocean movements to traffic flows.
- Finally, combine all this now with the billions of internet searches performed daily, the billions of status updates, wall posts, comments and likes generated on Facebook each day, the 400+ million tweets sent on Twitter per day and the 72 hours of video uploaded to YouTube every minute.

What Is Datafication?

- ✓ The term "datafication" was introduced by Kenneth Cukier and Victor Mayer-Schöenberger in 2013 to refer to transforming invisible processes into data that companies can use to optimize their business.
- ✓ Datafication is an information technology-driven sense-making process.
- ✓ Datafication is a current technological trend that aims to transform most aspects of a business into quantifiable data that can be tracked, monitored, and analyzed. It refers to the use of tools and processes to turn an organization into a data-driven enterprise.

For example, we create data every time we talk on the phone, SMS, tweet, email, use Facebook, watch a video, withdraw money from an ATM, use a credit card, or even walk past a security camera.

Datafication process is actively used:

- o Insurance: Data used to update risk profile development and business models.
- o Banking: Data used to establish trustworthiness and likelihood of a person paying back a loan.
- o Human resources: Data used to identify e.g. employees risk-taking profiles.
- o Hiring and recruitment: Data used to replace personality tests.
- o Social science research: Datafication replaces sampling techniques and restructures the manner in which social science research is performed.

Examples:

And here could be many examples of datification.

Let's say social platforms, Facebook or Instagram, for example, collect and monitor data information of our friendships to market products and services to us and surveillance services to agencies which in turn changes our behaviour; promotions that we daily see on the socials are also the result of the monitored data.

Netflix Case:

Netflix, an internet streaming media provider, is a bright example of datafication process. It provides services in more than 40 countries and 33 million streaming members. Originally, operations were more physical in nature with its core business in mail order-based disc rental (DVD and Blu-ray). Simply said, the operating model was that the subscriber creates and maintains the queue (an ordered list) of media content that they want to rent (for example, a movie). If you limit the total number of disks, the contents can be stored for a long time, as the subscriber wishes. However, to rent a new disk, the subscriber sends the previous one back to Netflix, which then forwards the next available disk to the subscribers queue. Thus, the business goal of the disk rental model is to help people fill their turn. The model has changed and now Netflix is actively transforming their service into a smart one, actively using datafication processes.

There are three areas of business where datafication can really make an impact:

1. **Analytics:** In today's data-driven world, analytics is king. By collecting and analyzing data, businesses can gain valuable insights into consumer behavior, trends, and preferences, allowing them to make informed decisions that drive growth and success.
2. **Marketing Campaigns:** Marketing campaigns can be supercharged with datafication, allowing companies to personalize ads and offers for specific customers based on their interests and behaviors.
3. **Forecasting**: Predictive analytics can help businesses forecast future trends and stay ahead of the competition by anticipating changes in consumer demand.

What Makes Datafication the Way Forward for Businesses?
Before you formulate a datafication strategy, here are four considerations to keep in mind:

1. The Role of data in Decision-making and Strategy Development
In the current business landscape, datafication has the potential to fundamentally change the way companies make decisions and formulate strategies. Data-driven insights can provide businesses with valuable information about their operations, customers, and market trends

2. The Potential Benefits of Datafication for Businesses.
Businesses that embrace datafication can benefit in numerous ways, including increased efficiency, reduced costs, and enhanced revenue.

3. The Impact of Datafication on Customer Experience and Engagement.
Datafication has dramatically transformed how companies interact with their customers, enabling them to provide more personalized experiences and relevant content.

4. The Competitive Advantage of Data-Driven Companies
Companies that are data-driven have a significant competitive advantage over their peers. By leveraging data to make better decisions, optimize their operations, and deliver more personalized experiences to customers, these companies can create a level of sophistication that is challenging for competitors to replicate.

Current Applications of Datafication

Datafication is no longer just a buzzword because of its numerous applications across multiple industries that include:

Human Resource Management
Companies can gather data from mobile phones, social media, and apps to identify potential talents and analyze their characteristics, including their personalities and risk-taking profiles.

Customer Relationship Management
Enterprises that use customer data also benefit from using datafication tools and strategies to understand their clients. They can craft appropriate triggers relevant to their target audiences' buying behaviors and personalities.

Financial Service Provision
Insurance agencies employ datafication to understand a person's risk profile and update their business models.

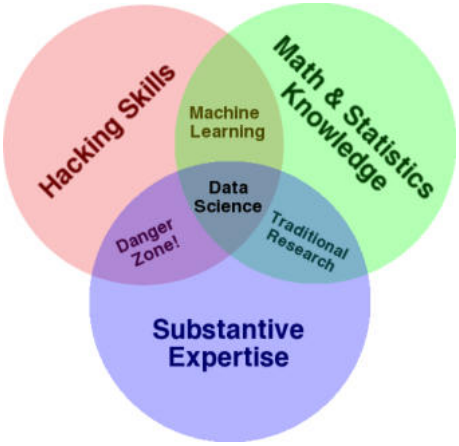
The Current Landscape (with a Little History)

Data Science is now widely accepted and adopted in various fields to solve different kinds of problems from decision making to automation through discovery, analysis, prediction and recommendation of the information.

According to Gartner, Data and Analytics trends for the future are:

- Dynamic storytelling
- Augmented Data Management
- Pervasive cloud management
- Convergence of data and analytics platforms

Driscoll then refers to Drew Conway's Venn diagram of data science from 2010, shown in Figure



Drew Conway's Venn diagram of data science

These basic things are:

- Math and Statistics
- Computer Programming
- Domain Knowledge

Data Science is in the middle of this Venn Diagram combining all these skills. The Data Science Venn Diagram gives the **visual representation** of how these areas work together in **Data Science**.

1. Hacking Skills

Hacking requires great **coding** skills. Coding is important because it helps you to gather and prepare the data because a lot of data is unstructured or present in unusual formats. You also require programming skills to apply statistics to your problems, handle the database, etc. One with hacking skills can apply very **complex** algorithms by computer programming.

2. Math and Statistics Knowledge

After collecting and preparing the data, now comes the part of extracting the insights from it. Mathematics is important for **analyzing** the data.

For analyzing the data, you will require several tools from mathematics such as **probability**, **algebra**, etc. It helps in the diagnosis of the problem by applying various mathematical and statistical approaches to your data.

3. Domain Expertise

For applying Data Science, you need to have an understanding of the right questions to ask to collect the data and discover insights from it.

Domain Expertise means the knowledge of the particular field in which you are working. It may be **business**, **healthcare**, **Finance**, **Education**, **etc**. You must know about the goals of that field, various methods, and constraints that you will be dealing with.

In the Data Science Venn Diagram, there are some areas that include the intersection of these skills which are **Machine Learning**, **Traditional Research**, and **Danger Zone**.

1. Machine Learning

According to the Data Science Venn Diagram, **Machine learning** involves the knowledge of Computer programming and Math but without any domain expertise.

2. Traditional Research
This area represents that you have the knowledge of **math**, **statistics** and you are an expert in your domain but do not know coding or programming.

3. Danger Zone
As the name suggests, it is the most dangerous area of the Data Science Venn diagram.
Danger Zone is the combination of coding and domain knowledge but without Math and Statistics. When Drew Conway proposed this Data Science Venn Diagram, he believed that this is the **rarest** case and is most unlikely to happen.

"Rise of the Data Scientist", which include:

- Statistics (traditional analysis you're used to thinking about)
- Data munging (parsing, scraping, and formatting data)
- Visualization (graphs, tools, etc.)

So major things shaping and accelerating the technology and software world that are helping businesses to make the best possible decisions in the shortest time are:

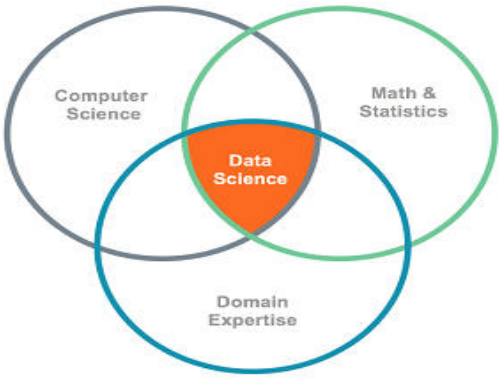
- Artificial Intelligence
- Cloud Computing
- Web/Mobile apps
- Agile framework

Almost all companies right from startups to huge enterprises have identified the need for Data science and/or Artificial Intelligence for solving the problems they are facing. Lot of POCs/MVPs are being created and results are shown to clients.

1. What is datafication?
Q2.Why is datafication important for the future of business?
Q3.What are some risks and challenges associated with datafication?
Q4.How can institutions ensure that students are comfortable with the use of location-based technology on campus?
Q5.Are there any privacy concerns related to using location-based technology in higher education?

Data science is the craft of turning data into action. Data is being generated and, perhaps more importantly, digitally captured at outstanding new levels. However, abundant data only represents potential value. It has to be mined, refined and harvested. Data science is the process of extracting information, understanding and learning from raw data to inform decision making in a proactive and systematic fashion that can be generalized. A key aspect of data science is the utilization of the scientific method to form and challenge hypotheses to validate conclusions about underlying patterns in data.

Data science is a powerful combination of various disciplines.



Computer Science Skills
- Programming
- Big data technologies

Math and Statistics Knowledge
- Machine learning
- Ensemble models
- Anomaly detection

Domain Expertise
- Business knowledge
- Expert systems
- User testing

Practicing data science requires the combining of a diverse set of skills. Data scientists need to be able to query and manipulate large swaths of data, so a strong *computer science* background is a must. Additionally, familiarity with *mathematics and statistics* help form a strong understanding of the algorithms commonly deployed and tuned. Combining a lot of computing power and sophisticated algorithms is called **data mining**. However, a major hazard of this approach is the potential to mistake noise for signal. *Domain expertise* is a helpful component in the verification of causal and logical relationships in models and conclusions.

In general, a data scientist needs to know more about statistics than an average programmer, more programming than an average statistician, and be able to apply both skills to solve business problems.

The overall objective of data science may seem straightforward, but implementation is a very complex process and involves a number of steps before the value of a data science product can be observed. Here's what that looks like:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. DS team evaluation
6. Stakeholder evaluation
7. Deployment

In the Modelling stage, a data scientist will look to apply statistical learning techniques (otherwise known as machine learning techniques or algorithms) to tease out details in the underlying raw data. **Machine learning** involves the utilization of statistical computing to understand tendencies, patterns, characteristics, attributes and structure in the underlying data, so as to inform decisions in some future state on new observations.

THE ROLE OF THE SOCIAL SCIENTIST IN DATA SCIENCE

Both LinkedIn and Facebook are social network companies. Oftentimes a description or definition of data scientist includes hybrid statistician, software engineer, and social scientist. This made sense in the context of companies where the product was a *social* product and still makes sense when we're dealing with human or user behavior.

Use of Data Science

1. Data science may detect patterns in seemingly unstructured or unconnected data, allowing conclusions and predictions to be made.
2. Tech businesses that acquire user data can utilise strategies to transform that data into valuable or profitable information.
3. Data Science has also made inroads into the transportation industry, such as with driverless cars. It is simple to lower the number of accidents with the use of driverless cars. For example, with driverless cars, training data is supplied to the algorithm, and the data is examined using data Science approaches, such as the speed limit on the highway, busy streets, etc.
4. Data Science applications provide a better level of therapeutic customisation through genetics and genomics research.

Data Science Jobs

Data Scientist

Data scientists are among the most recent analytical data professionals who have the technical ability to handle complicated issues as well as the desire to investigate what questions need to be answered.

data scientist may do the following tasks:

1. Discover patterns and trends in datasets to get insights.
2. Create forecasting algorithms and data models.
3. Improve the quality of data or product offerings by utilising machine learning techniques.
4. Distribute suggestions to other teams and top management.
5. In data analysis, use data tools such as R, SAS, Python, or SQL.
6. Top the field of data science innovations.

Data science offers you the opportunity to focus on and specialize in one aspect of the field. Here's a sample of different ways you can fit into this exciting, fast-growing field.

Data Scientist

- Job role: Determine what the problem is, what questions need answers, and where to find the data. Also, they mine, clean, and present the relevant data.
- Skills needed: Programming skills (SAS, R, Python), storytelling and data visualization, statistical and mathematical skills, knowledge of Hadoop, SQL, and Machine Learning.

The responsibilities you have to shoulder as a data scientist includes:

- Manage, mine, and clean unstructured data to prepare it for practical use.
- Develop models that can operate on Big Data
- Understand and interpret Big Data analysis
- Take charge of the data team and help them towards their respective goals
- Deliver results that have an impact on business outcomes

Data Analyst

- Job role: Analysts bridge the gap between the data scientists and the business analysts, organizing and analyzing data to answer the questions the organization poses. They take the technical analyses and turn them into qualitative action items.
- Skills needed: Statistical and mathematical skills, programming skills (SAS, R, Python), plus experience in data wrangling and data visualization.

As a data analyst, you will have to assume specific responsibilities, including:

- Collecting information from a database with the help of query
- Enable data processing and summarize results
- Use basic algorithms in their work like logistic regression, linear regression and so on
- Possess and display deep expertise in data munging, data visualization, exploratory data analysis and statistics

Data Engineer

- Job role: Data engineers focus on developing, deploying, managing, and optimizing the organization's data infrastructure and data pipelines. Engineers support data scientists by helping to transfer and transform data for queries.
- Skills needed: NoSQL databases (e.g., MongoDB, Cassandra DB), programming languages such as Java and Scala, and frameworks (Apache Hadoop).

Your responsibilities in this role are:

- Data Mining for getting insights from data
- Conversion of erroneous data into a useable form for data analysis
- Writing queries on data
- Maintenance of the data design and architecture
- Develop large data warehouses with the help of extra transform load (ETL)

Statistical Inference Definition

- ✓ Statistical Inference is defined as the procedure of analyzing the result and making conclusions from data based on random variation. The two applications of statistical inference are hypothesis testing and confidence interval.
- ✓ Statistical inference is the technique of making decisions about the parameters of a population that relies on random sampling.
- ✓ It enables us to assess the relationship between dependent and independent variables. The idea of statistical inference is to estimate the uncertainty or sample to sample variation.
- ✓ It enables us to deliver a range of value for the true value of something in the population.

The components used for making the statistical inference are:

- Sample Size
- Variability in the sample
- Size of the observed difference

The main types of statistical inference are:

- Estimation
- Hypothesis testing

Estimation

- ✓ Statistics from a sample are used to estimate population parameters.
- ✓ The most likely value is called a **point estimate**.
- ✓ There is **always** uncertainty when estimating.
- ✓ The uncertainty is often expressed as **confidence intervals** defined by a likely lowest and highest value for the parameter.

EX: could be a confidence interval for the number of bicycles a Dutch person owns:

"The average number of bikes a Dutch person owns is between 3.5 and 6."

Hypothesis Testing

Hypothesis testing is a method to check if a claim about a population is true. More precisely, it checks how likely it is that a hypothesis is true is based on the sample data.

examples of claims or questions that can be checked with hypothesis testing:

- 90% of Australians are left handed
- Is the average weight of dogs more than 40kg?
- Do doctors make more money than lawyers?

Importance of Statistical Inference

Inferential Statistics is important to examine the data properly. To make an accurate conclusion, proper data analysis is important to interpret the research results. It is majorly used in the future prediction for various observations in different fields. It helps us to make inference about the data. The statistical inference has a wide range of application in different fields, such as:

- Business Analysis
- Artificial Intelligence
- Financial Analysis
- Fraud Detection
- Machine Learning
- Share Market
- Pharmaceutical Sector

What is Statistical Inference Example

The Statistical Inference example is given below help you to understand the concepts clearly:

A bag containing 2 yellow balls, 3 red balls, and 5 black balls. Only one ball is drawn at random from the bag. What is the probability that the black ball is drawn?

Solution: Through statistical inference solution.
Total number of balls in a bag = 10

i.e 2 + 3+ 5 = 10
Number of black balls= 5

Probability of getting a black balls = Number of balck balls/ Total number of balls
= 5/10
= 1/2
Hence, the probability of getting black balls is 1/2.

Basic terminology of Statistics:

1. **Population** –
   It is actually a collection of set of individuals or objects or events whose properties are to be analyzed. In statistics, population is the entire set of items from which you draw data for a statistical study. It can be a group of individuals, a set of items, etc.
2. **Finite population**
   This is a type of population in which the number of elementary units is exactly quantifiable.
   Example- *Books in a university library.*
3. **Infinite population**
   In this type of population, The count of elementary units is not quantifiable to most certainty.
   Example- *Population of a country.*
4. **Real population**
   This is such a type of population that is mostly based on real-time data and the information is concrete and reliable. This population does not require approximation or hypothetical data.
   Example- *Employees working in a company.*
5. **Hypothetical population**
   This can be a finite or infinite imaginary population designed by a researcher.

Example- *Possible outcomes of a die if rolled 'n' times.*

- **Sample** –
  It is the subset of a population.

Sample
A part of the population drawn according to a rule or plan for concluding characteristics is called a sample.
Example-*Imagine an XYZ company that has around 50k employees. To do some analysis based on the information of these employees, It is practically difficult for researchers concerning time and money with all of 50k employees. The best possible way is to select 5k people (or any random number) from this population and collect the data from these employees to do the analysis.*

Sample size
The number of items in a sample is called a sample size. In the above example, Out of 50k employees, 5k was selected for analysis and that makes the sample size 5k.

Characteristics of the sample
A sample should follow certain characteristics to make it fit for data analysis.
1. Representativeness

A sample should represent the overall behavior of a population. Imagine the situation in the above example in which 5k employees are selected out of 50k employees.

2. Homogeneity

Homogeneity is nothing but the matching of behavior in multiple samples. If we derive multiple samples from a population, It is expected that all samples infer somewhat the same conclusions about the population.

Imagine if we want to calculate the mean salary of the 50 k employees and we have 3 samples each of a 5k sample size.

· ***Sample 1 has a mean salary of $40k***

· ***Sample 2 has a mean salary of 38k***

· ***Sample 3 has a mean salary of $41k***

We can say that these samples are homogeneous since all samples are giving approximately equal information regarding the salary of the employees.

3. Adequacy

The number of sampling units in a sample should be adequate for doing the research.

Samples are used when :

- The population is too large to collect data.
- The data collected is not reliable.
- The population is hypothetical and is unlimited in size.

Differences Between Population and Sample

- Now, try to understand what a sample and a population are, with the help of suitable examples.

Population	Sample
All residents of a country would constitute the Population set	All residents who live above the poverty line would be the Sample
All residents above the poverty line in a country would be the Population	All residents who are millionaires would make up the Sample
All employees in an office would be the Population	Out of all the employees, all managers in the office would be the Sample

Modeling

Data modeling is the process of producing a descriptive diagram of relationships between various types of information that are to be stored in a database.

What is a model?
Humans try to understand the world around them by representing it
in different ways. Architects capture attributes of buildings through blueprints and three-dimensional, scaled-down versions.

Types of Data Modeling

There are three main types of data models that organizations use.
1. Conceptual Model
It is a visual representation of database concepts and the relationships between them identifying the high-level user view of data.

2. Logical Model
This model further defines the structure of the data entities and their relationships. Usually, a logical data model is used for a specific project since the purpose is to develop a technical map of rules and data structures.

6. **Physical Model**

This is a schema or framework defining how data is physically stored in a database. It is used for database-specific 20odelling where the columns include exact types and attributes. A physical model designs the internal schema

Data Modeling Examples
The best way to picture a data model is to think about a building plan of an architect. An architectural building plan assists in putting up all subsequent conceptual models, and so does a data model.

7. **ER (Entity-Relationship) Model**

This model is based on the notion of real-world entities and relationships among them. It creates an entity set, relationship set, general attributes, and constraints.
Eg: an employee is an entity in an employee database.

2. Hierarchical Model
This data model arranges the data in the form of a tree with one root, to which other data is connected.

8. **Network Model**

This database model enables many-to-many relationships among the connected nodes. The data is arranged in a graph-like structure, and here 'child' nodes can have multiple 'parent' nodes.

9. **Relational Model**

This popular data model example arranges the data into tables. The tables have columns and rows, each 20odelling20g an attribute present in the entity.

10. **Object-Oriented Database Model**

This data model defines a database as an object collection, or recyclable software components, with related methods and features.

11. **Object-Relational Model**

This model is a combination of an object-oriented database model and a relational database model.

What is Statistical Modeling?

Statistical modeling is the process of describing the connections between variables in a dataset using mathematical equations and statistical approaches. In statistical modeling, we use a collection of statistical methods to investigate the connections between variables and uncover patterns in data.

(or)

Statistical modelling is an elaborate method of generating sample data and making real-world predictions using numerous statistical models and explicit assumptions. A mathematical link exists between random and non-random variables in this process.

Eg: statistical approaches such as regression analysis, we can determine the correlations between these factors and the number of passengers utilizing the railway route. For example, we might discover that the number of passengers is larger during rush hour and on weekdays, and fewer when it is raining.

There are three main types of statistical models, including:

- **Parametric:** Probability distributions with a finite number of parameters
- **Non-parametric:** The number and nature of parameters aren't fixed but flexible
- **Semi-parametric:** Have both parametric and non-parametric components

Where are statistical models used?

Statistical models are used in data science, machine learning, engineering, or operations research. These models have various real-world applications.

- **Time series analysis** involves investigating a series of data points that occur successively over time. It provides insights into factors that influence certain events from time to time.
- **Recommendation systems** predict a user's choice or preference for an item and the ratings they're likely to give.
- **Market segmentation** creates different market fragments based on potential buyers' needs, preferences, and priorities.
- **Association rule learning** enables the discovery of interesting relationships between variables in large databases.
- **Predictive Modelling** helps researchers predict the results or outcomes of an event, regardless of when it happens.
- **Scoring models** are based on logistic regression and decision trees. Investigators use them in combination with multiple algorithms to detect credit card fraud.
- **Clustering**, or a cluster model, groups items into a cluster so that there are more similarities within the group than other items across different groups.

Types of Statistical Models

There are several statistical models, each designed to solve a specific research issue or data format. Here are a few common types of statistical models and their applications:

1. **Linear regression models:** These models are used to represent the connection between a continuous result variable and one or more predictor variables. For example, depending on a person's height, age, and gender, a linear regression model may be used to estimate their weight.
2. **Logistic regression models:** Logistic regression models are used to represent the connection between a binary outcome variable (for example, yes/no) and one or more predictor variables. For example, depending on age, blood pressure, and cholesterol levels, a logistic regression model may be used to predict if a patient would have a heart attack.
3. **Time series models:** Time series models are used to model data that changes over time, such as stock prices, weather trends, or monthly sales numbers. These types of models may be applied to data to find trends, seasonal patterns, and other forms of temporal correlations.
4. **Multilevel models:** These models are used to model data having a hierarchical structure, such as pupils in schools or patients in hospitals. Multilevel models can be used to investigate how individual-level and group-level factors impact outcomes, as well as to account for the fact that people in the same group may be more similar to each other than those in different groups.
5. **Structural equation models:** These types of models are used to represent complicated interactions between several variables. Structural equation models can be used to evaluate ideas regarding causal links between variables and to quantify their strength and direction.
6. **Clustering models:** Clustering models are used to bring together comparable observations based on their similarities in terms of features. Clustering algorithms can be used to uncover patterns in data that would be difficult to detect using other approaches.

Statistical Modeling Techniques

statistics provides the framework and tools necessary for clear and effective scientific research. Statistics allows scientists to collect, analyze, and interpret data, enabling them to draw meaningful conclusions about the world around us.

Here are some of the techniques addressed under statistical modeling:

1. **Regression analysis:** Regression analysis is used to discover the connection between one or more independent variables and one or more dependent variables. It is used to forecast and determine the strength and direction of associations.
2. **Time series analysis:** Time series analysis is used to evaluate data that has been gathered over time. It is used to identify data trends, patterns, and seasonal fluctuations.
3. **Cluster analysis:** This technique is used to group comparable things or people together based on their characteristics. It's used to spot trends in data and categorize consumers or items.

4. **Survival analysis:** Survival analysis is used to assess time-to-event data, such as how long it takes for a patient to recover or how long it takes for a machine to break down. It is used to calculate the likelihood of an event occurring at a certain period.
5. **Decision trees:** Decision trees are used to simulate decisions and their repercussions. They are used to discover the most critical factors in a decision-making process and to find the best option based on the facts provided.
6. **Neural networks:** Neural networks are used to simulate complicated interactions between variables. They are used in image recognition, natural language processing, and predictive modeling, among other things.
7. **Factor analysis:** Factor analysis is used to reduce a large number of variables into a smaller number of components. It is used to find underlying dimensions or structures that explain the relationships between a group of variables.

Probability distributions

- ✓ Data Science has become one of the most popular interdisciplinary fields.
- ✓ It uses scientific approaches, methods, algorithms, and operations to obtain facts and insights from unstructured, semi-structured, and structured datasets.
- ✓ Organizations use these collected facts and insights for efficient production, business growth, and to predict user requirements.
- ✓ Probability distribution plays a significant role in performing data analysis equipping a dataset for training a model.

What Is Probability?

Probability denotes the possibility of something happening. It is a mathematical concept that predicts how likely events are to occur. The probability values are expressed between 0 and 1. The definition of probability is the degree to which something is likely to occur. This fundamental theory of probability is also applied to probability distributions.

What is Probability Distribution?
A Probability Distribution is a statistical method that determines all the probable values and possibilities that a random variable can deliver from a particular range. This range of values will have a lower bound and an upper bound, which we call the minimum and the maximum possible values.

Various factors on which plotting of a value depends are standard deviation, mean (or average), skewness, and kurtosis. All of these play a significant role in Data science as well.

General Properties of Probability Distributions
Probability distribution determines the likelihood of any outcome. The mathematical expression takes a specific value of x and shows the possibility of a random variable with p(x). Some general properties of the probability distribution are –

1. The total of all probabilities for any possible value becomes equal to 1.
2. In a probability distribution, the possibility of finding any specific value or a range of values must lie between 0 and 1.
3. Probability distributions tell us the dispersal of the values from the random variable. Consequently, the type of variable also helps determine the type of probability distribution.

Basic Terminology related to probability is as follows.

- **Experiment:** An activity whose outcomes are not known is an experiment. Every experiment has a few favorable outcomes and a few unfavorable outcomes. The historic experiments of Thomas Alva Edison had more than a thousand unsuccessful attempts before he could make a successful attempt to invent the light bulb.
- **Random Experiment:** A random experiment is an experiment for which the set of possible outcomes is known, but which particular outcome will occur on a particular execution of the experiment cannot be said prior to performing the experiment. Tossing a coin, rolling a die, and drawing a card from a deck are all examples of random experiments.
- **Trial:** The numerous attempts in the process of an experiment are called trials. In other words, any particular performance of a random experiment is called a trial. For example, tossing a coin is a trial.
- **Event:** A trial with a clearly defined outcome is an event. For example, getting a tail when tossing a coin is termed as an event.
- **Random Event:** An event that cannot be easily predicted is a random event. For such events, the probability value is very less.The formation of a rainbow during the rain is a random event.
- **Outcome:** This is the result of a trial. In the process of a sportsperson hitting a ball towards the goal post, there are two clear outcomes. He may either make the goal or miss the goal.
- **Possible Outcome:** The list of all the outcomes in an experiment can be referred to as possible outcomes. In tossing a coin, the possible outcomes are heads or tails.
- **Equally likely Outcomes:** An experiment in which each of the outcomes has an equal probability, such outcomes are referred to as equally likely outcomes. In the process of rolling a six-faced dice, the probability of getting any number is equal. P(any number)=16$\blacklozenge$(any number)=16
- **Sample Space:** It is the set of outcomes of all the trials in an experiment. On rolling a dice, the possible outcomes are 1, 2, 3, 4, 5, and 6. These outcomes make up the sample space. S = {1, 2, 3, 4, 5, 6}
- **Probable Event:** An event that can be predicted is called a probable event. We can calculate the probability of such events. The probability of a particular child being promoted to the next class can be calculated, hence, we can refer to this as a probable event.
- **Impossible Event:** An event that is not a part of the experiment, or which does not belong to the sample space of the outcomes of the experiment can be referred to as an impossible event. There is no snowfall in a temperate climatic region. Here, the snowfall can be referred to as an impossible event because the probability of occurrence of such an event is zero.

Common Data Types

Data analysts and data engineers have to deal with a broad spectrum of data, such as text, numerical, image, audio, voice, and many more. Each of these have a specific means to be represented and analyzed. Data in a probability distribution can either be discrete or continuous. Numerical data especially takes one of the two forms.

- **Discrete data:** They take specific values where the outcome of the data remains fixed. Like, for example, the consequence of rolling two dice or the number of overs in a T-20 match. In the first case, the result lies between 2 and 12. In the second case, the event will be less than 20.

  Continuous data:

It can obtain any value irrespective of bound or limit. Example: weight, height, any trigonometric value, age, etc.

Types of Probability(Statistical) distributions

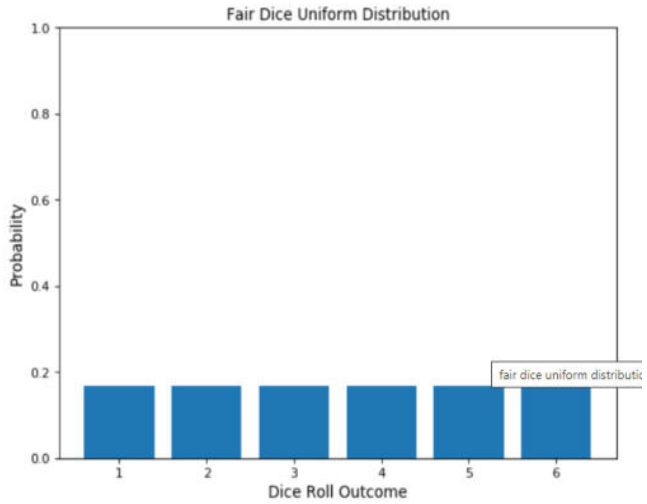
Depending on the type of data we use, we have grouped distributions into two categories, discrete distributions for discrete data (finite outcomes) and continuous distributions for continuous data (infinite outcomes).

Discrete distributions

Discrete uniform distribution: *All outcomes are equally likely*

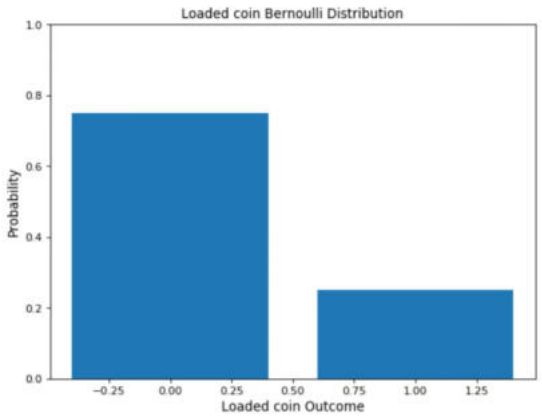
In statistics, uniform distribution refers to a statistical distribution in which all outcomes are equally likely. Consider rolling a six-sided die. You have an equal probability of obtaining all six numbers on your next roll, i.e., obtaining precisely one of 1, 2, 3, 4, 5, or 6, equaling a probability of 1/6, hence an example of a discrete uniform distribution.

As a result, the uniform distribution graph contains bars of equal height representing each outcome. In our example, the height is a probability of 1/6 (0.166667).



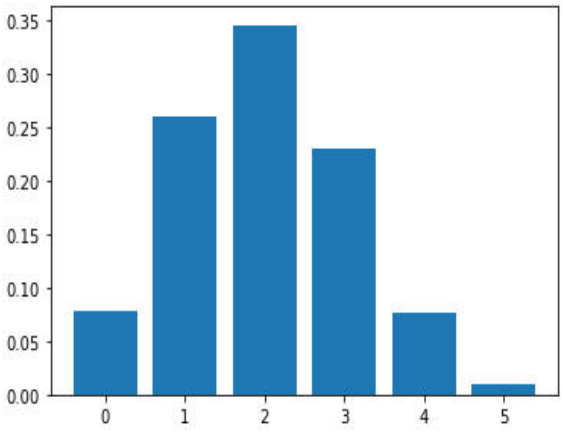
Uniform distribution is represented by the function U(a, b), where a and b represent the starting and ending values, respectively.

Bernoulli Distribution: *Single-Trial with Two Possible Outcomes*



The Bernoulli distribution is one of the easiest distributions to understand. It can be used as a starting point to derive more complex distributions. Any event with a single trial and only two possible outcomes follow a Bernoulli distribution. Flipping a coin or choosing between True and False in a quiz are examples of a Bernoulli distribution. They have a single trial and only two outcomes. Let's assume you flip a coin once; this is a single trail. The only two possible outcomes are either heads or tails. This is an example of a Bernoulli distribution.

Binomial Distribution

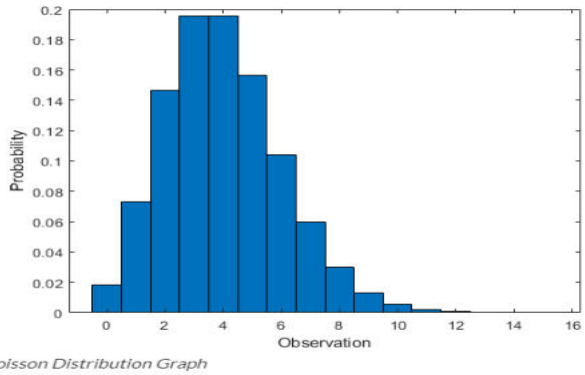


The binomial distribution is a discrete distribution with a finite number of possibilities. When observing a series of what are known as Bernoulli trials, the binomial distribution emerges. A Bernoulli trial is a scientific experiment with only two outcomes: success or failure.

Consider a random experiment in which you toss a biased coin six times with a 0.4 chance of getting head. If 'getting a head' is considered a 'success', the binomial distribution will show the probability of r successes for each value of r.
The binomial random variable represents the number of successes (r) in n consecutive independent Bernoulli trials.

