

RNS INSTITUTE OF TECHNOLOGY
Dr. VISHNUVARDHAN ROAD, CHANNASANDRA, BENGALURU -560 098

Department of Information Science and Engineering



VISION of the College

Building RNSIT into a World - Class Institution

MISSION of the College

To impart high quality education in Engineering, Technology and Management with a difference, enabling students to excel in their career by

1. Attracting quality Students and preparing them with a strong foundation in fundamentals so as *to achieve distinctions in various walks of life* leading to outstanding contributions.
2. Imparting value based, need based, and choice based and skill based professional education to the aspiring youth and *carving them into disciplined, World class Professionals with social responsibility.*
3. Promoting excellence in Teaching, Research and Consultancy that galvanizes academic consciousness among Faculty and Students.
4. Exposing Students to emerging frontiers of knowledge in various domains and make them suitable for Industry, Entrepreneurship, Higher studies, and Research & Development.
5. Providing freedom of action and choice for all the Stake holders with better visibility.

VISION of the Department

Building Information Technology Professionals by Imparting Quality Education and Inculcating Key Competencies.

MISSION of the Department

- Provide strong fundamentals through learner centric approach
- Instil technical, interpersonal, interdisciplinary skills and logical thinking for holistic development
- Train to excel in higher education, research, and innovation with global perspective
- Develop leadership and entrepreneurship qualities with societal responsibilities

PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

ISE Graduates within three-four years of graduation will have

- **PEO1:** Acquired the fundamentals of computers and applied knowledge of Information Science & Engineering and continue to develop their technical competencies by problem solving using programming.
- **PEO2:** Ability to formulate problems, attained the Proficiency to develop system/application software in a scalable and robust manner with various platforms, tools and frameworks to provide cost effective solutions.
- **PEO3:** Obtained the capacity to investigate the necessities of the software Product, adapt to technological advancement, promote collaboration and interdisciplinary activities, Protecting Environment and developing Comprehensive leadership.
- **PEO4:** Enabled to be employed and provide innovative solutions to real-world problems across different domains.
- **PEO5:** Possessed communication skills, ability to work in teams, professional ethics, social responsibility, entrepreneur and management, to achieve higher career goals, and pursue higher studies.

PROGRAM OUTCOMES (POs)

Engineering Graduates will be able to:

- ▮ **PO1: Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization for the solution of complex engineering problems
- ▮ **PO2: Problem analysis:** Identify, formulate, research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences and engineering sciences.
- ▮ **PO3: Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for public health and safety, and cultural, societal, and environmental considerations.
- ▮ **PO4: Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
- ▮ **PO5: Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools, including prediction and modeling to complex engineering activities, with an understanding of the limitations.
- ▮ **PO6: The engineer and society:** Apply reasoning informed by the contextual

knowledge to assess Societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

- ▮ **PO7: Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
- ▮ **PO8: Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
- ▮ **PO9: Individual and team work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
- ▮ **PO10: Communication:** Communicate effectively on complex engineering activities with the engineering community and with the society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
- ▮ **PO11: Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
- ▮ **PO12: Life-long learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

PROGRAM SPECIFIC OUTCOMES (PSOs)

ISE Graduates will have

- **PSO1 – Problem Solving Abilities:** Ability to demonstrate the fundamental and theoretical concepts, analyze the real time problems and develop customized software solutions by applying the knowledge of mathematics and algorithmic techniques.
- **PSO2 – Applied Engineering Skills:** Enable creative thinking, Ability to apply standard practices and strategies, technical skills in software design, development, integration of systems and management for improving the security, reliability and survivability of the infrastructure.
- **PSO3 – General Expertise and Higher Learning:** Ability to exchange knowledge effectively demonstrate the ability of team work, documentation skills, professional ethics, entrepreneurial skills and continuing higher education in the field of Information technology.

COURSE OBJECTIVES OF AI&ML

This course will enable students to

- CLO 1.** Gain a historical perspective of AI and its foundations
- CLO 2.** Become familiar with basic principles of AI toward problem solving
- CLO 3.** Familiarize with the basics of Machine Learning & Machine Learning process, basics of Decision Tree, and probability learning
- CLO 4.** Understand the working of Artificial Neural Networks and basic concepts of clustering algorithms

COURSE OUTCOMES OF AI&ML

After studying this course, students will be able to:

- CO 1.** Apply the knowledge of searching and reasoning techniques for different applications.
- CO 2.** Have a good understanding of machine learning in relation to other fields and fundamental issues and challenges of machine learning.
- CO 3.** Apply the knowledge of classification algorithms on various dataset and compare results
- CO 4.** Model the neuron and Neural Network, and to analyze ANN learning and its applications.
- CO 5.** Identifying the suitable clustering algorithm for different pattern

Module 2

Contents: Informed Search Strategies: Greedy best-first search, A*search, Heuristic functions. Introduction to Machine Learning , Understanding Data

Textbook 1: Chapter 3 - 3.5, 3.5.1, 3.5.2, 3.6

Textbook 2: Chapter 1 and 2

2.1 NEED FOR MACHINE LEARNING Business organizations use huge amount of data for their daily activities. Earlier, the full potential of this data was not utilized due to two reasons. One reason was data being scattered across different archive systems and organizations not being able to integrate these sources fully. Secondly, the lack of awareness about software tools that could help to unearth the useful information from data. Not anymore! Business organizations have now started to use the latest technology, machine learning, for this purpose.

Machine learning has become so popular because of three reasons:

1. High volume of available data to manage: Big companies such as Facebook, Twitter, and YouTube generate huge amount of data that grows at a phenomenal rate. It is estimated that the data approximately gets doubled every year
2. Second reason is that the cost of storage has reduced. The hardware cost has also dropped. Therefore, it is easier now to capture, process, store, distribute, and transmit the digital information.
3. Third reason for popularity of machine learning is the availability of complex algorithms now. Especially with the advent of deep learning, many algorithms are available for machine learning

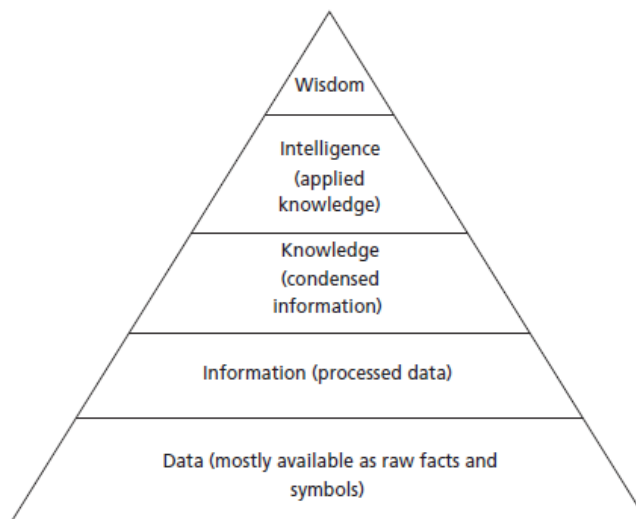


Figure 1.1: The Knowledge Pyramid

What is data? All facts are data. Data can be numbers or text that can be processed by a computer.

