INTRODUCTION TO MACHINE L E A R N I N G

Why Machine Learning is Popular

1. **High Data Volume**: Large companies like Facebook, Twitter, and YouTube generate huge amounts of data, which doubles annually.
2. **Reduced Storage Costs**: Declining hardware and storage costs make it easier to capture, store, and process vast amounts of data.
3. **Advanced Algorithms**: The development of complex algorithms, especially in deep learning, enables more powerful machine learning applications.

The Knowledge Pyramid

1. **Data**: Basic facts and raw numbers. Organizations store vast amounts of data from sources like databases and warehouses.

2. **Information**: Processed data revealing patterns or associations. For instance, analyzing sales data to determine the best-selling product.

3. **Knowledge**: Condensed information, such as historical patterns and future trends. Extracting knowledge from data is crucial for decision-making.

4. **Intelligence**: Applied knowledge. It represents actionable insights, such as strategies derived from knowledge.

5. **Wisdom**: The highest level, where intelligence evolves into maturity and sound judgment, typically exhibited by humans.

Objective of Machine Learning

Machine learning processes archival data to:

- Make better decisions.
- Design new products.
- Improve business processes.
- Build decision support systems.

What is Machine Learning?

Machine learning is a sub-field of AI that allows computers to learn without explicit programming. As Arthur Samuel defined: **"Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed."**

In conventional programming, we teach computers how to perform tasks step-by-step. However, real-world problems like image recognition or complex games require systems that can learn from data directly.

Evolution of Machine Learning

- Early systems like **expert systems** (e.g., MYCIN for medical diagnosis) relied on human rules and logic, but they didn't exhibit true intelligence.
- Machine learning evolved with **data-driven systems**, focusing on learning from data to automatically predict unknown outcomes.

Learning System

- **Human Learning (Fig. 1.2a)**: Humans make decisions based on experience.
- **Machine Learning (Fig. 1.2b)**: Machines create models from data patterns and use these models for prediction, akin to human experience.

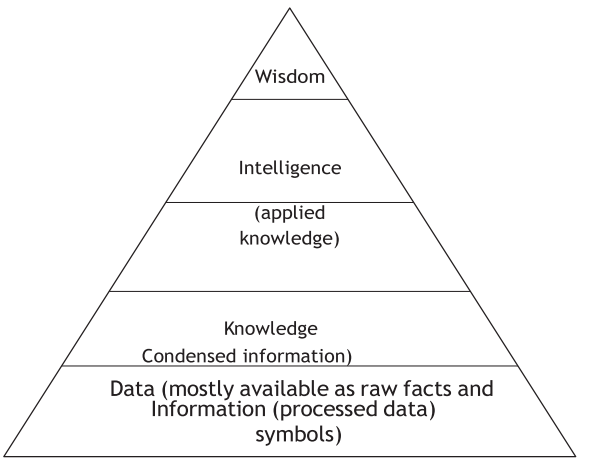


Figure 1.1: The Knowledge Pyramid

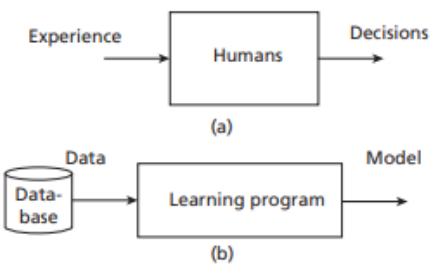


Figure 1.2: (a) A Learning System for Humans (b) A Learning Systemfor Machine Learning

Data Quality and Learning

- The **quality of data** directly impacts the **quality of experience** and, ultimately, the **quality of learning systems**.

Statistical Learning

- In **statistical learning**, the relationship between input **x** and output **y** is modeled as: **y = f(x)**
  - **f** is the **learning function** mapping inputs to outputs.
- In **machine learning**, this is referred to as the **mapping of input to output**.

