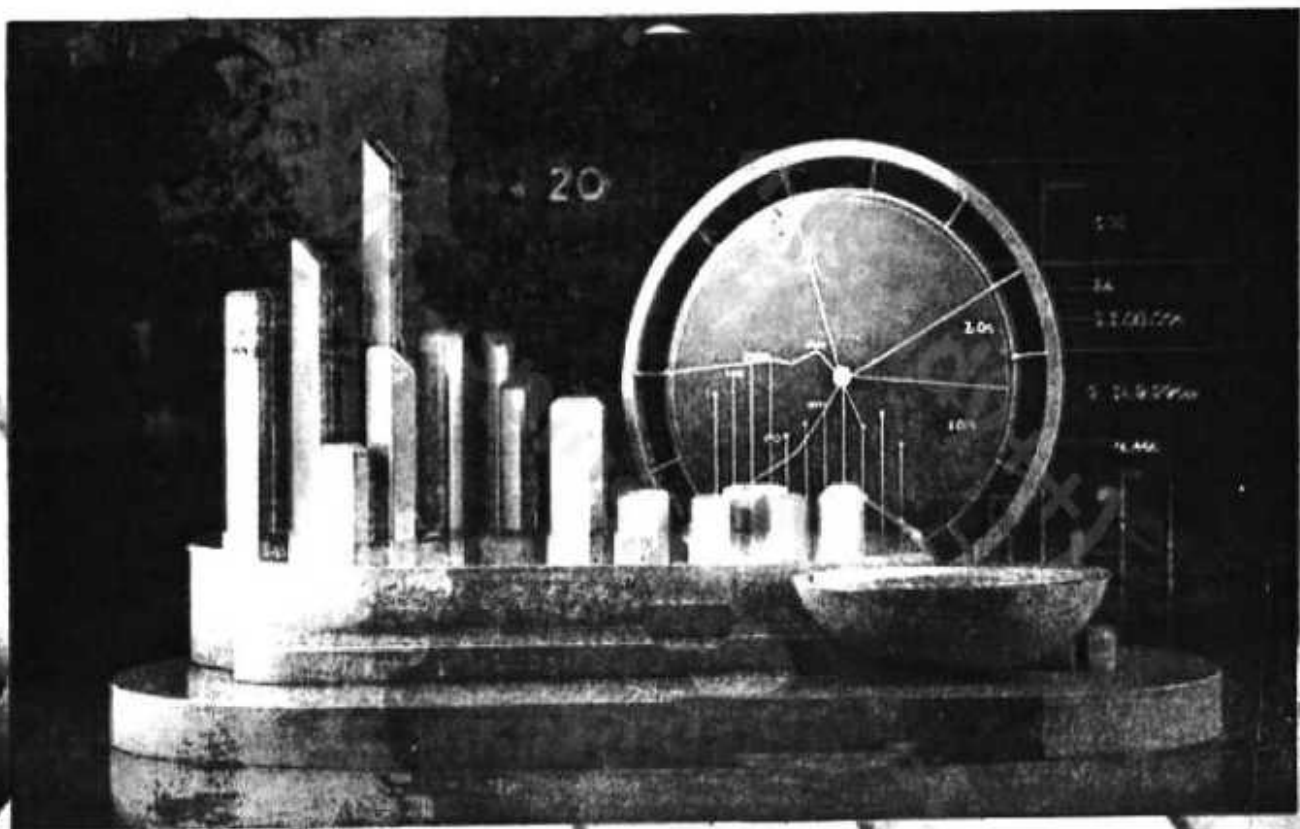




Learning Outcomes

At the end of this unit students will be able to:

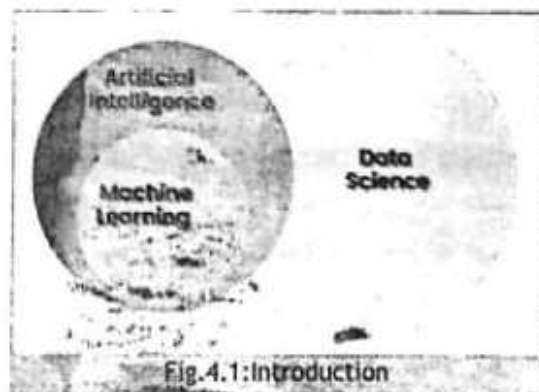
- understand and explain the scope of data science, Artificial Intelligence (AI), and Machine Learning (ML), including types of supervised and unsupervised learning models, and their applications to common real-world problems.
- understand and explain the types, uses, and methods of data visualizations and understand the benefits of visualizing data.
- to apply stages of the data science life cycle e.g. understanding a real-world business problem, data gathering, building model, interpreting results).



Introduction

Data science is a branch of knowledge, which joins together mathematics and statistics skills with the programming domain of computer science. This combination is useful in various fields of life e.g. if we want to predict the result of a football match, data science will help us. Currently, in the age of artificial intelligence, organization of available data into meaningful form is necessary, Data Science helps us to achieve the objective of getting insight from the data.

Consequently, we can find optimized solutions along with predictive modeling. In the previous grade we learned some basic concepts about data science. Now, we will learn some advanced topics based on our previous knowledge.



4.1 Data Science

Data Science refers to an interdisciplinary field of mathematics, statistics, and computer science. Main objective of data science techniques is to get meaningful insight from the available data. The available data can be structured as well as unstructured. This objective is achieved through the steps of data collection, data cleaning, data preprocessing and data visualization. Various programming languages like Python, R studio and MATLAB can be used for data visualization.



Fig.4.2: Data Science

4.1.1 Data Science and its scope

Data Science has wide scope in many disciplines of life and it is continuously expanding. We can briefly learn about its scope in some common fields of life as follows:

Healthcare: To predict the possibilities of some disease, based on X-Rays, Ultrasounds or other analytics of patient.

Sports: If we want to predict who is going to win a cricket match, based on the previous performance of teams, data science will help us.

Finance: In the banking and finance sector, data science is helpful in fraud detection. It is also helpful in developing trading and investment strategies.

Goods transportation: It helps in route optimization and minimize delivery times. Data driven approach helps to reduce cost. It also helps in real time tracking of transportation of goods.

Airline: It helps in efficient route planning to reduce fuel consumption. It helps to improve services based on customer experience and feedback.

Energy and utilities: It helps to optimize energy production and distribution. Realtime monitoring is useful for load distribution and facilitates immediate response for demand fluctuations.

Education: It is useful to enhance personalized learning based on the performance of learners. It also helps to predict students' performance based on previous data.

Communication Media: It is used for content creation and distribution. Various approaches like sentiment analysis, social media monitoring and recommender systems are a few examples of scope of data science in media and communication sector.

Entertainment: Data science plays important role when you get suggestions about your favorite songs while playing online computer games.

Business: Data science helps in intelligent decision making about the business to get optimal profits. Customer insight and sentiments are also analyzed to improve the business.

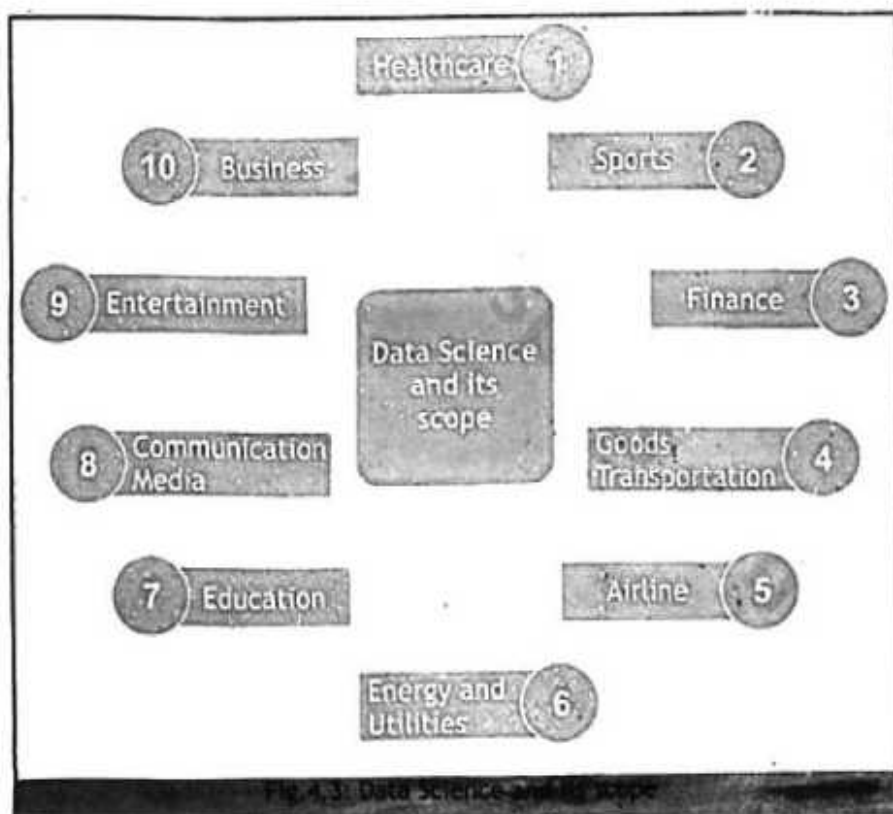


Fig. 4.3: Data Science and its scope



DO YOU KNOW?

A recommender system is a tool or technology. It is used to recommend products, services or information to the user based on their past behavior and interests.



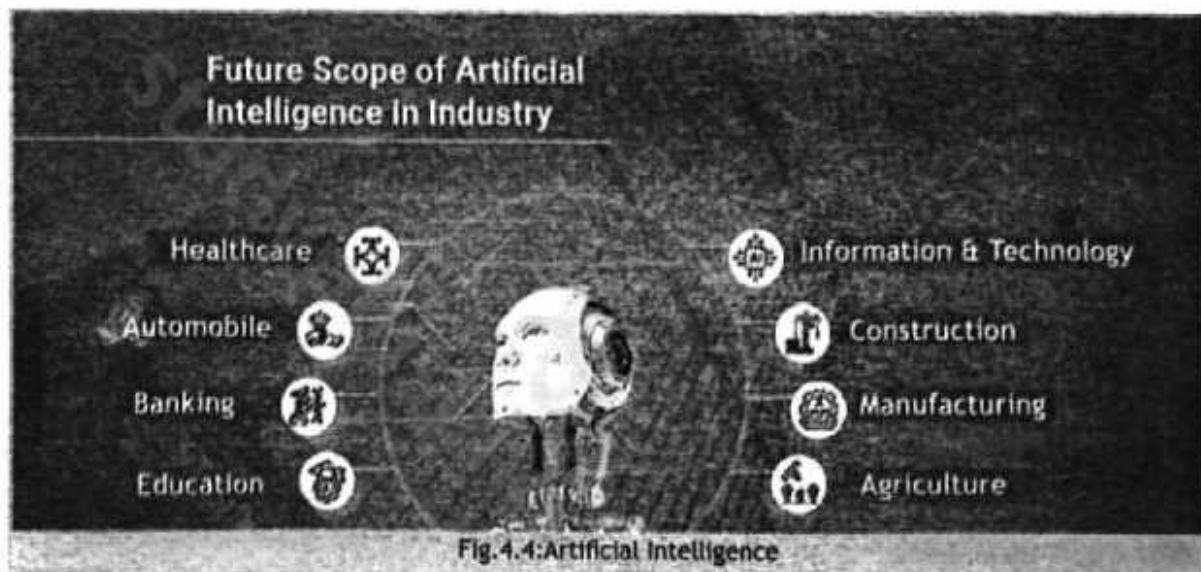
DO YOU KNOW?

Sentiment analysis is the term used to identify the sentiments of a customer by analyzing the review about the product. The sentiment can be positive, negative, or neutral. Sentiment analysis can be performed on reviews, text, opinions etc.

4.1.2 Artificial Intelligence and its scope

The term artificial intelligence is not new. In 1950 a British mathematician Alan Turing proposed a Turing Test, which measured the ability of machine to exhibit intelligent behavior. Now a days the modern robots are supposed to be smarter if they can pass the Turing Test. The term Artificial Intelligence is referred to the ability of a machine to exhibit human behavior like problem solving, understanding natural language and interacting with the environment intelligently.