Poisson Distribution: *The probability that an event May or May not occur*

Poisson distribution deals with the frequency with which an event occurs within a specific interval. Instead of the probability of an event, Poisson distribution requires knowing how often it happens in a particular period or distance. For example, a cricket chirps two times in 7 seconds on average. We can use the Poisson distribution to determine the likelihood of it chirping five times in 15 seconds.



Poisson Distribution Graph

The graph of Poisson distribution plots the number of instances an event occurs in the standard interval of time and the probability of each one.

A Poisson process is represented with the notation Po(λ), where λ represents the expected number of events that can take place in a period. The expected value and variance of a Poisson

process is λ. X represents the discrete random variable. A Poisson Distribution can be modeled using the following formula.

The main characteristics which describe the Poisson Processes are:

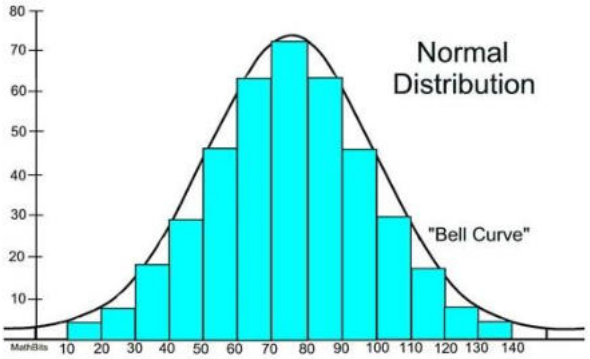
- The events are independent of each other.
- An event can occur any number of times (within the defined period).
- Two events can't take place simultaneously.

Continuous Distributions

Normal Distribution: *Symmetric Distribution of Values Around the Mean*

Normal distribution is the most used distribution in data science. In a normal distribution graph, data is symmetrically distributed with no skew. When plotted, the data follows a bell shape, with most values clustering around a central region and tapering off as they go further away from the center.

The normal distribution frequently appears in nature and life in various forms. For example, the scores of a quiz follow a normal distribution. Many of the students scored between 60 and 80 as illustrated in the graph below. Of course, students with scores that fall outside this range are deviating from the center.

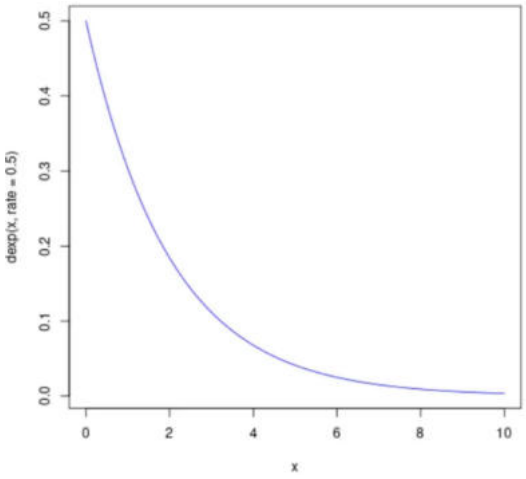


Normal Distribution Bell Curve Graph

Here, you can witness the "bell-shaped" curve around the central region, indicating that most data points exist there. The normal distribution is represented as N(μ, σ2) here, μ represents the mean, and σ2 represents the variance, one of which is mostly provided. The expected value of a normal distribution is equal to its mean.

Exponential Distribution

In a Poisson process, an exponential distribution is a continuous probability distribution that describes the time between events (success, failure, arrival, etc.).



For example, in physics, it is often used to measure radioactive decay; in engineering, to measure the time associated with receiving a defective part on an assembly line; and in finance, to measure the likelihood of the next default for a portfolio of financial assets.

The exponential distribution is commonly represented as Exp(λ), where λ is the distribution parameter, often called the rate parameter. We can find the value of λ by the formula = 1/μ, where μ is the mean. Here standard deviation is the same as the mean. Var (x) gives the variance = 1/λ2

Fitting a Model:

Model fitting is the measure of how well a machine learning model generalizes data similar to that with which it was trained. A **good model fit** refers to a model that accurately approximates the output when it is provided with unseen inputs.

Fitting refers to adjusting the parameters in the model to improve accuracy. The process involves running an algorithm on data for which the target variable ("labeled" data) is known to produce a machine learning model.

Overfitting and Underfitting

Overfitting negatively impacts the performance of the model on new data. It occurs when a model learns the details and noise in the training data too efficiently.
Reasons for Overfitting are as follows:

1. High variance and low bias
2. The model is too complex
3. The size of the training data

Techniques to reduce overfitting:

1. Increase training data.
2. Reduce model complexity.
3. Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
4. Ridge Regularization and Lasso Regularization
5. Use dropout for neural networks to tackle overfitting.

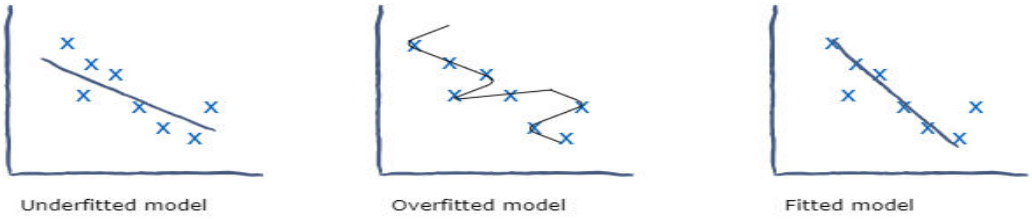
Underfitting happens when the machine learning model cannot sufficiently model the training data nor generalize new data. An underfit machine learning model is not a suitable model; this will be obvious as it will have a poor performance on the training data.

Reasons for Underfitting:

1. High bias and low variance
2. The size of the training dataset used is not enough.
3. The model is too simple.
4. Training data is not cleaned and also contains noise in it.

Techniques to reduce underfitting:

1. Increase model complexity
2. Increase the number of features, performing feature engineering
3. Remove noise from the data.
4. Increase the number of epochs or increase the duration of training to get better results.



Good Fit in a Statistical Model: Ideally, the case when the model makes the predictions with 0 error, is said to have a *good fit* on the data.

Important Frequently Asked Questions:

What is Data?
Data is "*facts*, such as numbers, words, measurements, observations and *statistics* collected together for reference or analysis."
Data — Information — Statistics
— Data is measurement of some kind that you are collecting. This is, 'raw unprocessed information'.

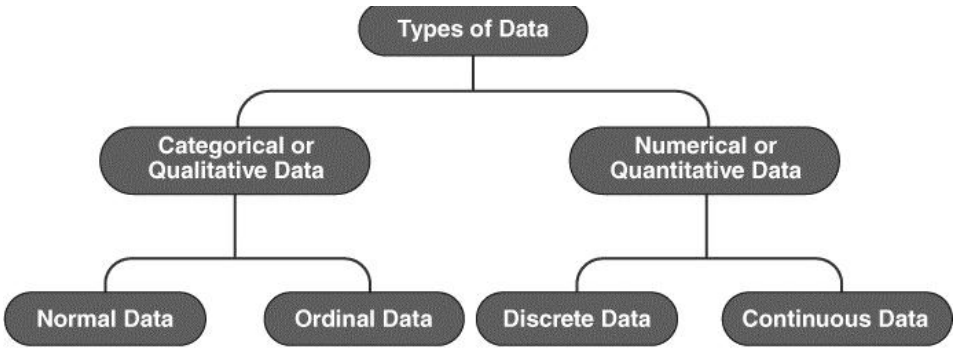
Why does Data Matter?
- ✓ Helps in understanding more about the data by identifying **relationships**.
- ✓ Helps in **predicting the future** or **forecast** based on the previous trend of data.
- ✓ Helps in determining **patterns** that may exist between data.
- ✓ Helps in detecting fraud by uncovering anomalies in the data.

Data matters a lot nowadays as we can infer important information from it.

Importance of Data

Today data is everywhere in every field. Whether you are a data scientist, marketer, businessman, data analyst, researcher, or you are in any other profession, you need to play or experiment with raw or structured data. This data is so important for us that it becomes important to handle and store it properly, without any error. While working on these data, it is important to know the **types of data** to process them and get the right results.

What are Types of Data in Data Science?



Qualitative or Categorical Data

Qualitative or Categorical Data is data that can't be measured or counted in the form of numbers Qualitative data tells about the perception of people. This data helps market researchers understand the customers' tastes and then design their ideas and strategies accordingly.

Nominal Data
Nominal Data is used to label variables without any order or quantitative value.
These data don't have any meaningful order; their values are distributed into distinct categories.

Examples of Nominal Data :

- ✓ Colour of hair (Blonde, red, Brown, Black, etc.)
- ✓ Marital status (Single, Widowed, Married)
- ✓ Nationality (Indian, German, American)
- ✓ Gender (Male, Female, Others)

Ordinal Data

Ordinal data have natural ordering where a number is present in some kind of order by their position on the scale.
ordinal data have some kind of order that is not present in nominal data.

Examples of Ordinal Data :

- ✓ When companies ask for feedback, experience, or satisfaction on a scale of 1 to 10
- ✓ Letter grades in the exam (A, B, C, D, etc.)
- ✓ Ranking of people in a competition (First, Second, Third, etc.)
- ✓ Economic Status (High, Medium, and Low)

Quantitative Data
Quantitative data can be expressed in numerical values, making it countable and including statistical data analysis. These kinds of data are also known as Numerical data. It answers the questions like “how much,” “how many,” and “how often.”

Discrete Data
The term discrete means distinct or separate. The discrete data contain the values that fall under integers or whole numbers.
These data are represented mainly by a bar graph, number line, or frequency table.

Examples of Discrete Data :

- ✓ Total numbers of students present in a class
- ✓ Cost of a cell phone
- ✓ Numbers of employees in a company
- ✓ The total number of players who participated in a competition

Continuous Data
Continuous data are in the form of fractional numbers. It can be the version of an android phone, the height of a person, the length of an object, etc.
continuous data stores the fractional numbers to record different types of data such as temperature, height, width, time, speed, etc.

Examples of Continuous Data :

- ✓ Height of a person
- ✓ Speed of a vehicle
- ✓ “Time-taken” to finish the work
- ✓ Wi-Fi Frequency

Quantitative and qualitative data provide different outcomes but are often used together to get the complete picture. Here are the differences between these two types of data:

Qualitative data	Quantitative data
Can't be measured.	Can be quantified and is measurable.
Can be quantified and is measurable.	The data is expressed as numbers and values.
The data describes qualities or characteristics.	The data is statistical and structured.
The data is nonstatistical and unstructured.	The data answers the questions "how much," "how many," or "how often"
The data can be collected using questionnaires, interviews, focus groups, or observation.	The data can be collected through instruments, tests, experiments, surveys, market reports, and metrics.
Examples include a person's name, hair color, and occupation.	Examples include age, height, and the number of visitors a website gets per day.

What is Data Science?
Data Science can be explained as the entire process of gathering actionable insights from raw data that involves various concepts that include statistical analysis, data analysis, machine learning algorithms, data modeling, preprocessing of data, etc.

1. Why is data science important for business?
Data Science assists businesses in monitoring, managing, and collecting performance metrics to improve decision-making throughout the organization. Trend analysis may help businesses make crucial decisions that will raise revenue, increase consumer involvement, and improve corporate performance.

2. What is the most important thing in data science?
The most crucial aspect of data science is that you can always learn new abilities, giving you a leg up on the competition with knowledge and experience.

3. What are the 3 main concepts of data science?
The three main concepts of data science are:

- Mathematical concepts like probability and linear algebra.
- SQL and NoSQL database languages.
- Concepts related to machine learning, such as supervised and unsupervised classification algorithms.

4. Why is data science the future?

- As more and more people connect to mobile devices, daily data quantities increase dramatically. Future growth will be impressive for technologies like IoT, AI, big data analytics, blockchain, and quantum computing.

Data Science vs Big Data: Basis of Information

What is the difference between big data and data science regarding where they get their information? The following are the basis of information for data science.

1. Internet users/traffic
2. Electronic apparatuses (sensors, RFID, etc.)
3. Live feeds and audio/video streams
4. Online message boards
5. Data produced by businesses (transactions, DB, spreadsheets, emails, etc.)
6. Information derived from system logs

The basis of information for big data are:

- Uses scientific methods to draw knowledge from large amounts of data
- Associated with data preparation, analysis, and filtering
- Identify intricate patterns in massive data and create models
- Programmers design working apps using developed models

Data Science vs Big Data: Application Areas

Application Areas of Data Science

1. Search engines use data science techniques to return the most relevant results for user queries quickly.
2. Data science algorithms are used across the board in digital marketing, from display banners to digital billboards. Rather than conventional advertisements, digital ads generally have higher click-through rates mostly because of this.
3. Recommender systems enhance the user experience. It also makes it easy to recognize suitable products from the billions of options available. This approach is used by many businesses to market their goods and ideas in line with what the customer wants and what information is pertinent. Based on the user's prior search results, recommendations are made.

Application Areas of Big Data

1. Big data is utilized in financial services. Retail banks, institutional investment banks, private finance management advisors, insurance companies, venture capitalists, and credit card companies use big data for their financial services. The major issue in these sectors is that multi-structured data is spread across in massive amounts in numerous dissimilar systems. With the help of big data, the problem can be resolved. Big data is applied in various ways, including customer, compliance, fraud, and operational analytics.
2. The main priorities for telecommunications service providers include expanding within existing subscriber bases, maintaining current consumers, and gaining new ones. The ability to aggregate and evaluate the vast amounts of user- and machine-generated data produced daily holds the key to solving these problems.
3. The key to remaining relevant and competitive is to understand your customers better. To do this, one must be able to examine the various data sources that businesses use daily, including blogs, consumer transaction data, social media, store-branded credit cards, and information from loyalty programs.

Data Science vs Big Data: Approach

1. Enhancing business agility
2. To become more competitive
3. Utilizing datasets for advantage in business
4. Identify reasonable metrics and ROI
5. To be sustainable
6. To understand markets better and attract new clients
7. Uses mathematics, statistics, and other tools
8. Modern methods and algorithms for data mining
9. Coding expertise (SQL, NoSQL) and Hadoop platforms
10. Acquiring, preparing, processing, publishing, preserving, or erasing data
11. Visualization of data, prediction

Probability Sampling Methods	**Non-probability Sampling Methods**
Probability Sampling is a sampling technique in which samples taken from a larger population are chosen based on probability theory.	Non-probability sampling method is a technique in which the researcher chooses samples based on subjective judgment, preferably random selection.
These are also known as Random sampling methods.	These are also called non-random sampling methods.
These are used for research which is conclusive.	These are used for research which is exploratory.
These involve a long time to get the data.	These are easy ways to collect the data quickly.
There is an underlying hypothesis in probability sampling before the study starts. Also, the objective of this method is to validate the defined hypothesis.	The hypothesis is derived later by conducting the research study in the case of non-probability sampling.

UNIT-I QUESTIONS:

1. What is Data Science and Big Data? Why Data Science is Hype now a days? Why not in the earlier.

2. What is Datafication? Describe the Current Landscape of Perspectives.

3. Express the statistical modeling and probability normal distribution

4. What is a model and mathematical model? What do you mean by fitting a model? Describe Overfitting.

5. Describe the Data Science Process.

6. What is Data Science? Explain the Life cycle of data science and its techniques with an example?

7. What is Big Data and its types? How Big Data and Data Science creates a hype in real world?

8. What is a model? Explain different types of Data Modeling with an example?

9. Briefly explain about probability distributions with an example?

UNIT - 2

DATA ANALYSIS AND ALGORITHMS: Exploratory Data Analysis (EDA), tools for EDA, The Data Science Process, role of data scientist's, case study. Algorithms: Machine Learning Algorithms, Three Basic Algorithms - Linear Regression - k-Nearest Neighbors (k-NN) - k-means – SVM, Naïve Bayes, Logistic Regression.

What is Exploratory Data Analysis?

Exploratory Data Analysis (EDA) is one of the techniques used for extracting vital features and trends used by machine learning and deep learning models in Data Science.

1. Exploratory Data Analysis (EDA) is an approach to analyzing and understanding data sets through statistical methods and visualizations.
2. EDA aims to uncover patterns, relationships, and insights in the data that might not be immediately obvious.
3. It is an iterative process that involves summarizing the main characteristics of a data set, creating visualizations to spot outliers and trends, and transforming the data to fit the requirements of more advanced models.
4. EDA is often the first step in the data analysis process and helps inform the development of more formal models.
5. It is an important step in the data science workflow as it provides insights into the quality of the data and any potential issues that need to be addressed before building more advanced models.
6. EDA is typically done using software tools such as R or Python, with popular libraries such as pandas, matplotlib, and seaborn.

Importance of EDA in Data Science

The Data Science field is now very important in the business world as it provides many opportunities to make vital business decisions by analyzing hugely gathered data.
A model built on such data results in sub-optimal performance.

Data set description

The dataset contains cases from the research carried out between the years 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

Attribute information:

1. Patient's age at the time of operation (numerical).
2. Year of operation (year — 1900, numerical).
3. A number of positive axillary nodes were detected (numerical).
4. Survival status (class attribute)
   1: the patient survived 5 years or longer post-operation.
   2: the patient died within 5 years post-operation.

Attributes 1, 2, and 3 form our features (independent variables), while attribute 4 is our class label (dependent variable).

Objective of Exploratory Data Analysis

The overall objective of exploratory data analysis is to obtain vital insights and hence usually includes the following sub-objectives:

- Identifying and removing data outliers
- Identifying trends in time and space
- Uncover patterns related to the target
- Creating hypotheses and testing them through experiments
- Identifying new sources of data

How to Perform Exploratory Data Analysis?

- Data specialists perform exploratory data analysis using popular scripting languages for statistics, such as Python and R. For effective EDA, data professionals also use a variety of BI (Business Intelligence) tools, including Qlik Sense, IBM Cognos, and Tableau.
- Python and R programming languages enable analysts to analyze data better and manipulate it using libraries and packages such as Plotly, Seaborn, or Matplotlib.
- BI tools, incorporating interactive dashboards, robust security, and advanced visualization features, provide data processors with a comprehensive view of data that helps them develop Machine Learning (ML) models.

The exploratory data analysis steps that analysts have in mind when performing EDA include:

- Asking the right questions related to the purpose of data analysis
- Obtaining in-depth knowledge about problem domains
- Setting clear objectives that are aligned with the desired outcomes.

Steps Involved In Exploratory Data Analysis (EDA)

The process of undertaking exploratory data analysis involves numerous steps:

- **Data Collection** – Data gathering is a crucial step in exploratory data analysis. It speaks of the method used to locate and transfer the information into our system. You can purchase trustworthy information from private companies or find it on various public websites. Websites like Kaggle, Github, the Deep Learning Repository, etc., are reliable sources for data acquisition.
- **Data Cleaning** – Data cleaning is the process of eliminating incorrect parameters and numbers from your dataset as well as other imperfections. Such abnormalities may unreasonably distort the data, which will hurt the outcomes. To clean data, actions like removing erroneous rows and columns, outliers, and missing values, and reformatting and re-indexing our data can be taken.
- **Missing Values** – The columns of the data contain some missing values. Three components of missing values dominate:
  - ✓ **These values are MCARs (Missing Fully at Random)** because they are completely random and independent of other discounts.
  - ✓ **These values are MAR (Missing at Random)**-dependent and depend on several further attributes.
  - ✓ **MNAR (Missing Not At Random)**: These values are missing for a purpose.
- **Outliners** – Two categories of outliers exist:

- ✓ **Outliers** in a single variable are statistics whose values deviate significantly from the normal distribution of values. In this case, only one variable is being taken into account.
- ✓ **Outliers with multiple variables**: These outliers rely on two variables' correlation. When charting data, one factor may not deviate significantly from the predicted range, but the values may be substantially different when the same variable is plotted alongside another variable.

- **Univariate Analysis** – You examine data with only one variable in univariate analysis. Your dataset's variables each correspond to a particular feature or column. You can accomplish this by locating precise mathematical values within the data using either graphical or non-graphical methods. Several visual techniques include:
  - ➢ Histograms are bar plots where the frequency of the data is shown as rectangle-shaped bars.
- **Box-plots**: In this case, the data is displayed as boxes.
  - ➢ **Bivariate Analysis** – In this case, you compare two variables. In this manner, you might discover how one property influences another. It is carried out using scatter plots, which show individual data points, or correlation matrices, which show the correlation as a color-coded graph. Boxplots provide a further option.

Exploratory Data Analysis is majorly performed using the following methods:

- Univariate visualization—provides summary statistics for each field in the raw data set
- Bivariate visualization—is performed to find the relationship between each variable in the dataset and the target variable of interest
- Multivariate visualization—is performed to understand interactions between different fields in the dataset
- Dimensionality reduction—helps to understand the fields in the data that account for the most variance between observations and allow for the processing of a reduced volume of data.

Eg: Retail

For example, an online store sells various types of footwear, such as sandals, sneakers, dress shoes, hiking boots, and formal shoes.

Exploratory data analysis can enable analysts to represent different sales trends graphically and visualize data related to best-selling product categories, buyer demographics and preferences, customer spending patterns, and units sold over a certain period.

Types of Exploratory Data Analysis

There are several types of exploratory data analysis (EDA)

1. Univariate Non-graphical
2. Univariate graphical
3. Multivariate Non-graphical
4. Multivariate graphical

Univariate Non-graphical

Univariate Non-graphical Exploratory Data Analysis (EDA) involves analyzing each variable in a dataset individually, without visualizations. The goal of univariate non-graphical EDA is to understand the distribution and characteristics of a single variable in the data.

Some common techniques used in univariate non-graphical EDA include:

- **Descriptive Statistics:** This involves calculating measures such as mean, median, mode, standard deviation, quartiles, and range to summarize the main characteristics of the data.
- **Frequency Tables:** This involves counting the number of occurrences of each unique value in a variable and creating a frequency table to summarize the data distribution.
- **Percentiles:** This involves dividing the data into 100 equal parts and calculating the value of each part, known as a percentile, which can be used to summarize the data distribution.
- **Z-Scores:** This involves standardizing the values in a variable by subtracting the mean and dividing by the standard deviation, which can be used to identify outliers in the data.

Univariate Graphical

Univariate Graphical Exploratory Data Analysis (EDA) involves visualizations to analyze each variable in a dataset individually. The goal of univariate graphical EDA is to understand the distribution and characteristics of a single variable in the data.

Some common visualizations used in univariate graphical EDA include:

- **Histograms:** A histogram is a bar graph representing a single variable's distribution by dividing the data into intervals (or bins) and counting the number of observations in each bin. Histograms can help identify the distribution's shape and highlight skewness, outliers, and patterns in the data.
- **Box Plots:** A box plot (or box-and-whisker plot) is a visualization that represents the distribution of a single variable by plotting the median, quartiles, and outliers in the data. Box plots can provide a compact distribution summary and highlight outliers and skewness in the data.
- **Density Plots:** A density plot is a smoothed histogram that estimates the probability density function of a single variable. Density plots can be used to visualize the distribution and highlight the shape of the data, including skewness and multi-modality.
- **Violin Plots:** A violin plot combines a box plot and a density plot that provides a more detailed visualization of the distribution of a single variable. Violin plots can provide information on the data's shape, spread, and skewness, as well as the presence of multiple modes.

Multivariate Non-Graphical

Multivariate Non-Graphical Exploratory Data Analysis (EDA) involves analyzing the relationship between multiple variables in a dataset without visualizations. The goal of multivariate non-graphical EDA is to understand how different variables in the data interact and influence one another.

Some common techniques used in multivariate non-graphical EDA include:

- **Correlation Matrix:** A correlation matrix is a table that shows the correlation between all pairs of variables in the data. Correlation measures the linear relationship between two variables and can be used to identify strongly or weakly associated variables.
- **Covariance Matrix:** A covariance matrix is a table that shows the covariance between all pairs of variables in the data. Covariance measures the joint variability between two variables and can be used to identify variables that change together.
- **Regression Analysis:** Regression analysis is a statistical method that involves fitting a line or curve to the data to model the relationship between a dependent variable and one or more independent variables. Regression analysis can estimate the strength and direction of the relationship between variables and make predictions about future values based on past observations.
- **ANOVA:** Analysis of Variance (ANOVA) is a statistical method that involves comparing the means of two or more data groups to determine if there is a significant difference between the groups. ANOVA can be used to identify variables that significantly affect the outcome and test hypotheses about the relationship between variables.

Multivariate Graphical

Multivariate Graphical Exploratory Data Analysis (EDA) involves analyzing the relationship between multiple variables in a dataset using visualizations. The goal of multivariate graphical EDA is to understand how different variables in the data interact and influence one another and to identify patterns and relationships that can inform further analysis and modeling.

Some common visualizations used in multivariate graphical EDA include:

- **Scatter Plots:** A scatter plot is a visualization that plots the values of two variables on a graph, with one variable on the x-axis and the other on the y-axis. Scatter plots can be used to visualize the relationship between variables, including the strength, direction, and shape of the relationship, as well as the presence of outliers and skewness.
- **Pair Plots:** A pair plot (or scatter plot matrix) is a visualization that plots all possible combinations of two variables in the data in a grid of scatter plots. Pair plots can provide a comprehensive overview of the relationships between all variables in the data, including the presence of correlations, outliers, and skewness.
- **Heat Maps:** A heat map is a visualization that represents the values of two or more variables in a grid of cells, with the color of each cell representing the value of a third variable. Heat maps can visualize the relationship between variables and identify patterns and correlations in the data.
- **Parallel Coordinates:** A parallel coordinates plot is a visualization that plots multiple variables in parallel lines on a graph, with each line representing an observation in the data. Parallel coordinates plots can be used to visualize the relationship between variables and to identify outliers and skewness in the data.

Exploratory Data Analysis Tools

Several tools can be used for Exploratory Data Analysis in data science, including:

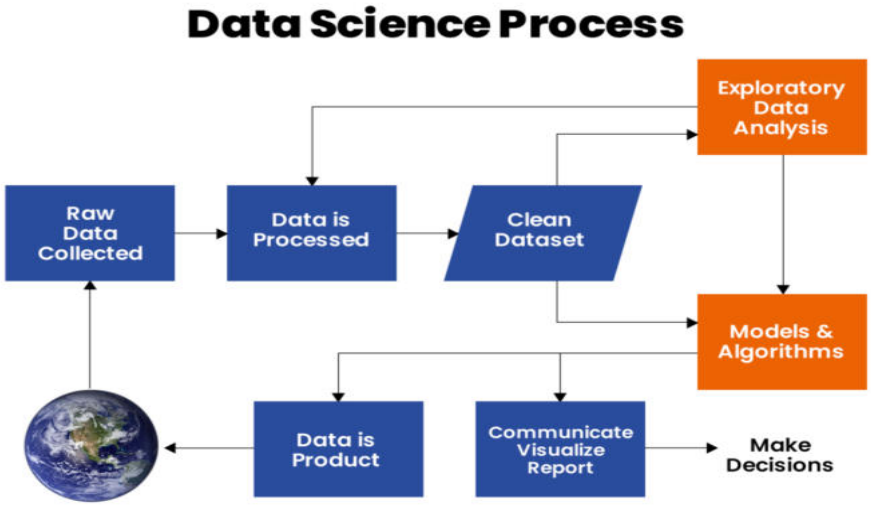
1. R: R is a widely used programming language for statistical computing and data analysis. It has many packages and libraries designed explicitly for EDA, including

the "tidyverse" collection of packages, which provides a comprehensive suite of tools for data manipulation, visualization, and analysis.

2. Python: Python is another widely used programming language for data science, with several libraries and packages specifically designed for EDA. Popular Python libraries for EDA include Pandas, Seaborn, Matplotlib, and Plotly, which provide data manipulation, visualization, and analysis tools.
3. Tableau: Tableau is a data visualization and business intelligence tool that provides an interactive and user-friendly interface for EDA. Tableau allows users to easily explore and visualize their data using a drag-and-drop interface and provides a wide range of visualizations and charts to support EDA.
4. QlikView: QlikView is a business intelligence and data visualization tool that provides interactive visualizations and dashboards for EDA. QlikView allows users to explore and analyze their data efficiently and offers a wide range of visualizations and charts to support EDA.

Exploratory Data Analysis for Data Science Process – Steps

Exploratory Data Analysis (EDA) plays a crucial role in data analytics and model-building. The following are the reasons for its importance and relevance:



1. **Understanding the Data:** EDA helps understand the data structure, variables, and relationships. It provides insights into the nature of the data, including the presence of outliers, skewness, and other data characteristics.
2. **Data Cleaning:** EDA helps to identify missing values, duplicate data, and other data quality issues. Cleaning the data is crucial for building accurate models, and EDA helps to identify areas that require cleaning.
3. **Data Visualization:** EDA allows for data visualization in various ways. This helps to understand the data's distribution, relationships, and patterns. Graphical representation of the data is a key component of EDA and provides insights into the data that might not be possible through other methods.
4. **Feature Selection:** EDA helps to identify the important features that should be used in the model-building process. It can reduce the number of features in the data and simplify the model, thus reducing the risk of overfitting.

5. **Model Validation:** EDA provides the basis for validating the model by allowing us to check if the model is correctly capturing the relationships in the data. It helps to identify the areas where the model is not working as expected and provides insights into how to improve the model.
6. **Communication:** EDA provides a way to communicate the insights and findings of the data analysis to stakeholders. By visualizing the data and summarizing the key insights, EDA provides a way to communicate complex data in a way that is easily understood by others.
   Overall, EDA is a crucial step in the data analytics and model-building process. It provides a foundation for building accurate models that can be used to make informed decisions.

Example Code of EDA in Python

EDA in Python

Here is an example of how Exploratory Data Analysis (EDA) can be performed using Python:

Import libraries
To start, you must import the necessary libraries, such as Pandas and Matplotlib.

- import pandas as pd
- import matplotlib.pyplot as plt

Load the data
Next, you will need to load the data into a Pandas DataFrame. For this example, we'll use the famous Iris dataset:

- iris = pd.read_csv('iris.csv')

Data Exploration
To explore the data, you can use various functions and methods provided by Pandas. For example, to see the first few rows of the data, you can use the head() method:

- print(iris.head())

Summary Statistics
To get summary statistics of the data, you can use the describe() method:

- print(iris.describe())

Univariate Analysis
To perform univariate analysis, you can plot histograms and box plots using the Matplotlib library. For example, to plot the histogram of the Sepal Length column, you can use the following code:

- plt.hist(iris['Sepal Length'])
- plt.xlabel('Sepal Length')

- plt.ylabel('Frequency')
- plt.title('Histogram of Sepal Length')
- plt.show()

Multivariate Analysis

You can use scatter plots and pair plots to perform multivariate analysis. For example, to plot a scatter plot between Sepal Length and Sepal Width, you can use the following code:

- plt.scatter(iris['Sepal Length'], iris['Sepal Width'])
- plt.xlabel('Sepal Length')
- plt.ylabel('Sepal Width')
- plt.title('Scatter Plot of Sepal Length vs Sepal Width')
- plt.show()

Data Scientist Roles and Responsibilities

A data scientist's job is to gather a large amount of data, analyze it, separate out the essential information, and then utilize tools like SAS, R programming, Python, etc. to extract insights that may be used to increase the productivity and efficiency of the business.

Data scientist roles and responsibilities include:

- ✓ Data mining or extracting usable data from valuable data sources
- ✓ Using machine learning tools to select features, create and optimize classifiers
- ✓ Carrying out preprocessing of structured and unstructured data
- ✓ Enhancing data collection procedures to include all relevant information for developing analytic systems
- ✓ Processing, cleansing, and validating the integrity of data to be used for analysis
- ✓ Analyzing large amounts of information to find patterns and solutions
- ✓ Developing prediction systems and machine learning algorithms
- ✓ Presenting results in a clear manner
- ✓ Propose solutions and strategies to tackle business challenges
- ✓ Collaborate with Business and IT teams

Roles & Responsibilities of a Data Scientist in terms of Field-wise

- **Management:** The Data Scientist plays an insignificant managerial role where he supports the construction of the base of futuristic and technical abilities within the Data and Analytics field in order to assist various planned and continuing data analytics projects.
- **Analytics:** The Data Scientist represents a scientific role where he plans, implements, and assesses high-level statistical models and strategies for application in the business's most complex issues.

- **Strategy/Design:** The Data Scientist performs a vital role in the advancement of innovative strategies to understand the business's consumer trends and management as well as ways to solve difficult business problems, for instance, the optimization of product fulfillment and entire profit.
- **Collaboration:** The role of the Data Scientist is not a solitary role and in this position, he collaborates with superior data scientists to communicate obstacles and findings to relevant stakeholders in an effort to enhance drive business performance and decision-making.
- **Knowledge:** The Data Scientist also takes leadership to explore different technologies and tools with the vision of creating innovative data-driven insights for the business at the most agile pace feasible.
- **Other Duties:** A Data Scientist also performs related tasks and tasks as assigned by the Senior Data Scientist, Head of Data Science, Chief Data Officer, or the Employer.

Data Scientist Skills

- Programming Skills – knowledge of statistical programming languages like R, Python, and database query languages like SQL, Hive, Pig is desirable. Familiarity with Scala, Java, or C++ is an added advantage.
- Statistics – Good applied statistical skills, including knowledge of statistical tests, distributions, regression, maximum likelihood estimators, etc. Proficiency in statistics is essential for data-driven companies.
- Machine Learning – good knowledge of machine learning methods like k-Nearest Neighbors, Naive Bayes, SVM, Decision Forests.
- Strong Math Skills (Multivariable Calculus and Linear Algebra) - understanding the fundamentals of Multivariable Calculus and Linear Algebra is important as they form the basis of a lot of predictive performance or algorithm optimization techniques.
- Data Wrangling – proficiency in handling imperfections in data is an important aspect of a data scientist job description.
- Experience with Data Visualization Tools like matplotlib, ggplot, d3.js., Tableau that help to visually encode data
- Excellent Communication Skills – it is incredibly important to describe findings to a technical and non-technical audience.
- Strong Software Engineering Background
- Hands-on experience with data science tools
- Problem-solving aptitude
- Analytical mind and great business sense
- Degree in Computer Science, Engineering or relevant field is preferred
- Proven Experience as Data Analyst or Data Scientist

Machine Learning Algorithms

Machine Learning is the science of making computers learn and act like humans by feeding data and information without being explicitly programmed.