- Today, organizations are accumulating vast and growing amounts of data with data sources such as flat files, databases, or data warehouses in different storage formats.
- Processed data is called information. This includes patterns, associations, or relationships among data. For example, sales data can be analyzed to extract information like which is the fast selling product.
- Condensed information is called knowledge. For example, the historical patterns and future trends obtained in the above sales data can be called knowledge. Unless knowledge is extracted, data is of no use. Similarly, knowledge is not useful unless it is put into action.
- Intelligence is the applied knowledge for actions. An actionable form of knowledge is called

intelligence.

- Computer systems have been successful till this stage. The ultimate objective of knowledge pyramid is wisdom that represents the maturity of mind that is, so far, exhibited only by humans.

2.2 MACHINE LEARNING EXPLAINED

Machine learning is an important sub-branch of Artificial Intelligence (AI). A frequently quoted definition of machine learning was by Arthur Samuel, one of the pioneers of Artificial Intelligence. He stated that “Machine learning is the field of study that gives the computers ability to learn without being explicitly programmed.”

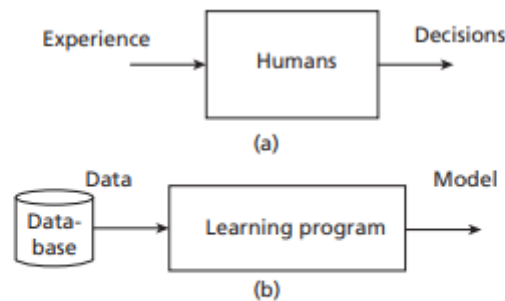


Figure 1.2: (a) A Learning System for Humans (b) A Learning System for Machine Learning

The aim of machine learning is to learn a model or set of rules from the given dataset automatically so that it can predict the unknown data correctly. As humans take decisions based on an experience, computers make models based on extracted patterns in the input data and then use these data-filled models for prediction and to take decisions. For computers, the learnt model is equivalent to human experience. This is shown in Figure 1.2

In summary, a model can be a formula, procedure or representation that can generate data decisions. The difference between pattern and model is that the former is local and applicable only to certain attributes but the latter is global and fits the entire dataset. For example, a model can be helpful to examine whether a given email is spam or not. The point is that the model is generated automatically from the given data.

Another pioneer of AI, Tom Mitchell's definition of machine learning states that, “**A computer program is said to learn from experience E , with respect to task T and some performance measure P , if its performance on T measured by P improves with experience E .**” The important components of this definition are experience E , task T , and performance measure P . For example, the task T could be detecting an object in an image. The machine can gain the knowledge of object using training dataset of thousands of images. This is called experience E . So, the focus is to use this experience E for this task of object detection T . The ability of the system to detect the object is measured by performance measures like precision and recall. Based on the performance measures, course correction can be done to improve the performance of the system.

2.3 MACHINE LEARNING IN RELATION TO OTHER FIELDS

Machine learning uses the concepts of Artificial Intelligence, Data Science, and Statistics primarily. It is the resultant of combined ideas of diverse fields.

2.3.1 Machine Learning and Artificial Intelligence: Machine learning is an important branch of AI, which is a much broader subject. The aim of AI is to develop intelligent agents. An agent can be a robot, humans, or any autonomous system

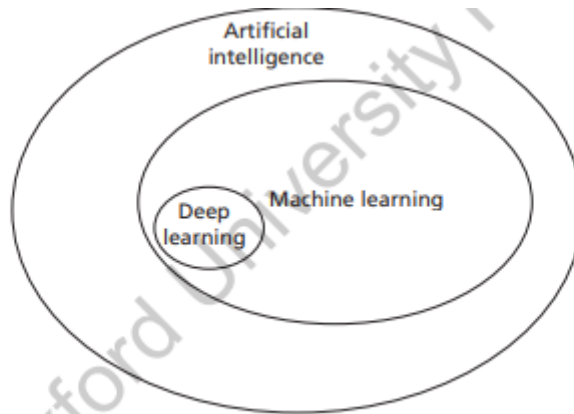


Figure 1.3: Relationship of AI with Machine Learning

The aim is to find relations and regularities present in the data. Machine learning is the subbranch of AI, whose aim is to extract the patterns for prediction. It is a broad field that includes learning from examples and other areas like reinforcement learning. The relationship of AI and machine learning is shown in Figure 1.3. The model can take an unknown instance and generate results. Deep learning is a subbranch of machine learning. In deep learning, the models are constructed using neural network technology. Neural networks are based on the human neuron models.

2.3.2 Machine Learning, Data Science, Data Mining, and Data Analytics

Data science is an ‘Umbrella’ term that encompasses many fields. Machine learning starts with data. Therefore, data science and machine learning are interlinked. Machine learning is a branch of data science. Data science deals with gathering of data for analysis. It is a broad field that includes

- i) **Big data** is a field of data science that Big Data deals with data’s following characteristics:
 - 1. Volume: Huge amount of data is generated by big companies like Facebook, Twitter, YouTube.
 - 2. Variety: Data is available in variety of forms like images, videos, and in different formats.
 - 3. Velocity: It refers to the speed at which the data is generated and processed.
- ii) **Data mining’s** original genesis is in the business. Like while mining the earth one Data Mining gets into precious resources, it is often believed that unearthing of the data produces hidden information that otherwise would have eluded the attention of the management.
- iii) **Data Analytics:** Another branch of data science is data analytics. It aims to extract useful Data Analytics knowledge from crude data. There are different types of analytics. Predictive data

analytics is used for making predictions. Machine learning is closely related to this branch of analytics and shares almost all algorithms.

- iv) **Pattern Recognition:** It is an engineering field. It uses machine learning algorithms to extract Pattern Recognition the features for pattern analysis and pattern classification. One can view pattern recognition as a specific application of machine learning

2.3.3 Statistics: Statistics requires knowledge of the statistical procedures and the guidance of a good statistician. It is mathematics intensive and models are often complicated equations and involve many assumptions. Statistical methods are developed in relation to the data being analysed. In addition, statistical methods are coherent and rigorous.