Like data science, the scope of artificial intelligence is wide. Here, we can discuss a few of them as follows:



Decision making: Artificial intelligence is very useful in optimal decision making, based on a data driven approach.

Personalized recommendations: It helps to provide personalized feed to customers, based on their past usage of machines.

Automation industry: It has played a vital role in automating the tedious jobs like car manufacturing and image & video analysis. The most important utilization of artificial intelligence is integration of internet in the devices of daily use. The entire industry of IoT and smart devices is based on automation.

Natural Language Processing (NLP): AI enabled the machines to understand and respond to the natural human languages. ChatGPT and chatbots are examples of it. Now you can switch ON your air conditioner with voice command rather than picking up the remote controller.

Robotics: Whether you watch some latest robot, talking and cooking like a human being or you get a small robot to clean your house, all this technology is due to significant advancement of artificial intelligence.

Healthcare: Like data science, artificial intelligence has played equally well in the field of healthcare. The examples are personalized treatment recommendations based on medical imaging.

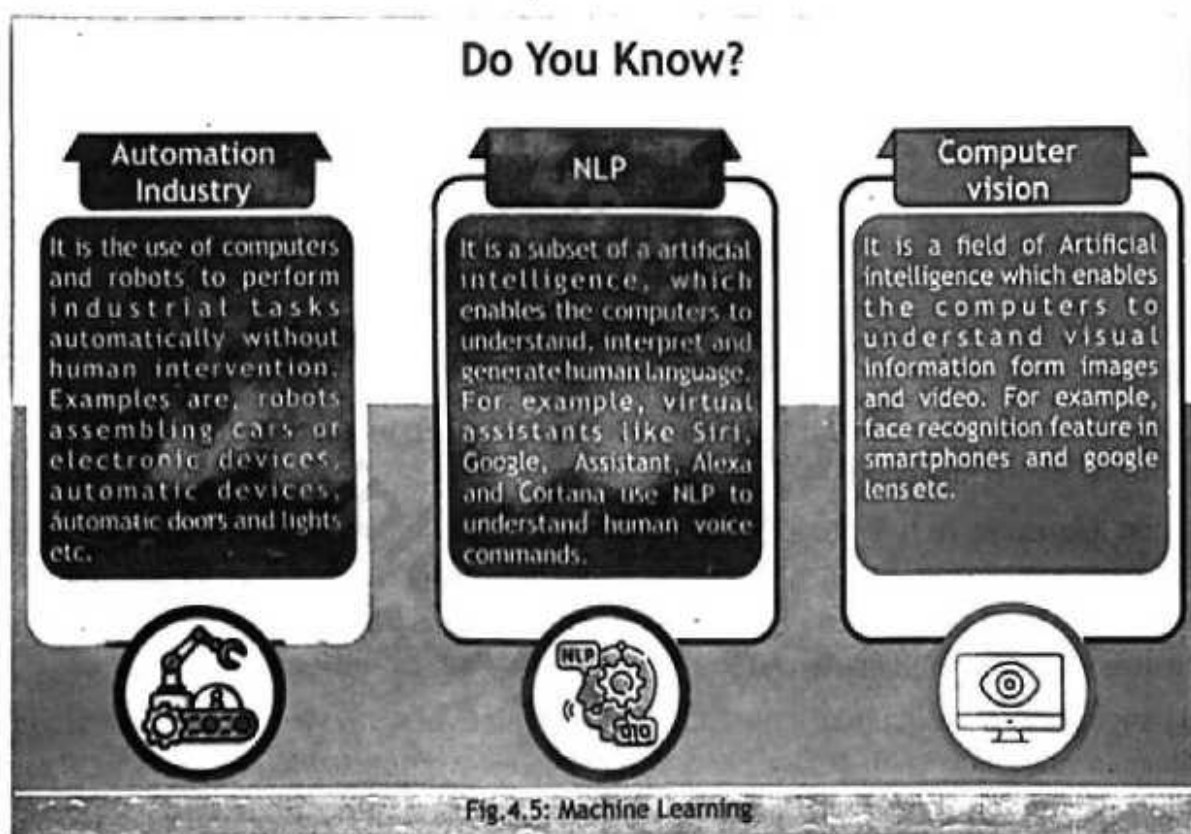
Computer vision: It is a specialized branch of AI that teaches computers to draw meaningful results from digital images, videos and other visuals.

Smart cities: Artificial Intelligence enables to develop efficient infrastructure and services, which can be used to manage energy consumption, traffic management and other necessary tasks for smart cities.

AI Agents: Artificial Intelligence provides bases to develop modern AI agents like Siri, Alexa, Google assistant, ChatGPT and Cortana.

4.1.3 Machine Learning and its scope

Machine learning is used in many fields of life, like healthcare, automation industry, to develop recommender systems, finance and banking, pattern recognition, NLP, computer vision, research and innovation. An example of machine learning is Automated fraud detection. It helps to identify fraudulent activity by finding anomalies such as sudden large transaction or unusual network traffic. In short, the entire scope of Data Science and Artificial Intelligence is based on the algorithms developed in the field of Machine learning.



4.1.4 Data Science, Artificial Intelligence and Machine Learning Skills
Artificial Intelligence and Machine Learning skills refer to develop systems that can learn and perform decision making. Data Science skill refers to extract insights from data and make informed decisions. There are many skills to achieve excellence in artificial intelligence. These skills include programming language, machine learning, and acquiring domain specific

knowledge. In programming languages, Python and R are the most used languages in AI. In machine learning algorithms, the knowledge of TensorFlow framework, deep learning, neural networks and NLP are very important.

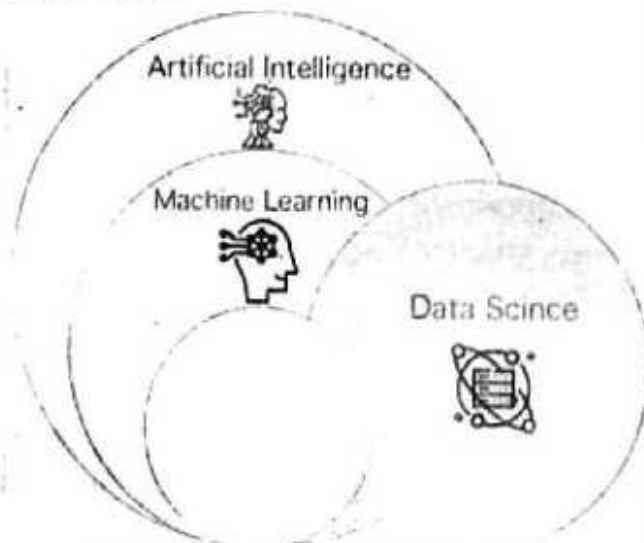


Fig.4.6: Data Science, Artificial Intelligence and Machine Learning

4.1.5 Machine Learning Models

Primarily, there are three models of machine learning.

- Supervised Machine Learning Model
- Unsupervised Machine Learning Model
- Reinforcement Machine Learning Model

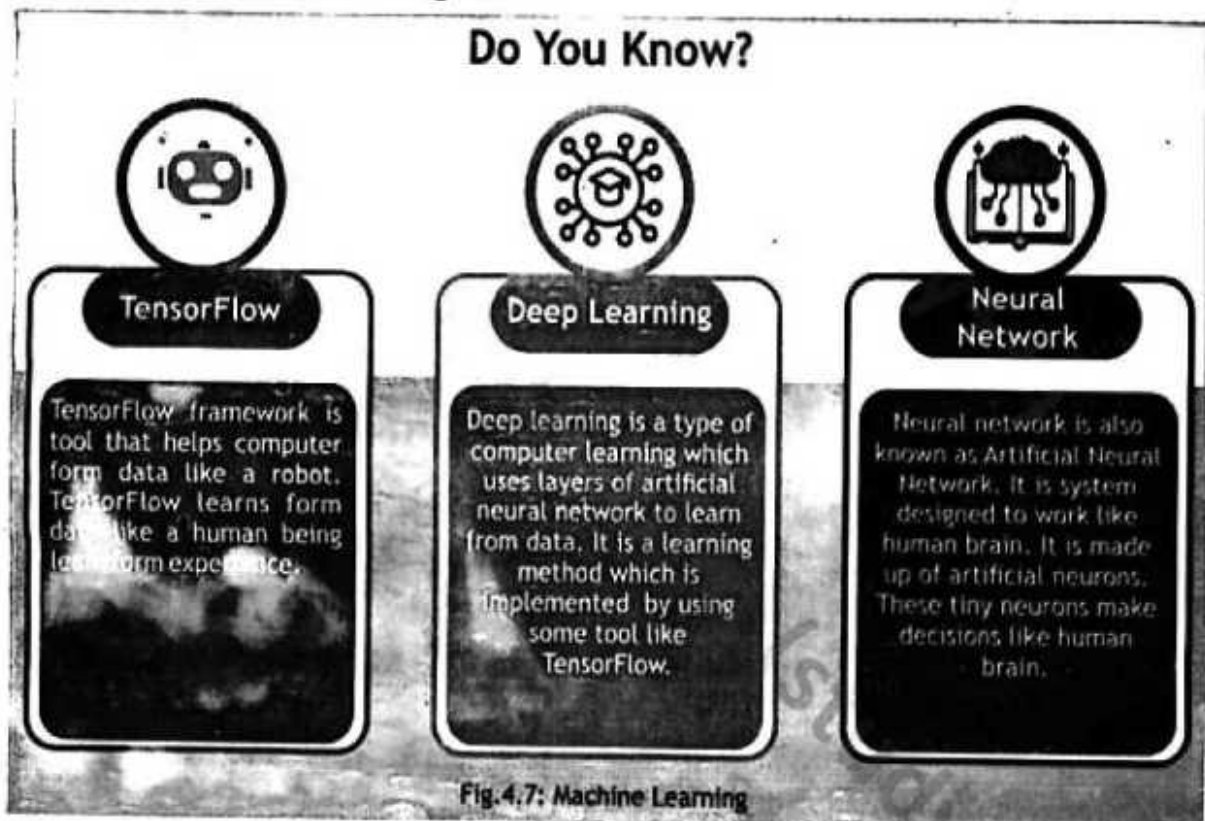
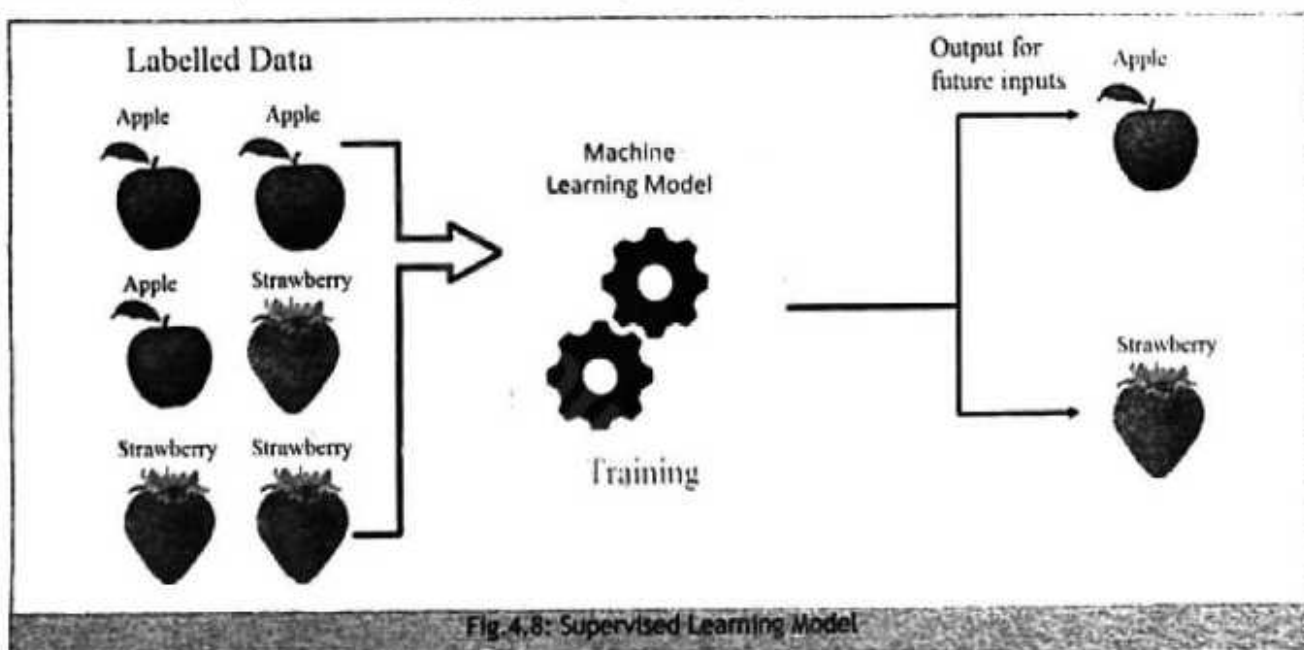