Machine learning algorithms are trained with training data. When new data comes in, they can make predictions and decisions accurately based on past data.

Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.

How Does Machine Learning Work?

Machine Learning is, undoubtedly, one of the most exciting subsets of Artificial Intelligence. It completes the task of learning from data with specific inputs to the machine. It's important to understand what makes Machine Learning work and, thus, how it can be used in the future.

Features of Machine Learning:

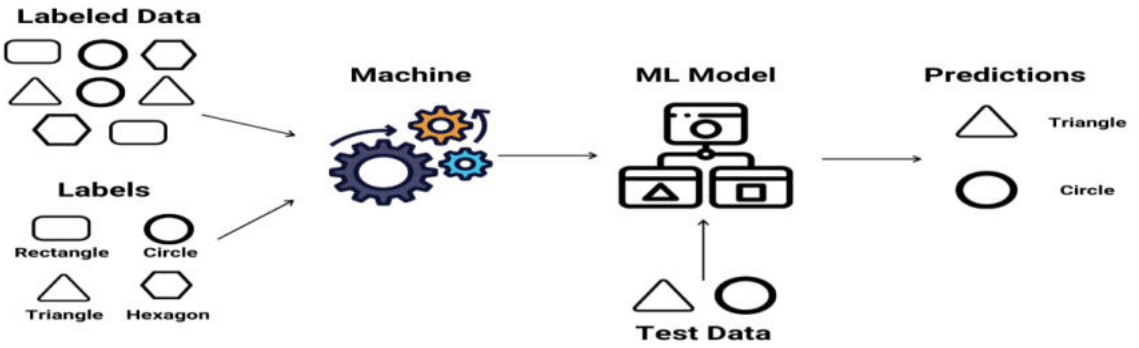
- Machine learning uses data to detect various patterns in a given dataset.
- It can learn from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.

There are two types of machine learning:

1. Supervised Learning
2. Unsupervised Learning

Supervised Learning

In supervised learning, we use known or labeled data for the training data. Since the data is known, the learning is, therefore, supervised, i.e., directed into successful execution. The input data goes through the Machine Learning algorithm and is used to train the model. Once the model is trained based on the known data, you can use unknown data into the model and get a new response.



The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is **spam filtering**.

Advantages of Supervised learning:

- With the help of supervised learning, the model can predict the output on the basis of prior experiences.
- In supervised learning, we can have an exact idea about the classes of objects.
- Supervised learning model helps us to solve various real-world problems such as **fraud detection, spam filtering**, etc.

Disadvantages of supervised learning:

- Supervised learning models are not suitable for handling the complex tasks.
- Supervised learning cannot predict the correct output if the test data is different from the training dataset.
- Training required lots of computation times.
- In supervised learning, we need enough knowledge about the classes of object.

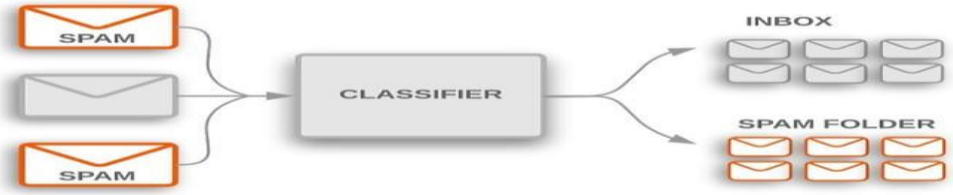
Types of supervised Machine learning Algorithms:

Supervised learning can be further divided into two types:

1. Classification
2. Regression

1. Classification - Supervised Learning

Classification is used when the output variable is categorical i.e. with 2 or more classes. For example, yes or no, male or female, true or false, etc.



In order to predict whether a mail is spam or not, we need to first teach the machine what a spam mail is. This is done based on a lot of spam filters - reviewing the content of the mail, reviewing the mail header, and then searching if it contains any false information.
Based on the content, label, and the spam score of the new incoming mail, the algorithm decides whether it should land in the inbox or spam folder.

Below are some popular Classification algorithms

- Random Forest
- Decision Trees
- Logistic Regression
- Support vector Machines

2. Regression - Supervised Learning

Regression is used when the output variable is a real or continuous value. In this case, there is a relationship between two or more variables i.e., a change in one variable is associated with a change in the other variable. For example, salary based on work experience or weight based on height, etc.

Regression is used when the output variable is a real or continuous value. In this case, there is a relationship between two or more variables i.e., a change in one variable is associated with a change in the other variable. For example, salary based on work experience or weight based on height, etc.

Let's consider two variables - humidity and temperature. Here, 'temperature' is the independent variable and 'humidity' is the dependent variable. If the temperature increases, then the humidity decreases.

Below are some popular Regression algorithms which come under supervised learning:

- Linear Regression
- Regression Trees
- Non-Linear Regression
- Bayesian Linear Regression
- Polynomial Regression

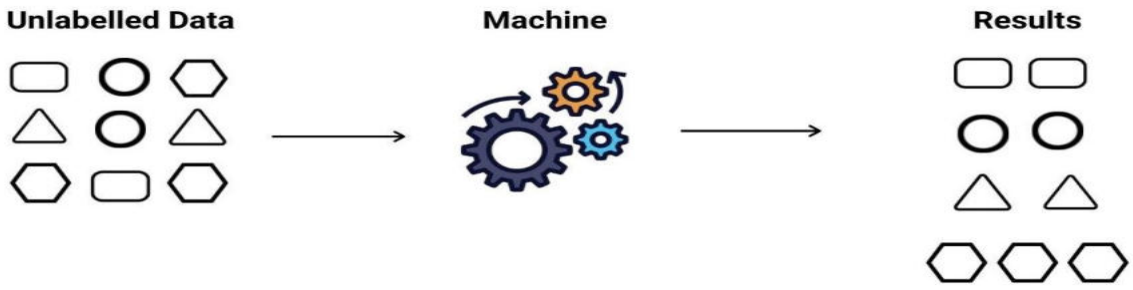
Real-Life Applications of Supervised Learning

- **Risk Assessment**
  Supervised learning is used to assess the risk in financial services or insurance domains in order to minimize the risk portfolio of the companies.
- **Image Classification**
  Image classification is one of the key use cases of demonstrating supervised machine learning. For example, Facebook can recognize your friend in a picture from an album of tagged photos.
- **Fraud Detection**
  To identify whether the transactions made by the user are authentic or not.
- **Visual Recognition**
  The ability of a machine learning model to identify objects, places, people, actions, and images.

What is Unsupervised Learning?

In Unsupervised Learning, the machine uses unlabeled data and learns on itself without any supervision. The machine tries to find a pattern in the unlabeled data and gives a response.

Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.



Use of Unsupervised Learning?

Below are some main reasons which describe the importance of Unsupervised Learning:

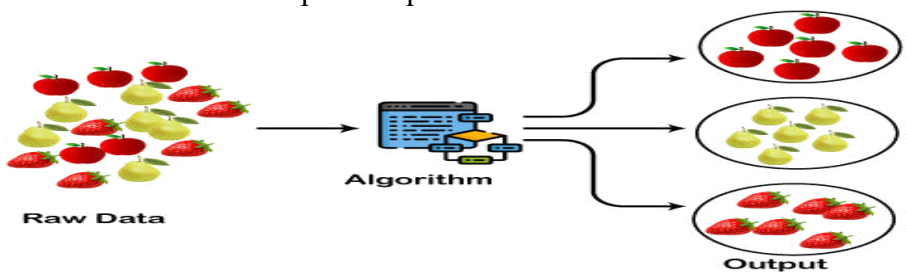
- Unsupervised learning is helpful for finding useful insights from the data.
- Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.
- Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.
- In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

Unsupervised learning can be further grouped into types:

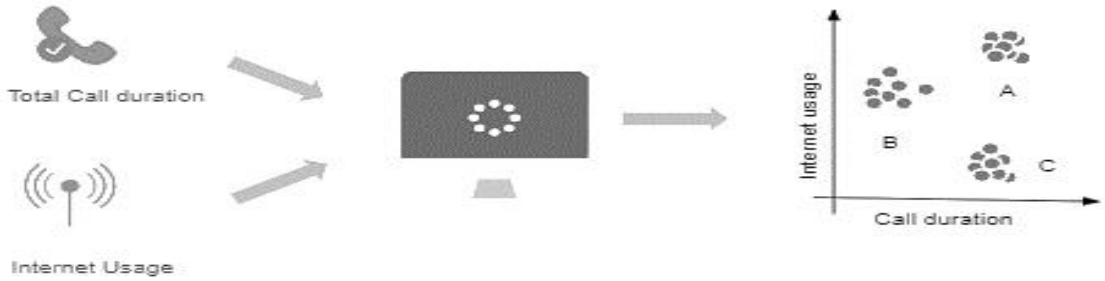
1. Clustering
2. Association

1. Clustering - Unsupervised Learning

Clustering is the method of dividing the objects into clusters that are similar between them and are dissimilar to the objects belonging to another cluster. For example, finding out which customers made similar product purchases.



Suppose a telecom company wants to reduce its customer churn rate by providing personalized call and data plans. The behavior of the customers is studied and the model segments the customers with similar traits. Several strategies are adopted to minimize churn rate and maximize profit through suitable promotions and campaigns.

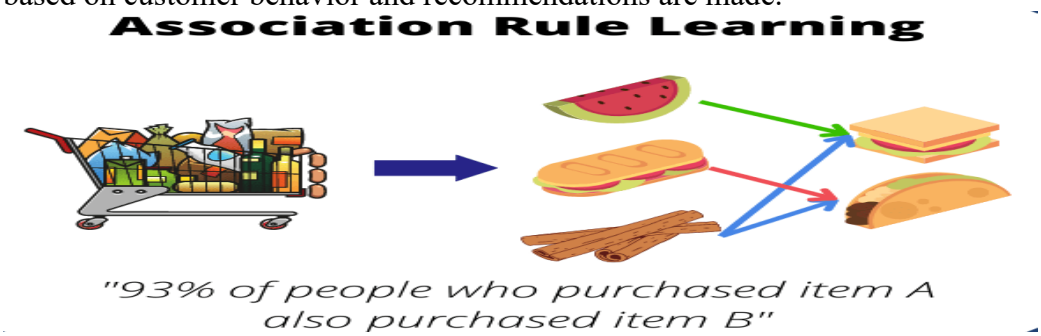


Group A customers use more data and also have high call durations. Group B customers are heavy Internet users, while Group C customers have high call duration. So, Group B will be given more data benefit plants, while Group C will be given cheaper called call rate plans and group A will be given the benefit of both.

2. Association - Unsupervised Learning

Association is a rule-based machine learning to discover the probability of the co-occurrence of items in a collection. For example, finding out which products were purchased together.

Let's say that a customer goes to a supermarket and buys bread, milk, fruits, and wheat. Another customer comes and buys bread, milk, rice, and butter. Now, when another customer comes, it is highly likely that if he buys bread, he will buy milk too. Hence, a relationship is established based on customer behavior and recommendations are made.



An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item. A typical example of Association rule is Market Basket Analysis.

Real-Life Applications of Unsupervised Learning

- **Market Basket Analysis**
  It is a machine learning model based on the algorithm that if you buy a certain group of items, you are less or more likely to buy another group of items.
- **Semantic Clustering**
  Semantically similar words share a similar context. People post their queries on websites in their own ways. Semantic clustering groups all these responses with the same meaning in a cluster to ensure that the customer finds the information they want quickly and easily. It plays an important role in information retrieval, good browsing experience, and comprehension.
- **Delivery Store Optimization**
  Machine learning models are used to predict the demand and keep up with supply. They are also used to open stores where the demand is higher and optimizing roots for more efficient deliveries according to past data and behavior.
- **Identifying Accident Prone Areas**
  Unsupervised machine learning models can be used to identify accident-prone areas and introduce safety measures based on the intensity of those accidents.

Unsupervised Learning algorithms:

Below is the list of some popular unsupervised learning algorithms:

- **K-means clustering**

- **KNN (k-nearest neighbors)**
- **Hierarchal clustering**
- **Anomaly detection**
- **Neural Networks**
- **Principle Component Analysis**
- **Independent Component Analysis**
- **Apriori algorithm**
- **Singular value decomposition**

Advantages of Unsupervised Learning

- Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.
- Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.

Disadvantages of Unsupervised Learning

- Unsupervised learning is intrinsically more difficult than supervised learning as it does not have corresponding output.
- The result of the unsupervised learning algorithm might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.

Major differences between Supervised and Unsupervised Learning

Parameters	**Supervised machine learning technique**	**Unsupervised machine learning technique**
Process	In a supervised learning model, input and output variables will be given.	In unsupervised learning model, only input data will be given
Input Data	Algorithms are trained using labeled data.	Algorithms are used against data which is not labeled
Algorithms Used	Support vector machine, Neural network, Linear and logistics regression, random forest, and Classification trees.	Unsupervised algorithms can be divided into different categories: like Cluster algorithms, K-means, Hierarchical clustering, etc.
Computational Complexity	Supervised learning is a simpler method.	Unsupervised learning is computationally complex
Use of Data	Supervised learning model uses training data to learn a link between the input and the outputs.	Unsupervised learning does not use output data.
Accuracy of Results	Highly accurate and trustworthy method.	Less accurate and trustworthy method.
Real Time Learning	Learning method takes place offline.	Learning method takes place in real time.
Number of Classes	Number of classes is known.	Number of classes is not known.

Parameters	**Supervised machine learning technique**	**Unsupervised machine learning technique**
Main Drawback	Classifying big data can be a real challenge in Supervised Learning.	You cannot get precise information regarding data sorting, and the output as data used in unsupervised learning is labeled and not known.

Difference between Parametric and Non-Parametric Methods are as follows:

Parametric Methods	**Non-Parametric Methods**
Parametric Methods uses a fixed number of parameters to build the model.	Non-Parametric Methods use the flexible number of parameters to build the model.
Parametric analysis is to test group means.	A non-parametric analysis is to test medians.
It is applicable only for variables.	It is applicable for both – Variable and Attribute.
It always considers strong assumptions about data.	It generally fewer assumptions about data.
Parametric Methods require lesser data than Non-Parametric Methods.	Non-Parametric Methods requires much more data than Parametric Methods.
Parametric methods assumed to be a normal distribution.	There is no assumed distribution in non-parametric methods.
Parametric data handles – Intervals data or ratio data.	But non-parametric methods handle original data.
Here when we use parametric methods then the result or outputs generated can be easily affected by outliers.	When we use non-parametric methods then the result or outputs generated cannot be seriously affected by outliers.
Parametric Methods can perform well in many situations but its performance is at peak (top) when the spread of each group is different.	Similarly, Non-Parametric Methods can perform well in many situations but its performance is at peak (top) when the spread of each group is the same.
Parametric methods have more statistical power than Non-Parametric methods.	Non-parametric methods have less statistical power than Parametric methods.

Parametric Methods	Non-Parametric Methods
As far as the computation is considered these methods are computationally faster than the Non-Parametric methods.	As far as the computation is considered these methods are computationally slower than the Parametric methods.
Examples: Logistic Regression, Naïve Bayes Model, etc.	Examples: KNN, Decision Tree Model, etc.

Regression

Regression is a tool that allows you to estimate how the dependent variable changes as the independent variable(s) change.
Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line.

Regression models can be used for many purposes:

- Evaluating the effect of an independent variable on a dependent variable.
- Forecasting future values of the dependent variable based on prior observations of both variables.

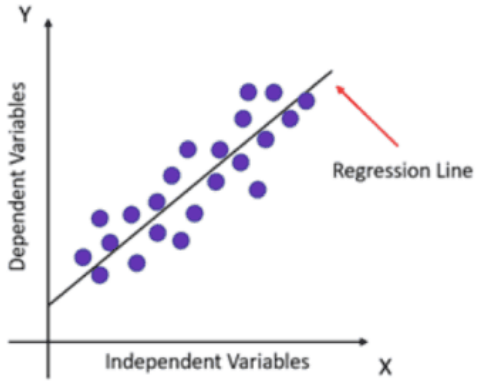
What is Regression Analysis?

- Regression analysis is one of the popular methods in data analysis that follows a controlled or supervised machine learning algorithm. It is an effective technique to identify and establish a relationship among variables in data.
- Regression analysis involves sorting out viable variables using mathematical strategies to draw highly accurate conclusions about those sorted variables.

What Is Simple Linear Regression?
Simple linear regression is a statistical method for establishing the relationship between two variables using a straight line. The line is drawn by finding the slope and intercept, which define the line and minimize regression errors.

The simplest form of simple linear regression has only one x variable and one y variable. The x variable is the independent variable because it is independent of what you try to predict the dependent variable. The y variable is the dependent variable because it depends on what you try to predict.

Linear Regression Equation



Linear regression can be expressed mathematically as:

$y= \beta 0+ \beta\ 1x+ \varepsilon$

Here,

- Y= Dependent Variable
- X= Independent Variable
- β 0= intercept of the line
- β1 = Linear regression coefficient (slope of the line)
- ε = random error

The last parameter, random error ε, is required as the best fit line also doesn't include the data points perfectly.

Linear Regression Equation

Let's consider a dataset that covers RAM sizes and their corresponding costs.

In this case, the dataset comprises two distinct features: memory (capacity) and cost. The more RAM, the more the purchase cost of RAMs.

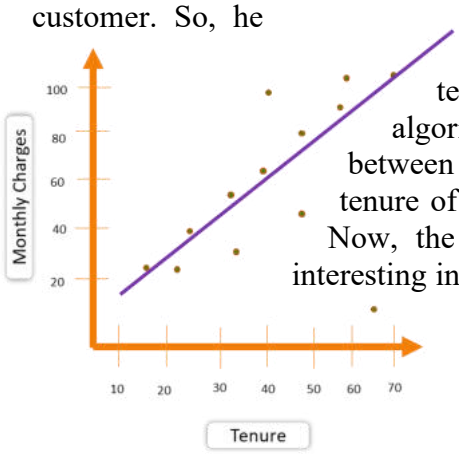
Ram Capacity	Cost
2 GB	$12
4 GB	$16
8 GB	$28
16 GB	$62

Dataset: RAM Capacity vs. Cost

If we plot RAM on the X-axis and its cost on the Y-axis, a line from the lower-left corner of the graph to the upper right represents the relationship between X and Y. On plotting these data points on a scatter plot, we get the following graph:

Mathematically these slant lines follow the following equation,
Y = m*X + b
Where X = dependent variable (target)
Y = independent variable
m = slope of the line (slope is defined as the 'rise' over the 'run')

Example:
Let's assume there is a telecom network called Neo. Its delivery manager wants to find out if there's a relationship between the monthly charges of a customer and the tenure of the customer. So, he collects all customer data and implements linear regression by taking monthly charges as the dependent variable and tenure as the independent variable. After implementing the algorithm, what he understands is that there is a relationship between the monthly charges and the tenure of a customer. As the tenure of the customer increases, the monthly charges also increase. Now, the best-fit line helps the delivery manager find out more interesting insights from the data. With this, he can predict the value of y for every new value of x.



What is Multivariate Regression?

Multivariate is a controlled or supervised Machine Learning algorithm that analyses multiple data variables. It is a continuation of multiple regression that involves one dependent variable and many independent variables. The output is predicted based on the number of independent variables.

Example: Consider the task of calculating blood pressure. In this case, height, weight, and amount of exercise can be considered independent variables. Here, we can use multiple linear regression to analyze the relationship between the three independent variables and one dependent variable, as all the variables considered are quantitative.
There are numerous areas where multivariate regression can be used. Let's look at some examples to understand multivariate regression better.

1. Praneeta wants to estimate the price of a house. She will collect details such as the location of the house, number of bedrooms, size in square feet, amenities available, or not. Basis these details price of the house can be predicted and how each variables are interrelated.
2. An agriculture scientist wants to predict the total crop yield expected for the summer. He collected details of the expected amount of rainfall, fertilizers to be used, and soil conditions. By building a Multivariate regression model scientists can predict his crop yield. With the crop yield, the scientist also tries to understand the relationship among the variables.

Some key points about MLR:

- For MLR, the dependent or target variable(Y) must be the continuous/real, but the predictor or independent variable may be of continuous or categorical form.
- Each feature variable must model the linear relationship with the dependent variable.
- MLR tries to fit a regression line through a multidimensional space of data-points.

- The equation for a model with two input variables can be written as:
- **$y = \beta0 + \beta1.x1 + \beta2.x2$**
- What if there are three variables as inputs? Human visualizations can be only three dimensions. In the machine learning world, there can be n number of dimensions. The equation for a model with three input variables can be written as:
- **$y = \beta0 + \beta1.x1 + \beta2.x2 + \beta3.x3$**

Assumptions for Multiple Linear Regression:

- ✓ A **linear relationship** should exist between the Target and predictor variables.
- ✓ The regression residuals must be **normally distributed**.
- ✓ MLR assumes little or **no multicollinearity** (correlation between the independent variable) in data.

- ✓ The dependent variable is nominal or ordinal. The nominal variables have two or more categories without any meaningful organization. Ordinal variables can also have two or more categories, but they have a structure and can be ranked.
- ✓ There can be single or multiple independent variables that can be ordinal, continuous, or nominal. Continuous variables are those that can have infinite values within a specific range.
- ✓ The dependent variables are mutually exclusive and exhaustive.
- ✓ The independent variables do not have a strong correlation among themselves.

Advantages of Multivariate Regression

1. Multivariate regression helps us to study the relationships among multiple variables in the dataset.
2. The correlation between dependent and independent variables helps in predicting the outcome.
3. It is one of the most convenient and popular algorithms used in machine learning.

Disadvantages of Multivariate Regression

- The complexity of multivariate techniques requires complex mathematical calculations.
- It is not easy to interpret the output of the multivariate regression model since there are inconsistencies in the loss and error outputs.
- Multivariate regression models cannot be applied to smaller datasets; they are designed for producing accurate outputs when it comes to larger datasets.

Implementation of Multiple Linear Regression model:
To implement MLR using Python, we have below problem:

Problem Description:

We have a dataset of **50 start-up companies**. This dataset contains five main information: **R&D Spend, Administration Spend, Marketing Spend, State, and Profit for a financial year**. Our goal is to create a model that can easily determine which company has a maximum profit, and which is the most affecting factor for the profit of a company.

Since we need to find the Profit, so it is the dependent variable, and the other four variables are independent variables.

What is the Classification Algorithm?

The 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. Such as, **Yes or No, 0 or 1, Spam or Not Spam, cat or dog,** etc. Classes can be called as targets/labels or categories.

In classification algorithm, a discrete output function(y) is mapped to input variable(x).

y=f(x), where y = categorical output

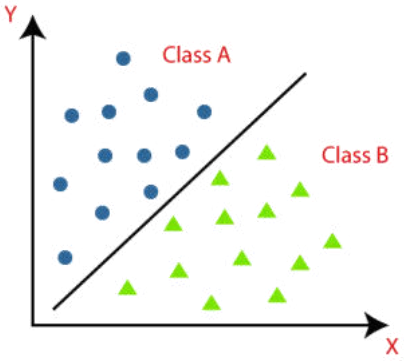
In simple words, classification is a type of pattern recognition in which classification algorithms are performed on training data to discover the same pattern in new data sets.
In simple words, classification machine learning algorithms allow us to assign a label or category to a piece of data.

These labels are known as Classes.

- It can be done by analyzing the properties of each instance (or object) fed into the system for classification.
- The entire process involves supervised learning, i.e., where the objects with known properties (or labels) are used to train a model for future predictions. Once trained, this model can then be used to classify new instances.

Classification in Machine Learning: Terminologies:

Classifier: An algorithm that maps the input variable into a specific class.
Feature: A metric or a measurable property of the scenario selected.
Initialize: Assigning the classifier used.
Classification Model: A model that categorizes the input data into two or more discrete groups.
Evaluate: Evaluation of the model by finding the accuracy score and classification report.
Classification algorithms can be better understood using the below diagram. In the below diagram, there are two classes, class A and Class B. These classes have features that are similar to each other and dissimilar to other classes.



There are two types of Classifications:

Binary Classifier: If the classification problem has only two possible outcomes, then it is called as Binary Classifier.

Examples: YES or NO, MALE or FEMALE, SPAM or NOT SPAM, CAT or DOG, etc.

Multi-class Classifier: If a classification problem has more than two outcomes, then it is called as Multi-class Classifier.
Example: Classifications of types of crops, Classification of types of music.

Learners in Classification Problems

There are two types of learners.

1. **Lazy Learners**
   It first stores the training dataset before waiting for the test dataset to arrive. When using a lazy learner, the classification is carried out using the training dataset's most appropriate data. Less time is spent on training, but more time is spent on predictions. Some of the examples are case-based reasoning and the KNN algorithm.
2. **Eager Learners**
   Before obtaining a test dataset, eager learners build a classification model using a training dataset. They spend more time studying and less time predicting. Some of the examples are ANN, naive Bayes, and Decision trees.

Types of ML Classification Algorithms:

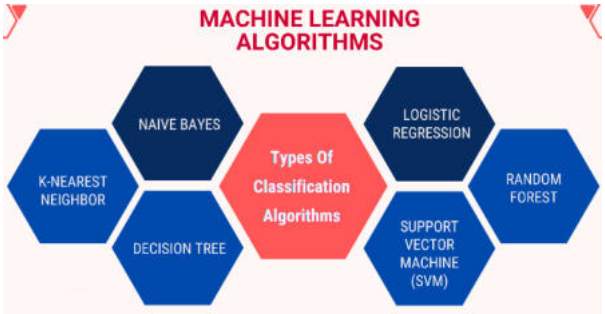
Classification Algorithms can be further divided into the Mainly two category:

1. Linear Models

- ✓ Logistic Regression
- ✓ Support Vector Machines

2. Non-linear Models

- ✓ K-Nearest Neighbours
- ✓ Kernel SVM
- ✓ Naïve Bayes
- ✓ Decision Tree Classification
- ✓ Random Forest Classification



What is Logistic Regression?

Logistic regression is a statistical method that is used for building machine learning models where the dependent variable is dichotomous: i.e. binary. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables. The independent variables can be nominal, ordinal, or of interval type.

The name "logistic regression" is derived from the concept of the logistic function that it uses. The logistic function is also known as the sigmoid function. The value of this logistic function lies between zero and one.

The goal of Logistic Regression is to discover a link between characteristics and the likelihood of a specific outcome. For example, when predicting whether a student passes or fails an exam based on the number of hours spent studying, the response variable has two values: pass and fail.

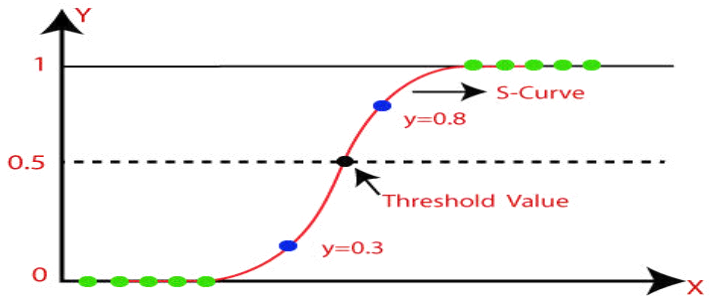
A Logistic Regression model is similar to a Linear Regression model, except that the Logistic Regression utilizes a more sophisticated cost function, which is known as the "Sigmoid function" or "logistic function" instead of a linear function.

It's called 'Logistic Regression' since the technique behind it is quite similar to Linear Regression. The name "Logistic" comes from the Logit function, which is utilized in this categorization approach.

Logistic Regression is considered a regression model also. This model creates a regression model to predict the likelihood that a given data entry belongs to the category labeled "1."

Logistic regression models the data using the sigmoid function, much as linear regression assumes that the data follows a linear distribution.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Example-1: There is a dataset given which contains the information of various users obtained from the social networking sites. There is a car making company that has recently launched a new SUV car. So the company wanted to check how many users from the dataset, wants to purchase the car.

- For this problem, we will build a Machine Learning model using the Logistic regression algorithm. The dataset is shown in the below image. In this problem, we will predict the **purchased variable (Dependent Variable)** by using **age and salary (Independent variables)**.

Example-2: In medicine, for example, a frequent application is to find out which variables have an influence on a disease. In this case, 0 could stand for "not diseased" and 1 for "diseased". Subsequently, the influence of age, gender and smoking status (smoker or not) on this particular disease could be examined.

Business example:

For an online retailer, you need to predict which product a particular customer is most likely to buy. For this, you receive a data set with past visitors and their purchases from the online retailer.

Medical example:
You want to investigate whether a person is susceptible to a certain disease or not. For this purpose, you receive a data set with diseased and non-diseased persons as well as other medical parameters.

Political example:
Would a person vote for party A if there were elections next weekend?

Type of Logistic Regression:
On the basis of the categories, Logistic Regression can be classified into three types:

1. Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.

2. Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"

3. Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Advantages of the Logistic Regression Algorithm

- Logistic regression performs better when the data is linearly separable
- It does not require too many computational resources as it's highly interpretable
- There is no problem scaling the input features—It does not require tuning
- It is easy to implement and train a model using logistic regression
- It gives a measure of how relevant a predictor (coefficient size) is, and its direction of association (positive or negative)

Steps in Logistic Regression: To implement the Logistic Regression using Python, we will use the same steps as we have done in previous topics of Regression. Below are the steps:

1. Data Pre-processing step
2. Fitting Logistic Regression to the Training set
3. Predicting the test result
4. Test accuracy of the result(Creation of Confusion matrix)
5. Visualizing the test set result.

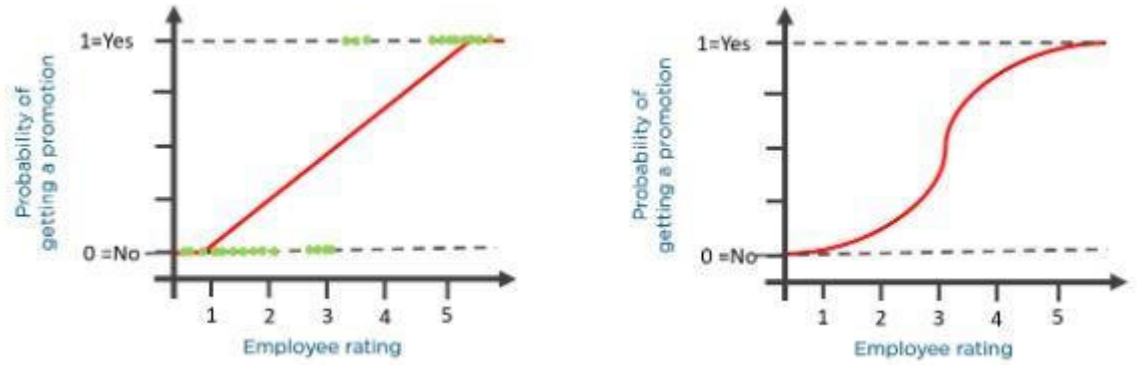
Applications of Logistic Regression

1. Using the logistic regression algorithm, banks can predict whether a customer would default on loans or not
2. To predict the weather conditions of a certain place (sunny, windy, rainy, humid, etc.)
3. Ecommerce companies can identify buyers if they are likely to purchase a certain product
4. Companies can predict whether they will gain or lose money in the next quarter, year, or month based on their current performance
5. To classify objects based on their features and attributes

Does the Logistic Regression Algorithm Work?
Consider the following example: An organization wants to determine an employee's salary increase based on their performance.

For this purpose, a linear regression algorithm will help them decide. Plotting a regression line by considering the employee's performance as the independent variable, and the salary increase as the dependent variable will make their task easier.

Now, what if the organization wants to know whether an employee would get a promotion or not based on their performance? The above linear graph won't be suitable in this case. As such, we clip the line at zero and one, and convert it into a sigmoid curve (S curve).



Based on the threshold values, the organization can decide whether an employee will get a salary increase or not.

Linear Regression vs. Logistic Regression

Linear Regression	Logistic Regression
Used to solve regression problems	Used to solve classification problems
The response variables are continuous in nature	The response variable is categorical in nature
It helps estimate the dependent variable when there is a change in the independent variable	It helps to calculate the possibility of a particular event taking place

It is a straight line	It is an S-curve (S = Sigmoid)

K-Nearest Neighbor(KNN) Algorithm for Machine Learning

K-Nearest Neighbors algorithm in Machine Learning (or KNN) is one of the most used learning algorithms due to its simplicity. KNN is a lazy learning, non-parametric algorithm. It uses data with several classes to predict the classification of the new sample point. KNN is non-parametric since it doesn't make any assumptions on the data being studied, i.e., the model is distributed from the data.

Working of KNN Algorithm

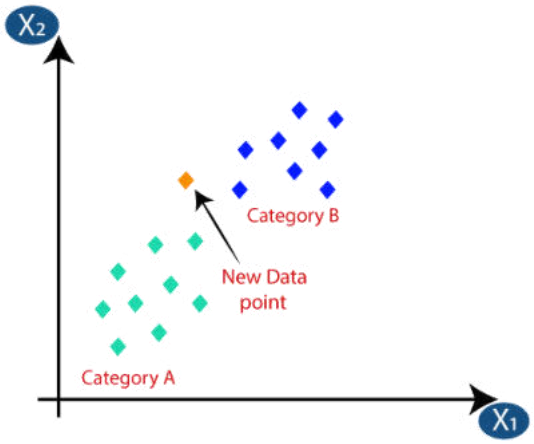
K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. We can understand its working with the help of following steps −

Step 1 − For implementing any algorithm, we need dataset. So during the first step of KNN, we must load the training as well as test data.
Step 2 − Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.
Step 3 − For each point in the test data do the following −

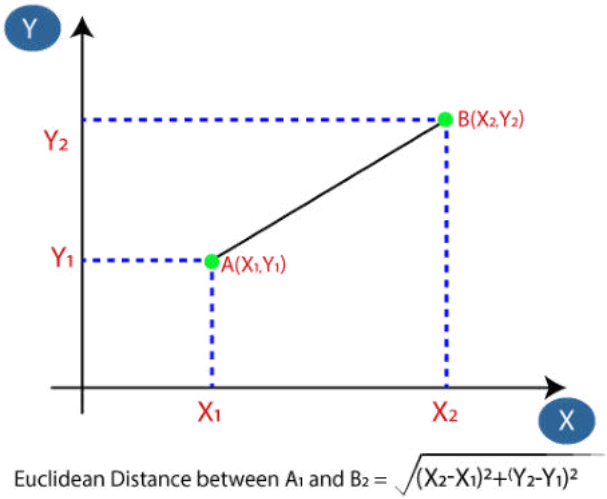
- **3.1** − Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.
- **3.2** − Now, based on the distance value, sort them in ascending order.
- **3.3** − Next, it will choose the top K rows from the sorted array.
- **3.4** − Now, it will assign a class to the test point based on most frequent class of these rows.