Model in Machine Learning

- A **model** is a summary of raw data, structured into a representation for decision-making.
- Models can be in different forms, such as:

1. **Mathematical equations**: e.g., linear regression formula
2. **Relational diagrams**: like **decision trees** or **graphs**
3. **Logical rules**: like **if/else** rules (e.g., rule-based spam filters)
4. **Clusters**: for grouping data (e.g., **k-means clustering**)

Patterns vs. Models

- **Pattern**: Local; applicable to specific parts of the data.
- **Model**: Global; fits the entire dataset.
  - **Example**: A model trained to detect spam can be used to predict whether an email is spam or not.

Tom Mitchell's Definition of Machine Learning

Tom Mitchell's famous definition:
"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."

- **Experience (E)**: Data used to learn (e.g., thousands of images to train an object detection model).
- **Task (T)**: The job the machine does (e.g., detecting objects in images).
- **Performance measure (P)**: How well the machine performs the task (e.g., precision, recall).

Example:

- **Task (T)**: Detecting an object in images.

- **Experience (E)**: Training data containing thousands of labeled images.
- **Performance measure (P)**: The system is evaluated by how accurately it detects objects, using metrics like **precision** and **recall**. Improvements can be made if the system underperforms.

Human vs. Machine Experience

- **Human Experience**: Gained through learning, observing, imitation, and trial & error.
  - Example: Learning how to ride a bike through practice and observation.
- **Machine Experience**: Gained through **data processing** and model building:

1. **Data Collection**: Gathering data from the environment (e.g., images, text, or numbers).
2. **Abstraction**: Extracting key features from the data (e.g., identifying basic features of an elephant: trunk, ears).
3. **Generalization**: Turning abstraction into an actionable form, like forming rules (heuristics) from past experiences.
   - Example: A self-driving car generalizes rules about stopping at red lights.
4. **Heuristics**: Actionable **"rules of thumb"** that guide decisions based on prior experience.
   - **Example**: A heuristic rule: If you see a red light, stop.
   - Heuristics can sometimes fail but are typically effective.

Evaluation & Course Correction

- **Heuristics**: Often work, but sometimes fail due to limitations (it's a general rule, not a certainty).
  - **Example**: If someone runs when sensing danger, it's an automatic response based on past experience (heuristics).
- **Evaluation**: Assesses the effectiveness of the model or heuristic. If the model underperforms, we use **evaluation measures** to **adjust** and **improve** it.

1.1 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.

1.3.1 Machine Learning and Artificial Intelligence

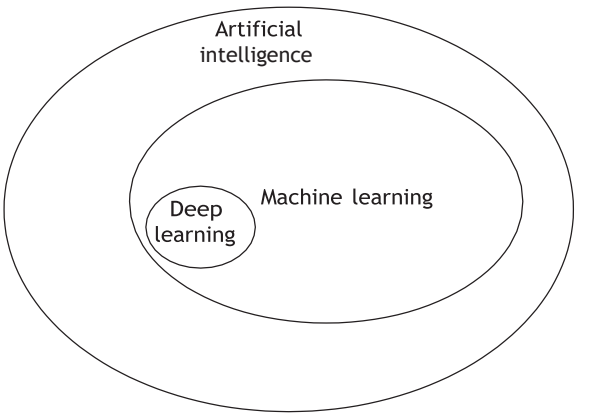


Figure 1.3: Relationship of AI with Machine Learning

Artificial Intelligence (AI) is a broad field focused on creating systems (called "intelligent agents") that can perform tasks autonomously, such as robots, humans, or other systems. The early goal of AI was ambitious: to create intelligent systems that could think and act like humans, focusing on logic and reasoning.

However, AI faced several challenges and periods of slow progress, called **AI winters**, where enthusiasm and funding declined. AI's resurgence came with the rise of **data-driven systems**—models that learn by finding patterns in data. This led to the development of **Machine Learning (ML)**, a key branch of AI.

Machine Learning aims to extract patterns from data to make predictions. Instead of explicitly programming systems for every possible scenario, ML algorithms "learn" from examples (training data) and can handle new, unseen situations. Machine learning includes various techniques like **reinforcement learning**, where agents learn by interacting with their environment.