Fig.4.7: Machine Learning

4.1.6 Supervised Learning Model

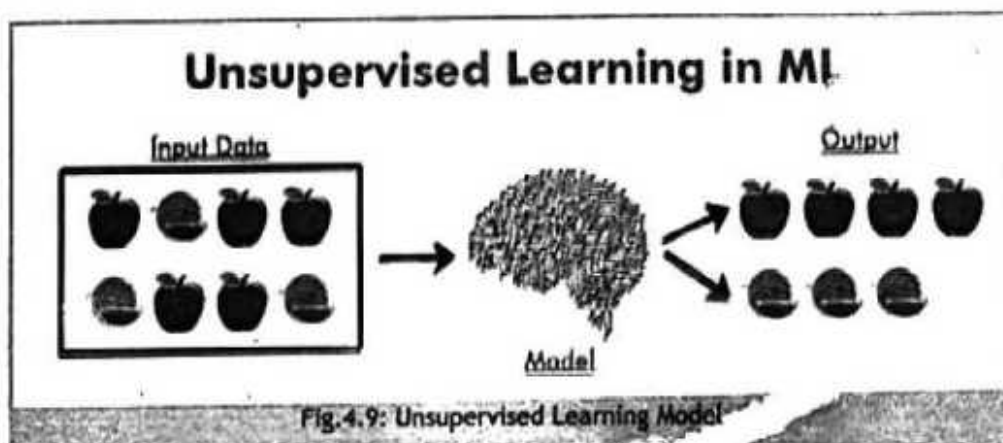
In this learning model a set of labeled data is provided to the machine/computer. The labeled data means that questions with their correct answers are provided to the machine. After reading/learning a lot of questions/answers the machine is trained. Now, only new questions without correct answers are provided to that machine and it provides us with the possible correct answers.

For example, if we provide the features(labels) of types of fruits i.e. shape, color and texture and give their names(answers) as well. After the machine is trained when we will give the features of a new fruit the supervised learning model will predict/provide the name of that fruit.



4.1.7 Unsupervised Learning Model

In unsupervised learning model, unlabeled data is provided to the machine/computer. The algorithm identifies the pattern in the data and divide it into groups based on similarities and differences. When the training is completed, the computer can identify new data and put it in the relevant category.



For example, if we provide various types of news to a computer then it will start dividing them into various groups, based on keywords. Suppose from a huge dataset the computer divides the news into various categories i.e. national politics, international politics, entertainment, medical, and technological news. When we give a new piece of information "Scientists successfully, saved the cancer patients at last stage" to the computer, it will categorize it into medical news.

4.1.8 Reinforcement Learning Model

In reinforcement learning the machines are trained with the help of positive or negative feedback. To explain reinforcement learning we will use the following key terms:

1. Agent (machine/computer)
2. Environment (where agent takes actions)
3. Actions (moves made by agent)
4. Rewards (points for helpful actions)
5. Penalties (points deducted for unhelpful actions)

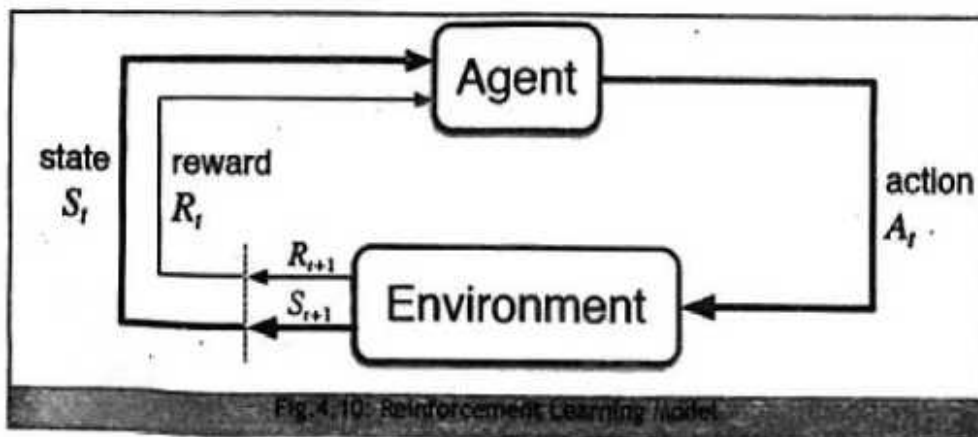


Fig. 4.10: Reinforcement Learning Model

The agent is provided with an environment where it must take certain actions, if the action is helpful towards the desired solution the agent will be rewarded with some points through feedback. If the actions are not helpful towards the desired solution the agent will be penalized (punished) by deducting the points. The agent improves its actions based on the feedback. This model is used to get optimal results.

For example, if you are learning to play chess and whenever you start your game by moving knights (name of a piece in chess game) you win, you will always try to start your game by moving knights. On the other hand, whenever you start your game by moving Queen piece, you lose the game. Therefore, you will always avoid moving Queen piece in the beginning of game in future. In this way, by winning and losing the game by moving certain pieces you have learned winning moves. In reinforcement learning this process is continuous until the model is trained and provides accurate results.

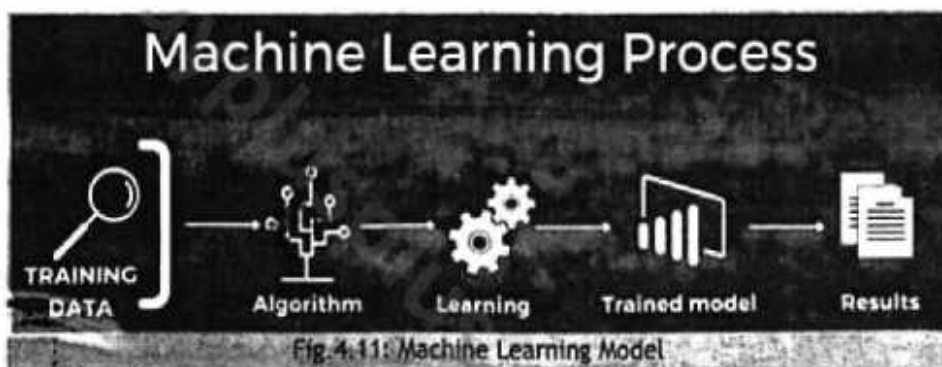
Fig. 4.10 shows reinforcement learning pictorially. The Agent (Chess Player) takes an action (moving Knight piece) at a certain time (in the beginning of the game). The Environment (Chess Board) is in certain state which changes after the Action A_t . Suppose the Agent moves Knight and

captures the Rook of the opponent player. The state of board changes to new state $S(t+1)$. The capturing opponent's rook is a rewarding move. In this way the Agent(Player) learns from his/her winning or losing moves and improves his/her performance by reinforcement learning.

4.1.9 Choose and appropriate machine learning model

As we learnt different models, how will we decide which machine learning model to use for a certain problem. First, we will define our problem, what output or solution we want? Do we want to identify some items, classify the given items, or forecast some future value. The next step is to analyze the given data.

If the given data has labels or questions/answers information, then we can use supervised learning model. If we don't have apparent labels and data have various groups which may have some similarities and some differences. We can apply unsupervised learning model. If in the given data, you identify some agent must achieve some goal then you will apply reinforcement learning model.



Machine Learning Process Steps:

Following is the brief description of Machine Learning Model:

Step 1: The preprocessed, cleaned data is split into Training data and Testing data. Sometimes we divide it into three parts and keep a small amount of data to validate the final results.

Step 2: A suitable algorithm is selected to train the machine. There are several algorithms like neural network, decision tree etc.

Step 3: The model is trained on the training set, it learns the pattern from data and develop a decision making ability.

Step 4: After the model is trained, test data and validation data is provided to the model and results are compared with actual values. The performance of the model can be improved by changing or improving the algorithm or by adjusting other necessary parameters which are beyond the scope of this book yet.

4.1.10 Artificial Intelligence, Machine Learning and Data Science(Similarities and Differences)

Artificial Intelligence is based on creating intelligent systems. Machine learning is a subset of AI that focuses on algorithms and statistical modeling to learn from data. Data science is used to get meaningful insight from the data to make informed decisions. There are many similarities and

differences among these three concepts. Some of these are as follows:

Similarities:

Data Driven Approach: All the three fields rely on data to provide optimal results. They use data to make predictions, recognize patterns and make decisions.

Use of Algorithm and Models: All the three fields work through various algorithms. For example, artificial intelligence relies on search algorithms. Machine learning relies on supervised/unsupervised/reinforcement learning algorithms. Data science relies on statistical and visualization algorithms.

Goal of Automation: The net goal of all the three fields is to automate various activities which were previously performed manually.

Differences:

Tools and techniques: They use various techniques, e.g. artificial intelligence uses neural networks, expert systems, and natural language processing. Machine learning uses supervised learning, unsupervised learning, reinforcement learning and deep learning techniques. Data science uses statistical analysis, data mining and data visualization techniques.

Application: The application of artificial intelligence is in speech and image recognition, autonomous vehicles, chatbots and virtual assistants. Application of machine learning is in recommendation system, fraud detection, predictive analysis, spam filtering, and personalized marketing. Applications of data science includes business intelligence, market analysis, healthcare analytics, risk management and scientific research.

Output: The output of artificial intelligence systems is that mimic human behavior. The output of machine learning is prediction and classification. The output of data science is insight, reports and visualization.

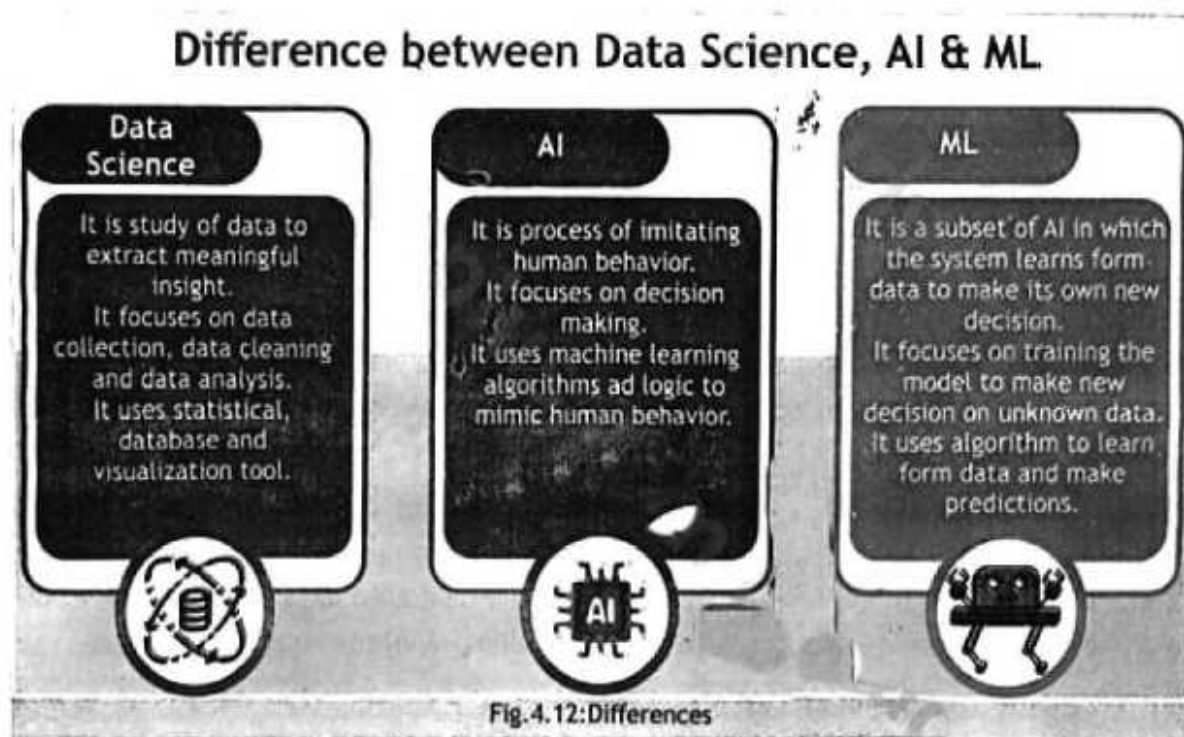


Fig. 4.12: Differences

4.1.11 Churn Prediction and Behavioral Segmentations

Churn Prediction is the technique to find out the possibility of a customer, to stop taking the services or products of a certain company. It is used in the fields of business and marketing. For example, a telecommunication company receives frequent complaints about their service from certain customers, then they can predict that these customers might stop using their service. These unhappy customers can switch to some other telecommunication company. Churn prediction is needed to run the business successfully. It is often observed that getting new customers is less beneficial than keeping the existing customers loyal to your business.