Step 4 − End
Example
The following is an example to understand the concept of K and working of KNN algorithm –

Suppose we have a new data point and we need to put it in the required category. Consider the below image:

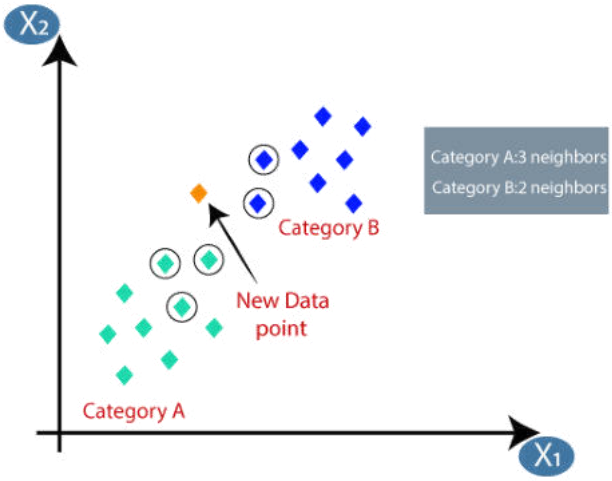


- Firstly, we will choose the number of neighbors, so we will choose the k=5.
- Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



Euclidean Distance between A_1 and $B_2 = \sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$

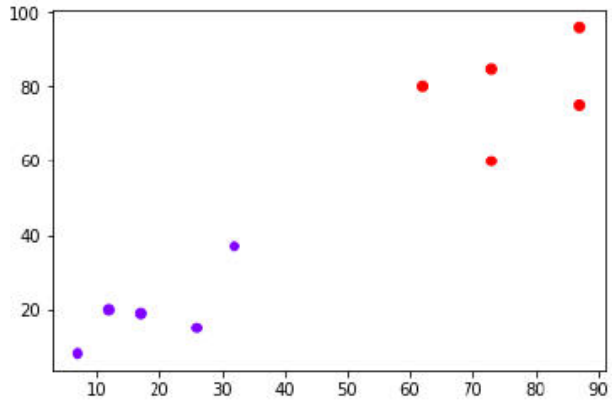
- By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:



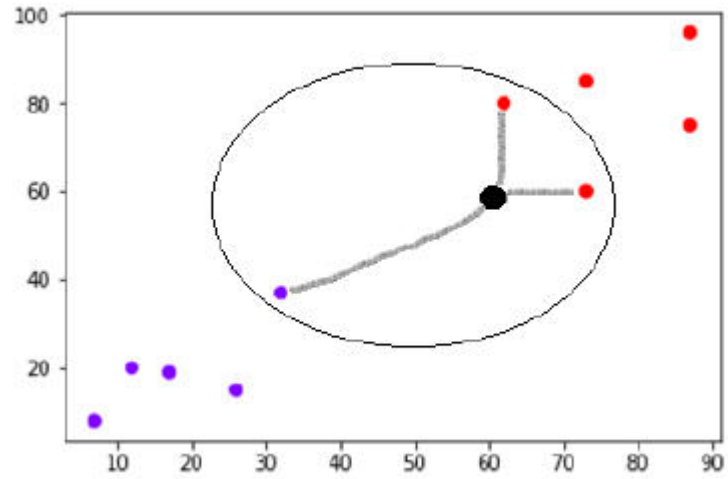
- As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

Example -2

Suppose we have a dataset which can be plotted as follows –



Now, we need to classify new data point with black dot (at point 60,60) into blue or red class. We are assuming K = 3 i.e. it would find three nearest data points. It is shown in the next diagram –



the three nearest neighbors of the data point with black dot. Among those three, two of them lies in Red class hence the black dot will also be assigned in red class.

Advantages of KNN Algorithm:

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

Why do we Need K-Nearest Neighbours Algorithm?

1. **K Nearest Neighbor** is one of the fundamental algorithms in machine learning. **Machine learning models** use a set of input values to predict output values. **KNN** is one of the simplest forms of machine learning algorithms mostly used for classification. It classifies the data point on how its neighbor is classified.

2. **KNN** classifies the new data points based on the similarity measure of the earlier stored data points. For example, if we have a dataset of tomatoes and bananas. KNN

will keep similar criteria like shape and color. Then, when a new object comes, it will check its similarity with the color (red or yellow) and shape.

K-Nearest Neighbors Classifiers and Model Example With Data Set

The K-NN algorithm using diagrams. But we didn't discuss how to know the distance between the new entry and other values in the data set.

Let's get started!

BRIGHTNESS	SATURATION	CLASS
40	20	Red
50	50	Blue
60	90	Blue
10	25	Red
70	70	Blue
60	10	Red
25	80	Blue

The table above represents our data set. We have two columns — **Brightness** and **Saturation**. Each row in the table has a class of either **Red** or **Blue**. Before we introduce a new data entry, let's assume the value of **K** is 5.

How to Calculate Euclidean Distance in the K-Nearest Neighbors Algorithm

Here's the new data entry:

BRIGHTNESS	SATURATION	CLASS
20	35	?

We have a new entry but it doesn't have a class yet. To know its class, we have to calculate the distance from the new entry to other entries in the data set using the Euclidean distance formula.

Here's the formula: $\sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$

Where:

- X_2 = New entry's brightness (20).
- X_1= Existing entry's brightness.
- Y_2 = New entry's saturation (35).
- Y_1 = Existing entry's saturation.

Let's do the calculation together. I'll calculate the first three.
Calculate the Distance of all the entries in the table:

For the first row, d1:

BRIGHTNESS	SATURATION	CLASS
40	20	Red

$d1 = \sqrt{(20 - 40)^2 + (35 - 20)^2}$
$= \sqrt{400 + 225}$
$= \sqrt{625}$
$= 25$

Here's what the table will look like after all the distances have been calculated:

BRIGHTNESS	SATURATION	CLASS	DISTANCE
40	20	Red	25
50	50	Blue	33.54
60	90	Blue	68.01
10	25	Red	10
70	70	Blue	61.03
60	10	Red	47.17
25	80	Blue	45

Let's rearrange the distances in ascending order:

BRIGHTNESS	SATURATION	CLASS	DISTANCE
10	25	Red	10
40	20	Red	25
50	50	Blue	33.54
25	80	Blue	45
60	10	Red	47.17
70	70	Blue	61.03
60	90	Blue	68.01

Since we chose 5 as the value of **K**, we'll only consider the first five rows. That is:

BRIGHTNESS	SATURATION	CLASS	DISTANCE
10	25	Red	10
40	20	Red	25
50	50	Blue	33.54
25	80	Blue	45
60	10	Red	47.17

the majority class within the 5 nearest neighbors to the new entry is **Red**. Therefore, we'll classify the new entry as **Red**.

Here's the updated table:

BRIGHTNESS	SATURATION	CLASS
40	20	Red
50	50	Blue
60	90	Blue
10	25	Red
70	70	Blue
60	10	Red
25	80	Blue
20	35	Red

How to Choose the Value of K in the K-NN Algorithm

There is no particular way of choosing the value **K**, but here are some common conventions to keep in mind:

- Choosing a very low value will most likely lead to inaccurate predictions.
- The commonly used value of **K** is 5.
- Always use an odd number as the value of **K**.

Advantages of K-NN Algorithm

- ✓ It is simple to implement.
- ✓ No training is required before classification.

Disadvantages of K-NN Algorithm

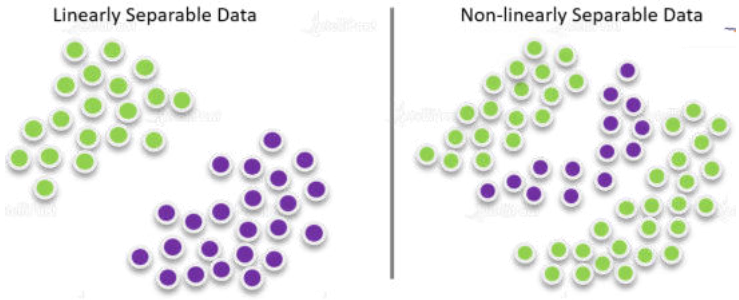
- ✓ Can be cost-intensive when working with a large data set.
- ✓ A lot of memory is required for processing large data sets.
- ✓ Choosing the right value of **K** can be tricky.

Application of K-Nearest Neighbours

1. KNN can be used in plagiarism checks to check if two documents are semantically identical.
2. ML in healthcare uses KNN to diagnose diseases based on subtle symptoms.
3. Netflix uses KNN in its recommendation systems.
4. KNN is used in face recognition software.
5. Political scientists and psychologists use KNN to classify people based on their behavior and response to stimuli.

Explain Support Vector Machines (SVM) Algorithm

Support Vector Machine or SVM algorithm is a simple yet powerful Supervised Machine Learning algorithm that can be used for building both regression and classification models. SVM algorithm can perform really well with both linearly separable and non-linearly separable datasets. Even with a limited amount of data, the support vector machine algorithm does not fail to show its magic.



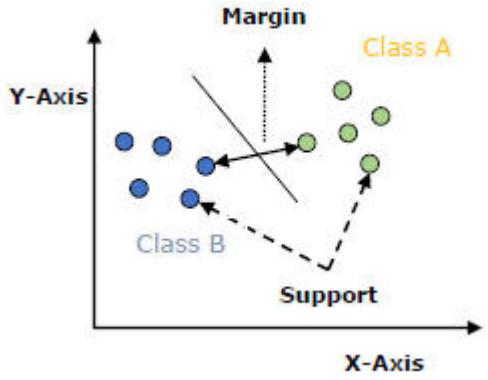
Types of SVM

Linear SVM : Linear SVM is used for data that are linearly separable i.e. for a dataset that can be categorized into two categories by utilizing a single straight line. Such data points are termed as linearly separable data, and the classifier is used described as a Linear SVM classifier.

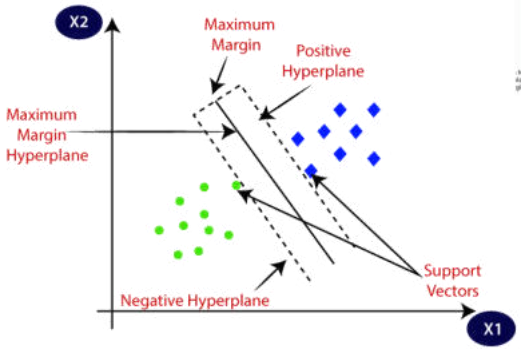
Non-linear SVM: Non-Linear SVM is used for data that are non-linearly separable data i.e. a straight line cannot be used to classify the dataset. Such data points are termed as non-linear data, and the classifier used is termed as a Non-linear SVM classifier.

Working of SVM

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).



(OR)



The followings are important concepts in SVM –

1. **Support Vectors** − Datapoints that are closest to the hyperplane is called support vectors. Separating line will be defined with the help of these data points.
2. **Hyperplane** − As we can see in the above diagram, it is a decision plane or space which is divided between a set of objects having different classes.
3. **Margin** − It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH) and it can be done in the following two steps −

- ✓ First, SVM will generate hyperplanes iteratively that segregates the classes in best way.
- ✓ Then, it will choose the hyperplane that separates the classes correctly.

SVM Parameters

SVM Parameters include the values, estimators, and various constraints used to implement ML algorithms. There are three types of SV parameters in a Neural Network:

- **Kernel**

Kernel transforms the input data into any first as per the user requirements. The Kernels used in SVM could be linear, polynomial, radial basis functions(RBFs), and non-linear hyperplanes, created using the polynomial and RBF functions.

- **Regularization**

The C parameters in Scikit-learn denote the error or penalty representing any miscalculation. You can maintain regularization by understanding the miscalculation and changing the decision boundary through tweaking the C parameters.

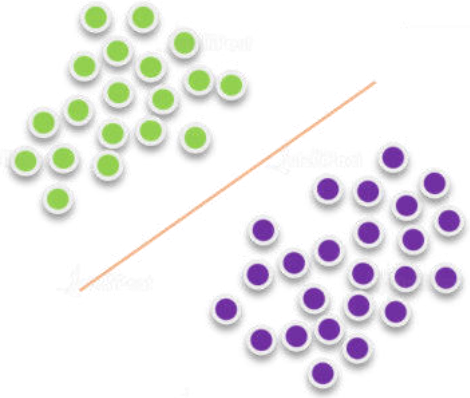
- **Gamma**

Gamma parameters determine their influence over a single training example. There are two types of gamma parameters, low meaning ‘far’ and high meaning ‘close’ values. The low or far values define a Gaussian function with a large variance. Whereas, high or close values define it with small variance.

Support Vector Machine Algorithm Example

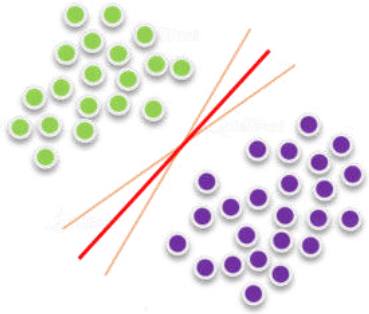
Support vector machine or SVM algorithm is based on the concept of 'decision planes', where hyperplanes are used to classify a set of given objects.

Let us start off with a few pictorial examples of support vector machine algorithms. As we can see in Figure 2, we have two sets of data. These datasets can be separated easily with the help of a line, called a **decision boundary**.



SVM Figure 2: Decision Boundary

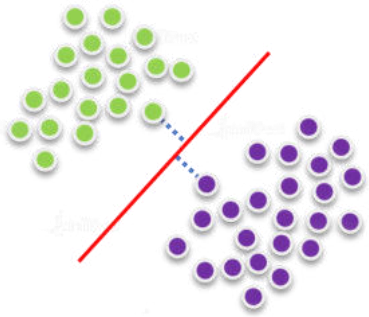
But there can be several decision boundaries that can divide the data points without any errors. For example, in Figure 3, all decision boundaries classify the datasets correctly. But how do



we pick the best decision boundary?

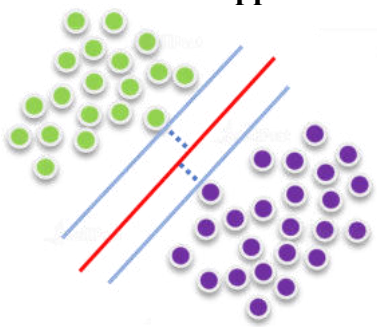
SVM Figure 3: Other Possible Decision Boundaries

Well, here's the tip: the best decision boundary is the one that has a maximum distance from the nearest points of these two classes, as shown in Figure 4.



SVM Figure 4: Maximum Distance from the Nearest Points

Also, remember that the nearest points from the optimal decision boundary that maximize the distance are called **support vectors**.



SVM Figure 5: Margin and Maximum Margin Classifier

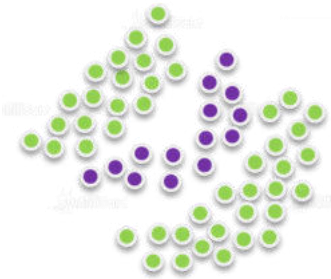
The region that the closest points define around the decision boundary is known as the **margin**.

That is why the decision boundary of a support vector machine model is known as the **maximum margin classifier** or the **maximum margin hyperplane**.

In other words, here's how a support vector machine algorithm model works:

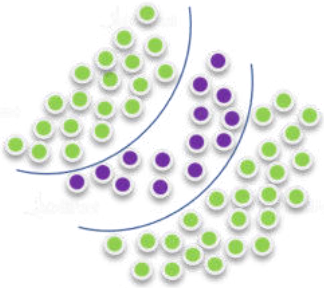
- First, it finds lines or boundaries that correctly classify the training dataset.
- Then, from those lines or boundaries, it picks the one that has the maximum distance from the closest data points.

Alright, in the above support vector machine example, the dataset was linearly separable. Now, the question, how do we classify non-linearly separable datasets as shown in Figure 6?



SVM Figure 6: Non-linearly Separable Dataset

Clearly, straight lines can't be used to classify the above dataset. That is where Kernel SVM comes into the picture.



SVM Figure 7: After Using Kernel Support Vector Classifier

Real-World Applications of SVM

SVM relies on supervised learning algorithms to perform classifications. It is a powerful method to classify unstructured data, make reliable predictions, and reduce redundant information.

Applications of SVM

SVM is mainly used to classify the unseen data and have various application in different fields:

1. **Face Detection**
   Classifies the images of people's faces in an environment from non-face by creating a square box around it.
2. **Bioinformatics**
   The Support vector machines are used for gene classification that allows researchers to differentiate between various proteins and identify biological problems and cancer cells.
3. **Text Categorization**
   Used in training models that are used to classify the documents into different categories based on the score, types, and other threshold values.
4. **Generalized Predictive Control(GPC)**
   Provides you control over different industrial processes with multivariable version and interactor matrix. GPC is used in various industries like cement mills, robotics, spraying, etc.
5. **Handwriting Recognization**
   SVM is widely used to recognize handwritten characters and test them against pre-existing data.
6. **Image Classification**
   Compared to the traditional query-based searching techniques, SVM has better accuracy when it comes to search and classifying the images based on various features.

Naïve Bayes Classifier Algorithm

1. Naïve Bayes algorithm is a supervised learning algorithm, which is based on **Bayes theorem** and used for solving classification problems.
2. It is mainly used in *text classification* that includes a high-dimensional training dataset.
3. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
4. **It is a probabilistic classifier, which means it predicts on the basis of the probability of an object**.
5. Some popular examples of Naïve Bayes Algorithm are **spam filtration, Sentimental analysis, and classifying articles**.

Why is it called Naïve Bayes?

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

- ✓ **Naïve**: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
- ✓ **Bayes**: It is called Bayes because it depends on the principle of Bayes' Theorem.

- ❖ The naive Bayes classifier is a powerful tool in machine learning, particularly in text classification, spam filtering, and sentiment analysis, among others.

- ❖ For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

- ❖ An NB model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes' Theorem:

- Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.

  The formula for Bayes' theorem is given as:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.
P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.
P(A) is Prior Probability: Probability of hypothesis before observing the evidence.
P(B) is Marginal Probability: Probability of Evidence.

Understanding Naive Bayes Classifier

Based on the Bayes theorem, the Naive Bayes Classifier gives the conditional probability of an event A given event B.

Let us use the following demo to understand the concept of a Naive Bayes classifier:

Shopping Example
Problem statement: To predict whether a person will purchase a product on a specific combination of day, discount, and free delivery using a Naive Bayes classifier.

Under the day, look for variables, like weekday, weekend, and holiday. For any given day, check if there are a discount and free delivery. Based on this information, we can predict if a customer would buy the product or not.

How Do Naive Bayes Algorithms Work?

Let's understand it using an example. Below I have a training data set of weather and corresponding target variable 'Play' (suggesting possibilities of playing). Now, we need to classify whether players will play or not based on weather condition.

Let's follow the below steps to perform it.

1. **Convert the data set into a frequency table**
   In this first step data set is converted into a frequency table
2. **Create Likelihood table by finding the probabilities**
   Create Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

3. **Use Naive Bayesian equation to calculate the posterior probability**
   Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of the prediction.

Naive Bayes uses for the following things:

- ✓ **Face Recognition**
  As a classifier, it is used to identify the faces or its other features, like nose, mouth, eyes, etc.
- ✓ **Weather Prediction**
  It can be used to predict if the weather will be good or bad.
- ✓ **Medical Diagnosis**
  Doctors can diagnose patients by using the information that the classifier provides. Healthcare professionals can use Naive Bayes to indicate if a patient is at high risk for certain diseases and conditions, such as heart disease, cancer, and other ailments.
- ✓ **News Classification**
  With the help of a Naive Bayes classifier, Google News recognizes whether the news is political, world news, and so on.

Advantages of Naive Bayes Classifier

The following are some of the benefits of the Naive Bayes classifier:

- It is simple and easy to implement
- It doesn't require as much training data
- It handles both continuous and discrete data
- It is highly scalable with the number of predictors and data points
- It is fast and can be used to make real-time predictions
- It is not sensitive to irrelevant features

What is Meant by the K-Means Clustering Algorithm?

K-Means Clustering Algorithm

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

In the K-means algorithm, every data sample from the dataset will follow two fundamental properties:

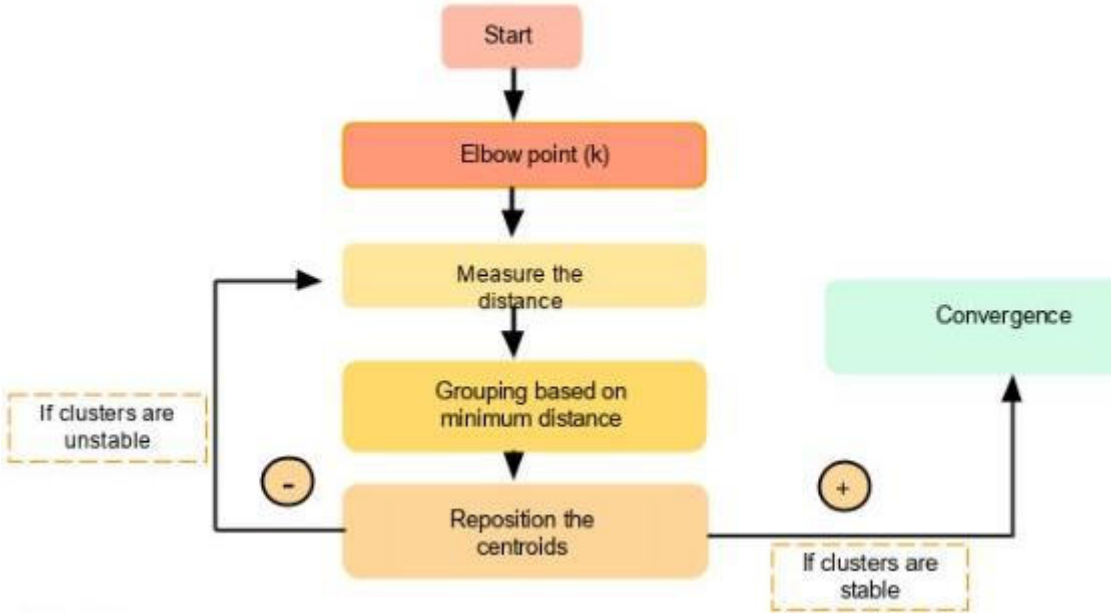
1. Each data sample belongs to at least one of the 'K' clusters. In simple terms, there can't be any sample that is not a part of any of the clusters.
2. No data sample will belong to more than one cluster. In simple terms, one sample cannot be present in two (or more) clusters at the same time.

What is K-Means Algorithm?

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

How Does K-Means Clustering Work?

The flowchart below shows how k-means clustering works:



K-Means Clustering Algorithm

Let's say we have x1, x2, x3…… x(n) as our inputs, and we want to split this into K clusters.

The steps to form clusters are:

Step 1: Choose K random points as cluster centers called centroids.

Step 2: Assign each x(i) to the closest cluster by implementing euclidean distance (i.e., calculating its distance to each centroid)

Step 3: Identify new centroids by taking the average of the assigned points.

Step 4: Keep repeating step 2 and step 3 until convergence is achieved

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

How does the K-Means Algorithm Work?

The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

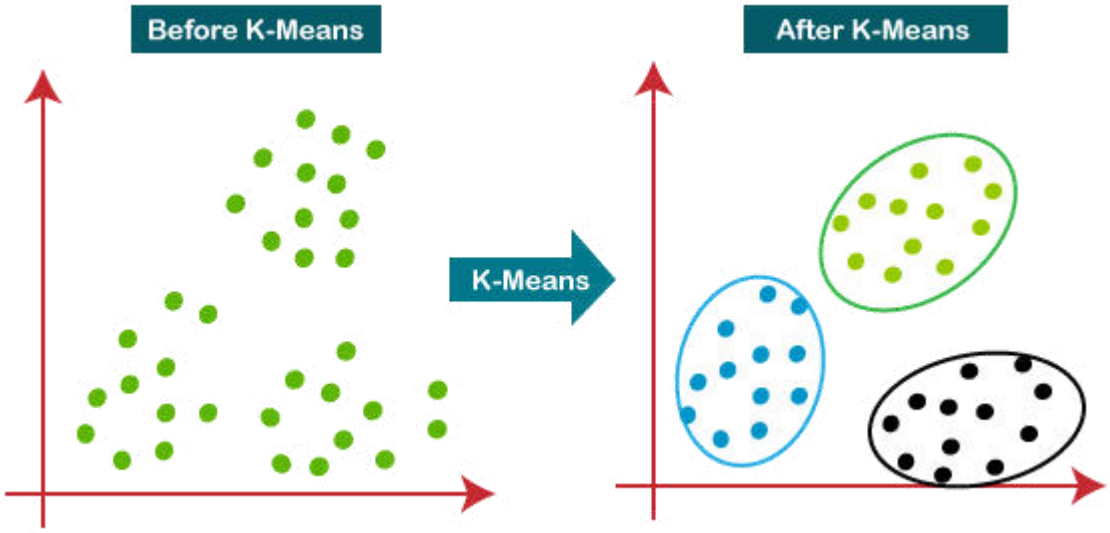
Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.



K-Means clustering is an unsupervised learning algorithm. There is no labeled data for this clustering, unlike in supervised learning. K-Means performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster. The term 'K' is a number. You need to tell the system how many clusters you need to create. For example, K = 2 refers to two clusters. There is a way of finding out what is the best or optimum value of K for a given data.

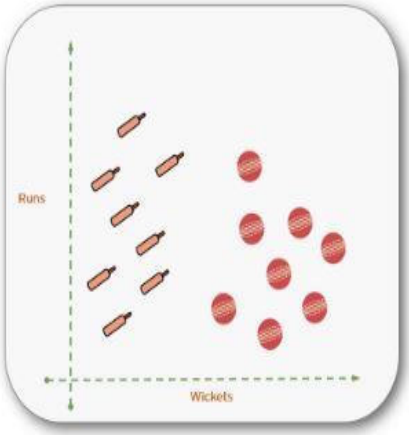
For a better understanding of k-means, let's take an example from cricket. Imagine you received data on a lot of cricket players from all over the world, which gives information on the runs scored by the player and the wickets taken by them in the last ten matches. Based on this information, we need to group the data into two clusters, namely batsman and bowlers.

Solution:

Assign data points

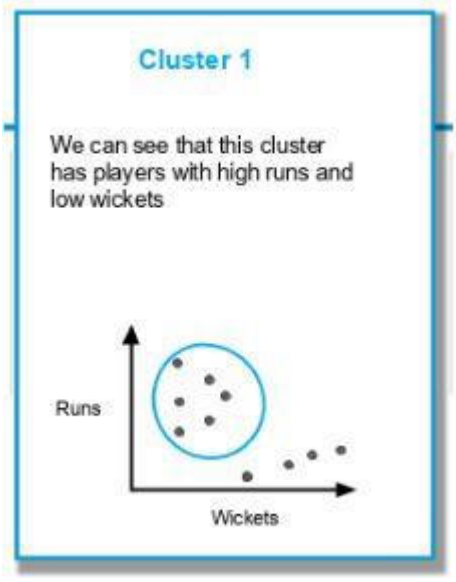
Here, we have our data set plotted on 'x' and 'y' coordinates. The information on the y-axis is about the runs scored, and on the x-axis about the wickets taken by the players.

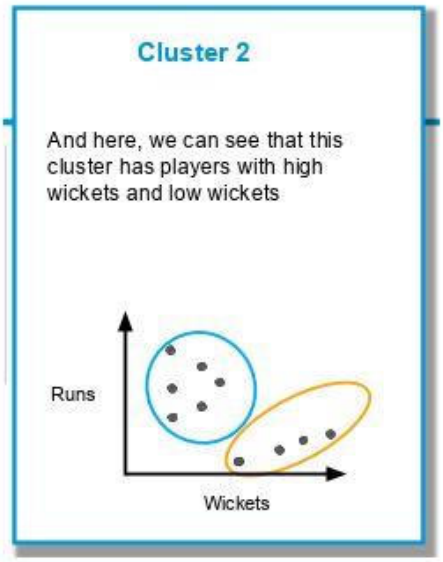
If we plot the data, this is how it would look:



Perform Clustering

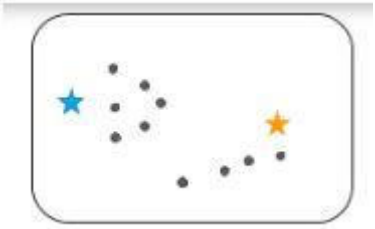
We need to create the clusters, as shown below:





Considering the same data set, let us solve the problem using K-Means clustering (taking K = 2).

The first step in k-means clustering is the allocation of two centroids randomly (as K=2). Two points are assigned as centroids. Note that the points can be anywhere, as they are random points. They are called centroids, but initially, they are not the central point of a given data set.



The next step is to determine the distance between each of the randomly assigned centroids' data points. For every point, the distance is measured from both the centroids, and whichever distance is less, that point is assigned to that centroid. You can see the data points attached to the centroids and represented here in blue and yellow.

The next step is to determine the actual centroid for these two clusters. The original randomly allocated centroid is to be repositioned to the actual centroid of the clusters.

This process of calculating the distance and repositioning the centroid continues until we obtain our final cluster. Then the centroid repositioning stops.

As seen above, the centroid doesn't need anymore repositioning, and it means the algorithm has converged, and we have the two clusters with a centroid.

Key Features of K-means Clustering

Find below some key features of k-means clustering;

1. It is very smooth in terms of interpretation and resolution.
2. For a large number of variables present in the dataset, K-means operates quicker than Hierarchical clustering.
3. While redetermining the cluster centre, an instance can modify the cluster.
4. K-means reforms compact clusters.
5. It can work on unlabeled numerical data.
6. Moreover, it is fast, robust and uncomplicated to understand and yields the best outcomes when datasets are well distinctive (thoroughly separated) from each other.

Limitations of K-means Clustering

The following are a few limitations with K-Means clustering;

1. Sometimes, it is quite tough to forecast the number of clusters, or the value of k.
2. The output is highly influenced by original input, for example, the number of clusters.
3. An array of data substantially hits the concluding outcomes.
4. In some cases, clusters show complex spatial views, then executing clustering is not a good choice.
5. Also, rescaling is sometimes conscious, it can't be done by normalization or standardization of data points, the output gets changed entirely.

Disadvantages of K-means Clustering

1. The algorithm demands for the inferred specification of the number of cluster/ centres.
2. An algorithm goes down for non-linear sets of data and unable to deal with noisy data and outliers.
3. It is not directly applicable to categorical data since only operatable when mean is provided.
4. Also, Euclidean distance can weight unequally the underlying factors.
5. The algorithm is not variant to non-linear transformation, i.e provides different results with different portrayals of data.

Applications of K-Means Clustering

K-Means clustering is used in a variety of examples or business cases in real life, like:

- Academic performance

- Diagnostic systems
- Search engines
- Wireless sensor networks

1. **Academic Performance**
   Based on the scores, students are categorized into grades like A, B, or C.
2. **Diagnostic systems**
   The medical profession uses k-means in creating smarter medical decision support systems, especially in the treatment of liver ailments.
3. **Search engines**
   Clustering forms a backbone of search engines. When a search is performed, the search results need to be grouped, and the search engines very often use clustering to do this.
4. **Wireless sensor networks**
   The clustering algorithm plays the role of finding the cluster heads, which collect all the data in its respective cluster.

Some More Applications of K-means Clustering

The concern of the fact is that the data is always complicated, mismanaged, and noisy. The conditions in the real world cast hardly the clear picture to which these types of algorithms can be applied.

Let's learn where we can implement k-means clustering among various

1. K-means clustering is applied in the **Call Detail Record (CDR) Analysis**. It gives in-depth vision about customer requirements and satisfaction on the basis of call-traffic during the time of the day and demographic of a particular location.
2. It is used in **the clustering of documents** to identify the compatible documents in the same place.
3. It is deployed **to classify the sounds** on the basis of their identical patterns and segregate malformation in them.
4. It serves as the **model of lossy images compression technique**, in the confinement of images, K-means makes clusters pixels of an image in order to decrease the total size of it.
5. It is helpful in the business sector for recognizing the portions of purchases made by customers, also to cluster movements on apps and websites.
6. In the field of insurance and fraud detection on the basis of prior data, it is plausible to cluster fraudulent consumers to demand based on their proximity to clusters as the patterns indicate.

Frequently Asked Questions

Q1. What is exploratory data analysis example?

An example of exploratory data analysis (EDA) could involve examining a dataset of customer demographics and purchase history for a retail business. EDA techniques may include calculating summary statistics, visualizing data distributions, identifying outliers, exploring relationships between variables, and performing hypothesis testing.

Q2. What are the four steps of exploratory data analysis?

The four steps of exploratory data analysis (EDA) typically involve:
1. Data Cleaning: Handling missing values, removing outliers, and ensuring data quality.
2. Data Exploration: Examining summary statistics, visualizing data distributions, and identifying patterns or relationships.
3. Feature Engineering: Transforming variables, creating new features, or selecting relevant variables for analysis.
4. Data Visualization: Presenting insights through plots, charts, and graphs to communicate findings effectively.

What are the differences between supervised and unsupervised learning?

Supervised Learning	Unsupervised Learning
• Uses known and labeled data as input • Supervised learning has a feedback mechanism • The most commonly used supervised learning algorithms are decision trees, logistic regression, and support vector machine	• Uses unlabeled data as input • Unsupervised learning has no feedback mechanism • The most commonly used unsupervised learning algorithms are k-means clustering, hierarchical clustering, and apriori algorithm

How can you avoid overfitting your model?