2.4 TYPES OF MACHINE LEARNING

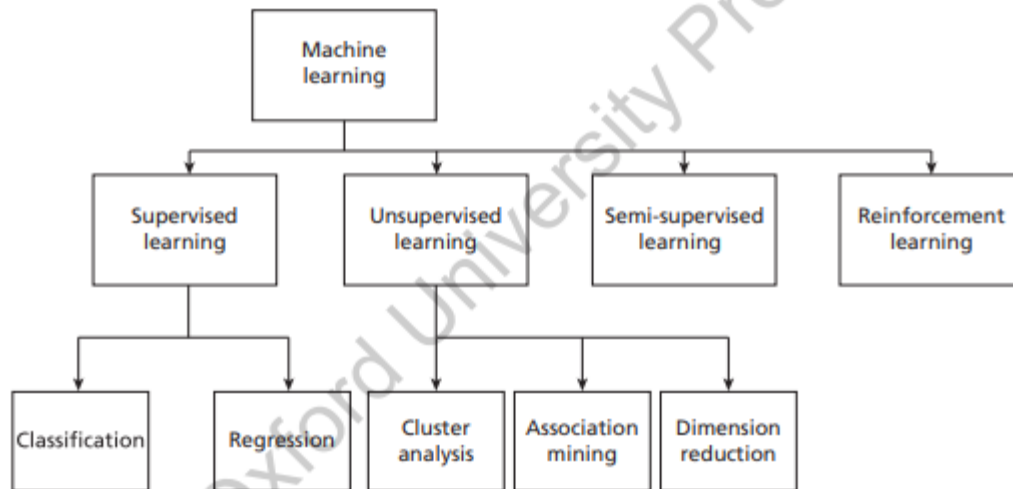


Figure 1.5: Types of Machine Learning

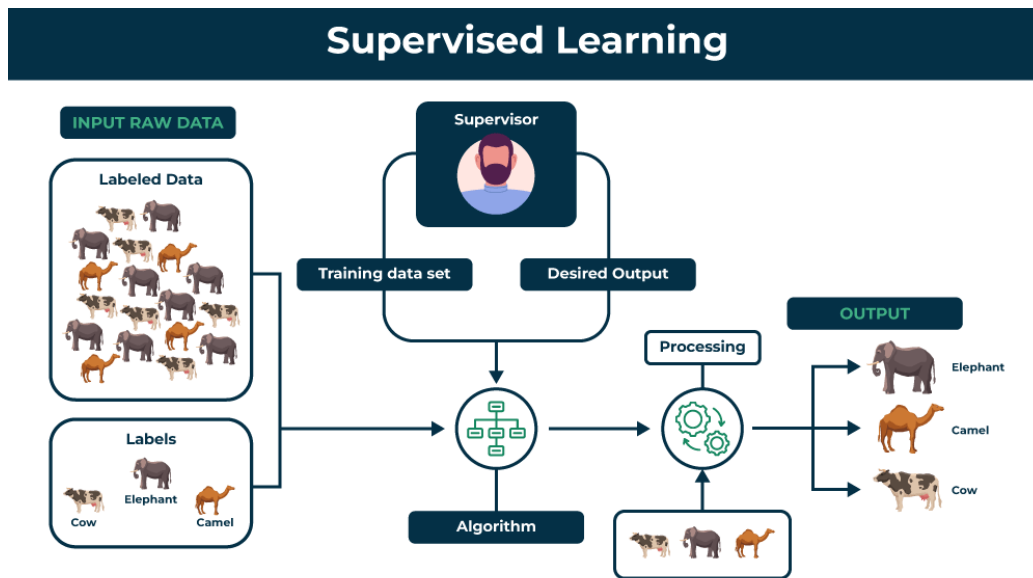
Types of Machine Learning

There are several types of machine learning, each with special characteristics and applications. Some of the main types of machine learning algorithms are as follows:

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Semi-Supervised Machine Learning
4. Reinforcement Learning

1. Supervised Machine Learning

Supervised learning is defined as when a model gets trained on a “**Labelled Dataset**”. Labelled datasets have both input and output parameters. In **Supervised Learning** algorithms learn to map points between inputs and correct outputs. It has both training and validation datasets labelled.



2. **Classification** : Classification deals with predicting **categorical** target variables, which represent discrete classes or labels. For instance, classifying emails as spam or not spam, or predicting whether a patient has a high risk of heart disease. Classification algorithms learn to map the input features to one of the predefined classes.

Some of the key algorithms of classification are: • Decision Tree • Random Forest • Support Vector Machines • Naïve Bayes • Artificial Neural Network and Deep Learning networks like CNN

3. Regression: **Regression**, on the other hand, deals with predicting **continuous** target variables, which represent numerical values. For example, predicting the price of a house based on its size, location, and amenities, or forecasting the sales of a product. Regression algorithms learn to map the input features to a continuous numerical value.

Here are some regression algorithms:

- Linear Regression
- Polynomial Regression
- Ridge Regression
- Lasso Regression
- Decision tree
- Random Forest

Advantages of Supervised Machine Learning

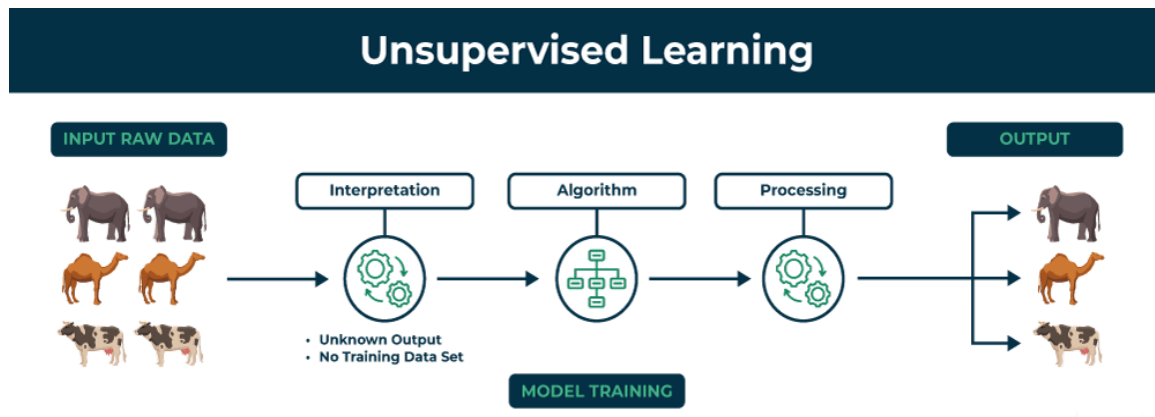
- **Supervised Learning** models can have high accuracy as they are trained on **labelled data**.
- The process of decision-making in supervised learning models is often interpretable.
- It can often be used in pre-trained models which saves time and resources when developing new models from scratch.

Disadvantages of Supervised Machine Learning

- It has limitations in knowing patterns and may struggle with unseen or unexpected patterns that are not present in the training data.
- It can be time-consuming and costly as it relies on **labeled** data only.
- It may lead to poor generalizations based on new data.

4. Unsupervised Machine Learning

Unsupervised Learning Unsupervised learning is a type of machine learning technique in which an algorithm discovers patterns and relationships using unlabeled data. Unlike supervised learning, unsupervised learning doesn't involve providing the algorithm with labeled target outputs. The primary goal of Unsupervised learning is often to discover hidden patterns, similarities, or clusters within the data, which can then be used for various purposes, such as data exploration, visualization, dimensionality reduction, and more.



Clustering: *Clustering* is the process of grouping data points into clusters based on their similarity. This technique is useful for identifying patterns and relationships in data without the need for labeled examples.

Some of the key clustering algorithms are: • k-means algorithm • Hierarchical algorithms

Association: *Association rule learning* is a technique for discovering relationships between items in a dataset. It identifies rules that indicate the presence of one item implies the presence of another item with a specific probability.

Dimensionality Reduction: Dimensionality reduction algorithms are examples of unsupervised algorithms. It takes a higher dimension data as input and outputs the data in lower dimension by taking advantage of the variance of the data. It is a task of reducing the dataset with few features without losing the generality.

Advantages of Unsupervised Machine Learning

- It helps to discover hidden patterns and various relationships between the data.
- Used for tasks such as **customer segmentation, anomaly detection, and data exploration.**
- It does not require labeled data and reduces the effort of data labeling.

Disadvantages of Unsupervised Machine Learning

- Without using labels, it may be difficult to predict the quality of the model's output.
- Cluster Interpretability may not be clear and may not have meaningful interpretations.
- It has techniques such as autoencoders and dimensionality reduction that can be used to extract meaningful features from raw data.