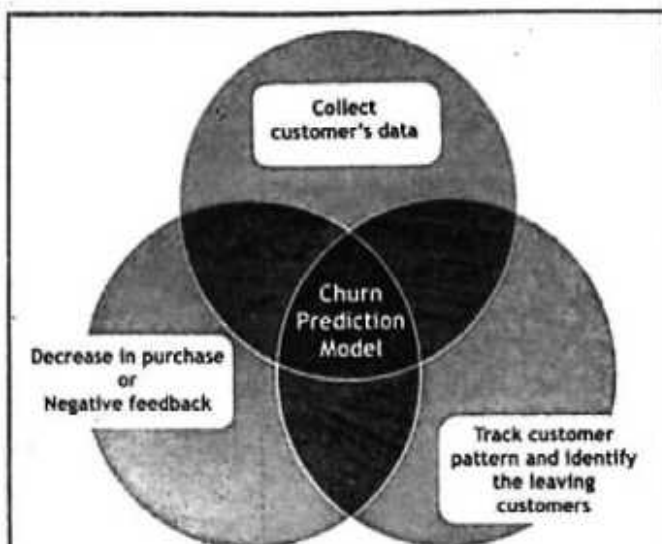


Fig. 4.13: Churn Prediction and Behavioral Segmentations

For example, a store is selling formal clothing. The store wants to predict which customers might stop buying in the future. The store will collect data about the customers, how often the customers buy cloth (purchase frequency), since how long are they buying (brand loyalty) any feedback they provide about their products etc. After applying various algorithms and visualizing the results the store might observe that some customers are not visiting the store as often as they were visiting a few months ago. Then the store might send some personalized offers to the customers which can reduce the possibility of losing the customers.

Behavioral segmentation is the process of dividing customers into groups based on their behavior. It is marketing strategy used to categorize customers according to their behavior towards some service or product. This approach enables the business to design marketing strategies and offerings more precisely. It enables the business to meet the specific needs of customers segments. The key aspects of behavioral segmentation are analyzing purchase frequency, brand loyalty, usage occasions and response to market stimuli.

4.2 Data Visualization

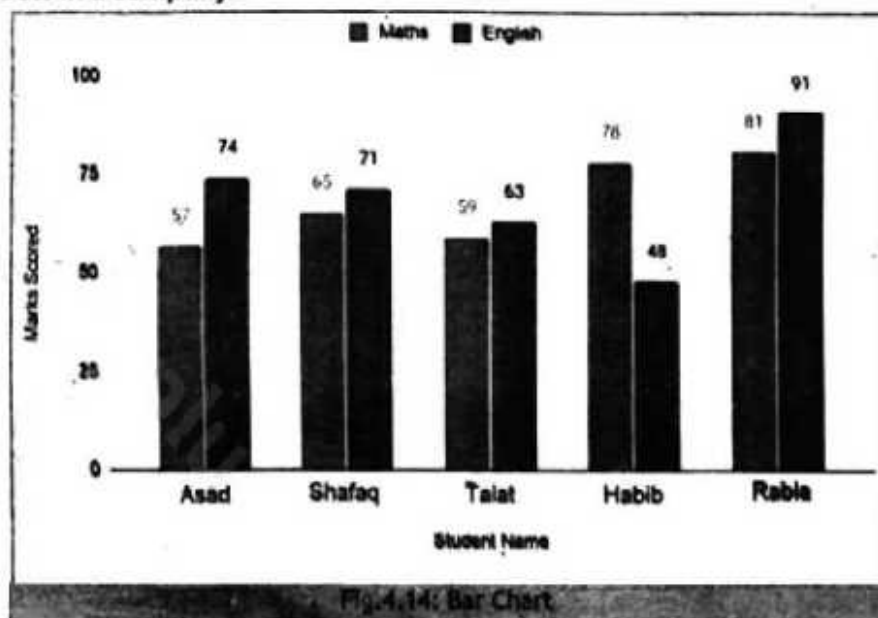
It is the process of creating graphical representation of data and information. It involves various visual elements like charts, graphs, and maps. This graphical representation is useful to make decisions in the field of science, arts and in daily life. It is a powerful tool that transforms raw data into meaningful and insightful information.

Data visualization through charts, graphs and plots, makes complex information simple to understand and helps in making meaningful decisions. Some interactive data visualization techniques are helpful to get optimal results from the available data. Data visualization have variety of techniques and types that helps to get meaningful information from data and use it for the improvement of business and services.

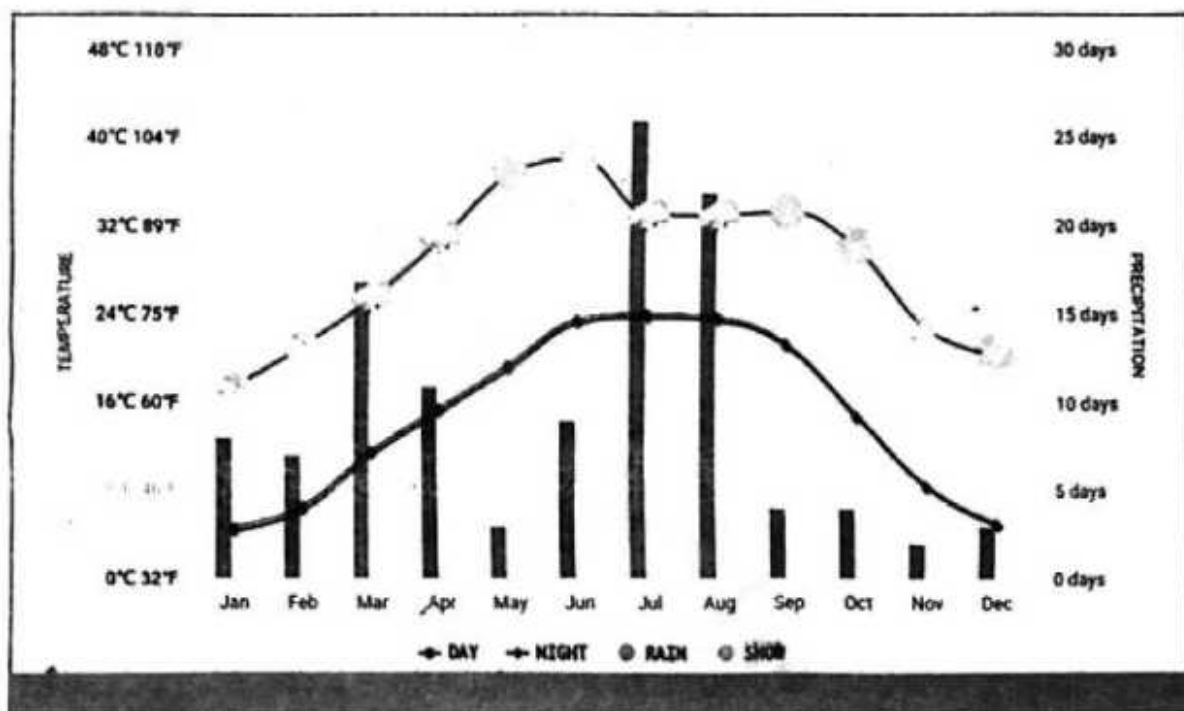
4.2.1 Methods of Data Visualization

Methods and types of data visualization are interrelated terms. However, methods of data visualization are the techniques to prepare data for visualization. These methods are explained briefly as follows:

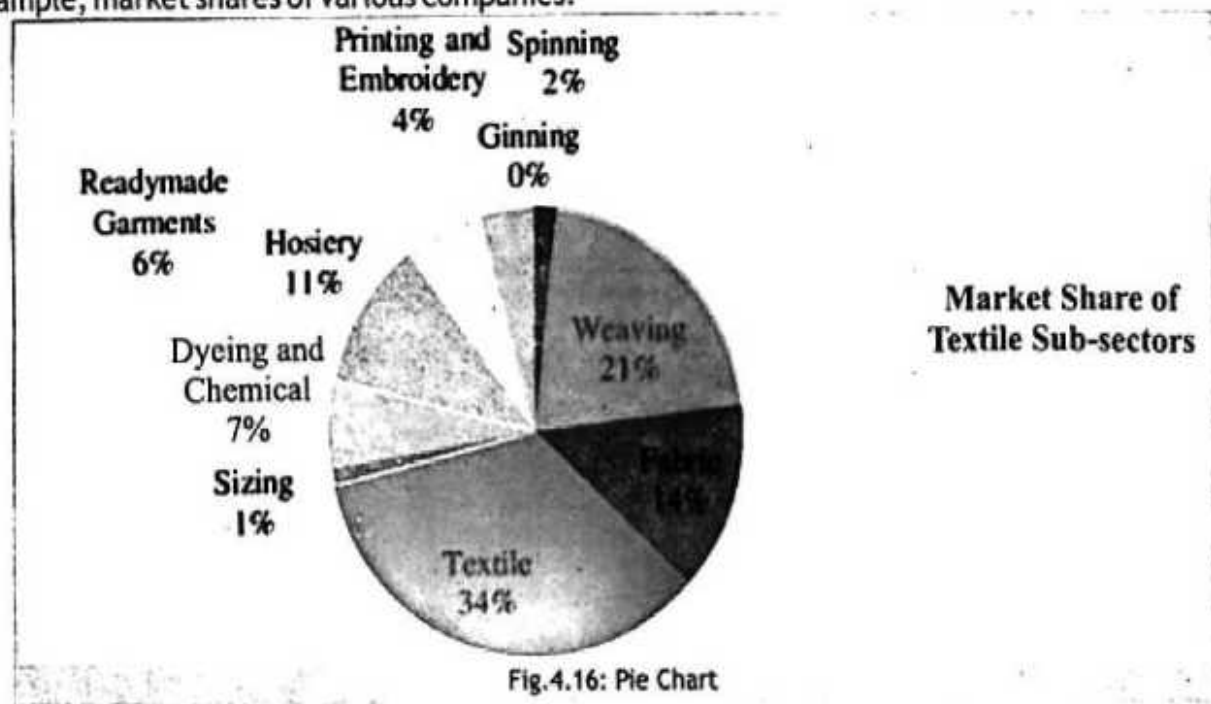
Bar Chart: This method of data visualization is used to show comparison among discrete values. For example, we can use Bar Chart to visualize marks of students in various subjects or sales data of various products in a company.