Overfitting refers to a model that is only set for a very small amount of data and ignores the bigger picture. There are three main methods to avoid overfitting:

1. Keep the model simple—take fewer variables into account, thereby removing some of the noise in the training data
2. Use cross-validation techniques, such as k folds cross-validation

3. Use regularization techniques, such as LASSO, that penalize certain model parameters if they're likely to cause overfitting

When do you use the Classification Technique over the Regression Technique?
Classification problems are mainly used when the output is the categorical variable (Discrete) whereas Regression Techniques are used when the output variable is Continuous variable.
In the Regression algorithm, we attempt to estimate the mapping function (f) from input variables (x) to numerical (continuous) output variable (y).

How is k-NN different from k-means clustering?
Ans. K-nearest neighbours is a classification algorithm, which is a subset of supervised learning. K-means is a clustering algorithm, which is a subset of unsupervised learning.
And K-NN is a Classification or Regression Machine Learning Algorithm while K-means is a Clustering Machine Learning Algorithm.
K-NN is the number of nearest neighbours used to classify or (predict in case of continuous variable/regression) a test sample, whereas K-means is the number of clusters the algorithm is trying to learn from the data.

What is a Linear Regression?
Ans. The linear regression equation is a one-degree equation with the most basic form being $Y = mX + C$ where m is the slope of the line and C is the standard error. It is used when the response variable is continuous in nature for example height, weight, and the number of hours. It can be a simple linear regression if it involves continuous dependent variable with one independent variable and a multiple linear regression if it has multiple independent variables.

What is Logistic Regression?
Ans. Logistic regression is a technique in predictive analytics which is used when we are doing predictions on a variable which is dichotomous(binary) in nature. For example, yes/no or true/false etc. The equation for this method is of the form $Y = eX + e – X$. It is used for classification based tasks. It finds out probabilities for a data point to belong to a particular class for classification.

Mention some drawbacks of the Linear Model
Ans. Here a few drawbacks of the linear model:

- The assumption regarding the linearity of the errors
- It is not usable for binary outcomes or count outcome
- It can’t solve certain overfitting problems
- It also assumes that there is no multicollinearity in the data.

What steps do you follow while making a decision tree?
Ans. The steps involved in making a decision tree are:

1. Determine the Root of the Tree Step
2. Calculate Entropy for The Classes Step
3. Calculate Entropy After Split for Each Attribute
4. Calculate Information Gain for each split
5. Perform the Split
6. Perform Further Splits Step
7. Complete the Decision Tree

What is 'Naive' in a Naive Bayes?
Ans. A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. Basically, it's "naive" because it makes assumptions that may or may not turn out to be correct.

Explain the SVM machine learning algorithm in detail.
Ans. SVM is an ML algorithm which is used for classification and regression. For classification, it finds out a muti dimensional hyperplane to distinguish between classes. SVM uses kernels which are namely linear, polynomial, and rbf. There are few parameters which need to be passed to SVM in order to specify the points to consider while the calculation of the hyperplane.

What are the various classification algorithms?
Ans. Different types of classification algorithms include logistic regression, SVM, Naive Bayes, decision trees, and random forest.

Unit 3

EXTRACTING MEANING FROM DATA: Feature Selection – User Retention, Feature Generation and Extraction - Feature Selection algorithms – Filters; Wrappers; Decision Trees; Entropy, Random Forests. Google's Hybrid approach to Social Research.

What is Data Extraction?

Data extraction is the process of retrieving data from a source. This can be done manually or through automated means. It can be used to retrieve data from a variety of sources, including databases, files, and web pages.

How do you extract data?
There are many ways to extract data. For example, extracting a list of contacts from an email, extracting information from a webpage, extracting financial data from accounting records, or extracting data from PDF documents.

There are two types of data extraction: manual and automated. Manual data extraction is a process in which data is manually collected from sources. Automated data extraction is a process in which data is collected from sources using software or other automated means.

What are the Challenges of Data Extraction?

The challenges of data extraction include the cost and time required to extract data, as well as the accuracy of the data. Data extraction can be a costly and time-consuming process, and the accuracy of the data depends on the quality of the data source.

1. Data quality
Data quality is one of the most important aspects in analytics. Many companies extract data from different sources to get a richer, more accurate picture of what is happening in their business, but this can come at a cost.

2. Lack of standardization
Information is everywhere, but it's not always in the format you need.

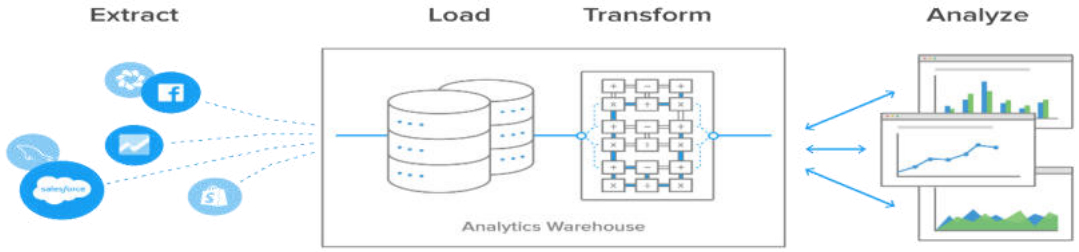
3. Lack of access
Finding the right data can be a daunting and costly process. There are many reasons why you might not be able to easily extract data from a source.

4. Incomplete data
The data extraction process is not always perfect. Some data may be missing due to errors or omissions during the extraction process.

The data extraction process

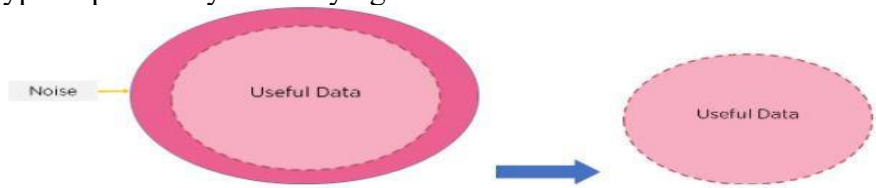
Whether the source is a database, a SaaS platform, Excel spreadsheet, web scraping, or something else, the process to extract information involves the following steps:

1. Check for changes to the structure of the data, including the addition of new tables and columns. Changed data structures have to be dealt with programmatically.
2. Retrieve the target tables and fields from the records specified by the integration's replication scheme.
3. Extract the appropriate data, if any.



What is Feature Selection?

Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data.
It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.



How to Choose a Feature Selection Model?

The process is relatively simple, with the model depending on the types of input and output variables.

Variables are of two main types:

- Numerical Variables: Which include integers, float, and numbers.
- Categorical Variables: Which include labels, strings, boolean variables, etc.

Based on whether we have numerical or categorical variables as inputs and outputs, we can choose our feature selection model as follows:

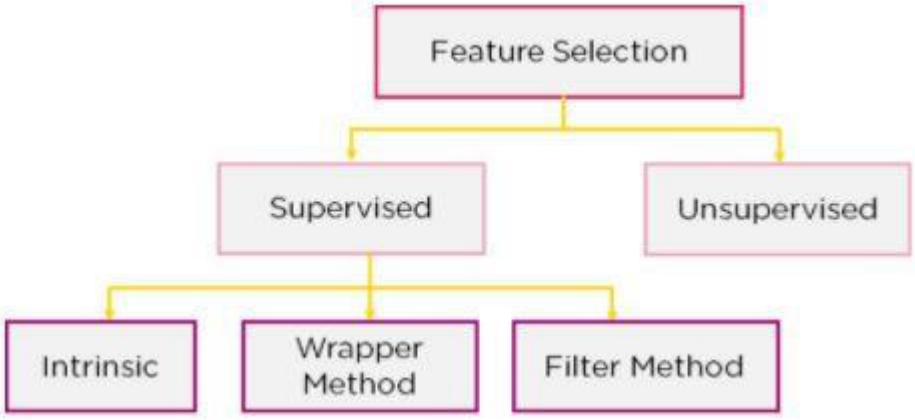
Input Variable	Output Variable	Feature Selection Model
Numerical	Numerical	Pearson's correlation coefficient Spearman's rank coefficient
Numerical	Categorical	ANOVA correlation coefficient (linear). Kendall's rank coefficient (nonlinear).

Categorical	Numerical	Kendall's rank coefficient (linear). ANOVA correlation coefficient (nonlinear).
Categorical	Categorical	Chi-Squared test (contingency tables). Mutual Information.

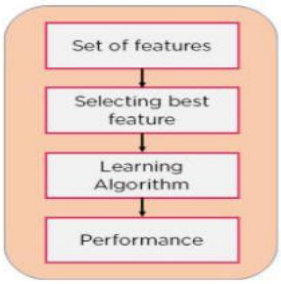
Feature Selection Models

Feature selection models are of two types:

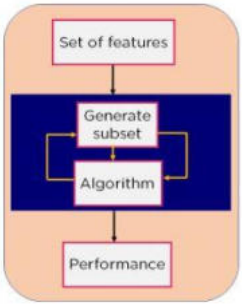
1. Supervised Models: Supervised feature selection refers to the method which uses the output label class for feature selection. They use the target variables to identify the variables which can increase the efficiency of the model
2. Unsupervised Models: Unsupervised feature selection refers to the method which does not need the output label class for feature selection. We use them for unlabelled data.



Filter Method: In this method, features are dropped based on their relation to the output, or how they are correlating to the output. We use correlation to check if the features are positively or negatively correlated to the output labels and drop features accordingly. Eg: Information Gain, Chi-Square Test, Fisher's Score, etc.



2. Wrapper Method: We split our data into subsets and train a model using this. Based on the output of the model, we add and subtract features and train the model again. It forms the subsets using a greedy approach and evaluates the accuracy of all the possible combinations of features. Eg: Forward Selection, Backwards Elimination, etc.



What is user retention?

User Retention is the continued use of a product, feature or app by your customers. While measuring retention for feature usage is relatively straightforward (using that feature at least once in a specified period of time), measuring retention for a product or complete app requires more definition.

It is important to lay out the usage factors that will be considered to be "product usage" before beginning your measurement efforts.

For example, does simply running the app or using the site count as usage or is there a higher threshold based on time, number of key features used, etc.

How to Measure User Retention?

Once you decide on a relevant time frame, you can measure retention by taking an active number of users back at the beginning of that period and subtracting it from the number of those users who are still using your app at the end.

To make the equivalent retention rate calculation, simply divide the number by the end number.

How do we define and measure retention?

Retention measures how many users return to your product over some specified time. There are three dimensions to consider in terms of measuring retention. Understanding different measures of retention and comparing them will help us find the appropriate retention measure for your product.

Dimension 1: Time

- N-day/week/month retention

N-day retention is the most classic way to calculate retention, which measures among users who first used the product at day 0, what proportion of them are still active at day N. Here "day" could be week or month.

Dimension 2: user status

In addition to time, user status is often another important dimension to consider. First, let's take a look at how to define user status. There are many ways to define status, and different companies/products often have their own ways to define user status. Here is one way user status can be defined:

- New user
- Churned user: x day inactive
- Inactive user: 0-x day inactive
- Reactive user: active after churned/inactive
- Active user: active users who are not new and not reactive

It is often important to calculate retention for different user statuses. For example, new user retention measures the proportion of new users who stay active.

Dimension 3: Action
When we say users use the product, we didn't define what we mean by "use". Should we define "use" as visiting the product page, staying for a certain amount of time, conducting certain actions, or purchasing a product.

FEATURE GENERATION VS FEATURE EXTRACTION

What is feature extraction/selection?

- Extraction: Getting useful features from existing data.
- Selection: Choosing a subset of the original pool of features.

What is Feature Extraction?
Feature extraction is a part of the dimensionality reduction process, in which, an initial set of the raw data is divided and reduced to more manageable groups. So when you want to process it will be easier. The most important characteristic of these large data sets is that they have a large number of variables.

Why Feature Extraction is Useful?
The technique of extracting the features is useful when you have a large data set and need to reduce the number of resources without losing any important or relevant information. Feature extraction helps to reduce the amount of redundant data from the data set.

Applications of Feature Extraction

- **Bag of Words-** Bag-of-Words is the most used technique for natural language processing. In this process they extract the words or the features from a sentence, document, website, etc. and then they classify them into the frequency of use. So in this whole process feature extraction is one of the most important parts.
- **Image Processing** –Image processing is one of the best and most interesting domain. In this domain basically you will start playing with your images in order to understand them. So here we use many many techniques which includes feature extraction as well and algorithms to detect features such as shaped, edges, or motion in a digital image or video to process them.
- **Auto-encoders:** The main purpose of the auto-encoders is efficient data coding which is unsupervised in nature. this process comes under unsupervised learning . So Feature extraction procedure is applicable here to identify the key features from the data to code by learning from the coding of the original data set to derive new ones.

What is Feature Selection?
"***It is a process of automatically or manually selecting the subset of most appropriate and relevant features to be used in model building***." Feature selection is performed by either including the important features or excluding the irrelevant features in the dataset without changing them.

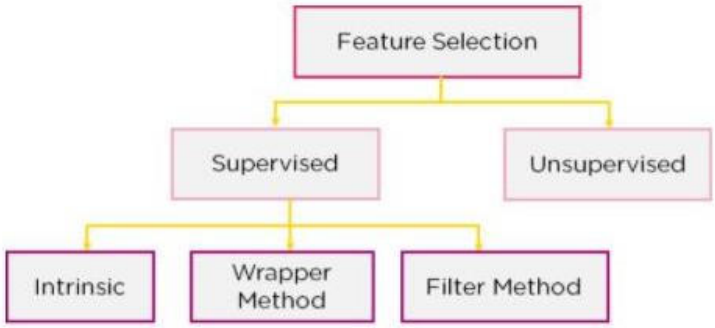
Below are some benefits of using feature selection in machine learning:

- **It helps in avoiding the curse of dimensionality.**
- **It helps in the simplification of the model so that it can be easily interpreted by the researchers.**
- **It reduces the training time.**
- **It reduces overfitting hence enhance the generalization.**

Feature Selection Techniques

There are mainly two types of Feature Selection techniques, which are:

- **Supervised Feature Selection technique**
  Supervised Feature selection techniques consider the target variable and can be used for the labelled dataset.
- **Unsupervised Feature Selection technique**
  Unsupervised Feature selection techniques ignore the target variable and can be used for the unlabelled dataset.

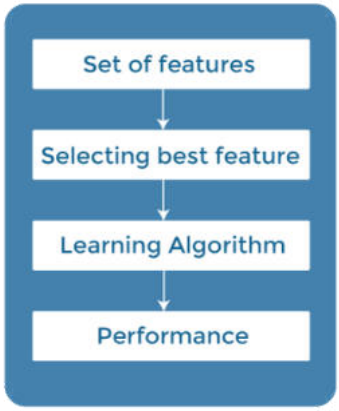


We can further divide the supervised models into three :

1. Filter Method: In Filter Method, features are selected on the basis of statistics measures. This method does not depend on the learning algorithm and chooses the features as a pre-processing step.

The filter method filters out the irrelevant feature and redundant columns from the model by using different metrics through ranking.

The advantage of using filter methods is that it needs low computational time and does not overfit the data.

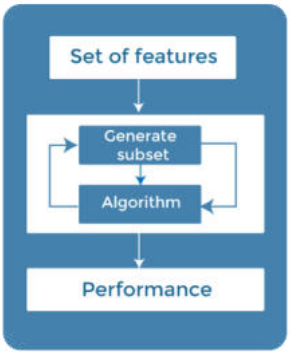


1. Wrapper Methods

In wrapper methodology, selection of features is done by considering it as a search problem, in which different combinations are made, evaluated, and compared with other combinations. It trains the algorithm by using the subset of features iteratively.

Some techniques of wrapper methods are:

- **Forward selection** - Forward selection is an iterative process, which begins with an empty set of features. After each iteration, it keeps adding on a feature and evaluates the performance to check whether it is improving the performance or not. The process continues until the addition of a new variable/feature does not improve the performance of the model.
- **Backward elimination** - Backward elimination is also an iterative approach, but it is the opposite of forward selection. This technique begins the process by considering all the features and removes the least significant feature. This elimination process continues until removing the features does not improve the performance of the model.
- **Exhaustive Feature Selection-** Exhaustive feature selection is one of the best feature selection methods, which evaluates each feature set as brute-force. It means this method tries & make each possible combination of features and return the best performing feature set.
- **Recursive Feature Elimination-** Recursive feature elimination is a recursive greedy optimization approach, where features are selected by recursively taking a smaller and smaller subset of features. Now, an estimator is trained with each set of features, and the importance of each feature is determined using *coef_attribute* or through a *feature_importances_attribute.*

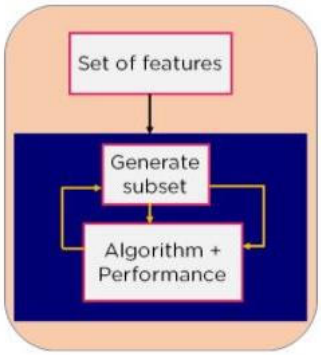


Missing Value Ratio:

The value of the missing value ratio can be used for evaluating the feature set against the threshold value. The formula for obtaining the missing value ratio is the number of missing values in each column divided by the total number of observations. The variable is having more than the threshold value can be dropped.

$$\text{Missing Value Ratio} = \frac{Number\ of\ Missing\ values * 100}{Total\ number\ of\ observations}$$

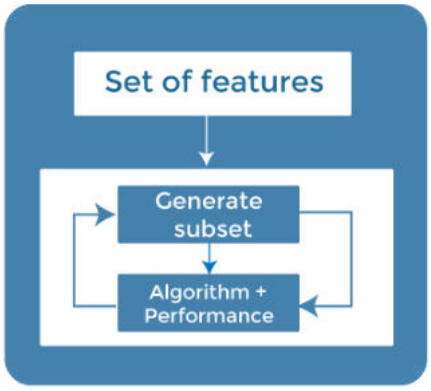
3. Intrinsic Method: This method combines the qualities of both the Filter and Wrapper method to create the best subset.



This method takes care of the machine training iterative process while maintaining the computation cost to be minimum. Eg: Lasso and Ridge Regression.

3. Embedded Methods

Embedded methods combined the advantages of both filter and wrapper methods by considering the interaction of features along with low computational cost. These are fast processing methods similar to the filter method but more accurate than the filter method.



These methods are also iterative, which evaluates each iteration, and optimally finds the most important features that contribute the most to training in a particular iteration. Some techniques of embedded methods are:

- **Regularization**- Regularization adds a penalty term to different parameters of the machine learning model for avoiding overfitting in the model. This penalty term is added to the coefficients; hence it shrinks some coefficients to zero. Those features with zero coefficients can be removed from the dataset. The types of regularization techniques are L1 Regularization (Lasso Regularization) or Elastic Nets (L1 and L2 regularization).
- **Random Forest Importance** - Different tree-based methods of feature selection help us with feature importance to provide a way of selecting features. Here, feature importance specifies which feature has more importance in model building or has a great impact on the target variable. Random Forest is such a tree-based method, which is a type of bagging algorithm that aggregates a different number of decision trees. It automatically ranks the nodes by their performance or decrease in the impurity (Gini impurity) over all the trees. Nodes are arranged as per the impurity values, and thus it

allows to pruning of trees below a specific node. The remaining nodes create a subset of the most important features.

How to Choose a Feature Selection Model?

How do we know which feature selection model will work out for our model? The process is relatively simple, with the model depending on the types of input and output variables.

Variables are of two main types:

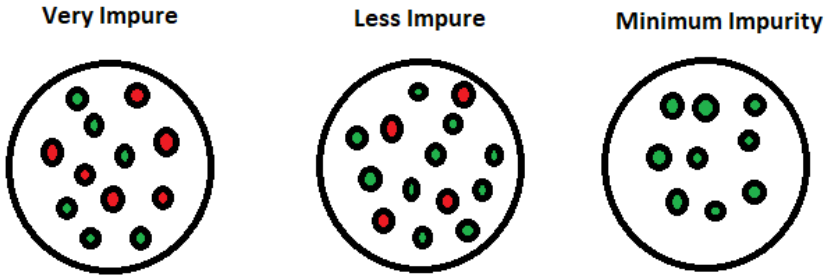
- Numerical Variables: Which include integers, float, and numbers.
- Categorical Variables: Which include labels, strings, boolean variables, etc.

Based on whether we have numerical or categorical variables as inputs and outputs, we can choose our feature selection model as follows:

Input Variable	Output Variable	Feature Selection Model
Numerical	Numerical	• Pearson's correlation coefficient • Spearman's rank coefficient
Numerical	Categorical	• ANOVA correlation coefficient (linear). • Kendall's rank coefficient (nonlinear).
Categorical	Numerical	• Kendall's rank coefficient (linear). • ANOVA correlation coefficient (nonlinear).
Categorical	Categorical	• Chi-Squared test (contingency tables). • Mutual Information.

Entropy

Entropy is an information theory metric that measures the impurity or uncertainty in a group of observations. It determines how a decision tree chooses to split data. The image below gives a better description of the purity of a set.



Consider a dataset with N classes. The entropy may be calculated using the formula below:

$E=-\sum_{i=1}^{N} p_i \log_2 p_i$

pipi is the probability of randomly selecting an example in class ii. Let's have an example to better our understanding of entropy and its calculation.

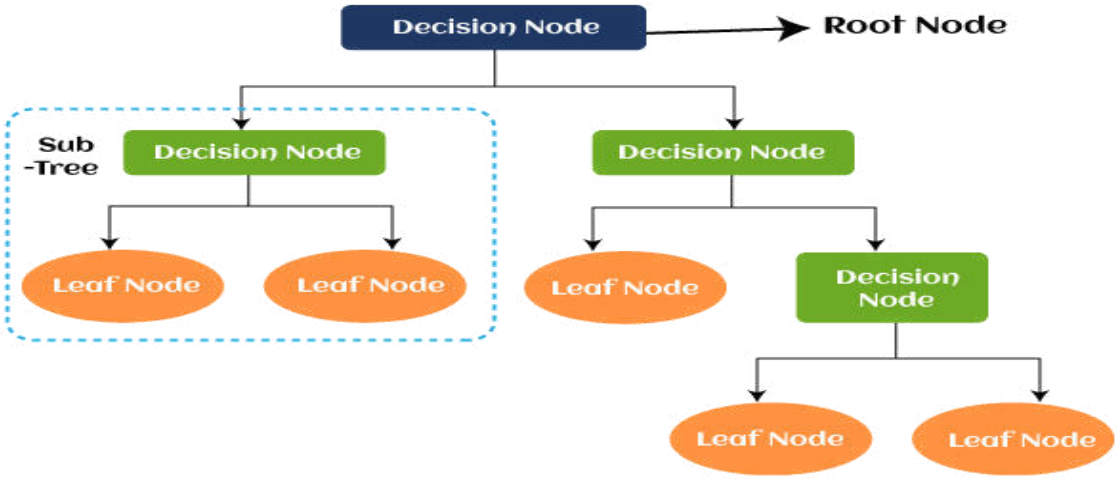
What is a Decision Tree

A decision tree is defined as the supervised learning algorithm used for classification as well as regression problems. However, it is primarily used for solving classification problems. Its structure is similar to a tree where internal nodes represent the features of the dataset, branches of the tree represent the decision rules, and leaf nodes as an outcome.

Decision Tree

Decision Tree, a supervised learning technique, is a hierarchical if-else statement which is nothing but a collection of rules or is also known as the splitting criteria that are based on comparison operators on the features.

A decision tree algorithm, which is a very widely used model and has a vast variety of applications, can be used for both regression and classification problems. An example of a binary classification categorizing a car type as a sedan or sports truck follows as below. The algorithm finds the relationship between the response variable and the predictors and expresses this relation in the form of a tree-structure.



Leaf Node: Leaf node is defined as the output of decision nodes, but if they do not contain any branch, it means the tree cannot be segregated further from this node.

Root Node: As the name suggests, a root node is the origin point of any decision tree. It contains the entire data set, which gets divided further into two or more sub-sets. This node includes multiple branches and is used to make any decision in classification problems.

Splitting: It is a process that divides the root node into multiple sub-nodes under some defined conditions.

Branches: Branches are formed by splitting the root node or decision node.

Pruning: Pruning is defined as the process of removing unwanted branches from the tree.
Parent Node: The root node in a decision tree is called the parent node.

Child Node: Except for the root node, all other nodes are called child nodes in the decision tree.

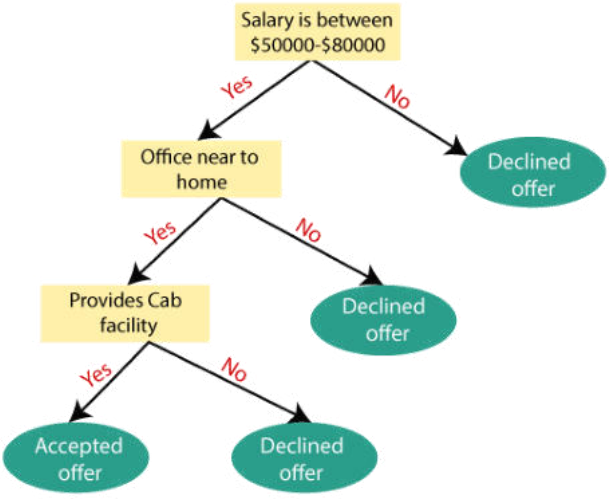
How does the Decision Tree algorithm Work?

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

- **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
- **Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**
- **Step-3:** Divide the S into subsets that contains possible values for the best attributes.
- **Step-4:** Generate the decision tree node, which contains the best attribute.
- **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Example: Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM). The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node (Cab facility) and one leaf node. Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer). Consider the below diagram:



Attribute Selection Measures

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as **Attribute selection measure or ASM.** By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are:

- **Information Gain**
- **Gini Index**

1. Information Gain:

- Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
- It calculates how much information a feature provides us about a class.
- According to the value of information gain, we split the node and build the decision tree.
- A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

1. Information Gain= Entropy(S)- [(Weighted Avg) *Entropy(each feature)

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

Where,

- **S= Total number of samples**
- **P(yes)= probability of yes**
- **P(no)= probability of no**

2. Gini Index:

- Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.
- An attribute with the low Gini index should be preferred as compared to the high Gini index.
- It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.
- Gini index can be calculated using the below formula:

Gini Index= $1- \sum_j P_j^2$

Pruning: Getting an Optimal Decision tree

Pruning is a process of deleting the unnecessary nodes from a tree in order to get the optimal decision tree.

A too-large tree increases the risk of overfitting, and a small tree may not capture all the important features of the dataset. Therefore, a technique that decreases the size of the learning tree without reducing accuracy is known as Pruning. There are mainly two types of tree **pruning** technology used:

- **Cost Complexity Pruning**
- **Reduced Error Pruning.**

Advantages of the Decision Tree

- It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
- It can be very useful for solving decision-related problems.
- It helps to think about all the possible outcomes for a problem.
- There is less requirement of data cleaning compared to other algorithms.

Disadvantages of the Decision Tree

- The decision tree contains lots of layers, which makes it complex.
- It may have an overfitting issue, which can be resolved using the **Random Forest algorithm.**
- For more class labels, the computational complexity of the decision tree may increase.

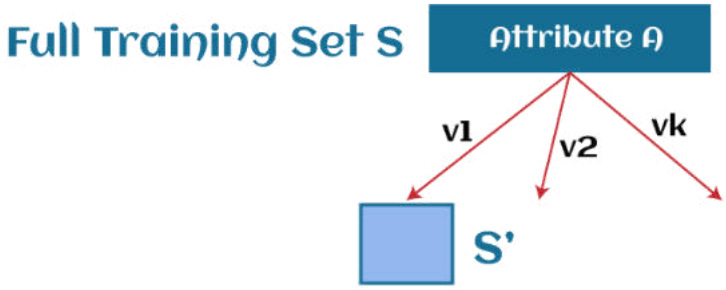
What is the information gain in Entropy?
Information gain is defined as the pattern observed in the dataset and reduction in the entropy.
Mathematically, information gain can be expressed with the below formula:
Information Gain = (Entropy of parent node)-(Entropy of child node)

Advantages of the Decision Tree:
- A decision tree can be easily understandable as it follows the same process of human thinking while making any decision.
- It is used to solve any decision-related problem in machine learning.
- It helps in finding out all the possible outcomes for a problem.
- There is less requirement for data cleaning compared to other algorithms.

How to build decision trees using information gain:

After understanding the concept of information gain and entropy individually now, we can easily build a decision tree. See steps to build a decision tree using information gain:

1. An attribute with the highest information gain from a set should be selected as the parent (root) node. From the image below, it is attributed A.

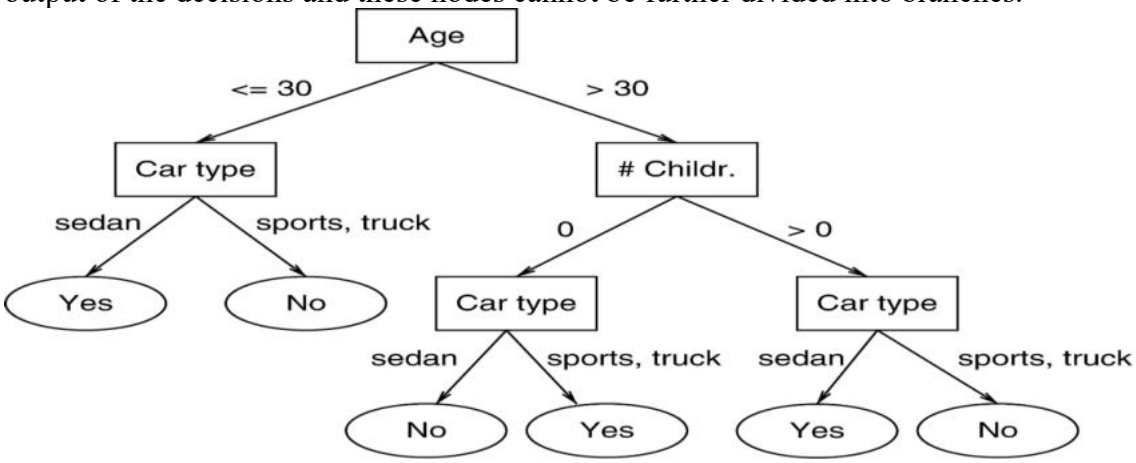


2. Build child nodes for every value of attribute A.
3. Repeat iteratively until you finish constructing the whole tree.

Use of Entropy in Decision Tree

In decision trees, heterogeneity in the leaf node can be reduced by using the cost function.

This flow-chart consists of the Root node, the Branch nodes, and the Leaf nodes. The root node is the original data, branch nodes are the decision rules whereas the leaf nodes are the output of the decisions and these nodes cannot be further divided into branches.



Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.*

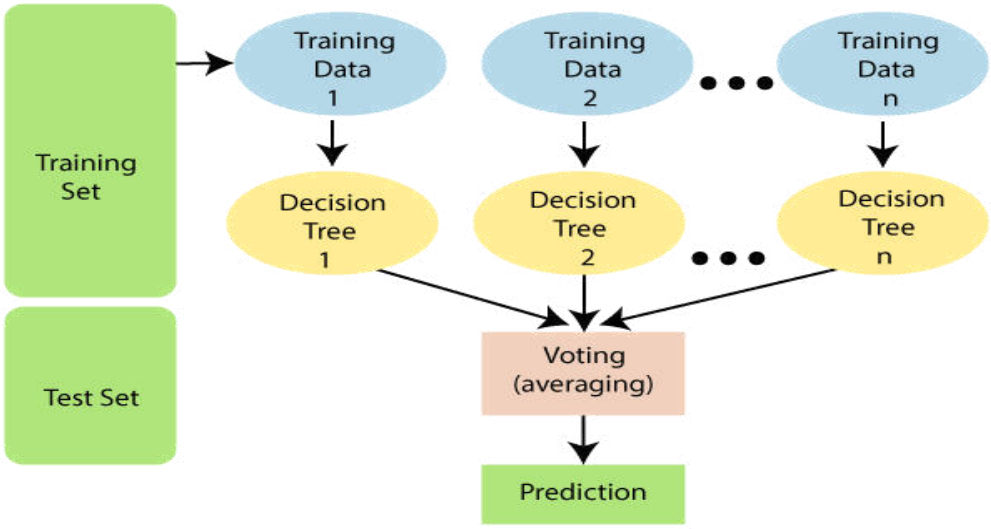
"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

There are different ways that the Random Forest algorithm makes data decisions, and consequently, there are some important related terms to know. Some of these terms include:

- **Entropy**
  It is a measure of randomness or unpredictability in the data set.
- **Information Gain**
  A measure of the decrease in the entropy after the data set is split is the information gain.
- **Leaf Node**
  A leaf node is a node that carries the classification or the decision.
- **Decision Node**
  A node that has two or more branches.
- **Root Node**
  The root node is the topmost decision node, which is where you have all of your data.

Working of Random Forest Algorithm

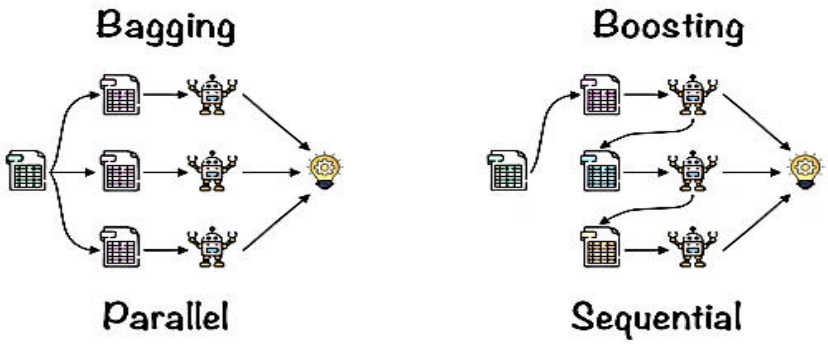


The following steps explain the working Random Forest Algorithm:

Step 1: Select random samples from a given data or training set.
Step 2: This algorithm will construct a decision tree for every training data.
Step 3: Voting will take place by averaging the decision tree.
Step 4: Finally, select the most voted prediction result as the final prediction result.

This combination of multiple models is called Ensemble. Ensemble uses two methods:

1. Bagging: Creating a different training subset from sample training data with replacement is called Bagging. The final output is based on majority voting.
2. Boosting: Combing weak learners into strong learners by creating sequential models such that the final model has the highest accuracy is called Boosting. Example: ADA BOOST, XG BOOST.