Relationship Between AI and Machine Learning:

- AI is the broader field aiming to create intelligent agents.
- ML is a subfield of AI that focuses on learning from data.
- **Deep Learning**, a subset of ML, uses **neural networks** inspired by the human brain to build models. These networks consist of layers of interconnected units ("neurons") that process information in a way that mimics how the brain works, and they are especially useful for tasks like image and speech recognition.

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

Data Science is an umbrella term that covers various fields related to working with data. It involves gathering, processing, analyzing, and drawing insights from data. **Machine learning** starts with data, which makes it closely linked to data science. Here's how machine learning connects to related fields:

Big Data:

Big data is part of data science and refers to massive volumes of data generated by companies like Facebook, Twitter, and YouTube. It deals with three key characteristics:

1. **Volume**: The sheer amount of data being generated.
2. **Variety**: Data comes in many forms—text, images, videos, etc.
3. **Velocity**: The speed at which data is generated and processed.

Big data is essential for machine learning because many algorithms rely on large datasets for training. For example, deep learning (a subfield of ML) uses big data for tasks like image recognition and language translation.

Data Mining:

Data mining originally came from business applications. It's like "mining" for valuable information hidden in large datasets. While data mining and machine learning overlap significantly, the distinction is:

- **Data Mining**: Focuses on discovering hidden patterns in data.
- **Machine Learning**: Uses those patterns to make predictions.

Data Analytics:

Another branch of data science is **data analytics**, which aims to extract useful insights from raw data. There are different types of analytics, such as **predictive analytics**, which forecasts future events based on past data. Machine learning plays a major role in predictive analytics since many of its algorithms are used to make predictions.

Pattern Recognition:

Pattern recognition is an engineering field that uses machine learning algorithms to detect and classify patterns. While it's often considered a specific application of machine learning, it has its own identity as a field, dealing with tasks like facial recognition or speech analysis.

These relations are summarized in Figure 1.4.

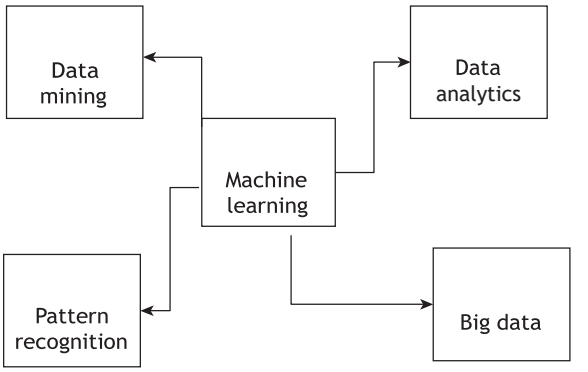


Figure 1.4: Relationship of Machine Learning with Other Major Fields

1.3.3 Machine Learning and Statistics

1. Statistics:

- **Definition**: A branch of mathematics focused on analyzing and interpreting data to uncover patterns and relationships.
- **Key Features**:
  - **Hypothesis-driven**: Starts with a hypothesis and tests it through experiments.
  - **Assumptions**: Requires strict assumptions (e.g., normal distribution, independence of variables).
  - **Mathematical Models**: Uses complex equations (e.g., regression, ANOVA) to explain data.
  - **Knowledge Required**: Strong statistical background needed for analysis and interpretation.
  - **Goal**: Primarily concerned with verifying relationships and patterns in data.

2. Machine Learning (ML):

- **Definition**: A branch of AI focused on building models that learn from data to make predictions or decisions without being explicitly programmed.
- **Key Features**:
  - **Data-driven**: Focuses on learning from data patterns for predictions.
  - **Less Assumptions**: Fewer restrictions on data (e.g., can handle non-normal data).
  - **Automation**: Emphasizes using tools and algorithms to automate the learning process.

- **Flexibility**: Works well with large, complex datasets; adaptable to different scenarios.
- **Goal**: Makes predictions based on learned patterns, often without needing detailed statistical knowledge.