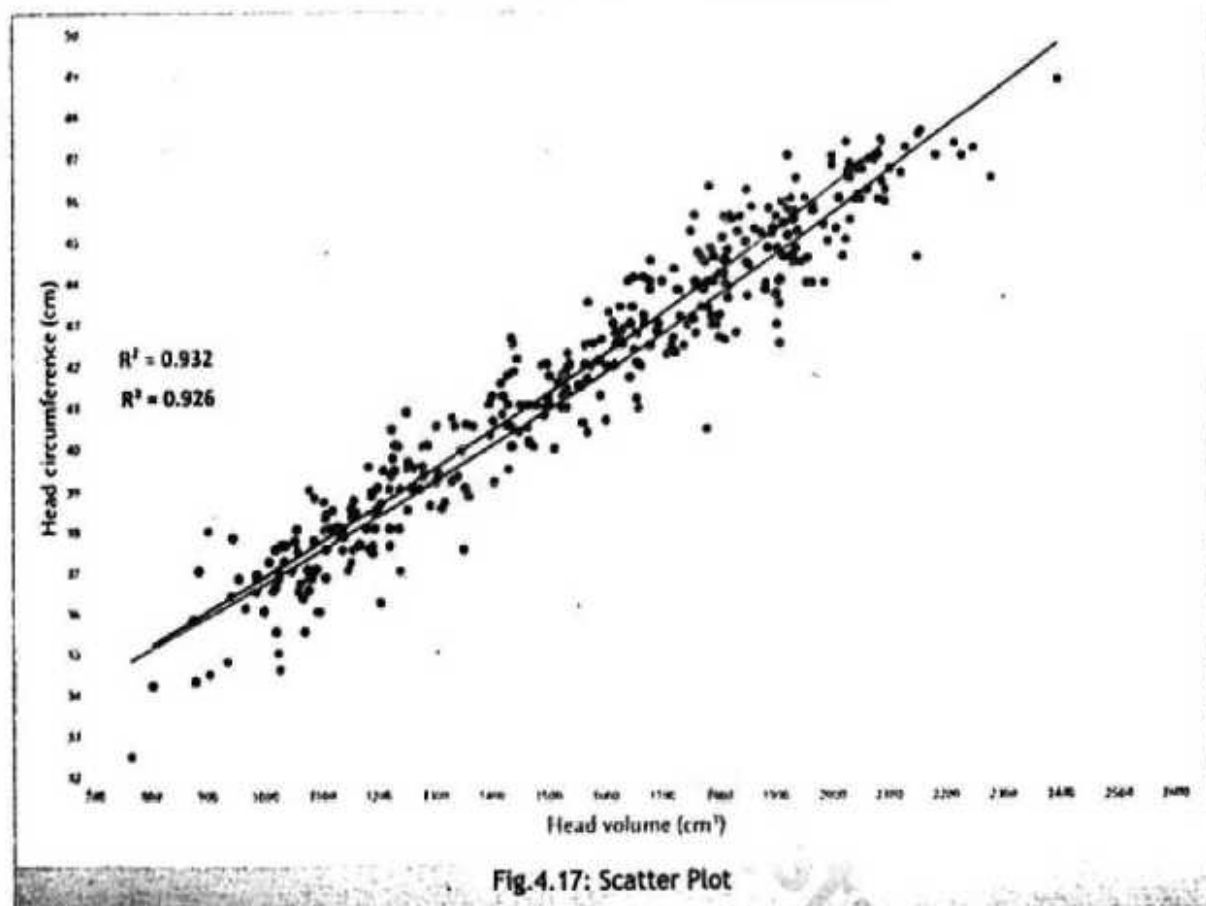
Line Chart: It is used to display trends of data over short or long period of time. For example, gold prices over the last five or ten years can be displayed by using a Line Chart. Similarly, stock prices, or change in temperature over a year can be presented by using a Line Chart.



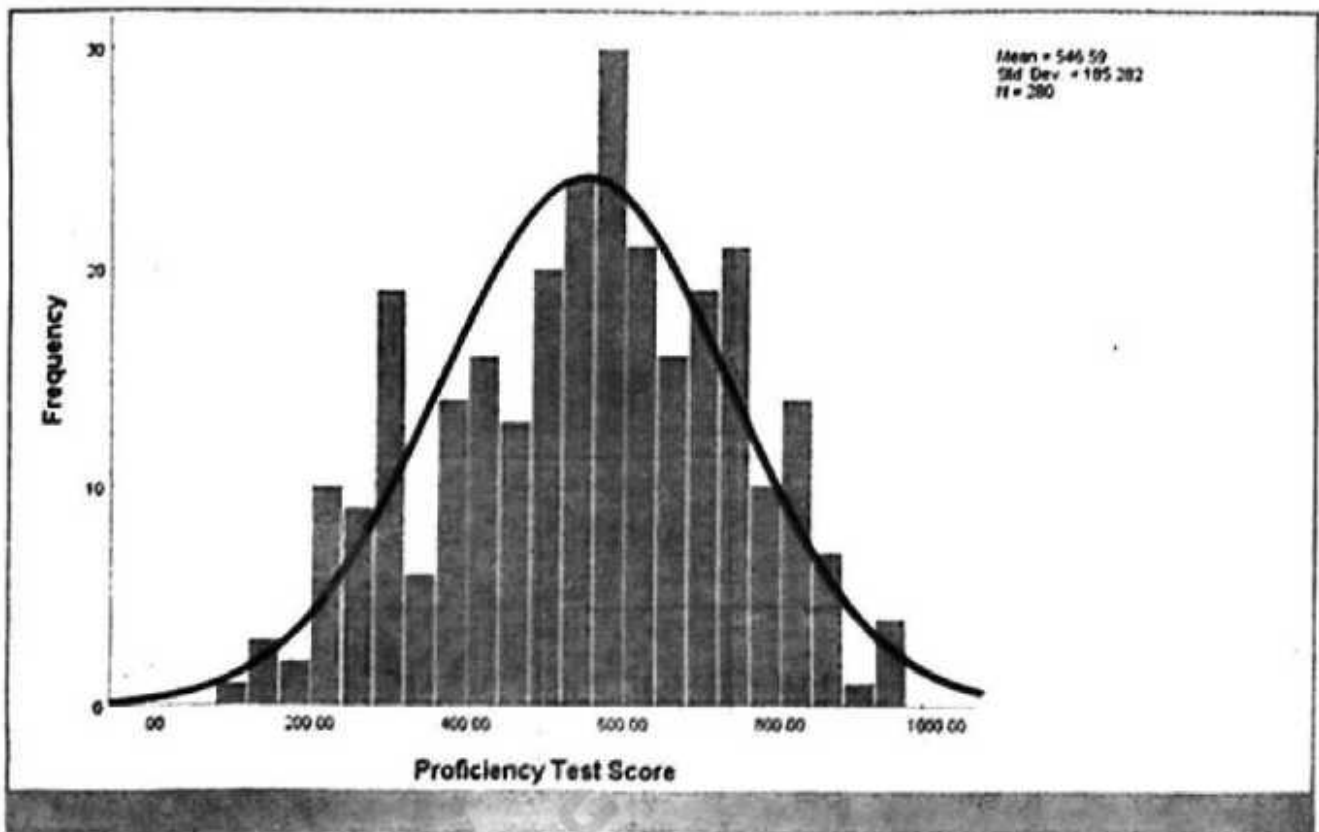
Pie Chart: It is used to represent proportions of various categories of data in a whole. For example, market shares of various companies.



Scatter Plot: It is used to display the relationship between numerical variables. For example, correlation between height and weight of a group of people or correlation between head volume and head circumference of male and female students of certain grades.



Histogram: It is used to show the distribution of a single numerical value. For example, distribution of score in a class.



Box Plot (Box-and-Whisker Plot): It is used to summarize the distribution of data with the help of minimum, first quartile, median, third quartile and maximum data. For example, analysis of salaries of various employees in a company.

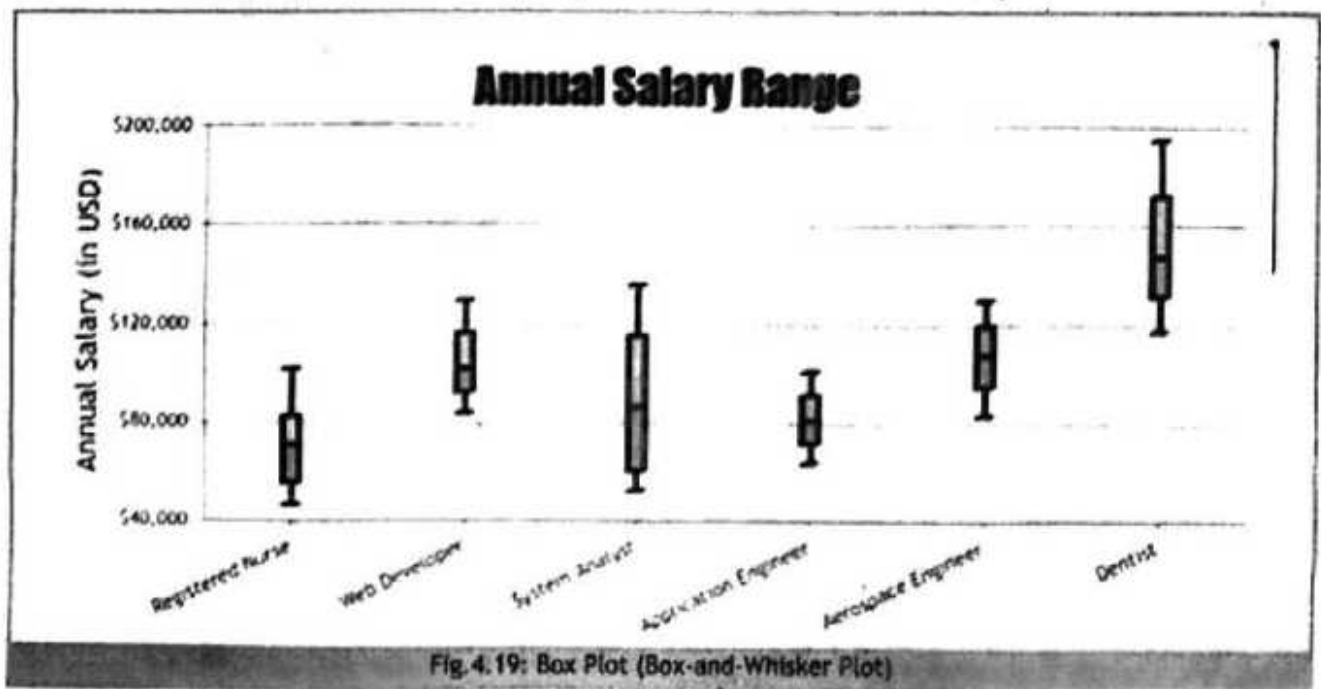
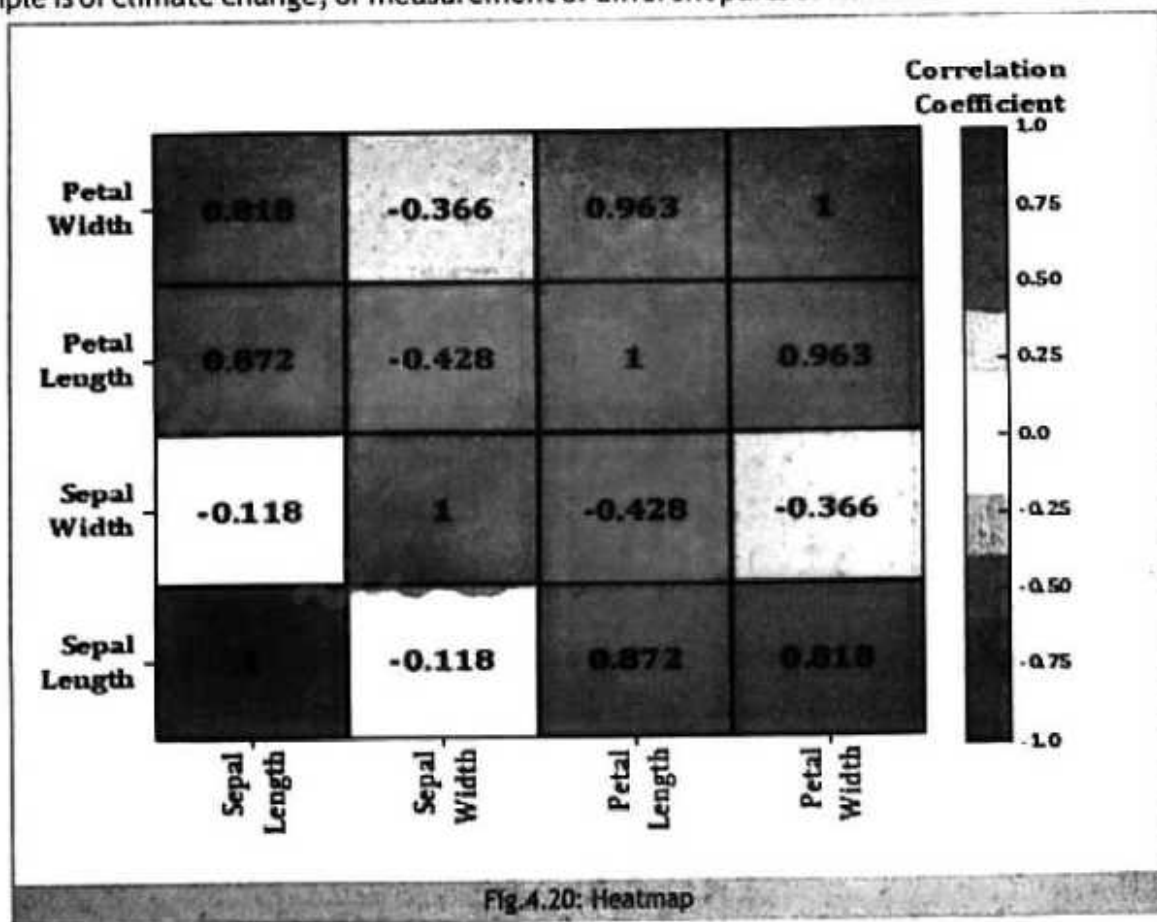
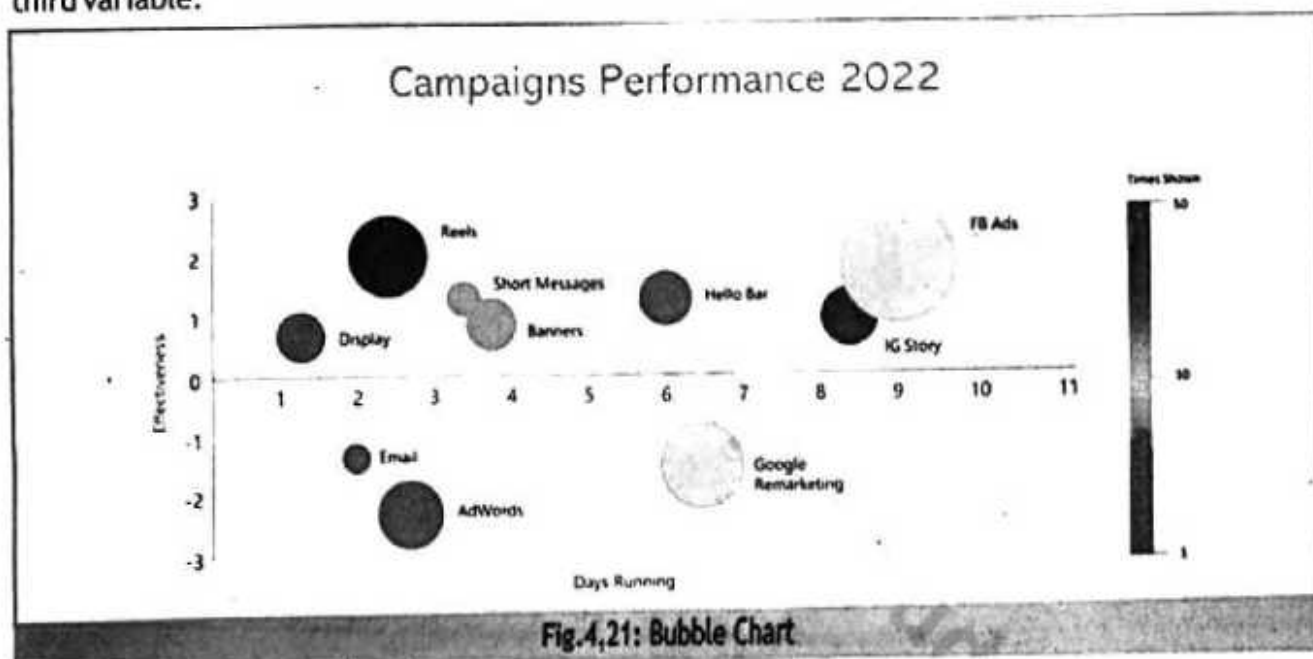