Bagging: From the principle mentioned above, we can understand Random forest uses the Bagging code. Now, let us understand this concept in detail. Bagging is also known as Bootstrap Aggregation used by random forest. The process begins with any original random data. After arranging, it is organised into samples known as Bootstrap Sample. This process is known as Bootstrapping.Further, the models are trained individually, yielding different results known as Aggregation. In the last step, all the results are combined, and the generated output is based on majority voting. This step is known as Bagging and is done using an Ensemble Classifier.

Essential Features of Random Forest

- Miscellany: Each tree has a unique attribute, variety and features concerning other trees. Not all trees are the same.
- Immune to the curse of dimensionality: Since a tree is a conceptual idea, it requires no features to be considered. Hence, the feature space is reduced.
- Parallelization: We can fully use the CPU to build random forests since each tree is created autonomously from different data and features.
- Train-Test split: In a Random Forest, we don't have to differentiate the data for train and test because the decision tree never sees 30% of the data.
- Stability: The final result is based on Bagging, meaning the result is based on majority voting or average.

Difference between Decision Tree and Random Forest

Decision Trees	Random Forest
• They usually suffer from the problem of overfitting if it's allowed to grow without any control.	• Since they are created from subsets of data and the final output is based on average or majority ranking, the problem of overfitting doesn't happen here.
• A single decision tree is comparatively faster in computation.	• It is slower.
• They use a particular set of rules when a data set with features are taken as input.	• Random Forest randomly selects observations, builds a decision tree and then the result is obtained based on majority voting. No formulas are required here.

How Google handles hybrid working

- Google recently announced their future hybrid strategy, and we should all pay attention to what it says. They've led the workplace revolution before, and could likely lead it again.
- They're going for an "office-first" hybrid strategy—aiming for roughly three days in the office, and two remote (whether that's home or elsewhere).
- The company is also introducing a number of new features in their physical workplaces, to optimise them for hybrid processes—including inflatable privacy balloons, "Campfire" meeting rooms, and modular furniture.
- Workplace flexibility and innovation will be needed as Google competes for talent with companies like Twitter, Shopify, and Spotify, who are all adopting "remote-first" policies.

The criteria that determine the future of work according to Google:

- **Provide more security or the feeling of security by spreading who is at the office at what time, in order to reduce the number of people physically present.** Google is thinking ahead to annual flu seasons and possible future pandemics and wants to reduce infection risks. Moreover, lower occupancy levels also benefit concentration.

- **Organisations can no longer demand that people come to the office five days a week.** Google wants to be able to respond more flexibly to the changing needs of employees: workstations that can be adapted for a particular team or project, personal heating and cooling systems at desks, outdoor meeting rooms in camp themes,

- **Companies must be able to cope with a certain mix of external and office workers.** For example, Google introduced "Campfire", a new meeting room where attendees sit in a circle, interspersed with large screens that are impossible to ignore.

Rationale for the hybrid approach in social research

A hybrid approach is particularly valuable for complex social issues where a single methodology fails to provide a complete picture. By integrating qualitative and quantitative data, researchers can achieve the following:

- **Validation and triangulation:** Comparing the results from both methodologies can strengthen the validity and reliability of the findings. Quantitative data can statistically confirm trends discovered through qualitative exploration, and qualitative data can explain why certain quantitative patterns appear.

- **Richer, more contextual insights:** Quantitative data answers questions about "what" and "how many," while qualitative data delves into the "why". Combining them reveals statistical trends and provides the context, motivations, and experiences behind the numbers.

- **Greater adaptability:** A hybrid approach allows for a dynamic and iterative research process. For example, unexpected findings from a quantitative survey can be explored in greater depth through qualitative interviews in a later phase

Common hybrid research designs

In social research, the integration of qualitative and quantitative data is intentional and happens at different stages of the research process. Common designs include:

- **Convergent parallel design:** In this design, quantitative and qualitative data are collected and analyzed independently and simultaneously. The results are then compared and integrated during the interpretation phase to see if they converge, diverge, or complement each other.

- **Explanatory sequential design:** The study begins with the collection and analysis of quantitative data. The quantitative results then guide the qualitative phase, which seeks to explain or elaborate on the initial numerical findings.

- **Exploratory sequential design:** This design starts with a qualitative phase to explore a topic in-depth. The insights and themes from the qualitative data are then used to inform and build a subsequent quantitative phase, such as developing a survey instrument.

- **Embedded design:** One research method (either qualitative or quantitative) is embedded within a larger, predominantly single-method study. The secondary method provides a supplementary data source to enrich the findings of the primary method.

Challenges of a hybrid approach

Despite its benefits, the hybrid approach is not without its challenges:

- **Complexity and resources:** Designing and executing a hybrid study can be complex and may require more time, funding, and a research team with expertise in both methodologies.

- **Data integration:** Effectively integrating and comparing two different types of data is a complex skill that requires a well-planned procedure to avoid conflicting or poorly connected insights.

- **Participant fatigue:** In sequential designs, involving participants in multiple phases of research may lead to a higher dropout rate or fatigue

Example: Understanding voter behavior

A social researcher could employ a hybrid approach to study voter behavior by:

- **Qualitative phase (exploratory):** Conduct focus groups or in-depth interviews with a small, diverse group of voters to understand their motivations, emotional drivers, and perception of political issues.
- **Quantitative phase (explanatory):** Use the insights from the qualitative phase to develop a large-scale survey. This survey can then be used to quantify which issues are most important to voters and identify any statistically significant voting patterns based on different demographics.
- **Integration:** Analyze the data by using the qualitative findings to explain the "why" behind the quantitative trends. For instance, if the survey reveals that a particular issue is highly important to a certain demographic, the interview data can provide detailed stories and rationale, adding depth and context.

UNIT :: 4

RECOMMENDATION ENGINES: Role of data in Building a User-Facing Data Product, Algorithmic ingredients of a Recommendation Engine, Bipartite graph, Nearest Neighbor algorithm and its problems. Dimensionality Reduction: Singular Value Decomposition - Principal Component Analysis - Exercise: build your own recommendation system.

Recommendation System

A recommendation system (or recommender system) is a class of machine learning that uses data to help predict, narrow down, and find what people are looking for among an exponentially growing number of options.

A recommendation system is an artificial intelligence or AI algorithm, usually associated with **machine learning**, that uses **Big Data** to suggest or recommend additional products to consumers. These can be based on various criteria, including past purchases, search history, demographic information, and other factors. Recommender systems are highly useful as they help users discover products and services they might otherwise have not found on their own.
Why the Recommendation system?

- Benefits users in finding items of their interest.
- Help item providers in delivering their items to the right user.
- Identity products that are most relevant to users.
- Personalized content.
- Help websites to improve user engagement.

What can be Recommended?

There are many different things that can be recommended by the system like movies, books, news, articles, jobs, advertisements, etc. Netflix uses a recommender system to recommend movies & web-series to its users.

How do User and Item matching is done?

In order to understand how the item is recommended and how the matching is done, let us a look at the images below;

SOCIAL WEBSITES	USER	ITEM
Amazon	Members	Product
Netflix	Members	Movies
Linkedin	Members	Members
Facebook	Members	Jobs

Benefits of Recommendation Systems

Companies implement recommender systems for a variety of reasons, including:

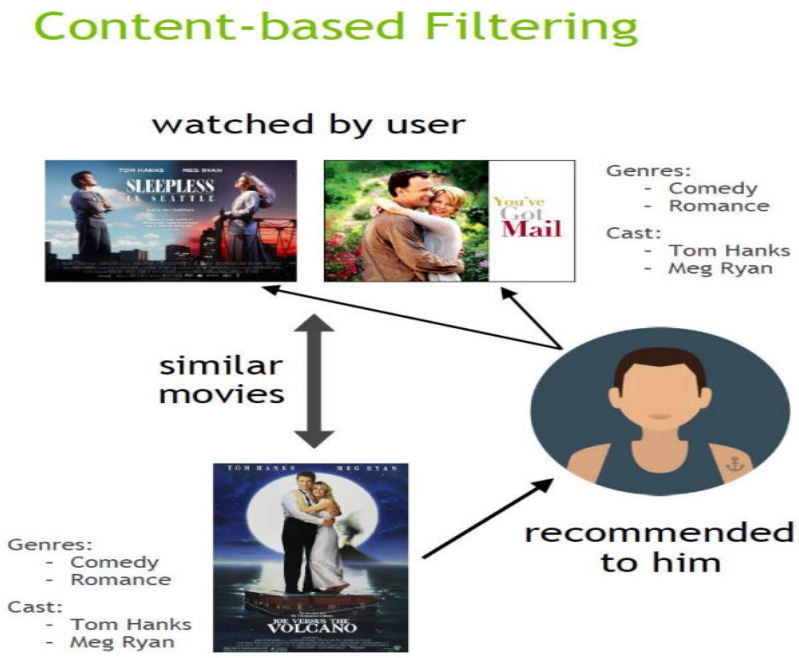
- Improving retention. By continuously catering to the preferences of users and customers, businesses are more likely to retain them as loyal subscribers or shoppers. When a customer senses that they're truly understood by a brand and not just having information randomly thrown at them, they're far more likely to remain loyal and continue shopping at your site.
- Increasing sales. Various research studies show increases in upselling revenue from 10-50% resulting from accurate 'you might also like' product recommendations. Sales can be increased with recommendation system strategies as simple as adding matching product recommendations to a purchase confirmation; collecting information from abandoned electronic shopping carts; sharing information on 'what customers are buying now'; and sharing other buyers' purchases and comments.
- Helping to form customer habits and trends. Consistently serving up accurate and relevant content can trigger cues that build strong habits and influence usage patterns in customers.
- Speeding up the pace of work. Analysts and researchers can save as much as 80% of their time when served tailored suggestions for resources and other materials necessary for further research.
- Boosting cart value. Companies with tens of thousands of items for sale would be challenged to hard code product suggestions for such an inventory.

Types of Recommendation Systems

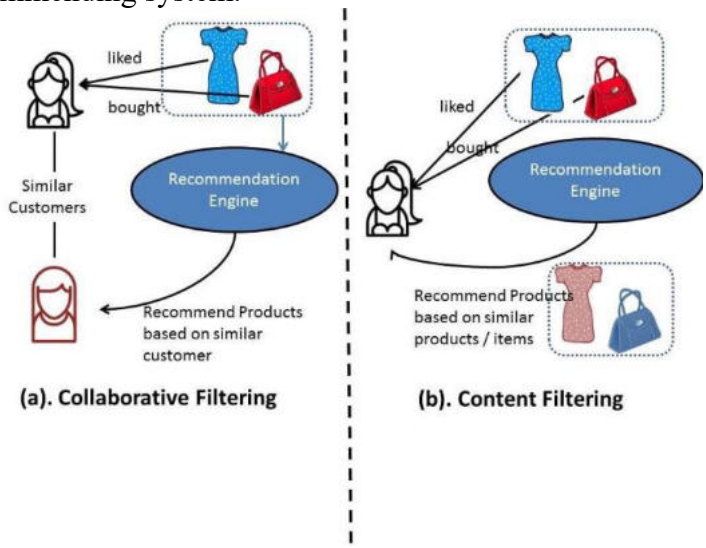
Collaborative filtering algorithms recommend items (this is the filtering part) based on preference information from many users (this is the collaborative part). This approach uses similarity of user preference behavior, given previous interactions between users and items, recommender algorithms learn to predict future interaction. These recommender systems build a model from a user's past behavior, such as items purchased previously or ratings given to those items and similar decisions by other users.

Content filtering, by contrast, uses the attributes or features of an item (this is the content part) to recommend other items similar to the user's preferences. This approach is based on similarity of item and user features, given information about a user and items they have interacted with (e.g. a user's age, the category of a restaurant's cuisine, the average review for a movie), model the likelihood of a new interaction. For example, if a content filtering recommender sees you liked the movies You've Got Mail and Sleepless in Seattle, it might

recommend another movie to you with the same genres and/or cast such as Joe Versus the Volcano.



Hybrid recommender systems combine the advantages of the types above to create a more comprehensive recommending system.



Some of the most popular examples of recommender systems include the ones used by Amazon, Netflix, and Spotify.

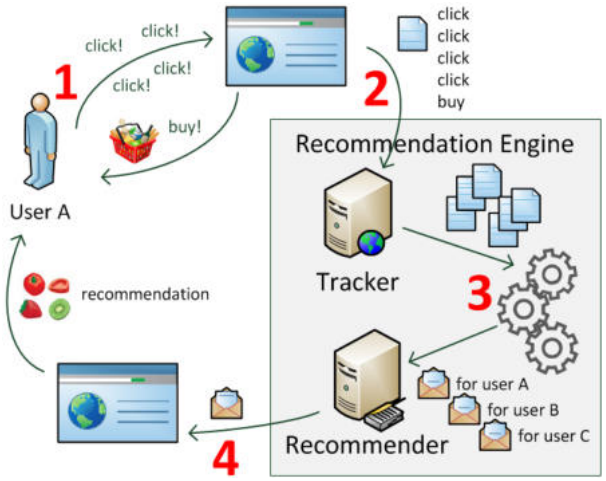
What are recommendation engines?

A recommendation engine filters the data using different algorithms and recommends the most relevant items to users. It first captures the past behavior of a customer and based on that, recommends products which the users might be likely to buy.

How Does a Recommendation Engine Work?

A recommendation engine works using a combination of data and machine learning technology. Data is crucial in the development of a recommendation engine – it is the building blocks from which patterns are derived. The more data it has, the more efficient and effective it will be in making relevant revenue-generating suggestions.

Recommendation engines need to know you better to be effective with their suggestion. Therefore, the information they *collect* and *integrate* is a critical aspect of the process. This can be information relating to **explicit interactions**, for example, information about your *past activity, your ratings, reviews and other information about your profile, such as gender, age, or investment objectives*. These can combine with **implicit interactions** *such as the device you use for access, clicks on a link, location, and dates*.



Recommendation engines complete a standard four-step process:

Step 1: Data collection
The first and most important step for creating a recommendation engine is to gather data. There are two main types of data to be collected:

Implicit Data: This includes information collected from activities such as web search history, clicks, cart events, search log, and order history.

Explicit Data: This is information gathered from customer input, such as reviews and ratings, likes and dislikes, and product comments.

Recommendation engines also use customer attribute data such as demographics (age, gender) and psychographics (interests, values) to identify similar customers, as well as feature data (genre, item type) to identify product similarity.

Step 2: Data storage Once the data is gathered, it needs to be stored. Over time, the amount of data will grow to be vast. This means ample, scalable storage must be available. Depending on the type of data you collect, different types of storage are available.

Step 3: Data analysis To be used, the data must then be drilled down into and analyzed. There are several different ways in which you can analyze data. These include:

Real-time analysis: Data is processed as it is created.
Batch analysis: Data is processed periodically.
Near-real-time analysis : Data is processed in minutes instead of seconds when you don't need it immediately.

Step 4: Data filtering
The final step is filtering. Different matrixes or mathematical rules and formulas are applied to the data depending on whether collaborative, content-based, or hybrid model recommendation filtering is being used. The outcome of this filtering is the recommendations.

Role of Data in Building a user-facing Data Product:

What is a Data Product?

A data product is an application or tool that uses data to help businesses improve their decisions and processes. Data products that provide a friendly user interface can use data science to provide predictive analytics, descriptive data modeling, data mining, machine learning, risk management, and a variety of analysis methods to non-data scientists.

(Or)

Definition of Data Products

A data product can be defined as a product or service that is generated, augmented, or enhanced by data, and that provides valuable insights or information to its users. Data products can take various forms, ranging from simple dashboards and reports to complex predictive models and machine-learning applications.

A data product is a combination of:

- **Data sets** which may be in a table, view, an ML model, or a stream. The data may be raw data or curated data integrated from multiple data sources. The data product must publish its data model.
- **Domain** model which adds a semantic layer. This layer abstracts the technical layout of the storage layer and instead exposes the business-friendly terms to the end-users. This layer also stores various calculations, metrics, and the transformation business logic.
- **Access** to data via APIs and other visualization options and with access control policies enforced.

Types of Data Products

Data products can take many forms and serve various purposes, depending on the specific needs and goals of an organisation.

Dashboards: Dashboards are visual displays of data that provide real-time insights and help users monitor key performance indicators (KPIs). They include charts, graphs, and other visualisations to make the data easily understandable.

Example: A sales dashboard showing revenue, leads, conversion rates, and customer acquisition costs.

Reports: Reports are structured presentations of data and analyses that help users understand specific trends, patterns, or issues. They can be generated regularly (e.g., weekly, monthly, quarterly) or on an ad-hoc basis.

Example: A quarterly financial report providing an overview of revenue, expenses, and profit margins.

Predictive Models: Predictive models use historical data and machine learning algorithms to forecast future outcomes or trends. These models can help organisations make proactive decisions and optimise processes.

Example: A churn prediction model that identifies customers at risk of cancelling their subscriptions, allowing a company to take action to retain them.

Recommender Systems: Recommender systems use data to suggest items or actions based on user preferences, behaviour, or other factors. These systems are widely used in industries such as e-commerce, content streaming, and advertising.

Example: An online shopping platform suggesting products to users based on their browsing history and past purchases.

Data APIs (Application Programming Interfaces): Data APIs allow developers to access and utilize data from various sources within their applications. These APIs enable the creation of new data-driven products and services and streamline data access for internal and external users.
Example: A weather data API provides real-time weather information for weather-related applications.

Data Visualisations: Data visualisations are graphical representations of data that help users understand complex information and identify trends, patterns, and relationships. They can take many forms, including charts, graphs, maps, and infographics.

Example: An interactive map displaying the spread of a disease outbreak over time and across geographic regions.

Examples of Successful Data Products
Many organisations have successfully leveraged data products to drive innovation and growth. Here are some examples:

Google Maps: Google Maps is a prime example of a data product that combines various data sources, such as satellite imagery, traffic data

Netflix: Netflix uses data-driven algorithms to recommend personalized content to its users, based on their viewing history and preferences.

Evaluating Data Product Success

To understand the impact and effectiveness of your data products, it is essential to evaluate their success using appropriate metrics and benchmarks. Different data products may require unique evaluation criteria, depending on their specific goals and objectives.

Some common metrics used to assess data product success include:

User adoption: The number of users who actively engage with the data product, indicating its relevance and value to the target audience.

User satisfaction: The level of satisfaction users express with the data product, often measured through surveys or user feedback.

Business impact: The extent to which the data product contributes to the organization's overall goals, such as increased revenue, cost savings, or improved operational efficiency.

Data accuracy: The degree to which the data product provides accurate and reliable information, reflecting the quality of the underlying data and analytics.

Data timeliness: The extent to which the data product delivers up-to-date information, ensuring users can access the most current insights.

There are lots of different factors that need to be taken into account to build a powerful artificial intelligence.

1. Accuracy

The classic recommender system takes a dataset of existing user ratings as input to predict the rating of other similar products or content. For example, for a video streaming service, the users would rate movies they have watched on a scale from 1 to 5 stars.

2. Coverage

The coverage represents the percentage of all the items that may be recommended. If you don't pay attention, the algorithm may have a very high accuracy metric, but actually, be excluding 90% of your products and only recommend 10% of them.

3. Popularity vs. Novelty

Popularity refers to the percentage of times that a product was bought or a content was consumed among all items: the greater, the most popular. Novelty is the opposite: it measures how different the item is from usually consumed items. Depending on your use case, you may want to bias your algorithm towards popular or novel results.

4. Serendipity

Serendipity is the ability to find valuable items not sought for, in our case unexpected but delightful recommendations.

5. Personalization

A good recommendation is subjective, this is why personalization is key. Each user should be suggested customized recommendations depending on their preferences to optimize the engagement of the users.

6. Diversity
Have you ever been to an e-commerce website, had a quick look at a pair of shoes, and had this pair recommended to you over and over again? To avoid these repetitive patterns, you should integrate some diversity into the recommended items.

7. Contextuality
Some factors like the time of the day or the device used have a big impact on the consumption behavior. For example, you might prefer to watch short news videos on your mobile during your lunch break, and rather watch long-form shows when you are back home after work.

8. Temporality
Things change over time: trending topics (elections, sports highlights, etc.) and seasonality (Christmas holiday season, summertime, etc.) but also users tastes.

9. Business rules
Depending on your type of business, you might want to bias your recommendations to conform to your constraints. For example for physical goods, you might want to prioritize high margin items.

10. Business objectives
Obviously, a recommendation is not an end in itself. Depending on your company's mission, they may serve other business goals such as sales or consumption lift, lifetime value, referrals, etc.

Algorithmic Ingredients of Recommendation engine:

- Neural Networks: A neural network is a type of machine learning algorithm that is similar to the brain. It is composed of interconnected neurons that can learn to recognize patterns. Neural networks are often used for prediction tasks, like recommender systems.

- K-NearestNeighbor (K-NN): The K-NN algorithm is often used for recommender systems because it is able to handle large amounts of data and can produce good predictions. The K-NN algorithm works by finding the k nearest neighbours of a given item. The neighbours are then used to vote on the rating of the item. The algorithm then uses the average of the votes to predict the rating of the item. The K-NN algorithm is often used for Recommender Systems because it is able to handle large amounts of data and can produce good predictions.

- Bayesian inference: Bayesian inference is a type of machine learning algorithm that is used to make better predictions. It is often used in recommender systems because it can handle large amounts of data. The Bayesian inference algorithm works by using a probability model to predict the rating of an item. The algorithm uses the past ratings of a user to build the probability model. This allows the algorithm to make better predictions about a user's preferences.

- Dimensionality reduction: Dimensionality reduction is a type of machine learning algorithm that is used to reduce the number of dimensions in a data set.

It is often used in Recommender Systems because it can help to reduce the amount of data that needs to be processed. The dimensionality reduction algorithm works by finding a lower dimensional representation of the data. This can be done by using techniques like Principal Component Analysis (PCA).These are just some of the machine learning algorithms that can be used in Recommender Systems. Each algorithm has its own strengths and weaknesses, and the best algorithm for a particular application will depend on the nature of the data.

Bipartite Graph:

A graph G=(V, E) is called a bipartite graph if its vertices V can be partitioned into two subsets V_1 and V_2 such that each edge of G connects a vertex of V_1 to a vertex V_2. It is denoted by K_{mn}, where m and n are the numbers of vertices in V_1 and V_2 respectively.

A bipartite graph is also known as a bigraph.

Draw the bipartite graphs K_2, 4 and K_3 ,4. Assuming any number of edges.

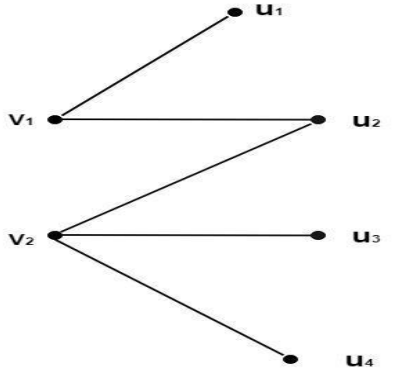


Fig:Bipartite Graph $K_{2,4}$

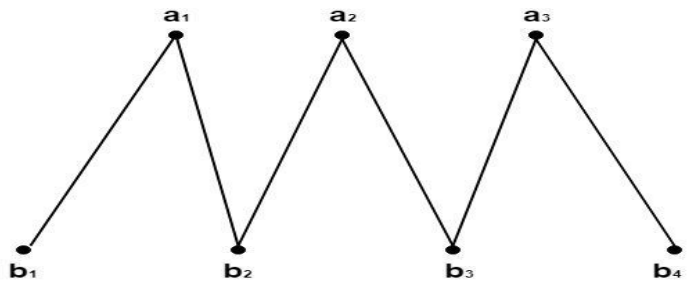


Fig:Bipartite Graph $K_{3,4}$

Properties of Bipartite Graphs

The properties of Bipartite Graphs are given below in brief. Knowing these will assist you further in learning the topic well.

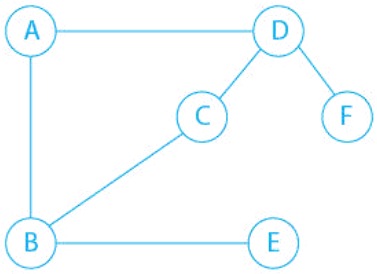
- The vertices present in one set are not connected to the vertex present in the same set.
- A Bipartite graph is a simple graph, which means that there are no self-loops or multiple edges between the same pair of vertices.
- It may or may not be connected.
- For a graph to be Bipartite, it should not contain any cycle of odd length.

Algorithm to Check if the Given Graph is a Bipartite Graph or Not

Here is an algorithm for checking whether the given graph is a Bipartite or not using the BFS approach.

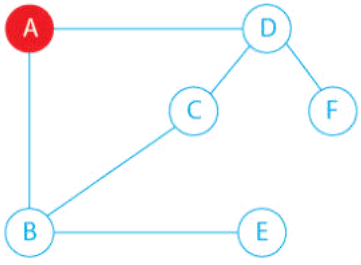
- **Step 1:** Start by picking an arbitrary vertex as the starting point and mark it as visited and add it to a queue.
- **Step 2:** Assign a color to the starting vertex (let's say color red).
- **Step 3:** While the queue is not empty, take the first vertex from the queue and look at its neighbors.
- **Step 4:** For each neighbor of the current vertex, check if it has already been visited or not.
- **Step 5:** If the neighbor has not been visited, mark it as visited, add it to the queue, and assign it the opposite color of its parent vertex. So, if the parent vertex is colored red, the neighbor will be colored blue, and vice versa.
- **Step 6:** If the neighbor has been visited and has the same color as its parent vertex, then the graph is not bipartite.
- **Step 7:** Repeat steps 3 to 6 until all vertices have been visited or until a non-bipartite condition is detected.
- **Step 8:** If all vertices have been visited without any non-bipartite condition, then the graph is bipartite.

Let us consider the following graph to understand the concept of the Bipartite Graph with 6 vertices named A, B, C, D, E, and F.

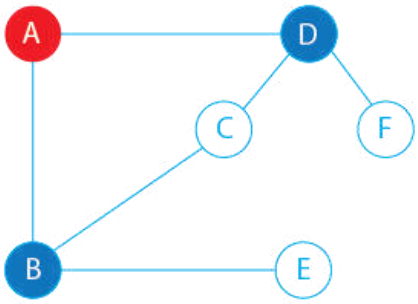


Let us dry-run the algorithm specified above on this graph.

- **Step 1:** Let us choose a random vertex (Here we have chosen A), and mark it as red. The graph will look like this.

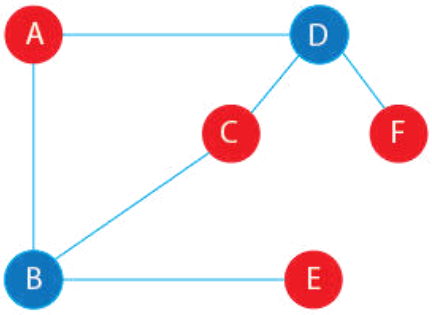


- **Step 2:** Next step is to find all the neighbors of the vertex A, which are B and D. Mark the neighbor nodes in Blue color (since the previous node was colored red) as shown in the below image.



- **Step 3:** Now select the neighbors of the current node (B and D). The selected neighbors are:
  C and E -> Neighbors of B
  F -> Neighbors of D

  Color the selected vertices red (opposite to the current nodes color, blue). The graph will look like this.



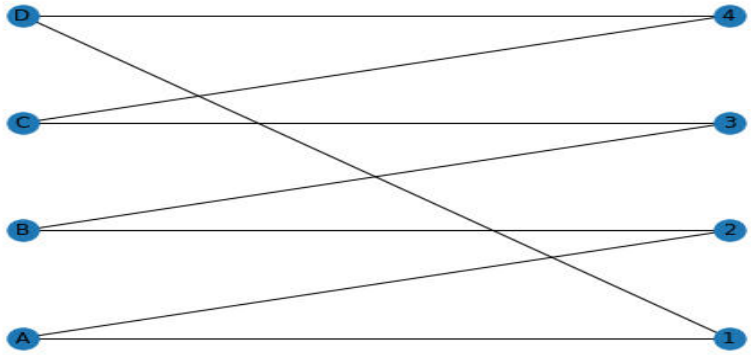
- **Step 4:** The algorithm will now search for the unvisited neighbors of the current nodes (C, E, and F). Since there are no unvisited neighbors, the Algorithm will stop and return **True** as the answer.

  The above graph is a bipartite graph, as we can see that no two consecutive nodes are colored with the same color.

Example:
Let us take a simple example of mapping students and dorm rooms. In this problem, students give a list of rooms they are willing to stay at. We represent students on side as nodes of a bipartite graph and rooms on the other side as nodes, and we put an edge between students and rooms as per this list.

We want to map students and rooms. Furthermore, we want to identify a subset where we match one student to exactly one other room (no roommates). The students give what rooms are acceptable, and many solutions are possible. Consider the below problem where we have

students (A, B, C and D) and we want to match them to rooms (1,2,3,4). The list of acceptable rooms for each student is given below.



Applications of Bipartite graphs?

Bipartite graphs have many applications in different fields, including:

- **Matching problems:** Bipartite graphs are commonly used to model matching problems, such as matching job seekers with job vacancies or assigning students to project supervisors. The bipartite structure allows for a natural way to match vertices from one set to vertices in the other set.
- **Social networks:** Bipartite graphs, where the nodes in one set represent users and the nodes in the other set reflect interests, groups, or communities, can be used to simulate social networks. The bipartite form makes it simple to analyse the connections between users and interests.
- **Web Search engine:** The query and click-through data of a search engine can be defined using a bipartite graph, where the two sets of vertices represent queries and web pages.

- Bipartite Graphs are used to solve the Matching Problems. For example, it can be used for assigning employees a number of tasks or assigning students to different courses.

Nearest Neighbor algorithm and its problems

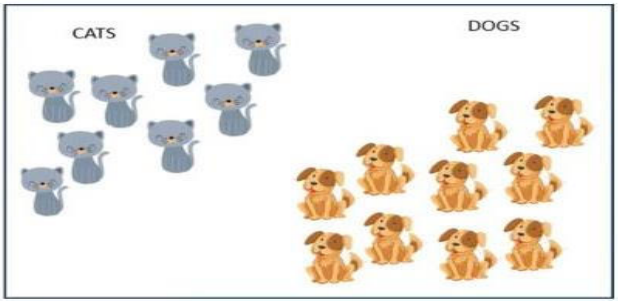
K Nearest Neighbors, or KNN—a popular supervised machine learning algorithm used for solving classification and regression problems. The main objective of the KNN algorithm is to predict the classification of a new sample point based on data points that are separated into several individual classes. It is used in text mining, agriculture, finance, and healthcare.

Why Do We Need the KNN Algorithm?

The KNN algorithm is useful when you are performing a pattern recognition task for classifying objects based on different features.

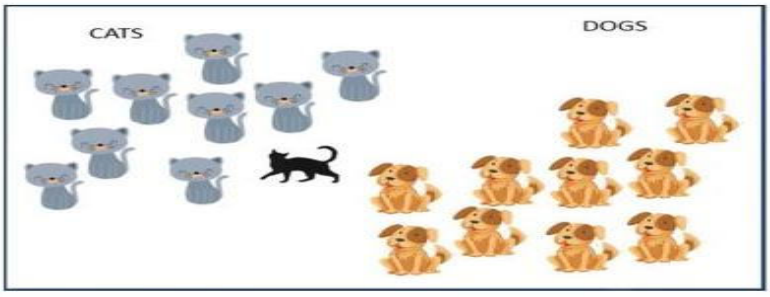
Suppose there is a dataset that contains information regarding cats and dogs. There is a new data point and you need to check if that sample data point is a cat or dog. To do this, you need to list the different features of cats and dogs.

Now, let us consider two features: claw sharpness and ear length. Plot these features on a 2D plane and check where the data points fit in.

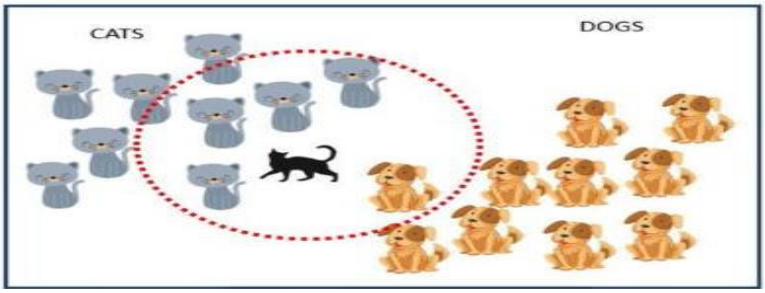


As illustrated above, the sharpness of claws is significant for cats, but not so much for dogs. On the other hand, the length of ears is significant for dogs, but not quite when it comes to cats.

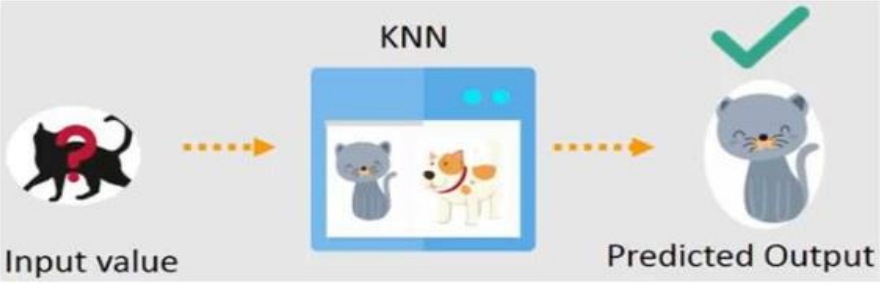
Now, if we have a new data point based on the above features, we can easily determine if it's a cat or a dog.



The new data point features indicate that the animal is, in fact, a cat.



Since KNN is based on feature similarity, we can perform classification tasks using the KNN classifier. The image below—trained with the KNN algorithm—shows the predicted outcome, a black cat.



What is KNN?

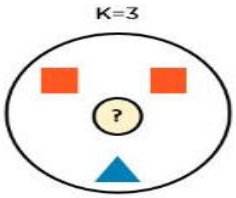
K-Nearest Neighbors is one of the simplest supervised machine learning algorithms used for classification. It classifies a data point based on its neighbors' classifications. It stores all available cases and classifies new cases based on similar features.