1.2 TYPES OF MACHINE LEARNING

What does the word 'learn' mean? Learning, like adaptation, occurs as the result of interaction of the program with its environment. There are four types of machine learning as shown in Figure 1.5.

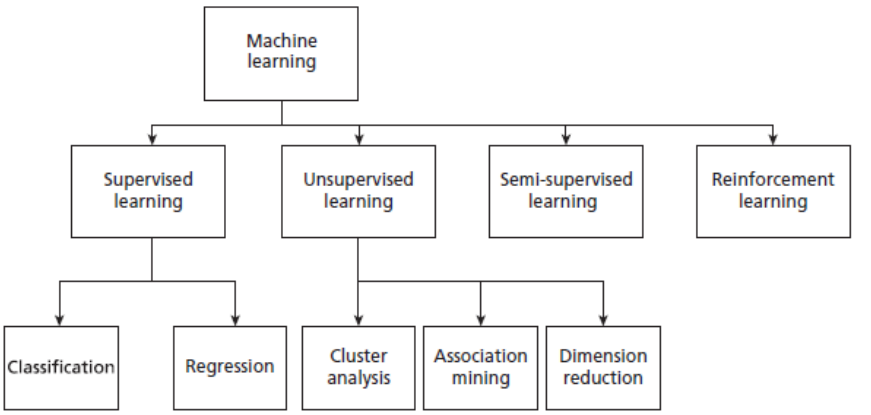


Figure 1.5: Types of Machine Learning

Before discussing the types of learning, it is necessary to discuss about data.

Labelled and Unlabelled Data: Data is a raw fact. Normally, data is represented in the form of a table. Data also can be referred to as a data point, sample, or an example. Each row of the table represents a data point. Features are attributes or characteristics of an object. Normally, the columns of the table are attributes. Out of all attributes, one attribute is important and is called a label. Label is the feature that we aim to predict. Thus, there are two types of data – labelled and unlabelled.

Labelled Data To illustrate labelled data, let us take one example dataset called Iris flower dataset or Fisher's Iris dataset. The dataset has 50 samples of Iris – with four attributes, length and width of sepals and petals. The target variable is called class. There are three classes – Iris setosa, Iris virginica, and Iris versicolor.

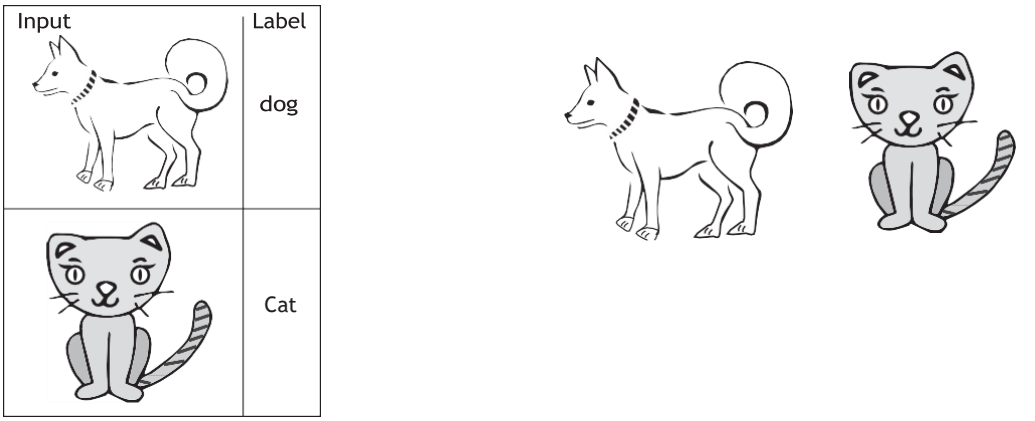
The partial data of Iris dataset is shown in Table 1.1.