Fig. 4.19: Box Plot (Box-and-Whisker Plot)

Heatmap: It is used to represent data in matrix form. Data values are shown by using color intensity. For example, images received from the space to analyze weather forecast. Another example is of climate change, or measurement of different parts of flowers.



Bubble Chart: It is similar to scatter plot, but it has a third variable which is represented by the size of bubble. For example, points in a 2D space where area of each point is proportional to the third variable.



4.2.2 Types of Data Visualization

There are different ways to represent data graphically to make it easier to understand. Some common types of data visualization are as follows:

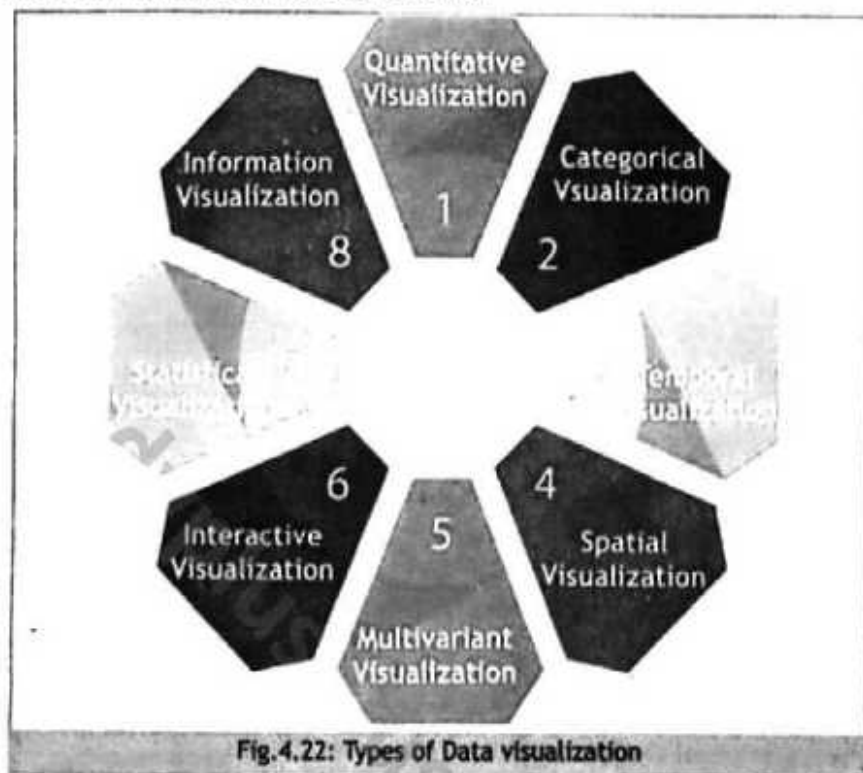


Fig.4.22: Types of Data visualization

Quantitative Visualization

Quantitative visualization is used to represent numerical data. It focuses on quantities or numbers to show measurable data. It is used to display data that can be measured or counted. For example, a bar chart showing sales figures of a company over several months effectively communicates quantitative information.

Categorical Visualization

Categorical visualization is used to represent data that falls into distinct categories. It helps to show proportions or parts of a whole. This type is ideal for displaying nominal or ordinal data. For example, a pie chart showing the market share of different companies in a specific industry.

Temporal Visualization

Temporal visualization is used to display data that changes over time. It is used for time-series data. Line graphs are usually used to represent temporal data. For example, a line graph.

Spatial Visualization

Spatial visualization is used to represent data related to physical locations or spaces, visualizing geographic or spatial data. For example, a heatmap showing population density across different regions.

Multivariate Visualization

Multivariate visualization is used to represent data involving more than two variables or dimensions. Scatter plots and heatmaps are effective tools to represent such data. For example, a scatter plot matrix showing relationships between income, age, and spending habits.

Interactive Visualization

Interactive visualization allows users to interact with the data, through digital platforms. These digital platforms can be dashboards and filters. A dashboard helps the user to filter and manipulate data visualizations and explore different trends and insights.

Statistical Visualization

Statistical visualization is used to present data to show statistical properties, such as distribution or correlation. Histograms, box plots, and scatter plots are popular choices.

Information Visualization

Information visualization is used to present complex data sets in an easier way, often for abstract or conceptual data. Network diagrams, tree maps, and word clouds are effective tools. For example, a network diagram showing relationships between entities, such as a social network.

4.2.3 Uses of Data Visualization

Like Artificial Intelligence, data science and machine learning, data visualization is useful in almost all the fields of life. Some of them are as follows:

Business Intelligence: Data visualization helps to make data driven, well informed decisions. It is used to find market trends and helps to track and improve performance.



Healthcare: It helps to visualize the impact of various diseases affecting the patient. It is helpful to track disease and visualize the spread of disease.

Education: Data visualization is very helpful to teach data literacy skills, concept building, creative thinking, and critical thinking.



Science and Research: It is useful to visualize complex findings, very huge and complex data such as complex scientific data received from satellite in the form of photographs.

Sports and gaming: It is useful to visualize performance of players whether they are playing football in a ground or chess players playing in online tournaments. It is also helpful in sports broadcasting, and other predictions.





Finance: It is helpful to analyze market trends, to track portfolio performance and to identify investment opportunities.

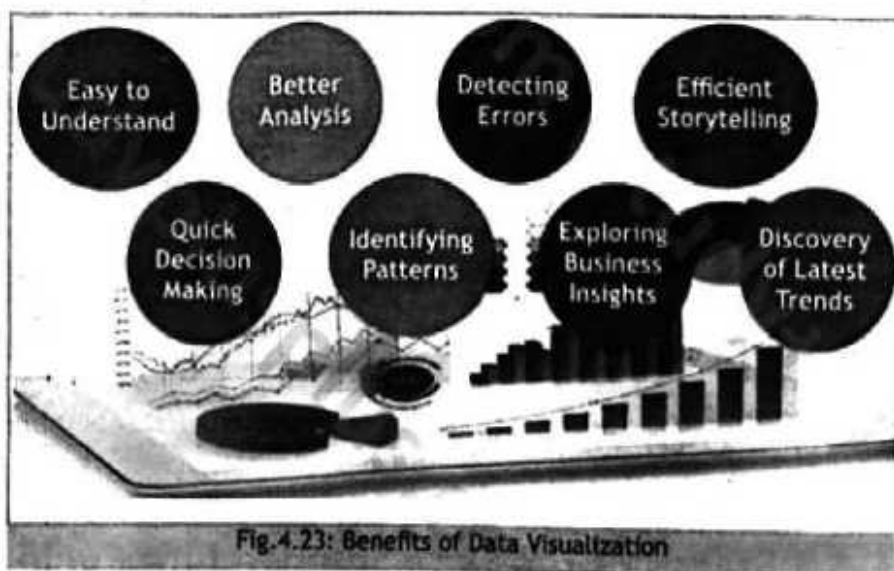


Entertainment: It helps the entertainment industry, to visualize movie performance data to predict future trends. It helps in content optimization by visualizing audiences insights and trends.

4.2.4 Advantages/Benefits of Data Visualization

Data visualization has many advantages and benefits, some of them are as follows:

- It helps non-technical stakeholders to understand complex data and make decisions regarding their business and services.
- It helps the decision-makers understand the existing data and make well informed strategic decisions.
- It improves the performance of related business or services.
- It helps in better communication, understanding, and getting better insight from well-organized real-time data.
- Data visualization saves time and identify trends quickly.
- It saves time and enable us to make quick and accurate decisions as compared to manual analysis.



4.2.3 Database and Machine Learning

Database is collection of related data stored in the form of tables. Various computer software is used to create databases to store data in organized form so that it can be managed, retrieved and stored efficiently. Traditionally databases have a key role in the development of modern technology of computer science. Oracle, Ms-Access and MySQL are some well-known database

management software used to create and manage databases.