3. **Semi-supervised Learning** : There are circumstances where the dataset has a huge collection of unlabelled data and some labelled data. Labelling is a costly process and difficult to perform by the humans. Semi-supervised algorithms use unlabelled data by assigning a pseudo-label. Then, the labelled and pseudo-labelled dataset can be combined.
4. **Reinforcement Learning**: Reinforcement learning mimics human beings. Like human beings use ears and eyes to perceive the world and take actions, reinforcement learning allows the agent to interact with the environment to get rewards. The agent can be human, animal, robot, or any independent program. The rewards enable the agent to gain experience. The agent aims to maximize the reward.

2.5 CHALLENGES OF MACHINE LEARNING

Some of the challenges are listed below:

1. Problems – Machine learning can deal with the ‘well-posed’ problems where specifications are complete and available. Computers cannot solve ‘ill-posed’ problems
2. Huge data – This is a primary requirement of machine learning. Availability of a quality data is a challenge. A quality data means it should be large and should not have data problems such as missing data or incorrect data.
3. High computation power – With the availability of Big Data, the computational resource requirement has also increased. Systems with Graphics Processing Unit (GPU) or even Tensor Processing Unit (TPU) are required to execute machine learning algorithms. Also, machine learning tasks have become complex and hence time complexity has increased, and that can be solved only with high computing power.
4. Complexity of the algorithms – The selection of algorithms, describing the algorithms, application of algorithms to solve machine learning task, and comparison of algorithms have become necessary for machine learning or data scientists now. Algorithms have become a big topic of discussion and it is a challenge for machine learning professionals to design, select, and evaluate optimal algorithms.
5. Bias/Variance – Variance is the error of the model. This leads to a problem called bias/ variance tradeoff. A model that fits the training data correctly but fails for test data, in general lacks generalization, is called overfitting. The reverse problem is called underfitting where the model fails for training data but has good generalization. Overfitting and underfitting are great challenges for machine learning algorithm

2.6 MACHINE LEARNING PROCESS The emerging process model for the data mining solutions for business organizations is CRISP-DM. Since machine learning is like data mining, except for the aim, this process can be used for machine learning. CRISP-DM stands for Cross Industry Standard Process – Data Mining. This process involves six steps.

The steps are listed below in Figure 1.11.

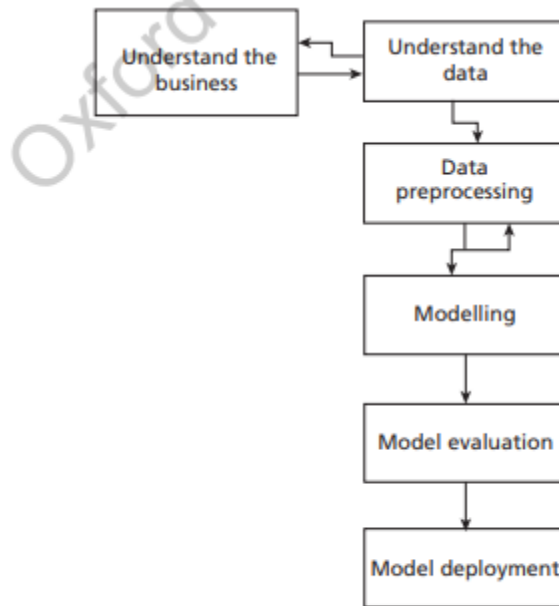


Figure 1.11: A Machine Learning/Data Mining Process

1. Understanding the business – This step involves understanding the objectives and requirements of the business organization. Generally, a single data mining algorithm is enough for giving the solution. This step also involves the formulation of the problem statement for the data mining process.
2. Understanding the data – It involves the steps like data collection, study of the characteristics of the data, formulation of hypothesis, and matching of patterns to the selected hypothesis.
3. Preparation of data – This step involves producing the final dataset by cleaning the raw data and preparation of data for the data mining process. The missing values may cause problems during both training and testing phases. Missing data forces classifiers to produce inaccurate results. This is a perennial problem for the classification models. Hence, suitable strategies should be adopted to handle the missing data.
4. Modelling – This step plays a role in the application of data mining algorithm for the data to obtain a model or pattern.
5. Evaluate – This step involves the evaluation of the data mining results using statistical analysis and visualization methods. The performance of the classifier is determined by evaluating the accuracy of the classifier. The process of classification is a fuzzy issue. For example, classification of emails requires extensive domain knowledge and requires domain experts. Hence, performance of the classifier is very crucial.
6. Deployment – This step involves the deployment of results of the data mining algorithm to improve the existing process or for a new situation.

2.7 MACHINE LEARNING APPLICATIONS

Machine Learning technologies are used widely now in different domains. Machine learning applications are everywhere! One encounters many machine learning applications in the day-to-day life. Some applications are listed below:

1. Sentiment analysis – This is an application of natural language processing (NLP) where the words of documents are converted to sentiments like happy, sad, and angry which are captured by emoticons effectively. For movie reviews or product reviews, five stars or one star are automatically attached using sentiment analysis programs.
2. Recommendation systems – These are systems that make personalized purchases possible. For example, Amazon recommends users to find related books or books bought by people who have the same taste like you, and Netflix suggests shows or related movies of your taste. The recommendation systems are based on machine learning.
3. Voice assistants – Products like Amazon Alexa, Microsoft Cortana, Apple Siri, and Google Assistant are all examples of voice assistants. They take speech commands and perform tasks. These chatbots are the result of machine learning technologies.
4. Technologies like Google Maps and those used by Uber are all examples of machine learning which offer to locate and navigate shortest paths to reduce time.

Table 1.4: Applications' Survey Table

S.No.	Problem Domain	Applications
1.	Business	Predicting the bankruptcy of a business firm
2.	Banking	Prediction of bank loan defaulters and detecting credit card frauds
3.	Image Processing	Image search engines, object identification, image classification, and generating synthetic images
4.	Audio/Voice	Chatbots like Alexa, Microsoft Cortana. Developing chatbots for customer support, speech to text, and text to voice
5.	Telecommunication	Trend analysis and identification of bogus calls, fraudulent calls and its callers, churn analysis
6.	Marketing	Retail sales analysis, market basket analysis, product performance analysis, market segmentation analysis, and study of travel patterns of customers for marketing tours
7.	Games	Game programs for Chess, GO, and Atari video games
8.	Natural Language Translation	Google Translate, Text summarization, and sentiment analysis
9.	Web Analysis and Services	Identification of access patterns, detection of e-mail spams, viruses, personalized web services, search engines like Google, detection of promotion of user websites, and finding loyalty of users after web page layout modification
10.	Medicine	Prediction of diseases, given disease symptoms as cancer or diabetes. Prediction of effectiveness of the treatment using patient history and Chatbots to interact with patients like IBM Watson uses machine learning technologies.
11.	Multimedia and Security	Face recognition/identification, biometric projects like identification of a person from a large image or video database, and applications involving multimedia retrieval
12.	Scientific Domain	Discovery of new galaxies, identification of groups of houses based on house type/geographical location, identification of earthquake epicenters, and identification of similar land use

2.8 The various Vs of big data

Big data is best described with the six Vs: volume, variety, velocity, value, veracity and variability

1. Volume

Volume is an obvious feature of big data and is mainly about the relationship between size and processing capacity. This aspect changes rapidly as data collection continues to increase. Just like the IT capacity for storage and processing.

2. Variety

The V of variety describes the wide variety of data that is being stored and still needs to be processed and analyzed. New types of data from social networks and mobile devices, among others, complement existing types of structured information. For example: audio and video files, photos, GPS data, medical files, instrument measurements, graphics, web documents, bonus cards and internet search behavior. Unstructured data such as voice and social media make processing and categorizing data extra complicated.