Table 1.1: Iris Flower Dataset

S.No.	Length ofPetal	Width ofPetal	Length ofSepal	Width ofSepal	Class
1.	5.5	4.2	1.4	0.2	Setosa
2.	7	3.2	4.7	1.4	Versicolor

3.	7.3	2.9	6.3	1.8	Virginica

A dataset need not be always numbers. It can be images or video frames. Deep neural networks can handle images with labels. In the following Figure 1.6, the deep neural network takes images ofdogs and cats with labels for classification. In unlabelled data, there are no labels in the dataset.



(a) (b)

Figure 1.6: (a) Labelled Dataset (b) Unlabeled Dataset

1.4.1 Supervised Learning

Supervised algorithms use labelled dataset. As the name suggests, there is a supervisor or teachercomponent in supervised learning. A supervisor provides labelled data so that the model is constructed and generates test data.

In supervised learning algorithms, learning takes place in two stages. In layman terms, during thefirst stage, the teacher communicates the information to the student that the student is supposed tomaster. The student receives the information and understands it. During this stage, the teacher has noknowledge of whether the information is grasped by the student.

This leads to the second stage of learning. The teacher then asks the student a set of questionsto find out how much information has been grasped by the student. Based on these questions, the student is tested, and the teacher informs the student about his assessment. This kind of learningis typically called supervised learning.

Supervised learning has two methods:

1. Classification
2. Regression

Classification

Classification is a type of supervised learning.

- **Supervised Learning:** The algorithm learns from labeled data, where we know the correct answers.
- **Independent Variables:** These are the input features, also called attributes.
- **Dependent Variable (Label):** This is the target we want to predict, and it's in the form of discrete categories or labels (e.g., dog or cat).

How Classification Works:

1. **Training Stage:**

o The algorithm is given a dataset that includes both the features (input) and their correct labels (output).

o The algorithm learns from this data and creates a model.

2. **Testing Stage:**

o The model is tested on new, unseen data (input), and it predicts the label (output).

o For example, if you input an image of a dog or cat that the model hasn't seen before, the model will assign the correct label based on what it has learned.

Example:

In the Iris dataset, if you input data like (6.3, 2.9, 5.6, 1.8, ?), the model will predict the missing label. This process of assigning a label to new data is called **classification**.

Applications of Classification:

Image Recognition: Classifying images of animals, plants, or even medical conditions like cancer.

Types of Classification Models:

Classification models can be grouped into two categories:

1.**Generative Models:** Focus on how the data is generated and its distribution (e.g., Naïve Bayes).

2.**Discriminative Models:** Focus only on distinguishing between different classes (e.g., Support Vector Machines).

Key Classification Algorithms:

- **Decision Tree**
- **Random Forest**
- **Support Vector Machines (SVM)**
- **Naïve Bayes**

• **Artificial Neural Networks (ANN)** and **Deep Learning** (e.g., Convolutional Neural Networks - CNN)

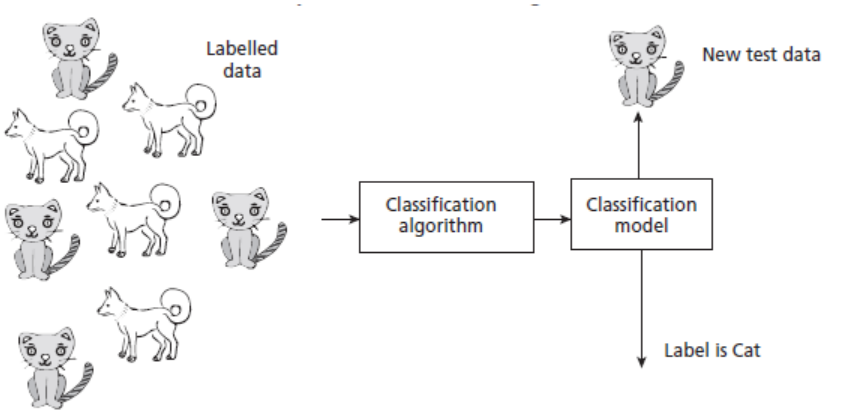


Figure 1.7: An Example Classification System

Regression Models

Regression is another type of **supervised learning**, similar to classification, but instead of predicting categories (labels), it predicts **continuous values**, like numbers.