Apparently, database and machine learning seem to be very distinct disciplines of computer science domain. However, they are interrelated closely. To understand relationship between these two disciplines, the simplified information is given as follows:

Characteristics	Database	Machine Learning
Data storage	In database data is stored in structured format such as table.	Algorithms of machine learning require the same structured data to train their models. Datasets are organized in rows and columns.
Data preprocessing	Database stores data after normalization, therefore data is clean without any duplication, data redundancy and other errors.	Machine learning also require clean data. If the data is not clean the performance/result of the algorithm is affected.
Data retrieval	As data is stored in the tables in an organized manner, the retrieval of data through SQL queries is very easy, fast and reliable.	Data is organized in the form of Dataset; therefore, it can be retrieved by using simple and easy algorithm of machine learning.
Data Integration	Data in a database is kept integrated. It is stored after removing duplicated values and other possible errors. Therefore, integrity of data in the database is assured.	For comprehensive model training integrated data is required. In this way Data provided by the database is equally useful for machine learning to train its algorithms.
Scalability	DBMS can handle various sizes of databases ranges from student database of 50 students to the database of entire country or even more.	Machine learning requires scalable data storage solutions. In this way it can implement big data applications. Number of records in a dataset varies from hundreds to thousands and even in millions in a single dataset.
Realtime Data Processing	Database provides Realtime data storage and retrieval support.	Machine learning model requires Realtime data to train model for online data.
Analytical Tools	Database provides various tools for data analysis, e.g. SQL server and other software etc.	Database analytical tools are helpful for machine learning to provide fast and accurate solutions.

4.3 Stages of the Data Science Life Cycle

Data science life cycle is an iterative process. It has various stages, each has certain steps to be taken to move forward. It requires a structured approach to solve data driven problems. The stages of data science life cycles are as follows:

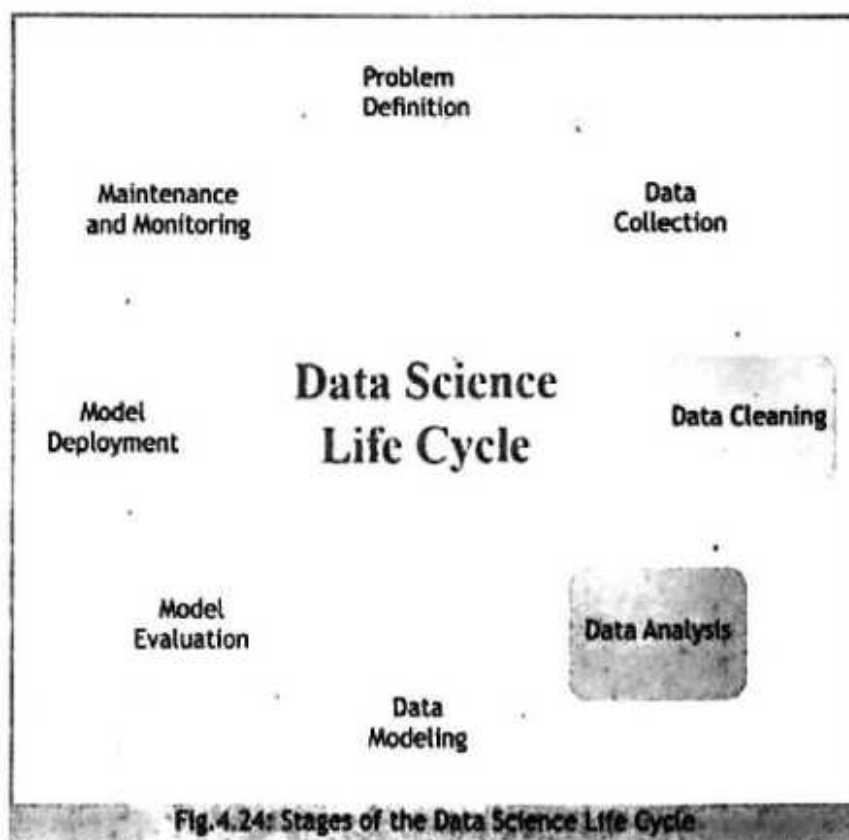


Fig. 4.24: Stages of the Data Science Life Cycle

Problem Definition:

At this stage the business problem or research question is defined with clear objectives and outcomes. The scope and limitations of the problem are determined. Preliminary research and literature reviews help the stakeholders to understand their needs and expectations.

Key performance indicators (KPI) are defined to measure the success of proposed solutions. KPIs are measurable targets that help us track progress and success of a process.

For example, increasing sales by 15% in two months can be set as a KPI of some business problem. Another example of KPI is improve the speed of website loading by 20% in 30 days. By setting KPIs, we can focus on finding the right solutions and track our progress. The outcome of this stage is a hypothesis or well-defined problem.

Data Collection:

At this stage, data is gathered by using various reliable and useful sources. The method of data collection depends on the kind and complexity of problems. For simple problems, we can observe and write down what we see, while more complex problems we might need surveys, experiments, or even using online services through APIs (Application Programming Interface). APIs are small programs that act like messengers between two or more software.

These are like waiters in a restaurant who receive an order from us and bring back the food as response to our order. The data collected should be relevant and accurate, as the quality of the data directly affects the results at later stages

Data Cleaning:

It is also called data preprocessing, this step is about fixing the data to make sure it's ready for analysis. In this step we remove errors, such as wrong information or duplicate values. If some values are missing, which might cause problems, we fill in with estimated values.

During this process, data is organized into an easily workable format. Sometimes, new features or attributes are created from existing data. For example, if we have dates, we might create a new column for the day of the week. Clean and well-organized data is essential for making good predictions later on.

Data Analysis:

After cleaning the data, we have to analyze it. Statistics or machine learning techniques are used for data analysis. These techniques help in finding patterns, trends, or important details. By looking at the data closely, we can test early ideas or assumptions about the data.

For example, we might want to know if a certain factor is causing an outcome, like is the weather affecting sales of certain products? At this stage, visualization tools like graphs or charts can be used to understand what's happening in the data.

Data Modeling:

In this stage, the data is organized into a structured model. It helps to create diagrams that show the relationships between different entities (like customers and the products they purchase) and their attributes (like names, prices, or dates).

We also focus on how to store and retrieve this data efficiently by setting up databases or other systems. Once the model is ready, it is tested using historical/previous data to see if it performs well. The goal is to create a structure that can be easily used for predictions or other analysis.

Model Evaluation:

At this stage we evaluate the performance of our machine learning model. This is done by measuring how accurate and reliable the model is, and whether it meets the performance goals we set earlier.

For example, we can test how often the model makes the right predictions or how fast it delivers results. If the model doesn't perform as expected, we might need to go back to the data modeling stage to make improvements. This stage also includes making sure that the data and the model are secure and respect privacy rules, especially if sensitive information is involved.

Model Deployment:

Once the model is working well, it is time to deploy it for use. This stage is called deployment, and it means the model is set up in the system or infrastructure where it will be used. It means to integrate the model on a website, application, or database.

Sometimes, deployment is done in small steps (partial deployment) to test if everything works fine before launching it completely. Other times, the model is deployed all at once. After deployment, the model starts working with real data, providing insights or making predictions.

Maintenance and Monitoring:

After the model is deployed, it is important to keep an eye on how it performs. This is a continuous process where the model's accuracy, speed, and reliability are monitored over time. If any issues are detected, like the model becoming less accurate or slow, adjustments can be made.

Feedback from users can also help improve the model. This ongoing maintenance ensures that the model remains useful and effective as conditions change, like if new data becomes available or the problem evolves.

4.3.1 Case Study: Understanding a real-world business problem and its formulation to data science problem

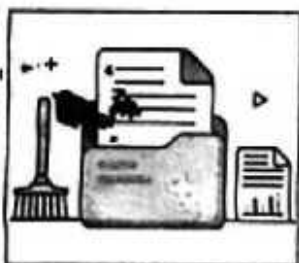
We will learn to apply the data science life cycle in real-world business problem. Suppose we have a task to optimize the inventory of a smart phones.

Problem definition: Suppose you own a super store and you want to ensure that you want to have good enough stock of popular smart phones. The objective is that you don't run out of stock to meet the demand of your customers. Keeping in mind that you also don't want to make unnecessary purchases that would be difficult to sell.

Data Collection: The first step is to gather data related to the smartphone stock and sales. The data includes:

- Current stock levels for each smartphone model (how many phones of each type are in stock).
- Sales data showing which models are selling fast and which are slow-moving.
- Supplier information about new models and upcoming promotions or price changes.

This data can be collected from several sources, like the store's sales system, purchase invoices, supplier's record, and even from customers' feedback on popular models. Gathering this data helps the superstore understand the current state of their inventory and the demand trends.



Data Cleaning: After collecting the data, the next step is to clean it to ensure accuracy. This involves:

Removing duplicates: such as cases where the same model is listed multiple times with different spellings or slight variations.

Fixing errors: in the stock levels and sales data, like incorrect quantities.

Handling missing values: such as missing prices or sales figures, by cross-referencing supplier information or older data.

By cleaning the data, the you ensure that the inventory and sales data is reliable and ready for analysis. For example, if there are missing prices for certain models, they can be updated with information from suppliers.



DO YOU KNOW?

Investigator: An investigator is a person who conducts the statistical enquiry.

Enumerators: To collect information for analysis, an investigator needs the help of some people. These people are known as enumerators.

Data Analysis: Now that the data is clean, the next step is to analyze it. The goal of this analysis is to understand which smartphone models are in high demand and which are not selling well. Key questions to answer include:



- Which models sell the fastest?
- Which models have stayed in stock for too long?
- Are there seasonal patterns in sales?

After answering these questions, you can identify the smartphones that need to be reordered more frequently and the ones that are less popular. Visualization tools like graphs or charts can help show sales trends over time, making it easier to see which models are consistently popular.

Data Modeling: At this stage, you will build a model to keep track of inventory levels and predict future demand. The model includes:

- Attributes for each smartphone like model, price, stock level, and sales history.
- Relationships between stock levels and sales rates. For example, when a phone sells quickly, the stock level drops, and the system signals the need for a reorder.

The model can use historical sales data to predict future demand, it helps to avoid running out of popular phones. In this way you will prevent overstocking on less popular models. The model is tested using past data to ensure it works accurately before it is put into use.

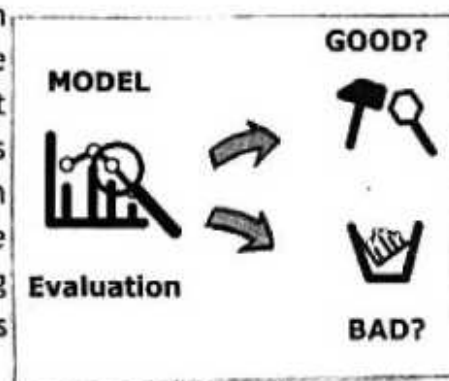
Model Evaluation:

After completing the model you will evaluate it to see how well it predicts inventory needs. Key performance measures include:

- **Accuracy:** How well does the model predict when a phone will go out of stock?
- **Efficiency:** Does the system notify the staff, in time to reorder popular models before they run out?
- **Cost-effectiveness:** Is the system helping the store avoid over-purchasing models that don't sell well?

If the model doesn't perform as expected, adjustments can be made. For example, if the model predicts too many unnecessary purchases, the store might fine-tune it to be more careful in its predictions. Security of sales data is also evaluated to ensure it is handled responsibly.

Model Deployment: Once the model is ready, it is deployed in the store's inventory management system. This could involve setting up alerts when stock levels of popular smartphones get low. Integrating the model into the store's existing sales software. The deployment may be done in phases, starting with a few popular models, or it may be used across the entire inventory at once. After deployment, the model starts working in real-time, sales tracking and adjusting stock levels accordingly.



Model Evaluation: Monitor the performance of the model by comparing the actual sales records with the predicted sale data. Adjust the model as needed to improve the accuracy of this model.

Maintenance and Monitoring:

After the model is deployed, you will continue to monitor its performance. This ongoing stage includes:

- Checking if the machine learning model correctly alerts staff when stock is low for popular models.
- Ensuring the store is not over-ordering less popular or even popular phones.
- Collecting feedback from store staff to see if the system is easy to use and helping them manage the inventory better.

Regular updates are made to improve the model, such as incorporating new data on sales trends or adjusting reorder levels based on changes in customer preferences. The system is continuously refined to ensure that the superstore never runs out of stock for popular models while avoiding excess inventory.

By following the above data science life cycle, the superstore successfully developed a system to manage its smartphone inventory more efficiently. The system helps the store keep popular smartphones in stock while preventing over-purchasing of less popular models, ensuring customer demand is met without wasting resources.