How to Choose the Factor 'K'?

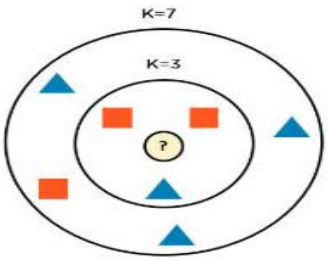
A KNN algorithm is based on feature similarity. Selecting the right K value is a process called parameter tuning, which is important to achieve higher accuracy.

There is not a definitive way to determine the best value of K. It depends on the type of problem you are solving, as well as the business scenario. The most preferred value for K is five. Selecting a K value of one or two can be noisy and may lead to outliers in the model, and thus resulting in overfitting of the model. The algorithm performs well on the training set, compared to its true performance on unseen test data.

Consider the following example below to predict which class the new data point belongs to. If you take K=3, the new data point is a red square.



But, if we consider K=7, the new data point is a blue triangle. This is because the amount of red squares outnumbers the blue triangles.



To choose the value of K, take the square root of n (sqrt(n)), where n is the total number of data points. Usually, an odd value of K is selected to avoid confusion between two classes of data.

Advantages & Disadvantages of KNN Algorithm

Advantages

- ✓ It is very easy to understand and implement
- ✓ It is an instance-based learning(lazy learning) algorithm.
- ✓ KNN does not learn during the training phase hence new data points can be added with affecting the performance of the algorithm.
- ✓ It is well suited for small datasets.
- ✓ Disadvantages
- ✓ It fails when variables have different scales.
- ✓ It is difficult to choose K-value.
- ✓ It leads to ambiguous interpretations.
- ✓ It is sensitive to outliers and missing values.
- ✓ Does not work well with large datasets.
- ✓ It does not work well with high dimensions.

What is Dimensionality Reduction

dimensionality reduction, a principal component analysis in machine learning.

Dimensionality reduction is defined as a method of reducing variables in a training dataset used to develop machine learning models. This article explains the core principles of dimensionality reduction and its key techniques with examples.

Why Dimensionality Reduction is Important

Dimensionality reduction brings many advantages to your machine learning data, including:

- Fewer features mean less complexity
- You will need less storage space because you have fewer data
- Fewer features require less computation time
- Model accuracy improves due to less misleading data
- Algorithms train faster thanks to fewer data
- Reducing the data set's feature dimensions helps visualize the data faster
- It removes noise and redundant features

Benefits Of Dimensionality Reduction

For AI engineers or data professionals working with enormous datasets, doing data visualisation, and analysing complicated data, dimension reduction is helpful.

1. It aids in data compression, resulting in less storage space being required.
2. It speeds up the calculation.
3. It also aids in removing any extraneous features.

Disadvantages Of Dimensionality Reduction

1. We lost some data during the dimensionality reduction process, which can impact how well future training algorithms work.
2. It may need a lot of processing power.
3. Interpreting transformed characteristics might be challenging.
4. The independent variables become harder to comprehend as a result.

Dimensionality Reduction Methods and Approaches

So now that we've established how much dimensionality reduction benefits machine learning, what's the best method of doing it? We have listed the principal approaches you can

take, subdivided further into diverse ways. This series of approaches and methods are also known as Dimensionality Reduction Algorithms.

- **Feature Selection.**

Feature selection is a means of selecting the input data set's optimal, relevant features and removing irrelevant features.

- Filter methods. This method filters down the data set into a relevant subset.
- Wrapper methods. This method uses the machine learning model to evaluate the performance of features fed into it. The performance determines whether it's better to keep or remove the features to improve the model's accuracy. This method is more accurate than filtering but is also more complex.
- Embedded methods. The embedded process checks the machine learning model's various training iterations and evaluates each feature's importance.

- **Feature Extraction.**

This method transforms the space containing too many dimensions into a space with fewer dimensions. This process is useful for keeping the whole information while using fewer resources during information processing. Here are three of the more common extraction techniques.

Dimensionality Reduction Examples

Dimensionality reduction methods are key to several real-life applications, including text categorization, image retrieval, face recognition, intrusion detection, neuroscience, gene expression analysis, email categorization, etc.

What is Singular Value Decomposition?

The Singular Value Decomposition of a matrix is a factorization of the matrix into three matrices. Thus, the singular value decomposition of matrix A can be expressed in terms of the factorization of A into the product of three matrices as $A = UDV^T$

Here, the columns of U and V are orthonormal, and the matrix D is diagonal with real positive entries.

Mathematics behind SVD

The SVD of mxn matrix A is given by the formula :

$A = UDV^T$

where:

- U: *mxn* matrix of the orthonormal eigenvectors of .
- V^T: transpose of a *nxn* matrix containing the orthonormal eigenvectors of A^{T}A.
- W: a *nxn* diagonal matrix of the singular values which are the square roots of the eigenvalues of .

Principal Component Analysis

The Principal Component Analysis is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss. It helps to find the most significant features in a dataset and makes the data

easy for plotting in 2D and 3D. PCA helps in finding a sequence of linear combinations of variables.

What is a Principal Component?

The Principal Components are a straight line that captures most of the variance of the data. They have a direction and magnitude. Principal components are orthogonal projections (perpendicular) of data onto lower-dimensional space.

Applications of PCA in Machine Learning

- PCA is used to visualize multidimensional data.
- It is used to reduce the number of dimensions in healthcare data.
- PCA can help resize an image.
- It can be used in finance to analyze stock data and forecast returns.
- PCA helps to find patterns in the high-dimensional datasets.
-

How does Principal Component Analysis Work?

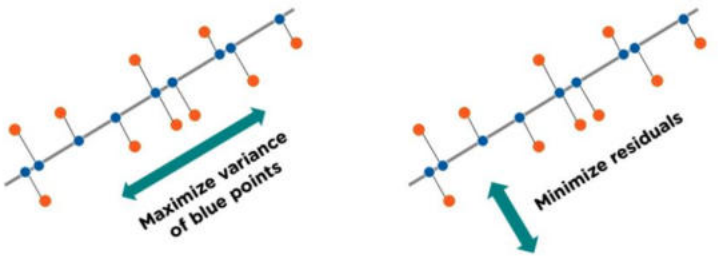
PCA works by considering the variance of each attribute because the high attribute shows the good split between the classes, and hence it reduces the dimensionality. Some real-world applications of PCA are ***image processing, movie recommendation system, optimizing the power allocation in various communication channels.*** It is a feature extraction technique, so it contains the important variables and drops the least important variable.

The PCA algorithm is based on some mathematical concepts such as:

- Variance and Covariance
- Eigenvalues and Eigen factors

Some common terms used in PCA algorithm:

- **Dimensionality:** It is the number of features or variables present in the given dataset. More easily, it is the number of columns present in the dataset.
- **Correlation:** It signifies that how strongly two variables are related to each other. Such as if one changes, the other variable also gets changed. The correlation value ranges from -1 to +1. Here, -1 occurs if variables are inversely proportional to each other, and +1 indicates that variables are directly proportional to each other.
- **Orthogonal:** It defines that variables are not correlated to each other, and hence the correlation between the pair of variables is zero.
- **Eigenvectors:** If there is a square matrix M, and a non-zero vector v is given. Then v will be eigenvector if Av is the scalar multiple of v.
- **Covariance Matrix:** A matrix containing the covariance between the pair of variables is called the Covariance Matrix.



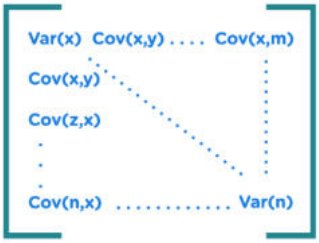
1. Normalize the data

Standardize the data before performing PCA. This will ensure that each feature has a mean = 0 and variance = 1.

$$Z = \frac{x - \mu}{\sigma}$$

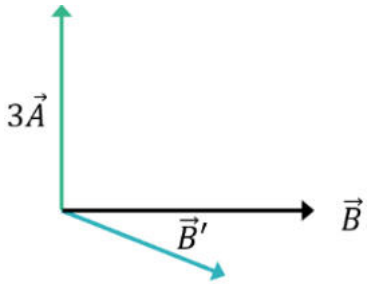
2. Build the covariance matrix

Construct a square matrix to express the correlation between two or more features in a multidimensional dataset.



3. Find the Eigenvectors and Eigenvalues

Calculate the eigenvectors/unit vectors and eigenvalues. Eigenvalues are scalars by which we multiply the eigenvector of the covariance matrix.



4. Sort the eigenvectors in highest to lowest order and select the number of principal components.

Some properties of these principal components are given below:

- The principal component must be the linear combination of the original features.
- These components are orthogonal, i.e., the correlation between a pair of variables is zero.
- The importance of each component decreases when going to 1 to n, it means the 1 PC has the most importance, and n PC will have the least importance.

Steps for PCA algorithm

1. **Getting the dataset**
   Firstly, we need to take the input dataset and divide it into two subparts X and Y, where X is the training set, and Y is the validation set.
2. **Representing data into a structure**
   Now we will represent our dataset into a structure. Such as we will represent the two-dimensional matrix of independent variable X. Here each row corresponds to the data items, and the column corresponds to the Features. The number of columns is the dimensions of the dataset.
3. **Standardizing the data**
   In this step, we will standardize our dataset. Such as in a particular column, the features with high variance are more important compared to the features with lower variance.
   If the importance of features is independent of the variance of the feature, then we

will divide each data item in a column with the standard deviation of the column. Here we will name the matrix as Z.

4. **Calculating the Covariance of Z**
   To calculate the covariance of Z, we will take the matrix Z, and will transpose it. After transpose, we will multiply it by Z. The output matrix will be the Covariance matrix of Z.
5. **Calculating the Eigen Values and Eigen Vectors**
   Now we need to calculate the eigenvalues and eigenvectors for the resultant covariance matrix Z. Eigenvectors or the covariance matrix are the directions of the axes with high information. And the coefficients of these eigenvectors are defined as the eigenvalues.
6. **Sorting the Eigen Vectors**
   In this step, we will take all the eigenvalues and will sort them in decreasing order, which means from largest to smallest. And simultaneously sort the eigenvectors accordingly in matrix P of eigenvalues. The resultant matrix will be named as P*.
7. **Calculating the new features Or Principal Components**
   Here we will calculate the new features. To do this, we will multiply the P* matrix to the Z. In the resultant matrix Z*, each observation is the linear combination of original features. Each column of the Z* matrix is independent of each other.
8. **Remove less or unimportant features from the new dataset.**
   The new feature set has occurred, so we will decide here what to keep and what to remove. It means, we will only keep the relevant or important features in the new dataset, and unimportant features will be removed out.

Applications of Principal Component Analysis

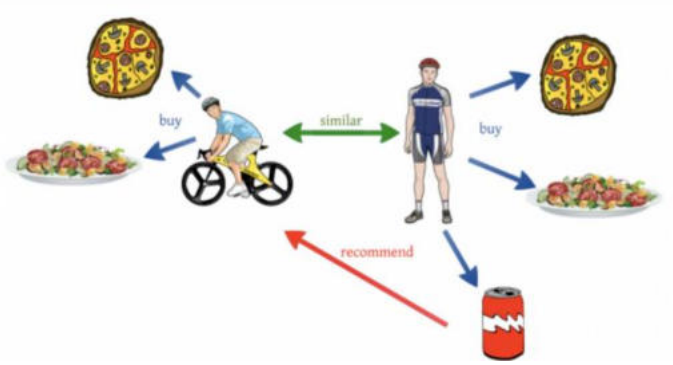
- PCA is mainly used as the dimensionality reduction technique in various AI applications such **as computer vision, image compression, etc.**
- It can also be used for finding hidden patterns if data has high dimensions. Some fields where PCA is used are Finance, data mining, Psychology, etc.

Types of Recommendation System

There are two primary types of recommendation systems, each with different sub-types. Depending on goals, audience, the platform, and what you're recommending, these different approaches can be employed individually, though generally, the best results come from using them in combination:

1 — Collaborative Filtering

It primarily makes recommendations based on inputs or actions from other people (rather than only the user for whom a recommendation is being made).

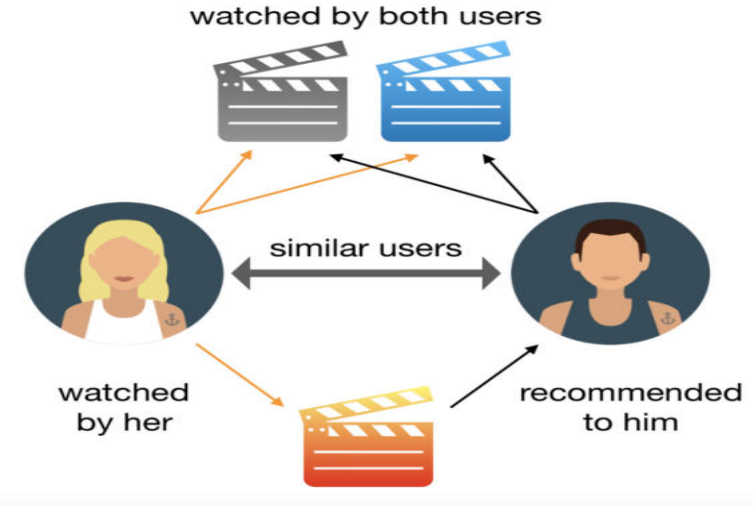


Variations on this type of recommendation system include:

- **By User Similarity:** This strategy involves creating user groups by comparing users' activities and providing recommendations that are popular among other members of the group.
- **By Association:** This is a specific type of the one mentioned above, otherwise known as "Users who looked at X also looked at Y." Implementing this type of recommendation system is a matter of looking at purchasing sequences or purchasing groups, and showing similar content.

2 — Content-Based

Content-based systems make recommendations based on the user's purchase or consumption history and generally become more accurate the more actions (inputs) the user takes.



More specific types of content-based recommendation systems include:

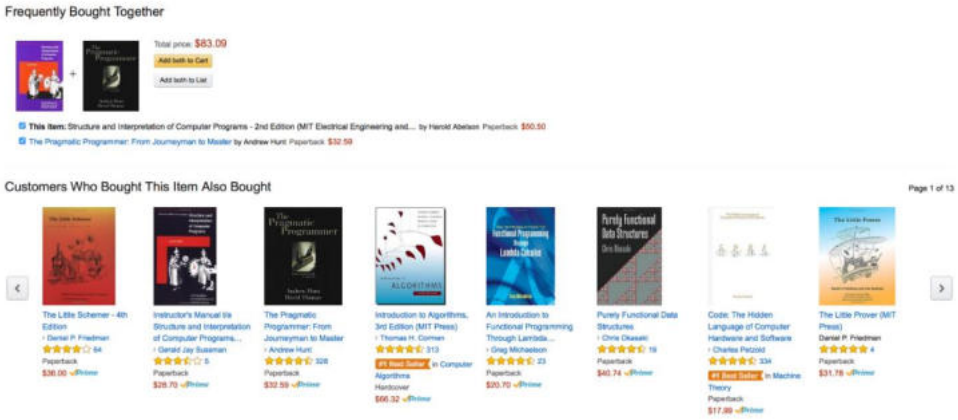
- **By Content Similarity:** As the most basic type of content-based recommendation system, this strategy involves recommending content that is close based on its metadata.
- **By Latent Factor Modeling:** Going one step further than the content similarity approach, the crux of this strategy is inferring individuals' inherent interests by assuming that previous choices are indicative of certain tastes or hobbies.
- **By Topic Modeling:** This is a variant of the Latent Factor Modeling strategy, whereby instead of considering users' larger actions, one would infer interests by analyzing unstructured text to detect particular topics of interest. It is particularly interesting for use cases with rich but unstructured textual information (such as news articles).

- **By Popular Content Promotion:** This involves highlighting product recommendations based on the product's intrinsic features that may make it interesting to a wide audience: price, feature, popularity, etc.

What is Recommendation system?

A recommendation system is an extensive class of web applications that involves predicting the user responses to the options.

You could have seen below image example for amazon recommendation system,



How to design a recommendation system?

machine learning (ML) is commonly used in building recommendation systems, it doesn't mean it's the only solution. There are many ways to build a recommendation system? simpler approaches, for example, we may have very few data, or we may want to build a minimal solution fast etc..

Assume that, for simpler video recommendation,In such that case, based on videos a user has watched, we can simply suggest same authors videos or same publications videos.

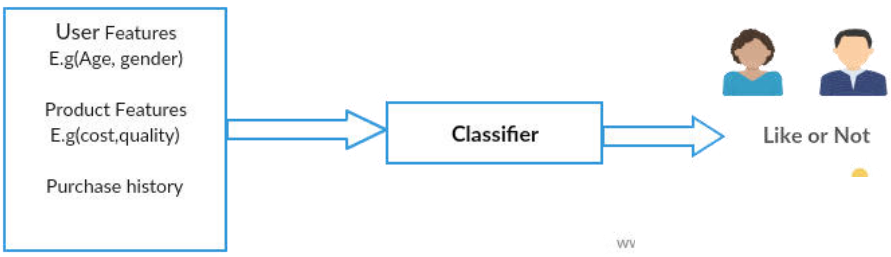
1. *popularity based*
2. *classification based*
3. *collaborative filtering*
4. **Popularity based:**

Easiest way to build a recommendation system is popularity based, simply over all the products that are popular, So how to identify popular products, which could be identified by which are all the products that are bought most,

Example, In shopping store we can suggest popular dresses by purchase count.

2.Classification based

Second way to build a recommendation system is classification model , In that use feature of both users as well as products in order to predict whether this product liked or not by the user. When new users come, our classifier will give a binary value of that product liked by this user or not, In such a way that we can recommend a product to the user .



Classifier

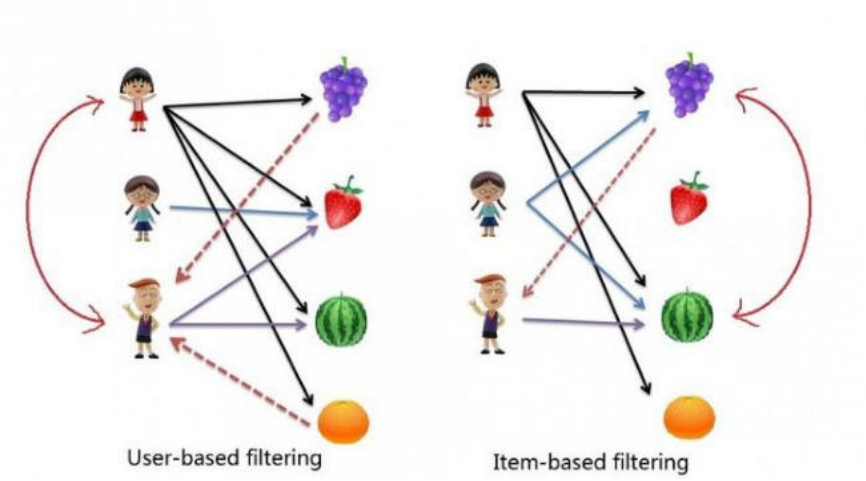
In above example using user features like Age, gender and product features like cost, quality and product history, based on this input our classifier will give a binary value user may like or not , based on that boolean we could recommend product to a customer

Collaborative filtering:
collaborative filtering models which are based on assumption that people like things similar to other things they like, and things that are liked by other people with similar taste.
collaborative filtering models are two types,

I.Nearest neighbor

II.Matrix factorization
let me explain each method of collaborative filtering in a nutshell,

N**earest neighbor collaborative filtering:**
In these type of recommendation systems are recommending based on nearest neighbors, nearest neighbor approach used to find out either similar users or similar products,
It can be looked at two ways,
i.User based filtering
ii.Item based filtering



Nearest neighbor collaborative filtering

User-based collaborative filtering:
Find the users who have similar taste of products as the current user , similarity is based on purchasing behavior of the user, so based on the neighbor purchasing behavior we can recommend items to the current user.

Item-based collaborative filtering :
Recommend Items that are similar to the item user bought,similarity is based on co-occurrences of purchases
Item A and B were purchased by both users X and Y then both are similar.

Matrix factorization:

It is basically model based collaborative filtering and matrix factorization is the **important technique in recommendation system.**
let me give an abstractive explanation for matrix factorization,
When a user gives feed back to a certain movie they saw (say they can rate from one to five), this collection of feedback can be represented in a form of a matrix. Where each row represents each users, while each column represents different movies. Obviously the matrix will be sparse since not everyone is going to watch every movies, (we all have different taste when it comes to movies).

Hybrid Recommendation systems:
Hybrid Recommendation systems are combining collaborative and content-based recommendation can be more effective. Hybrid approaches can be implemented by making content-based and collaborative-based predictions separately and then combining them.

UNIT 5

DATA VISUALIZATION: Types of data visualization, plots, graphs and summary statistics, Data for visualization, Technologies for visualization. Social Network Analysis- Data Engineering – MapReduce, Pregel, Hadoop, Next Generation Data Scientists- Applications of Data Science- Recent trends and development in Data Science.

INTRODUCTION:

A picture is worth more than thousands of words. People like to see pictures rather than read words. That's why visualization matters in all data science project lifecycle steps. From data understanding to model validation, data visualization plays an important role.

What is Data Visualization?

In simple terms, Data Visualization (DataViz) is the process of generating graphical representations of data for various purposes. These graphical representations are commonly known as plots or charts in data science terminology.

Data visualization convert large and small data sets into visuals, which is easy to understand and process for humans.

Data visualization tools provide accessible ways to understand outliers, patterns, and trends in the data.

In the world of Big Data, the data visualization tools and technologies are required to analyze vast amounts of information.

Why Use Data Visualization?

1. To make easier in understand and remember.
2. To discover unknown facts, outliers, and trends.
3. To visualize relationships and patterns quickly.
4. To ask a better question and make better decisions.
5. To competitive analyze.
6. To improve insights.

Why is Data Visualization Important in Data Science?

There are many reasons for data visualization in data science. Data visualization benefits include communicating your results or findings, monitoring the model's performance at the evaluation stage, hyperparameter tuning, identifying trends, patterns and correlation between dataset features, data cleaning such as outlier detection, and validating model assumptions.

What Makes Data Visualization Effective?

To get the most out of data visualization, you should consider the following things. These are the fundamentals of data visualization.

- **Clarity:** Data should be visualized in a way that everyone can understand.
- **Problem domain:** When presenting data, the visualizations should be related to the business problem.
- **Interactivity:** Interactive plots are useful to compare and highlight certain things within the plot.
- **Comparability:** We can compare the thighs easily with good plots.

- **Aesthetics:** Quality plots are visually aesthetic.
- **Informative:** A good plot summarizes all relevant information.

Importance of Data Visualization in Data Science

Earlier, I mentioned the importance of data visualization in data science. Here are some more details.

1. Data cleaning
Data visualization plays an important role in data clearing. Good examples are detecting outliers and removing multicollinearity. We can create scatterplots to detect outliers and generate heatmaps to check multicollinearity.

2. Data Exploration
Before building any model, we need to do some exploratory data analysis to identify dataset characteristics. For example, we can create histograms for continuous variables to check for normality in the data. We can create scatterplots between two features to check whether they are correlated. Likewise, we can create a bar chart for the label column with two or more classes to identify class imbalance.

3. Evaluation of modeling outputs
We can create a confusion matrix and learning curve to measure the performance of a model during training. Plots are also useful in validating model assumptions. For example, we can create a residuals plot and histogram for the distribution of residuals to validate the assumptions of a linear regression model.

4. Identifying trends
Time and seasonal plots are useful in time series analysis to identify certain trends over time.

5. Presenting results
As a data scientist, you need to present your findings to the company or other related persons who do not have more knowledge in the subject domain. So, you need to explain everything in plain English. You can use informative plots that summarize your findings.

Types of Data Visualization
The two basic types of data visualization are static visualization and interactive visualization.

STATIC VISUALIZATION
Static visualization refers to a method of displaying data that tells a specific story and focuses on only a single data relationship. A common example of static visualization is an engaging single-page layout like an infographic.

INTERACTIVE VISUALIZATION
Interactive visualizations, for the most part, only exist within software or web applications. This model allows users to select specific data points in order to present findings and create customized visual stories to compare against each other, thereby creating the opportunity for stakeholders to choose from a selection of insights to determine the best path forward, rather than deciding based on a single insight.

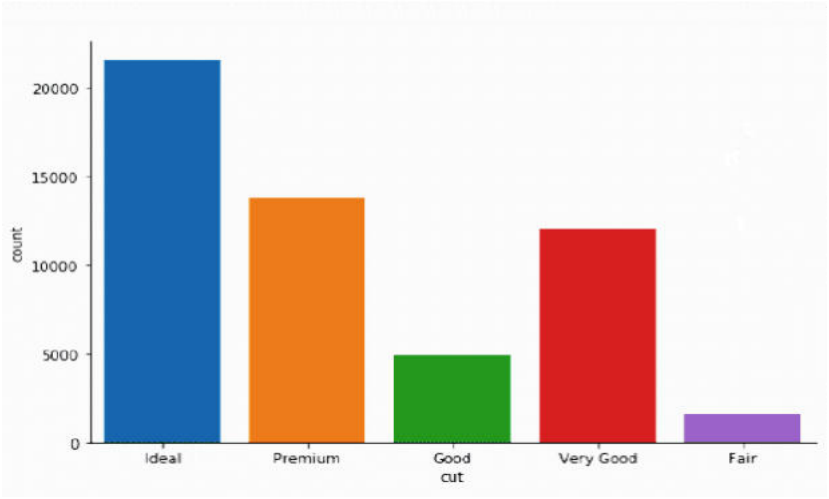
Different Types of Data Visualization

There are many data visualization types. The following are the commonly used data visualization charts.

These data plot types for visualization are sometimes called graphs or charts depending on the context.

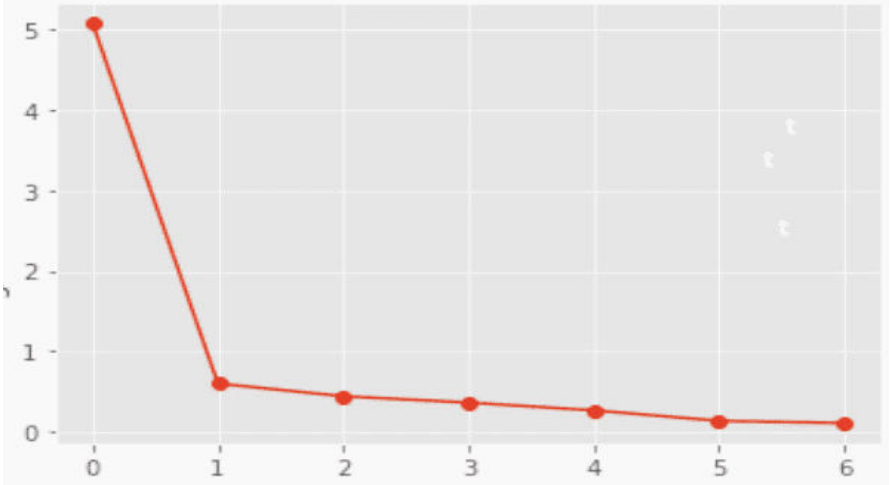
Bar plot

A bar plot is used to plot the frequency of occurring categorical data. Each category is represented by a bar. The bars can be created vertically or horizontally. Their heights or lengths are proportional to the values they represent.



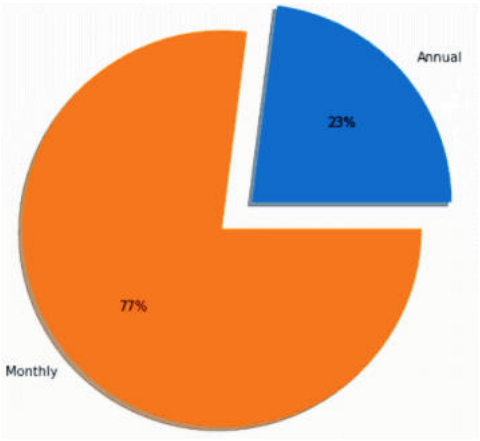
Line plot

A line plot is created by connecting a series of data points with straight lines. The number of periods is on the x-axis.



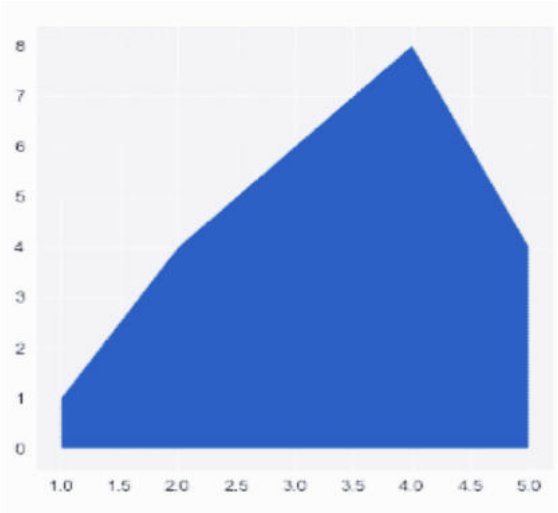
Pie chart

A categorical variable pie chart includes each category's values as slices whose sizes are proportional to the quantity they represent. It is a circular graph made with slices equal to the number of categories.



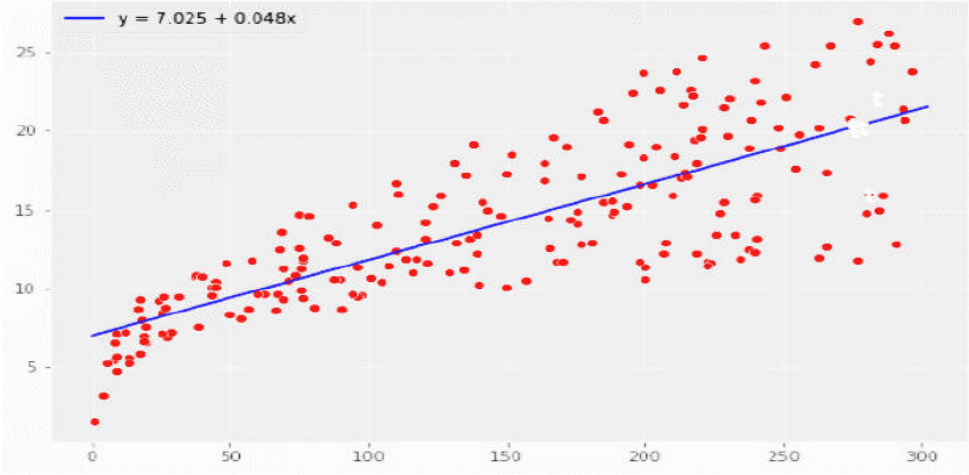
Area plot

The area plot is based on the line chart. We get the area plot when we cover the area between the line and the x-axis.



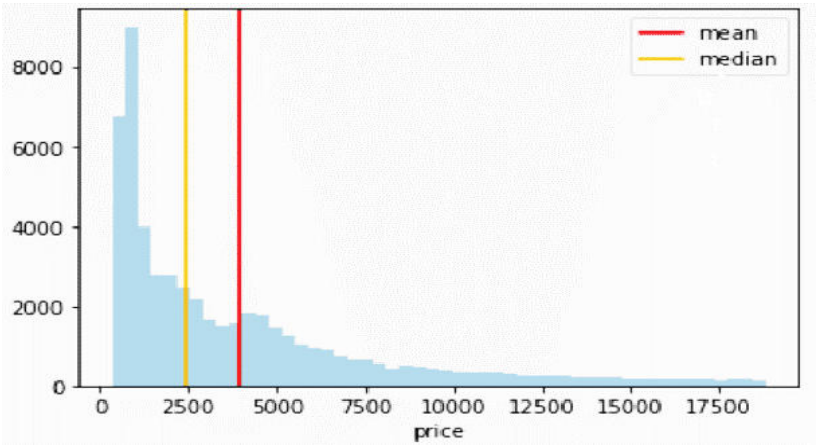
Scatter plot

Scatter plots are created to see whether there is a relationship (linear or non-linear and positive or negative) between two numerical variables. They are commonly used in regression analysis.



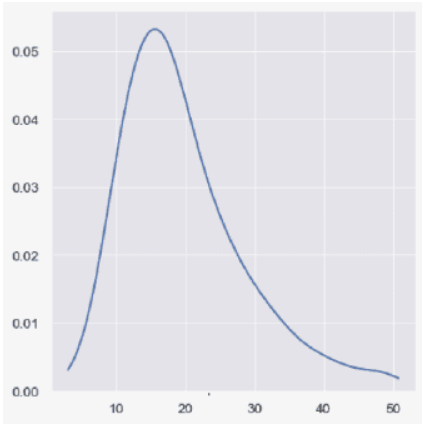
Histogram

A histogram represents the distribution of numerical data. Looking at a histogram, we can decide whether the values are normally distributed (a bell-shaped curve), skewed to the right or skewed left. A histogram of residuals is useful to validate important assumptions in regression analysis.



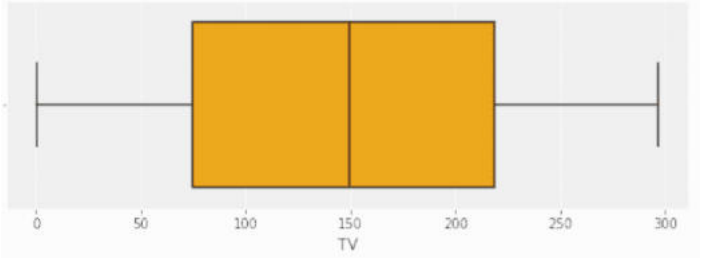
Distribution plot

A distribution plot is used to visualize data distribution. Example: Probability distribution plot or density curve.



Box and whisker plot

This plot is used to plot the variation of the values of a numerical feature. You can get the values' minimum, maximum, median, lower and upper quartiles.



Data Visualization Process/Workflow

The data visualization process or workflow includes the fowling key steps.

1. Develop your research question

This may be a business problem or any other related problem that could be solved with a data-driven approach. You should note all the objectives and outcomes plus required resources such as datasets, open-source software libraries, etc.

2. Get or create your data

The next step is collecting data. You can use existing datasets if they're relevant to your research question.