Key Difference:

- **Regression:** Predicts continuous values (e.g., product sales, house prices).
- **Classification:** Predicts labels or categories (e.g., dog or cat).

How Regression Works:

In a regression model, we are trying to find a relationship between the **independent variable(s) (x)** and the **dependent variable (y)**.

For example, in Figure 1.8, the **independent variable (x)** is the number of weeks, and the **dependent variable (y)** is product sales. The regression model fits a line to the data, which can be used to predict future sales. This line is written as: shown in fig 1.8

Sales (y)=0.66×Week (x)+0.54

- Here, 0.66 and 0.54 are **regression coefficients** that the model learns from the data.
- If you want to predict the sales for the 8th week, you can substitute x=8x = 8x=8 into the formula and calculate the predicted sales (y).

Example:

For the 8th week: Sales=0.66×8+0.54

Sales=0.66×8+0.54 This gives a predicted value of sales for week 8.

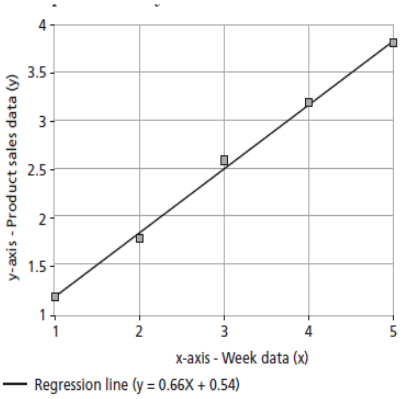


Figure 1.8: A Regression Model of the Form $y = ax + b$

Similarities Between Regression and Classification:

- Both are **supervised learning** methods, meaning they require a labeled training dataset.
- Both involve a **training stage** (where the model learns from data) and a **testing stage** (where the model is used to make predictions on new data).

Main Difference:

- **Regression:** Predicts **numbers** (continuous values).
- **Classification:** Predicts **categories** (discrete values, like class labels).

One of the most common regression algorithms is **linear regression**, which fits a straight line to the data.

Unsupervised learning

is a type of learning where there is **no supervisor or teacher** guiding the process. Instead, the algorithm **learns by itself** using trial and error.

How Unsupervised Learning Works:

- In this method, the algorithm is given data without any labels.
- The algorithm looks at the data and tries to find **patterns or groupings** on its own.
- The goal is to **group similar objects** together based on their characteristics.

Example of Unsupervised Learning:

Clustering

- **Clustering** is a common unsupervised learning technique.

- It groups objects into different **clusters**, where each cluster contains objects that are similar to each other.
- The objects in one cluster are different from those in other clusters.

For example, if you have a set of images of dogs and cats, a clustering algorithm will automatically group them into two clusters: one for dogs and one for cats, without needing any labels to tell it which is which.

Applications of Clustering:

- **Image Segmentation:** Grouping parts of an image, like separating a region of interest (e.g., identifying a tumor in a medical image).
- **Gene Analysis:** Finding groups of similar genes in a database.

In summary, unsupervised learning helps the algorithm **discover patterns** in data without any explicit instructions. **Cluster analysis** and **dimensional reduction** are key types of unsupervised learning.

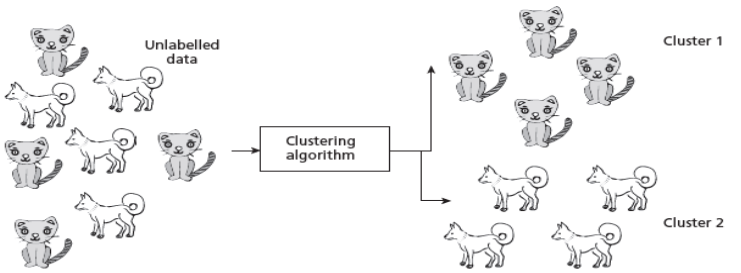


Figure 1.9: An Example Clustering Scheme

Some of the key clustering algorithms are:

- k-means algorithm
- Hierarchical algorithms

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 varianceof the data. It is a task of reducing the dataset with few features without losing the generality. The differences between supervised and unsupervised learning are listed in the followingTable 1.2.