3. Velocity

Velocity is a measurement of the temporary value of data. Big data is rapidly changing. Therefore, we need to process structured and unstructured data streams quickly to take advantage of geolocation data, perceived hypes and trends, and **real time** available market and customer information. Velocity involves the condition that you need to process your data within minutes or seconds to get the results you're looking for.

4. **Value:** This V describes what value you can get from which data and how big data gets better results from stored data

5. Veracity

Veracity shows the quality and origin of data, allows it to be considered questionable, conflicting or impure, and provides information about matters you are not sure how to deal with. In short: the truth and authenticity of the data, and what can you do with it? In a sense, it is a hygiene factor. By showing the veracity of your data, you show that you have taken a critical look at it.

6. Variability

Finally, variability: to what extent, and how fast, is the structure of your data changing? And how often does the meaning or shape of your data change?

2.9 Data Source

A DATA SOURCE CAN BE ANYTHING –

- STRUCTURED DATA
- SEMI-STRUCTURED DATA
- UNSTRUCTURED DATA

Structured Data:

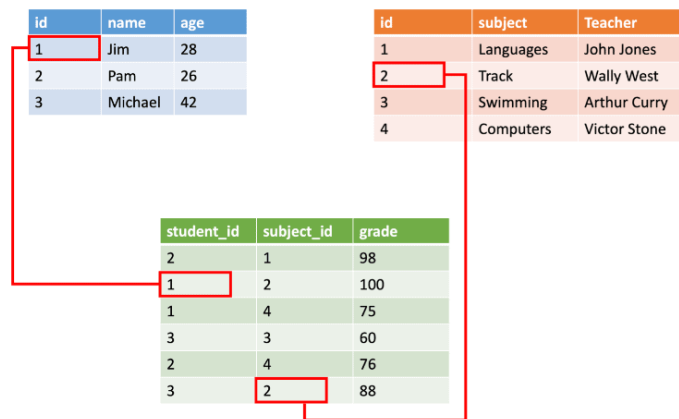
- Structured data is generally tabular data that is represented by columns and rows in a database.
- Databases that hold tables in this form are called relational databases.
- The mathematical term “relation” specifies a formed set of data held as a table.
- In structured data, all row in a table has the same set of columns.

A Structured Data Can Be Any One Of The Following –

- A. Record Data : A dataset is a collection of measurements taken from a process. The measurements can be arranged in the form of matrix. Rows in the matrix represent an object and can be called as entities, cases or records. Columns of dataset are called attributes, features or fields.
- B. Graph Data : It involves relationships among objects. For example, a web page can refer to

another web page. This can be modeled as a graph.

- C. Data Matrix: Variation of record type because it consists of numeric attributes
- D. Ordered Data – Ordered data objects involve attributes that have an implicit order among them.
 - Sequence Data: Sequential data without timestamps. Data involves sequence of words or letters.
 - Spatial Data: It has attributes such as position or areas. For example, maps are spatial data where points are related by location
 - Temporal Data: It is the data whose attributes are associated with time. Eg: customer purchasing patterns during festival time is sequential data over time.
- SQL (Structured Query Language) programming language used for structured data.



Semi-structured data:

- Semi-structured data is information that doesn't consist of Structured data (relational database) but still has some structure to it.
- Semi-structured data consists of documents held in JavaScript Object Notation (JSON) format. It also includes key-value stores and graph databases.

```
## Document 1 ##
{
  "customerID": "103248",
  "name":
  {
    "first": "AAA",
    "last": "BBB"
  },
  "address":
  {
    "street": "Main Street",
    "number": "101",
    "city": "Acity",
    "state": "NY"
  },
  "ccOnFile": "yes",
  "firstOrder": "02/28/2003"
}
```

Unstructured Data

- Unstructured data is information that either is not organized in a pre-defined manner or does not have a pre-defined data model.
- Unstructured information is a set of text-heavy but may contain data such as numbers, dates, and facts as well.
- **Videos, audio, and binary** data files might not have a specific structure. They're assigned to **unstructured** data.



2.10 Data Storage and Representation

Once the dataset is assembled, it must be stored in a structure that is suitable for data analysis. The goal of data storage management is to make data available for analysis. Some of them are listed below:

- Flat Files
- Database System
- World Wide Web
- XML
- Data Stream
- RSS
- JSON

Flat Files These are the simplest and most commonly available data source. It is also the cheapest way of organizing the data. These flat files are the files where data is stored in plain ASCII or EBCDIC format. Minor changes of data in flat files affect the results of the data mining algorithms. Hence, flat file is suitable only for storing small dataset and not desirable if the dataset becomes larger.

Some of the popular spreadsheet formats are listed below:

- CSV files – CSV stands for comma-separated value files where the values are separated by commas. These are used by spreadsheet and database applications. The first row may have attributes and the rest of the rows represent the data.
- TSV files – TSV stands for Tab separated values files where values are separated by Tab.

Both CSV and TSV files are generic in nature and can be shared. There are many tools like Google Sheets and Microsoft Excel to process these files.

Database Systems Types Are

1. **Transactional Database** : is a collection of transactional records. Each record is a transaction. A transaction may have a time stamp, identifier and a set of items. They are created for performing associational analysis that indicates the correlation among the items
2. **Time Series Database** : Time related information like log files where data is associated with timestamp. Observing sales of product continuously may yield a time series data
3. **Spatial Database**: contain spatial information in a raster or vector format. Eg: images can be stored as a raster data

World Wide Web (WWW) It provides a diverse, worldwide online information source. The objective of data mining algorithms is to mine interesting patterns of information present in WWW.

XML (eXtensible Markup Language) It is both human and machine interpretable data format that can be used to represent data that needs to be shared across the platforms.

Data Stream It is dynamic data, which flows in and out of the observing environment. Typical characteristics of data stream are huge volume of data, dynamic, fixed order movement, and real-time constraints.

RSS (Really Simple Syndication) It is a format for sharing instant feeds across services.

JSON (JavaScript Object Notation) It is another useful data interchange format that is often used for many machine learning algorithms.

2.11 Big Data Analytics and types of analytics

What is big data analytics?

Big data analytics is the process of collecting, examining, and analyzing large amounts of data to discover market trends, insights, and patterns that can help companies make better business decisions. This information is available quickly and efficiently so that companies can be agile in crafting plans to maintain their competitive advantage.

For example, big data analytics is integral to the modern health care industry. As you can imagine, thousands of patient records, insurance plans, prescriptions, and vaccine information need to be managed. It comprises huge amounts of structured and unstructured data, which can offer important insights when analytics are applied. Big data analytics does this quickly and efficiently so that health care providers can use the information to make informed, life-saving diagnoses.

Types of big data analytics (+ examples)

There are **four main types of big data analytics** that support and inform different business decisions.

1. Descriptive analytics

Descriptive analytics refers to data that can be easily read and interpreted. This data helps create reports and visualize information that can detail company profits and sales.

Example: During the pandemic, a leading pharmaceuticals company conducted data analysis on its offices and research labs. Descriptive analytics helped them identify unutilized spaces and departments that were consolidated, saving the company millions of dollars.

2. Diagnostics analytics

Diagnostics analytics helps companies understand why a problem occurred. Big data technologies and tools allow users to mine and recover data that helps dissect an issue and prevent it from happening in the future.