- **Data science:** It is a branch of knowledge, which joins together mathematics and statistics skills with the programming domain of computer science. It refers to an interdisciplinary field of mathematics, statistics, and computer science.
- **Turing Test:** It measures the ability of machine to exhibit intelligent behavior.
- **Artificial Intelligence:** is referred to the ability of a machine to exhibit human behavior like problem solving, understanding natural language and interacting with the environment intelligently. **Personalized recommendations:** It helps to provide personalized feed to customers, based on their past usage of machines.
- **Automation industry:** It has played a vital role in automating the tedious jobs like car manufacturing and image & video analysis. The most important utilization of artificial intelligence is integration of internet in the devices of daily use. The entire industry or IoT and smart devices is based on automation.
- **Natural Language Processing (NLP):** AI enabled the machines to understand and respond to the natural human languages. ChatGPT and chatbots are examples of it. Now you can switch ON your air conditioner with voice command rather than picking up the remote controller.
- **Robotics:** Whether you watch some latest robot, talking and cooking like a human being or you get a small robot to clean your house, all this technology is due to significant advancement of artificial intelligence.
- **Computer vision:** It is a specialized branch of AI that teaches computers to draw meaningful results from digital images, videos and other visuals.
- **Smart cities:** Artificial Intelligence enables to develop efficient infrastructure and services, which can be used to manage energy consumption, traffic management and other necessary tasks for smart cities.
- **AI Agents:** Artificial Intelligence provides bases to develop modern AI agents like Siri, Alexa, Google assistant, ChatGPT and Cortana to generate human language. For example, virtual assistants like Siri, Google Assistant, Alexa and Cortana use NLP to understand human voice commands.
- **Deep learning:** is a type of computer learning which uses layers of artificial neural network to learn from data. It is a learning method which is implemented by using some Neural network.
- **Neural Network:** It is system designed to work like human brain. It is made up of artificial neurons. These tiny neurons make decisions like human brain.
- **Supervised Learning Model:** In this learning model a set of labeled data is provided to the machine/computer. The labeled data means that questions with their correct

answers are provided to the machine. After reading/learning a lot of questions/answers the machine is trained. Now, only new questions without correct answers are provided to that machine and it provides us with the possible correct answers.

- **Unsupervised learning model:** In this model unlabeled data is provided to the machine/computer. The algorithm identifies the pattern in the data and divide it into groups based on similarities and differences. When the training is completed, the computer can identify new data and put it in the relevant category.
- **Reinforce learning model:** In this model the machines are trained with the help of positive or negative feedback. The agent is provided with an environment where it must take certain actions, if the action is helpful towards the desired solution the agent will be rewarded with some points through feedback. If the actions is not helpful towards the desired solution the agent will be penalized(punished) by deducting the points.
- **Churn Prediction** is the technique to find out the possibility of a customer, to stop taking the services or products of a certain company. It is used in the fields of business and marketing.
- **Behavioral segmentation:** is the process of dividing customers into groups based on their behavior. This approach enables the business to design marketing strategies and promotional offers more precisely. It enables the business to meet the specific needs of customers segments.
- **Data Visualization:** It is the process of creating graphical representation of data and information. It involves various visual elements like charts, graphs, and maps.
- **Methods of Data Visualization:** Methods of Data Visualization are Bar Chart, Line Chart, Pie Chart, Scatter Plot, Histogram, Box Plot, Heatmap, Bubble Chart etc.
- **Quantitative visualization** is used to represent numerical data. It focuses on quantities or numbers to show measurable data. It is used to display data that can be measured or counted.
- **Categorical visualization** is used to represents data that falls into distinct categories. It helps to proportions or parts of a whole. This type is ideal for displaying nominal or ordinal data.
- **Temporal visualization** is used to display data that changes over time. It is used for time-series data. Line graphs are usually used to represent temporal data
- **Spatial visualization:** is used to represent data related to physical locations or spaces, visualizing geographic or spatial data. For example, a heatmap showing population density across different regions. Multivariate visualization: is used to represents data involving more than two variables or dimensions. Scatter plots and heatmaps: are effective tools to represent such data. For example, a scatter plot matrix showing relationships between income, age, and spending habits.

Interactive Visualization Interactive visualization allows users to interact with the data, through digital platforms. These digital platforms can be dashboards and filters. A dashboard helps the user to filter and manipulate data visualizations and explore different trends and insights is

Statistical Visualization Statistical visualization is used to present data to show statistical properties, such as distribution or correlation. Histograms, box plots, and scatter plots are popular choices.

Information Visualization Information visualization is used to present complex data sets in an easier way, often for abstract or conceptual data. Network diagrams, tree maps, and word clouds are effective tools. For example, a network diagram showing relationships between entities, such as a social network.

Business Intelligence: Data visualization helps to make data driven, well informed decisions. It is used to find market trends and helps to track and improve performance.

Database is collection of related data stored in the form of tables. Various computer software is used to create databases to store data in organized form so that it can be managed, retrieved and stored efficiently.

Data science life cycle is an iterative process. It has various stages; each has certain steps to be taken to move forward. It requires a structured approach to solve data driven problems.

Problem Definition: At this stage the business problem or research question is defined with clear objectives and outcomes. The scope and limitations of the problem are determined.

Data Collection: At this stage, data is gathered by using various available, reliable and useful sources. The way data is collected depends on the kind of problem. For simple problems, we can observe and write down what we see, while more complex problems might need surveys, experiments etc.

Data Cleaning: Also called data preprocessing, this step is about fixing the data to make sure it's ready for analysis. Any errors, such as wrong information or duplicates, are corrected or removed. Missing values, which might cause problems, are either filled in with estimated values or taken out altogether.

➤ **Data Analysis:** After cleaning the data, we have to analyze it. This is done using tools like statistics or machine learning techniques, which help in finding patterns, trends, or important details.

➤ **Data Modeling:** In this stage, the data is organized into a model, which is a structure that represents the data and how different parts relate to each other. This could involve creating diagrams that show the relationships between different entities (like customers and the products they purchase) and their attributes.

Exercise

Q1. Select the best answer for the following MCQs.

1. Which of the following is the primary benefit of integrating Mathematics and Statistics with Computer Science in Data Science?
 - a) Improved data visualization
 - b) Better forecasting
 - c) Increased accuracy
 - d) Better decision making
2. Which of the following best describes the relationship between Data Science and Artificial Intelligence?
 - a) Data Science is a subset of Artificial Intelligence
 - b) Artificial Intelligence is a tool used in Data Science
 - c) Data Science and Artificial Intelligence are unrelated
 - d) Data Science enables Artificial Intelligence
3. The Turing Test, proposed by Alan Turing in 1950, measures a machine's ability to exhibit intelligent behavior. Which of the following is the fundamental assumption underlies this test?
 - a) Humans are better
 - b) Machines are equal
 - c) Intelligence levels vary
 - d) Machines copy humans
4. Which of the following should be considered critically while developing AI-powered chatbots and virtual assistants?
 - a) User experience
 - b) Data security
 - c) Contextual awareness
 - d) Emotional intelligence
5. What ethical consideration arises from the integration of Artificial Intelligence (AI) into daily life devices?
 - a) Job displacement due to automation.
 - b) Increased energy consumption.
 - c) Improved customer service.
 - d) Enhanced data security.
6. Which of the following fields of Artificial Intelligence (AI) enables smartphones to recognize faces and unlock devices?
 - a) NLP
 - b) Computer vision
 - c) Deep learning
 - d) Neural networks
7. A company wants to develop a system that categorizes customer feedback into positive, negative, or neutral. Which learning model would be most suitable?
 - a) Supervised learning
 - b) Unsupervised learning
 - c) Reinforcement learning
 - d) Deep learning

8. In a Reinforcement Learning model, what is the primary function of rewards and penalties provided as feedback to the agent?
- Labeling data
 - Evaluating performance
 - Improving action choices
 - Classifying outcomes
9. Which stage of the data science life cycle ensures the model's accuracy, reliability, and compliance with privacy rules?
- Model Deployment
 - Model Evaluation
 - Data Analysis
 - Maintenance and Monitoring
10. Which of the following is the key characteristic of the "Data Cleaning" stage in the data science life cycle?
- Data collection
 - Error removal and data organization
 - Pattern identification
 - Model deployment

Q2. Write answers of the following short response questions.

- Describe how data science helps businesses make informed decisions and provide two industry examples.
- Identify three ways data science contributes to machine learning and artificial intelligence.
- Differentiate between supervised learning and unsupervised learning.
- Describe an everyday example that illustrates reinforcement learning.
- Write the appropriate machine learning model for each of the following scenarios:

Sr. No.	Scenario	Suitable Machine Learning Model
1	You have a basket of mixed fruits (apple and banana) and you want a robot/machine to sort them.	
2	You are given a task to learn how to ride a bicycle to participate in some sports event.	
3	You have pile of Lego blocks of different colors, and you want your computer to group them by colors regardless of their shapes.	
4	You have a book with pictures, and you want to teach your sibling to recognize them.	
5	You want to train a toy robot to find its way out of maze.	
6	Your parents want you to clean your messy room if you want to attend the birthday party of your friend.	
7	You have a set of shapes (square, triangle, circle) and you want to teach a computer to recognize them.	
8	You have a book collection without specific categories, and you want your sibling to arrange them according to size, choice or ease of access.	
9	You are given a task to find the similarity in various flavors of ice cream.	
10	You have to unlock some rewards in your favorite video game.	

Q3. Write answers of the following extended response questions.

1. Analyze the interrelationship between Data Science, Machine Learning, and Artificial Intelligence.
2. Identify any three types of data visualization, give their applications as well.
3. Discuss the way data visualization can be used to communicate data uncertainty, provide two specific examples.
4. Describe the key considerations in selecting appropriate visualizations for different types of data and analyses?
5. Explain the uses of data visualization in detail
6. What are the potential consequences of poor data quality on model performance?



Activity-1

Predicting Pet Preferences:

Objective: To help students understand the usage of data science, machine learning and artificial intelligence by analyzing and predicting pet preferences in their daily lives.

Make the groups of students, to collect data from younger kids on their preferences about pet. The students will collect the age and gender of the younger kids and write down their preferred pet, which they want to have.

Ask the students to calculate the mean median and mode of the collected data after cleaning data and removing errors or duplication from their collected data.

Create a bar graph or pie graph to display the results and ask them to apply ML and AI on it.

Guide them to use a simple algorithm like decision tree to predict the pet preference for new data. For example, take the age and gender of a new kid which was not part of the survey, apply the algorithm and predict that which pet would be preferred by this kid. Compare the results with the actual choice of the kid.

To apply artificial intelligence concepts, ask them to imagine virtual pet assistant, that uses AI to suggest a pet based on the kids' preferences.

Outcome: This activity will help students with the implementation of machine learning, data science and artificial intelligence in daily life.



Activity-2

Visualize Favorite Foods:

Objective: To help students understand significance and application of data visualization to understand and compare data. They will learn how to find the opinion of people if they want to start some business somewhere.

Make the groups in the class, each group may contain 3 to 4 students. Ask them to conduct a survey of students' favorite food like biryani, nihari, pizza, ice cream etc. Each group will design different questions for the survey but overall it will be about food preferences.

Calculate the frequency of each food, like how many students like biryani, nihari, pizza, ice cream etc.

Ask the students to clean the data and create a graph best suitable for the collected data.

Let them visualize their data and check which food is more popular.

Let them compare the results of one group with the results of another group.

They will understand the differences in the results based on the differences in the questions.

Outcome: This activity will help students understand how data visualization is helpful to start some food business in some area. They will also learn that their results may vary depending upon their questions and methods of data collection.



Activity-3

Churn Prediction:

Objective: To help students understand how churn prediction can be helpful for any business.

This activity can be assigned to the students on an individual basis. Ask them to choose the business being conducted by some of their family member or someone in their friends' circle.

Students will collect the data about the products or services provided by the business they chose.

They will collect data about the customers as well.

By following the method explained in the case study given in this unit, they will make predictions about the customers who are likely to stop taking the product or services from that business.

Students will design affordable business offerings for the potential customers who are likely to leave that business.

They will compare the results of their predictions. And improve their model if they can.

Outcome: This activity will help students understand usage of churn predictions.