3. Clean your data

Real-world data are messy. So, you need to clean them before using them for visualization. You can identify missing values and outliers and treat them accordingly. You can perform feature selection and remove unnecessary features from the data. You can create a new set of features based on the original features.

4. Choose a chart type

The chart type depends on many factors. For example, it depends on the feature type (numerical or categorical). It also depends on the type of visualization you need. Let's say you have two numerical features. If you want to find their distributions, you can create two histograms for each feature.

5. Choose your tool

You can use open-source data visualization tools such as matplotlib, seaborn, plotty and ggplot. You can also use API-based software such as Matlab, Minitab, SPSS, etc.

6. Prepare data

You can extract relevant features. You can do feature standardization if the values of the features are not on the same scale. You can apply data preprocessing steps such as PCA to reduce the dimensionality of the data. That will allow you to visualize high-dimensional data in 2D and 3D plots!

7. Create a chart

This is the final step. Here. You define the title and names for the axes. You should also choose a proper chart background to ensure the content is easily readable.

Tools and Software for Data Visualization

There are multiple tools and software available for data visualization.

1. Python provides open-source libraries such as

- Matplotlib
- Seaborn
- Plotty
- Bokeh
- Altair

2. R provides open-source libraries such as

- Ggplot2
- Lattice

3. Other data visualization libraries

- IBM SPSS
- Minitab
- Matlab for data visualization
- Tableau
- Microsoft Power BI are popular among data scientists.

Data Visualization Techniques in Data Science

Some of the main data visualization techniques in data science are univariate analysis, bivariate analysis and multivariate analysis.

1. Univariate Analysis

In univariate analysis, as the name suggest, we analyze only one variable at a time. In other words, we analyze each variable separately. Bar charts, pie charts, box plots and histograms are common examples of univariate data visualization. Bar charts and pie charts are created

for categorical variables, while box plots and histograms are created for numerical variables.

2. Bivariate Analysis

In bivariate analysis, we analyze two variables at a time. Often, we see whether there is a relationship between the two variables. The scatter plot is a classic example of bivariate data visualization.

3. Multivariate Analysis

In multivariate analysis, we analyze more than two variables simultaneously. The heatmap is a classic example of multivariate data visualization. Other examples are cluster analysis and principal component analysis (PCA).

Advantages and Disadvantages of Data Visualization

Advantages

There are many advantages of data visualization. Data visualization is used to:

- Communicate your results or findings with your audience
- Tune hyperparameters
- Identify trends, patterns and correlations between variables
- Monitor the model's performance
- Clean data
- Validate the model's assumptions

Disadvantages

There are also some disadvantages of data visualization.

- We need to download, install and configure software and open-source libraries. The process will be difficult and time-consuming for beginners.
- Some data visualization tools are not available for free. We need to pay for those.
- When we summarize the data, we'll lose the exact information.

Examples of Data Visualization in Data Science

Here are some popular data visualization examples.

1. **Weather reports:** Maps and other plot types are commonly used in weather reports.
2. **Internet websites:** Social media analytics websites such as Social Blade and Google Analytics use data visualization techniques to analyze and compare the performance of websites.
3. **Astronomy:** NASA uses advanced data visualization techniques in its reports and presentations.
4. **Geography**
5. **Gaming industry**

Data Visualization Tools/Technologies/software's:

Data visualization allows you to interact with data. **Google**, **Apple**, **Facebook**, and **Twitter** all ask better a better question of their data and make a better business decision by using data visualization.

1. Tableau

Tableau is a data visualization tool. You can create graphs, charts, maps, and many other graphics.
A tableau desktop app is available for visual analytics. If you don't want to install tableau software on your desktop, then a server solution allows you to visualize your reports online and on mobile.

2. Infogram
Infogram is also a data visualization tool. It has some simple steps to process that:

1. First, you choose among many templates, personalize them with additional visualizations like maps, charts, videos, and images.
2. Then you are ready to share your visualization.
3. Infogram supports team accounts for journalists and media publishers, branded designs of classroom accounts for educational projects, companies, and enterprises.

JupyteR
A web-based application, JupyteR, is one of the top-rated data visualization tools that enable users to create and share documents containing visualizations, equations, narrative text, and live code. JupyteR is ideal for data cleansing and transformation, statistical modeling, numerical simulation, interactive computing, and machine learning.

Google Charts
One of the major players in the data visualization market space, Google Charts, coded with SVG and HTML5, is famed for its capability to produce graphical and pictorial data visualizations. Google Charts offers zoom functionality, and it provides users with unmatched cross-platform compatibility with iOS, Android, and even the earlier versions of the Internet Explorer browser.

Zoho Reports
Zoho Reports, also known as Zoho Analytics, is a comprehensive data visualization tool that integrates Business Intelligence and online reporting services, which allow quick creation and sharing of extensive reports in minutes. The high-grade visualization tool also supports the import of Big Data from major databases and applications.

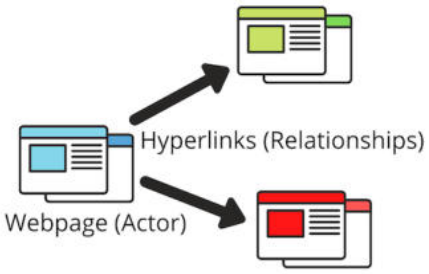
Visual.ly
Visual.ly is one of the data visualization tools on the market, renowned for its impressive distribution network that illustrates project outcomes. Employing a dedicated creative team for data visualization services, Visual.ly streamlines the process of data import and outsource, even to third parties.

RAW
RAW creates the missing link between spreadsheets and vector graphics on its home page.

What Is Social Network Analysis?

Social network analysis (SNA), also known as network science, is a field of data analytics that uses networks and graph theory to understand social structures. SNA techniques can also be applied to networks outside of the societal realm.

social network analysis (SNA), will give you valuable tools to gain insight on a variety of data sources.



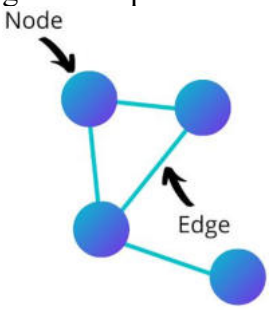
In order to build SNA graphs, we need two key components: actors and relationships. We commonly use SNA techniques with the internet. Web pages often link to other sites — either on their own website or an external page. These links can be considered relationships between actors (web pages) and this is a key component of search engine architecture.

Defining Our Terms

NODES AND EDGES

Nodes can represent a variety of actors. For example, in internet networks nodes can represent web pages while in social networks nodes can represent people. While nodes can represent a variety of things, each node always has a relationship with another thing.

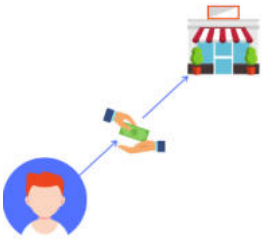
Edges can represent a variety of relationships. In internet networks, edges can represent hyperlinks and in social networks edges can represent connections.



EDGE DIRECTION

There are two types of edges: directed and undirected. It will be necessary to decipher what type of edge your data contains when building a network graph.

Directed edges are applied from one node to another with a starting node and an ending node.



Undirected edges are the opposite of directed edges. These relationships are reciprocated by both parties without a clear starting node or ending node.

EDGE WEIGHT

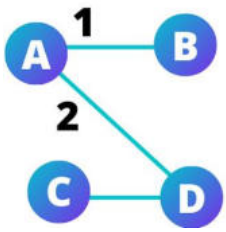
An edge's weight is the number of times that edge appears between two specific nodes. For example, if person A buys a coffee from a coffee shop three times, the edge connecting person A and the coffee shop will have a weight of three.

CENTRALITY MEASURES

Centrality is a collection of metrics used to quantify how important and influential a specific node is to the network as a whole.

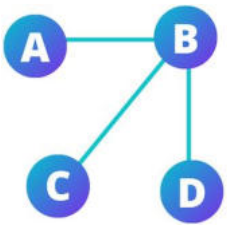
Degree

A node's degree is the number of edges the node has. In an undirected network, there's only one measure for degree. For example, if node A has edges connecting it to node B and node D, then node A's degree is two.



Closeness

Closeness measures how well connected a node is to every other node in the network. A node's closeness is the average number of hops required to reach every other node in the network. A hop is the path of an edge from one node to another. For example, node A is connected to node B, and node B is connected to node C. For node A to reach node C it would take two hops.



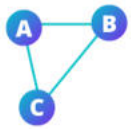
Betweenness

Betweenness measures the importance of a node's connections in allowing nodes to reach other nodes (in a hop).

NETWORK-LEVEL MEASURES

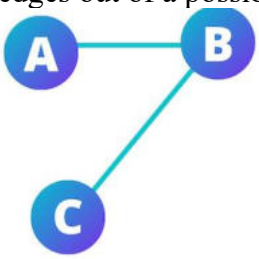
Network Size

Network size is the number of nodes in the network. The size of a network does not take into consideration the number of edges. For example, a network with nodes A, B, and C has a size of three.



Network Density

Network density is the number of edges divided by the total possible edges. For example, a network with node A connected to node B, and node B connected to node C, the network density is 2/3 because there are two edges out of a possible three.



Types of Social Networks Analysis

Social networks are the networks that depict the relations between people in the form of a graph for different kinds of analysis. The graph to store the relationships of people is known as Sociogram. All the graph points and lines are stored in the matrix data structure called Sociomatrix. The relationships indicate of any kind like kinship, friendship, enemies, acquaintances, colleagues, neighbors, disease transmission, etc.

Social Network Analysis (SNA) is the process of exploring or examining the social structure by using graph theory. It is used for measuring and analyzing the structural properties of the network. It helps to measure relationships and flows between groups, organizations, and other connected entities. We need specialized tools to study and analyze social networks.

Basically, there are two types of social networks:

- Ego network Analysis
- Complete network Analysis

1. Ego Network Analysis

Ego network Analysis is the one that finds the relationship among people. The analysis is done for a particular sample of people chosen from the whole population. This sampling is done randomly to analyze the relationship. The attributes involved in this ego network analysis are a person's size, diversity, etc.

2. Complete Network Analysis

Complete network analysis is the analysis that is used in all network analyses. It analyses the relationship among the sample of people chosen from the large population. Subgroup analysis, centrality measure, and equivalence analysis are based on the complete network analysis.

What is Network Analysis?

Network analysis is the study of social relations among a set of actors. It is a field of study -- a set of phenomena or data which we seek to understand.

What is Data Engineering?

Data engineering is the complex task of making raw data usable to data scientists and groups within an organization. Data engineering encompasses numerous specialties of data science.
In addition to making data accessible, data engineers create raw data analyses to provide predictive models and show trends for the short- and long-term. Without data engineering, it would be impossible to make sense of the huge amounts of data that are available to businesses.
Data engineers focus on the applications and harvesting of big data.

Data Engineering vs Data Science

Parameters	Data Engineering	Data Science
Data role	It is the role of the "architect" of the data."	It is the role of "the builder of the architect's plan."
Dependent on	It depends on managers, non-technical executives, and stakeholders to	It is dependent on the data that the engineer provides.

	understand the organization's requirements.	
Skills required	We require skills in the following areas: Data Warehousing, ETL, Advanced Programming, Hadoop, SQL, Data Infrastructure and Pipelining, Machine Learning, etc.	Statistical analysis, data visualization, data mining, machine learning, and artificial intelligence, as well as R or Python and SAS are the skills that are required to succeed in this field.
Deals with	Handles raw data in a variety of ways.	In this role, you will deal with data that the data engineers have manipulated.
Storytelling skills	You don't need to have any storytelling skills to convey the message.	For the presentation of the analysis to be successful, the Data Scientist needs to be a good storyteller.
Responsibility	Assumes responsibility for ensuring that data is accurate.	Provides communication between stakeholders and customers to create a connection between them.
Tools and Programming languages used	To process data, the following tools are used: MySQL, Hive, Oracle, Cassandra, PostgreSQL, MongoDB, and Sqoop.	Many programming languages are being used, including Python, R, SAS, SPSS, and Julia, as well as different visualization methods.
Decision making	There is no say in the decision-making process.	A company considers data scientists' analysis when making decisions.

Data Engineers

Your responsibilities in this role are:

- Data Mining for getting insights from data
- Conversion of erroneous data into a useable form for data analysis
- Writing queries on data
- Maintenance of the data design and architecture
- Develop large data warehouses with the help of extra transform load (ETL)

What is MapReduce in Hadoop

MapReduce is a Java-based, distributed execution framework within the Apache Hadoop Ecosystem. **MapReduce** is a software framework and programming model used for processing huge amounts of data. **MapReduce** program work in two phases, namely, Map and Reduce.

Map tasks deal with splitting and mapping of data while Reduce tasks shuffle and reduce the data.

How MapReduce Organizes Work?

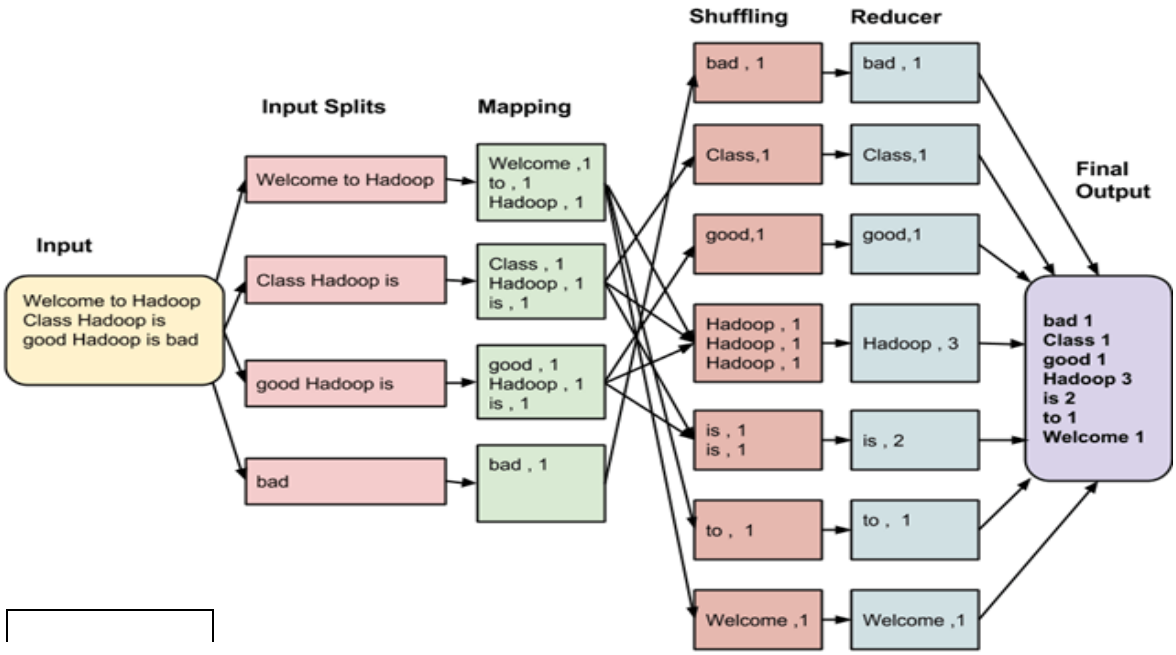
There are two types of tasks:

1. **Map tasks** (Splits & Mapping)
2. **Reduce tasks** (Shuffling, Reducing)

Steps in Map Reduce

- The map takes data in the form of pairs and returns a list of <key, value> pairs. The keys will not be unique in this case.
- Using the output of Map, sort and shuffle are applied by the Hadoop architecture. This sort and shuffle acts on these list of <key, value> pairs and sends out unique keys and a list of values associated with this unique key <key, list(values)>.
- An output of sort and shuffle sent to the reducer phase. The reducer performs a defined function on a list of values for unique keys, and Final output <key, value> will be stored/displayed.

Consider you have following input data for your MapReduce in Big data Program
Welcome to Hadoop Class
Hadoop is good
Hadoop is bad



The final output of the MapReduce task is

bad	1
Class	1
good	1
Hadoop	3
is	2
to	1
Welcome	1

The data goes through the following phases of MapReduce in Big Data

Input Splits:
An input to a MapReduce in Big Data job is divided into fixed-size pieces called **input splits** Input split is a chunk of the input that is consumed by a single map

Mapping
This is the very first phase in the execution of map-reduce program. In this phase data in each split is passed to a mapping function to produce output values. In our example, a job of mapping phase is to count a number of occurrences of each word from input splits (more details about input-split is given below) and prepare a list in the form of <word, frequency>

Shuffling
This phase consumes the output of Mapping phase. Its task is to consolidate the relevant records from Mapping phase output. In our example, the same words are clubed together along with their respective frequency.

Reducing
In this phase, output values from the Shuffling phase are aggregated. This phase combines values from Shuffling phase and returns a single output value. In short, this phase summarizes the complete dataset.

What is Hadoop?
Hadoop is an open-source software platform that uses basic programming principles to process enormous data sets across clusters of computers. Hadoop is built to scale from a single server to tens of thousands of computers.

Doug Cutting and Mike Cafarella created the open-source search engine Nutch, which gave rise to Hadoop. The two hoped to design a way to return web search results faster by sharing data and calculations across numerous computers so that multiple activities could be completed at the same time in the early days of the Internet.

While the platform is designed in Java, hadoop for data science can be programmed in a variety of languages including Python, C++, Perl, Ruby, and others.
After Google published a research paper that also explained its Google File System, Big Data concepts such as MapReduce became popular.

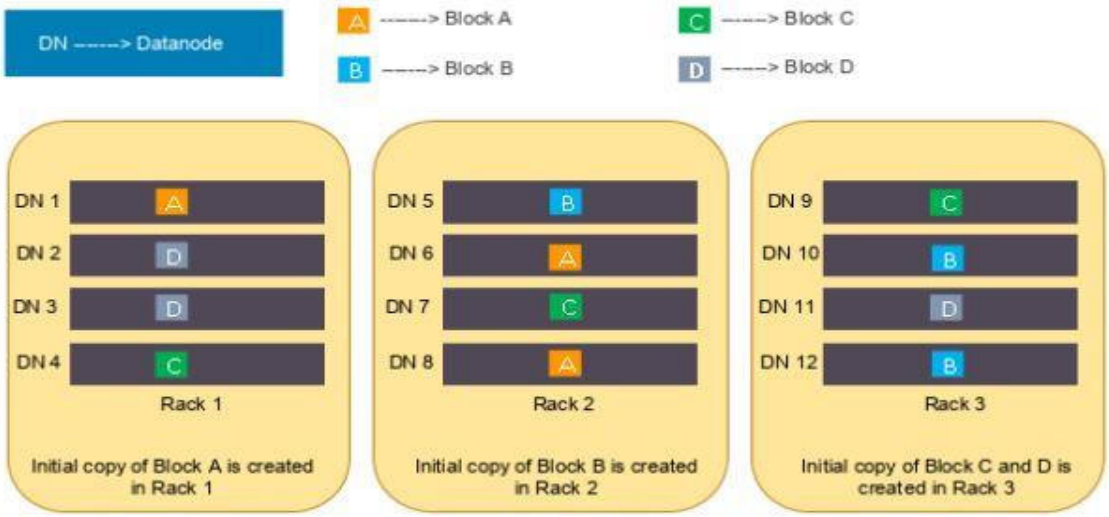
Hadoop Architecture

The Hadoop architecture is a package of the file system, MapReduce engine and the HDFS (Hadoop Distributed File System). The MapReduce engine can be MapReduce/MR1 or YARN/MR2.

A Hadoop cluster consists of a single master and multiple slave nodes. The master node includes Job Tracker, Task Tracker, NameNode, and DataNode whereas the slave node includes DataNode and TaskTracker.

HDFS in Hadoop Architecture divides large data into different blocks. Replicated three times by default, each block contains 128 MB of data. Replications operate under two rules:

1. Two identical blocks cannot be placed on the same DataNode
2. When a cluster is rack aware, all the replicas of a block cannot be placed on the same rack



In this example, blocks A, B, C, and D are replicated three times and placed on different racks. If DataNode 7 crashes, we still have two copies of block C data on DataNode 4 of Rack 1 and DataNode 9 of Rack 3.

There are three components of the Hadoop Distributed File System:

1. NameNode (a.k.a. masternode): Contains metadata in RAM and disk
2. Secondary NameNode: Contains a copy of NameNode's metadata on disk
3. Slave Node: Contains the actual data in the form of blocks

HDFS has a **Master-slave architecture**. The daemon called NameNode runs on the master server. It is responsible for Namespace management and regulates file access by the client. DataNode daemon runs on slave nodes. It is responsible for storing actual business data. Internally, a file gets split into a number of data blocks and stored on a group of slave machines. Namenode manages modifications to file system namespace. These are actions like the opening, closing and renaming files or directories. NameNode also keeps track of mapping of blocks to DataNodes. This DataNodes serves read/write request from the file system's client. DataNode also creates, deletes and replicates blocks on demand from NameNode.

Who uses Hadoop?

Hadoop is used by a number of tech powerhouses. Here's a quick round-up of who uses Hadoop for what:

- Hadoop is used by **eBay** for search optimization.
- Hadoop is utilised at **Facebook** to store copies of internal log and dimension data sources, as well as a source for reporting, analytics, and machine learning.
- **LinkedIn's** People You May Know functionality is powered by Hadoop.
- **Opower** use Hadoop to recommend ways for customers to save money on their energy costs.
- **Orbitz** analyses every element of visitors' sessions on its websites using Hadoop to discover user preferences.
- Hadoop is used by **Spotify** for content creation as well as data collection, reporting, and analysis.

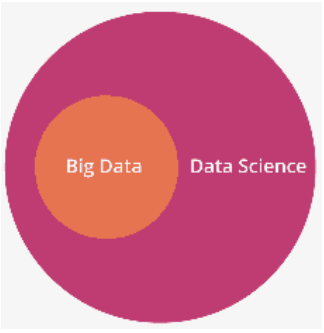
- Hadoop is used by **Twitter** to store and process tweets as well as log files.

Hadoop for Data Science

Data science is a broad topic. It is derived from a variety of disciplines like mathematics, statistics, and programming.

Data Scientists are skilled at extracting, analysing, and predicting information from large amounts of data. It is a broad phrase that encompasses practically all data-related technologies.

Hadoop's primary job is storage of Big Data. It also enables users to store various types of data, including structured and unstructured data.



However, data science differs from big data in that the former is a discipline that encompasses all data operations. As a result, Big Data is now considered a subset of Data Science. It is not required to understand Big Data because Data Science contains a sea of knowledge.

Hadoop skills, on the other hand, will undoubtedly add to your expertise, allowing you to handle massive amounts of data with ease. Understanding the use of hadoop in data science will also enhance your market value by a significant amount, giving you a competitive advantage over rivals.

Anatomy of Hadoop

Hadoop has a four-part design that supports two primary functions. The modules are as follows:

- Hadoop Common — Useful utilities and tools that the other modules refer to.
- Hadoop Distributed File System (HDFS) is a high-throughput file storage system developed by Hadoop.
- Hadoop YARN is a distributed process allocation job-scheduling framework.
- Hadoop MapReduce is a YARN-based parallel processing tool.

How Hadoop Improves on Traditional Databases

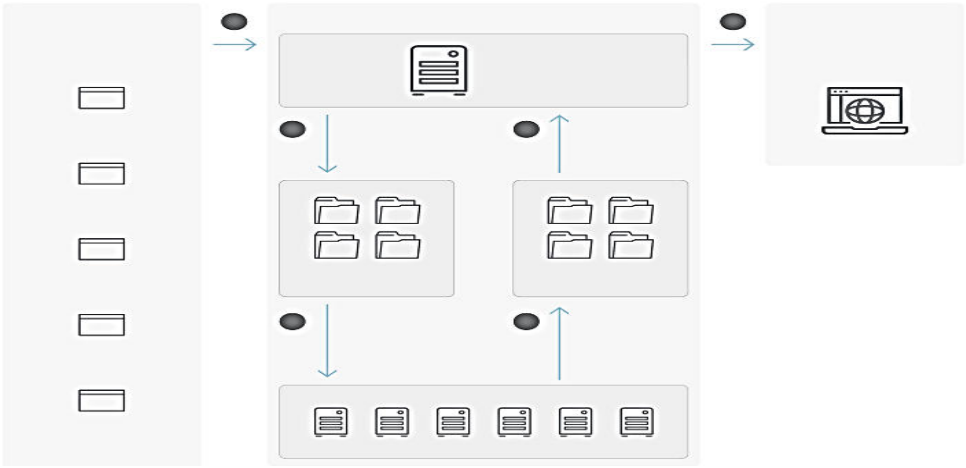
Hadoop solves two key challenges with traditional databases:

1. Capacity: Hadoop stores large volumes of data.

By using a distributed file system called an HDFS (Hadoop Distributed File System), the data is split into chunks and saved across clusters of commodity servers. As these commodity servers are built with simple hardware configurations, these are economical and easily scalable as the data grows.

2. Speed: Hadoop stores and retrieves data faster.

Hadoop uses the MapReduce functional programming model to perform parallel processing across data sets. So, when a query is sent to the database, instead of handling data sequentially, tasks are split and concurrently run across distributed servers. Finally, the output of all tasks is collated and sent back to the application, drastically improving the processing speed.



Where to use HDFS

- **Very Large Files:** Files should be of hundreds of megabytes, gigabytes or more.
- **Streaming Data Access:** The time to read whole data set is more important than latency in reading the first. HDFS is built on write-once and read-many-times pattern.
- **Commodity Hardware:**It works on low cost hardware.

Where not to use HDFS

- **Low Latency data access:** Applications that require very less time to access the first data should not use HDFS as it is giving importance to whole data rather than time to fetch the first record.
- **Lots Of Small Files:**The name node contains the metadata of files in memory and if the files are small in size it takes a lot of memory for name node's memory which is not feasible.
- **Multiple Writes:**It should not be used when we have to write multiple times.

What is Prequel?

It makes it easy for B2B companies to send data to their customers. Specifically, It helps companies sync data directly to their customer's data warehouse, on an ongoing basis. It exposes a RESTful API to configure sources and destinations, initiate data transfers, monitor activity, and more.

Prequel is a tool in the **Big Data as a Service** category of a tech stack.

Data from any source can be put to analysis enable reduction in costs, time reductions, product development and offerings that are optimized and finally smart decisions. Big data combined with powerful analytics helps

1. Determine the causes of failures, defects and issues in near-real time.
2. Generate coupons at sale points based on customer behavior
3. Recalculate risk portfolios quickly
4. Detect fraud behavior before the organization is affected.

Prequel Integrations

MySQL, PostgreSQL, Amazon S3, Google BigQuery, and **Amazon Redshift** are some of the popular tools that integrate with Prequel. Here's a list of all **7** tools that integrate with Prequel.

Prequel's Features

- Integrate with every data warehouse

- Start sending data in minutes, not months
- Expand revenue, land enterprise deals
- Get started in under an hour
- Enterprise grade security & compliance

Prequel Alternatives & Comparisons

WHAT ARE SOME ALTERNATIVES TO PREQUEL?
Google BigQuery
Run super-fast, SQL-like queries against terabytes of data in seconds, using the processing power of Google's infrastructure. Load data with ease. Bulk load your data using Google Cloud Storage or stream it in. Easy access. Access BigQuery by using a browser tool, a command-line tool, or by making calls to the BigQuery REST API with client libraries such as Java, PHP or Python.

- **Amazon Redshift**
  It is optimized for data sets ranging from a few hundred gigabytes to a petabyte or more and costs less than $1,000 per terabyte per year, a tenth the cost of most traditional data warehousing solutions.

- **Snowflake**
  Snowflake eliminates the administration and management demands of traditional data warehouses and big data platforms. Snowflake is a true data warehouse as a service running on Amazon Web Services (AWS)—no infrastructure to manage and no knobs to turn.

- **Amazon EMR**
  It is used in a variety of applications, including log analysis, data warehousing, machine learning, financial analysis, scientific simulation, and bioinformatics.

- **Stitch**
  Stitch is a simple, powerful ETL service built for software developers. Stitch evolved out of RJMetrics, a widely used business intelligence platform. When RJMetrics was acquired by Magento in 2016, Stitch was launched as its own company.

The next generation data scientist
The next generation of data scientist will maintain a breadth of hard technical skills such as mathematics, statistics, probability theory, machine learning, coding, data visualization, and data storytelling. Coding is important, so a good foundation in writing code along with good coding practices like agile software development techniques, code reviews, debugging, and version control are particularly valuable.

The data scientist role has exploded in popularity over the past decade as organisations have increasingly turned to data-driven decision making to stay competitive.
However, it was only in the early 2010s where the field started to take off at an astronomical pace. There are a few factors that have contributed to this exponential growth:

- **Exponential proliferation of data:** 95+% of all data has been created in the past 10 years.
- **Ongoing advancement of technology**: Today's smartphones are almost a thousand times faster than the mid-'80s Cray-2 Supercomputer, and several

multiples faster than the computer onboard NASA's Perseverence Rover currently exploring Mars .

- **Increased sophistication of learning algorithms/architectures:** e.g. New transformer-based deep learning architectures that have powered the latest generation of human-like natural language processing capabilities such as ChatGPT
- **Increasing importance of data-driven decision making:** Companies like Alphabet, (Google) and Amazon have built arguably the biggest and most successful companies by putting data at the heart of their business model.

Applications of Data Science

Data science has found its applications in almost every industry.

1. Healthcare

Healthcare companies are using data science to build sophisticated medical instruments to detect and cure diseases.

2. Gaming

Video and computer games are now being created with the help of data science and that has taken the gaming experience to the next level.

3. Image Recognition

Identifying patterns in images and detecting objects in an image is one of the most popular data science applications.

4. Recommendation Systems

Netflix and Amazon give movie and product recommendations based on what you like to watch, purchase, or browse on their platforms.

5. Logistics

Data Science is used by logistics companies to optimize routes to ensure faster delivery of products and increase operational efficiency.

6. Fraud Detection

Banking and financial institutions use data science and related algorithms to detect fraudulent transactions.

7. Internet Search

When we think of search, we immediately think of Google. Right? However, there are other search engines, such as Yahoo, Duckduckgo, Bing, AOL, Ask, and others, that employ data science algorithms to offer the best results for our searched query in a matter of seconds.

8. Speech recognition

Speech recognition is dominated by data science techniques. We may see the excellent work of these algorithms in our daily lives. Have you ever needed the help of a virtual speech assistant like Google Assistant, Alexa, or Siri? Well, its voice recognition technology is operating behind the scenes, attempting to interpret and evaluate your words and delivering useful results from your use.

9. Targeted Advertising

If you thought Search was the most essential data science use, consider this: the whole digital marketing spectrum. From display banners on various websites to digital billboards at airports, data science algorithms are utilised to identify almost anything. This is why digital advertisements have a far higher CTR (Call-Through Rate) than traditional marketing.

10. Airline Route Planning

As a result of data science, it is easier to predict flight delays for the airline industry, which is helping it grow. It also helps to determine whether to land immediately at the destination or to

make a stop in between, such as a flight from Delhi to the United States of America or to stop in between and then arrive at the destination.

- **Fintech:** Data science can help create credit reports and financial profiles, run accelerated underwriting and create predictive models based on historical payroll data.
- **Healthcare:** Data science can identify and predict disease, and personalize healthcare recommendations.
- **Transportation:** Data science can optimize shipping routes in real-time.
- **Sports:** Data science can accurately evaluate athletes' performance.
- **Government:** Data science can prevent tax evasion and predict incarceration rates.
- **E-commerce:** Data science can automate digital ad placement.
- **Gaming:** Data science can improve online gaming experiences.
- **Social media:** Data science can create algorithms to pinpoint compatible partners.

Recent trends and development in Data Science

Businesses are becoming more productive and increasing their return on investment. Today's trends include data analytics, artificial intelligence, big data, and data science. Business organizations are adopting data-driven models to simplify their processes and make decisions based on the insights derived from data analytics.

1. Artificial Intelligence

According to CMO, 47% of digitally mature organizations reported that they have a defined AI strategy in place. Artificial Intelligence or AI has been around for quite a long time. It has been used to make interaction with technology and collecting customer data easier over the decades. Due to its high processing speed and data access, it is now deeply rooted in your routine lifestyle.

2. Cloud Services

As humongous data is generated daily, it becomes a challenge to find solutions for low-cost storage and cheap power. This is where cloud computing and services come as a savior. Cloud services aim at storing large amounts of data for a low cost to efficiently tackle the issues encountered regarding storage in data science.

Growth of predictive analytics

By analyzing data of more than 100 million subscribers, Netflix was able to influence more than 80% of content watched by its users, thanks to accurate data insights.

Predictive analytics is all about predicting future trends and forecasts with the help of statistical tools and techniques leveraging past and existing data. With predictive analytics, organizations can make insightful business decisions that will help them grow.

AutoML

Automated Machine Learning, or AutoML, is one of the latest trends that is driving the democratization of data science. A huge part of a data scientist's job is spent on data cleansing and preparation, and each of these tasks are repetitive and time-consuming. AutoML ensures that these tasks are automated, and it involves building models, creating algorithms and neural networks.

TinyML
TinyML is a type of ML which shrinks deep learning networks so that it can be fit on any hardware. Its versatility, tiny-form factor, and cost-effectiveness make it one of the most exciting trends in the field of data science, with which a number of applications can be built.

Applications of TinyML:
- object recognition and classification
- gesture recognition
- keyword spotting
- machine monitoring
- audio detection

Augmented Consumer Interfaces
The near future might have an AI-agent in the form of an interface to help you with your shopping. You might be buying your products in VR, getting an idea about the product via audio or through an augmented consumer interface. Augment consumer interfaces can take multiple forms, it could be AR on mobile or a communication interface such as a Brain-Computer Interface (BCI).

AI as a Service (AIaaS)
It refers to businesses that offer out-of-the-box AI solutions which allows the clients to implement and scale AI techniques at a low cost.

IoT
IoT refers to a network of various objects such as people or devices that have unique IP addresses and an internet connection. These objects are designed in such a way to communicate with each other with the help of internet access.

Big Data
Big Data refers to humongous amounts of data that may be either structured or unstructured. These sets of data are too large to be quickly processed with the help of traditional techniques, and hence advanced techniques need to be employed for the same.