Example: A clothing company's sales have decreased even though customers continue to add items to their shopping carts. Diagnostics analytics helped to understand that the payment page was not working properly for a few weeks.

3. Predictive analytics

Predictive analytics looks at past and present data to make predictions. With artificial intelligence (AI), machine learning, and data mining, users can analyze the data to predict market trends.

Example: In the manufacturing sector, companies can use algorithms based on historical data to predict if or when a piece of equipment will malfunction or break down.

4. Prescriptive analytics

Prescriptive analytics provides a solution to a problem, relying on AI and machine learning to gather data and use it for risk management.

Example: Within the energy sector, utility companies, gas producers, and pipeline owners identify factors that affect the price of oil and gas in order to hedge risks.

2.12 Big Data Analysis Framework

Big data framework is a layered architecture. A 4 layer architecture has the following layers:

1. Data connection layer: It has data ingestion mechanism and data connectors. Data ingestion means taking raw data and importing it into appropriate data structures. It performs the tasks of ETL process.
2. Data Management layer: It performs preprocessing of data. The purpose of this layer is to allow parallel execution of queries and read, write and data management tasks. Schemes such as data-in place, where the data is not moved at all, or constructing data repositories such as data warehouses and pull data on-demand mechanisms.
3. Data Analytic layer: It has many functionalities such as statistical tests, machine learning algorithms to understand, and construction of machine learning models.
4. Presentation layer: It has mechanisms such as dashboards, and applications that display the results of analytical engines and machine learning algorithms

2.12.1 Data Collection

It is often estimated that most of the time is spent for collection of good quality data. 'Good Data' is one that has the following properties:

- Timeliness- Data should be relevant and not stale or obsolete data
- Relevancy- All the necessary information should be available and there should be no bias in data
- Knowledge about the data: data should be understandable and interpretable, and should be self-sufficient as desired by the domain knowledge engineer

Data Source can be classified as open/public data, social media data and multimodal data

1. Open or public data source eg: Government census data
2. Social Media – data from platforms like Twitter, facebook, instagram etc
3. Multimodal data- data that involves many modes such as text, video, audio and mixed types
Eg: WWW has huge distributed data, image archives

2.12.2 Data Preprocessing

In addition to incorrect data entry, dirty data can be generated due to the improper methods in data management and data storage. Some dirty data types are explained below:

- Inaccurate data – It is possible that a data value can be correct, but not accurate. At times, it is practical to examine against other files or fields to find out if the data value is accurate based on the context it is used. Still, accuracy can often only be validated by manual verification.
- Inconsistent data – Unchecked data redundancy leads to data inconsistencies. Each organization is affected with inconsistent and repetitive data. This is particularly typical with customer data.
- Incomplete data – Data with missing values is the main type of incomplete data.
- Duplicate data – Duplicate data may occur due to repeated submissions, improper data joining or user error.

- Outlier data- may be legitimate data and sometimes are of interest to the data mining algorithms. These errors come during data collection stage. These must be removed so that ML algorithms yield better results as the quality of the results is determined by the quality of input data.
- Data with missing values: Missing data need to be handled during data cleaning stage

Following procedures are used to solve the problem of missing data:

- a) Ignore the tuple – not effective when percentage of missing values are high
- b) Fill in the values manually- not feasible for larger sets
- c) A global constant can be used to fill in the missing attributes
- d) Attribute value may be filled by the attribute values eg: avg income can replace a missing values
- e) Use the most possible values to fill in the missing value

Removal of noisy or outlier data:

The binning method can be used for smoothing the data.

Mostly data is full of noise. Data smoothing is a data pre-processing technique using a different kind of algorithm to remove the noise from the data set. This allows important patterns to stand out.

Unsorted data for price in dollars

Before sorting: 8 16, 9, 15, 21, 21, 24, 30, 26, 27, 30, 34

First of all, sort the data

After Sorting: 8, 9, 15, 16, 21, 21, 24, 26, 27, 30, 30, 34

Smoothing the data by equal frequency bins

Bin 1: 8, 9, 15, 16

Bin 2: 21, 21, 24, 26,

Bin 3: 27, 30, 30, 34

Smoothing by bin means

For Bin 1:

$$(8 + 9 + 15 + 16 / 4) = 12$$

(4 indicating the total values like 8, 9, 15, 16)

Bin 1 = 12, 12, 12, 12

For Bin 2:

$$(21 + 21 + 24 + 26 / 4) = 23$$

Bin 2 = 23, 23, 23, 23

For Bin 3:

$$(27 + 30 + 30 + 34 / 4) = 30$$

Bin 3 = 30, 30, 30, 30

Smoothing by bin boundaries

Bin 1: 8, 8, 8, 15

Bin 2: 21, 21, 25, 25

Bin 3: 26, 26, 26, 34

Data Integration and Data Transformations

Data integration in data mining refers to the process of combining data from multiple sources into a single, unified view. This can involve cleaning and transforming the data, as well as resolving any inconsistencies or conflicts that may exist between the different sources. The goal of data integration is to make the data more useful and meaningful for the purposes of analysis and decision making. Data transformation in data mining refers to the process of converting raw data into a format that is suitable for analysis and modeling. The goal of data transformation is to prepare the data for data mining so that it can be used to extract useful insights and knowledge.

Normalization: Data normalization involves converting all data variables into a given range. Techniques that are used for normalization are:

a) **Min-Max Normalization:**

- This transforms the original data linearly.
- Suppose that: min_A is the minima and max_A is the maxima of an attribute, P
- Where v is the value you want to plot in the new range.
- v' is the new value you get after normalizing the old value.

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

https://T4Tutorials.com

- Suppose we have to normalize the following data set, **200, 300, 400, 600, 1000** to a new range [0, 1], then using min-max normalization
- min = 200, max = 1000, new_minA = 0, new_maxA = 1

$$v'_1 = \frac{200 - 200}{1000 - 200} (1 - 0) + 0 = 0$$

$$v'_2 = \frac{300 - 200}{1000 - 200} (1 - 0) + 0 = \frac{100}{800} = 0.125$$

$$v'_3 = \frac{400 - 200}{1000 - 200} (1 - 0) + 0 = \frac{200}{800} = 0.25$$

$$v'_4 = \frac{600 - 200}{1000 - 200} (1 - 0) + 0 = \frac{400}{800} = 0.5$$

$$v'_5 = \frac{1000 - 200}{1000 - 200} (1 - 0) + 0 = 1$$

The normalized data set is: **0, 0.125, 0.25, 0.5, 1**

- b) **Z-Score** normalization: helps in the normalization of data. If we normalize the data into a simpler form with the help of z score normalization, then it's very easy to understand by our brains.

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

The diagram shows the formula $Z = \frac{x - \mu}{\sigma}$ with red arrows pointing to the components: 'Score' points to x , 'Mean' points to μ , and 'SD' (Standard Deviation) points to σ .

Example 2.3: Consider the mark list $V = \{10, 20, 30\}$, convert the marks to z-score.

Solution: The mean and Sample Standard deviation (σ) values of the list V are 20 and 10, respectively. So the z-scores of these marks are calculated using Eq. (2.2) as:

$$\text{z-score of } 10 = \frac{10 - 20}{10} = -\frac{10}{10} = -1$$

$$\text{z-score of } 20 = \frac{20 - 20}{10} = \frac{0}{10} = 0$$

$$\text{z-score of } 30 = \frac{30 - 20}{10} = \frac{10}{10} = 1$$

Hence, the z-score of the marks 10, 20, 30 are -1, 0 and 1, respectively.

- c) **Data Reduction:** Data reduction reduces data size but produce the same results. Different ways of data reduction are data aggregation, feature selection and dimensionality reduction.

2.14 Descriptive Statistics

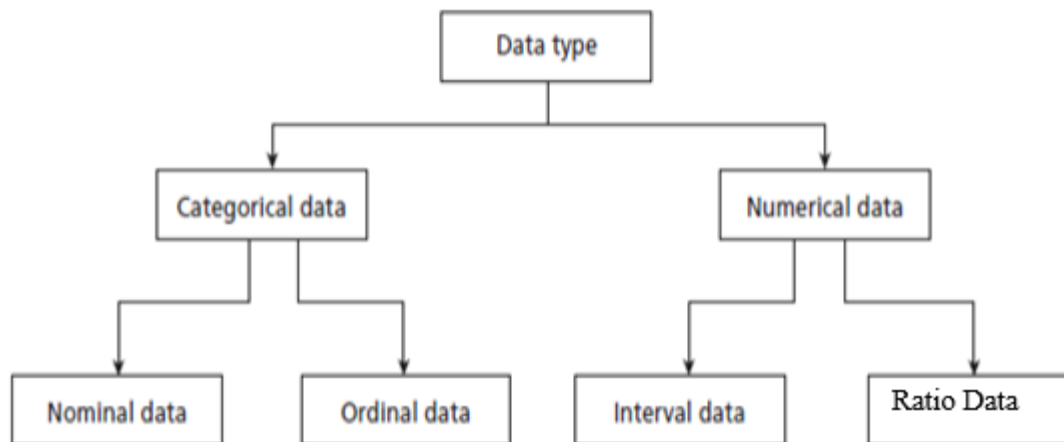
Descriptive statistics is a branch of statistics that summarizes dataset and describe data. Descriptive analysis and data visualization techniques help to understand the nature of the data which further helps to determine the kinds of machine learning or data mining tasks that can be applied to the data. This step is known as Exploratory Data Analysis (EDA).

Dataset and Data types






A dataset can be assumed to be collection of data objects, each data object may be records, patterns, vectors, events, cases or observations. These records contain attribute which is associated with a value.

Numerical Data:

Numerical data can be characterized by continuous or discrete data. Continuous data can assume any value within a range whereas discrete data has distinct values.



Categorical Data: Categorical data refers to **a form of information that can be stored and identified based on their names or labels**. It is a type of qualitative data that can be grouped into categories instead of being measured numerically. This data type is made up of categorical variables that show things like a person's gender, hometown, and so on.

Points	Nominal Data	Ordinal Data	
Meaning	Nominal data are those items which are distinguished by a simple naming system. They are data with no numeric value, such as profession. The nominal data just name a thing without applying it to an order related to other numbered items.	Ordinal data is data which is placed into some kind of order by their position on the scale. For example, they may indicate superiority. However, you cannot do arithmetic with ordinal numbers because they only show sequence.	
Are they categorical?	Yes, nominal data are also called categorical data.	Ordinal variables are “in between” categorical and quantitative variables.	
The level of quantitative value	Without any type of quantitative value.	We can assign numbers to ordinal data but we cannot do arithmetic with ordinal numbers.	
Key Points	<ul style="list-style-type: none"> Nominal data cannot be quantified. It also cannot be assigned to any type of order. The values are only allocated to distinct categories. Those categories have no meaningful order. 	<ul style="list-style-type: none"> Ordinal data is placed into some kind of order. Ordinal numbers only show sequence. We can assign numbers to ordinal data. We cannot do arithmetic with ordinal numbers. We don't know if the differences between the values are equal. 	
Examples	<ul style="list-style-type: none"> Gender (Women, Men) Religion (Muslin, Buddhist, Christian) Hair color (Blonde, Brown, Brunette, Red, etc.) Marital status (Married, Single, Widowed) Ethnicity (Hispanic, Asian) Eye color (Blue, Green, Brown). 	<ul style="list-style-type: none"> The first, second and third person in a competition. Education level: the elementary school, high school, college. Customer rating of the sales experience on a scale of 1-10. Letter grades: A, B, C, and etc. Economic status: low, medium and high. Customer level of satisfaction: very satisfied, satisfied, neutral, dissatisfied, very dissatisfied. 	

A **ratio scale** is a quantitative scale where there is a true zero and equal intervals between neighboring points. Unlike on an interval scale, a zero on a ratio scale means there is a total absence of the variable you are measuring. Eg: Height, Age, Weight, Length, area, and population are examples of ratio scales.

Interval data is measured along a numerical scale that has equal distances between adjacent values. These distances are called “intervals”. There is no true zero on an interval scale, which is what distinguishes it from a ratio scale. On an interval scale, zero is an arbitrary point, not a complete absence of the variable. Common examples of interval scales include standardized tests, such as the SAT, and psychological inventories, Temperature in Fahrenheit or Celsius.

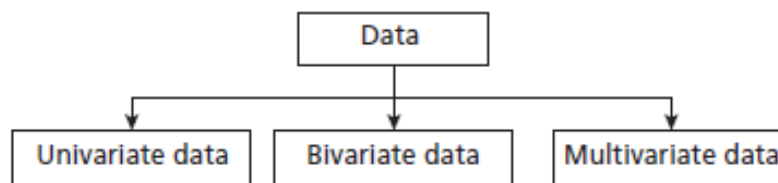


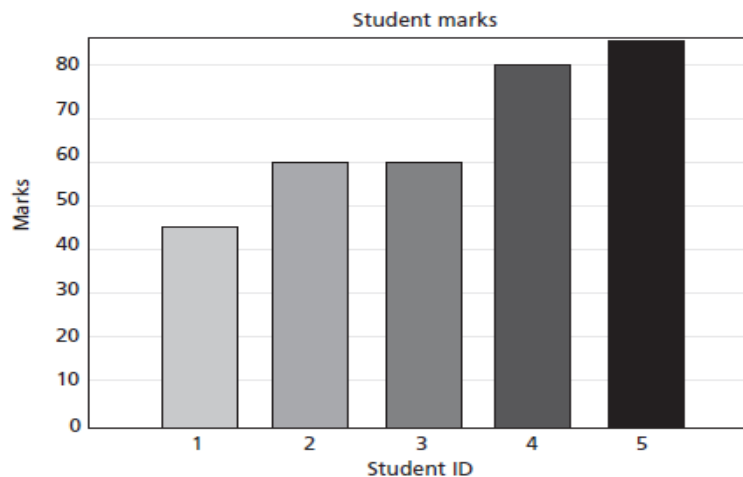
Figure 2.2: Types of Data Based on Variables

2.15 Univariate Data Analysis and Visualization

2.15.1 Data Visualization

Bar Chart A Bar chart (or Bar graph) is used to display the frequency distribution for variables. Bar charts are used to illustrate discrete data. The charts can also help to explain the counts of nominal data. It also helps in comparing the frequency of different groups.

The bar chart for students' marks {45, 60, 60, 80, 85} with Student ID = {1, 2, 3, 4, 5} is shown below in Figure 2.3.



Pie Chart These are equally helpful in illustrating the univariate data. The percentage frequency distribution of students' marks {22, 22, 40, 40, 70, 70, 70, 85, 90, 90} is below in Figure 2.4.

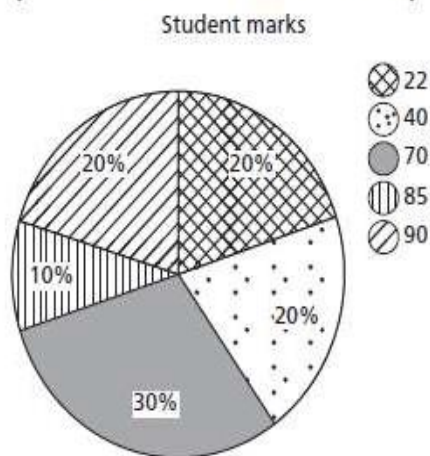


Figure 2.4: Pie Chart

It can be observed that the number of students with 22 marks are 2. The total number of students are 10. So, $2/10 \times 100 = 20\%$ space in a pie of 100% is allotted for marks 22 in Figure 2.4.

Histogram It plays an important role in data mining for showing frequency distributions. The histogram for students' marks {45, 60, 60, 80, 85} in the group range of 0–25, 26–50, 51–75, 76–100 is given below in Figure 2.5. One can visually inspect from Figure 2.5 that the number of students in the range 76–100 is 2.

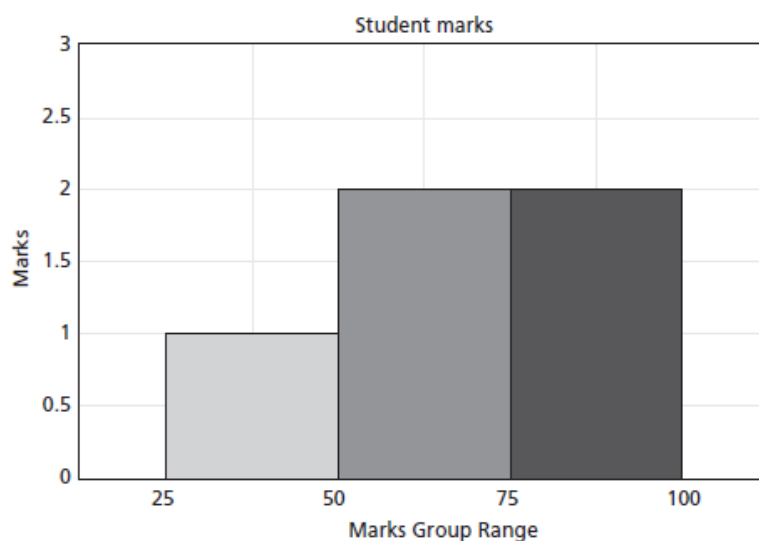


Figure 2.5: Sample Histogram of English Marks

Dot Plots These are similar to bar charts. They are less clustered as compared to bar charts, as they illustrate the bars only with single points. The dot plot of English marks for five students with ID as {1, 2, 3, 4, 5} and marks {45, 60, 60, 80, 85} is given in Figure 2.6. The advantage is that by visual inspection one can find out who got more marks.

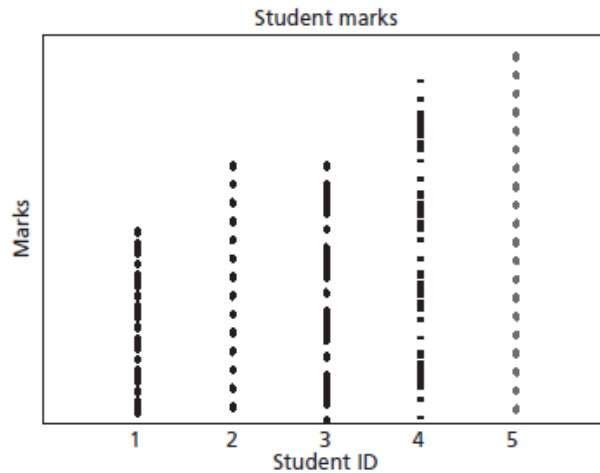
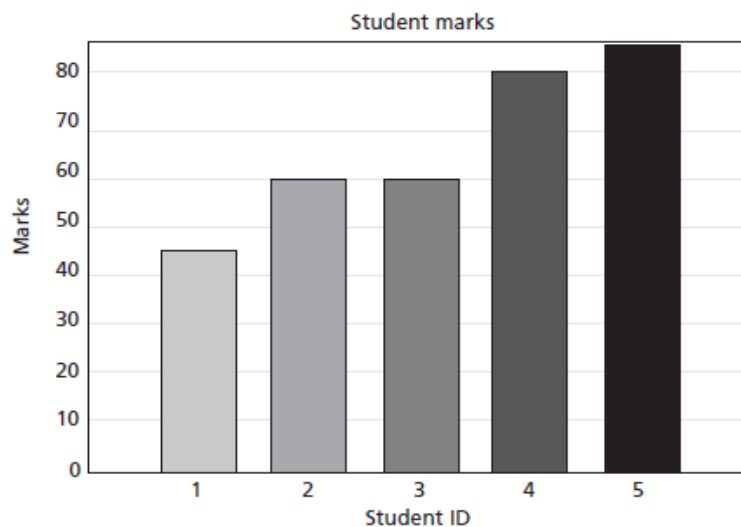


Figure 2.6: Dot Plots

Bar Chart A Bar chart (or Bar graph) is used to display the frequency distribution for variables. Bar charts are used to illustrate discrete data. The charts can also help to explain the counts of nominal data. It also helps in comparing the frequency of different groups.

The bar chart for students' marks {45, 60, 60, 80, 85} with Student ID = {1, 2, 3, 4, 5} is shown below in Figure 2.3.



2.15.2 Central Tendency

Measures of central tendency help you find the middle, or the average, of a dataset. The 3 most common measures of central tendency are the mode, median, and mean.

- **Mode:** the most frequent value.
- **Median:** the middle number in an ordered dataset.
- **Mean:** the sum of all values divided by the total number of values.

Mean Formula:

Arithmetic Mean:

$$\bar{X} = \frac{\sum X}{N}$$

- \bar{X} = mean of N numbers
- $\sum X$ = sum of each value in the population
- N = number of values in the population

Weighted Mean:

Weighted mean gives different importance to all items as the item importance varies.

$$\bar{x} = \frac{\sum_{i=1}^n (x_i * w_i)}{\sum_{i=1}^n w_i}$$

Geometric Mean:

$$\text{Geometric mean} = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{N}} = \sqrt[N]{x_1 \times x_2 \times \dots \times x_N}$$

Mean is extreme sensitive to noise. Even small changes in the input affect the mean drastically. Hence, often the top 2% is chopped off and then the mean is calculated for a larger dataset.

Median Formula:

Median – The middle value in the distribution is called median. If the total number of items in the distribution is odd, then the middle value is called median. If the numbers are even, then the average value of two items in the centre is the median. It can be observed that the median is the value where x_i is divided into two equal halves, with half of the values being lower than the median and half higher than the median. A median class is that class where $(N/2)^{\text{th}}$ item is present.

In the continuous case, the median is given by the formula:

$$\text{Median} = L_1 + \frac{\frac{N}{2} - cf}{f} \times i \quad (2.7)$$

Mode:

The **mode** or modal value of a data set is the most frequently occurring value. It's a measure of central tendency that tells you the most popular choice or most common characteristic of your sample.