Table 1.2: Differences between Supervised and Unsupervised Learning

S.No.	Supervised Learning	Unsupervised Learning
1.	There is a supervisor component	No supervisor component
2.	Uses Labelled data	Uses Unlabelled data
3.	Assigns categories or labels	Performs grouping process such that similar objectswill be in one cluster

1.4.2 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.

1.4.3 Reinforcement Learning

Reinforcement learning is a type of machine learning where an **agent** learns by interacting with its **environment**. The agent performs actions and receives feedback in the form of **rewards** or **penalties**, and its goal is to maximize the total reward over time.

Key Concepts:

- **Agent:** The learner or decision-maker.
- **Environment:** The world the agent interacts with.
- **Action:** What the agent can do.
- **Reward:** Feedback given to the agent based on its actions (positive for good actions, negative for bad actions).

The agent learns through **trial and error**, improving its strategy (called a **policy**) over time to make better decisions.

Example of Reinforcement Learning:

Consider a robot learning to walk:

- The robot (agent) takes steps (actions) in its environment.
- If the robot falls, it receives a **penalty** (negative reward). If it moves forward without falling, it receives a **reward** (positive reward).
- Over time, the robot adjusts its movements to maximize its forward motion and minimize falling, effectively learning how to walk.

In summary, reinforcement learning is about learning from **experience** to make better decisions in the future by maximizing rewards.

Consider the following example of a Grid game as shown in Figure 1.10.

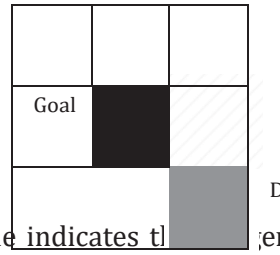


Figure 1.10: A Grid game

In this grid game, the gray tile indicates the danger, black is a block, and the tile with diagonallines is the goal. The aim is to start, say from bottom-left grid, using the actions left, right, top andbottom to reach the goal state.

To solve this sort of problem, there is no data. The agent interacts with the environment toget experience. In the above case, the agent tries to create a model by simulating many paths and finding rewarding paths. This experience helps in constructing a model.

1.5 CHALLENGES OF MACHINE LEARNING

Machine learning allows computers to solve certain types of problems much better than humans, especially tasks involving **computation**. For instance, computers can quickly calculate the square root of large numbers or win games like chess and Go against professional players.

However, humans are still better than machines at tasks like **recognition**, though modern machine learning systems, especially **deep learning**, are improving rapidly. For example, machines can recognize human faces instantly. But there are still challenges in machine learning, mainly due to the need for **high-quality data**.

Key Challenges in Machine Learning:

1. **Well-Posed Vs Ill-Posed Problems:**

- Machine learning works well with **well-posed problems**, where the problem is clearly defined and has enough information to find a solution.

- In **ill-posed problems**, there may be multiple possible answers, making it hard to find the correct one. For example, in a simple dataset (as shown in Table 1.3), several models could fit the data (e.g., multiplication or division). To solve such problems, more data is needed to narrow down the correct model.

Table 1.3: An Example

Input (x_1, x_2)	Output (y)

1, 1	1
2, 1	2
3, 1	3
4, 1	4
5, 1	5

Can a model for this test data be multiplication? That is, *y = x1 * x2*. Well! It is true! But, this is equally true that y may be *y = x1 / x2 or y = x1 ^ x2.* So, there are three functions that fit the data.

This means that the problem is ill-posed. To solve this problem, one needs more example to check the model. Puzzles and games that do not have sufficient specification may become an ill-posed problem and scientific computation has many ill-posed problems.

2. **Need for Huge, Quality Data:**

o Machine learning requires **large amounts of high-quality data**. The data must be complete, without missing or incorrect values. Poor-quality data can lead to inaccurate models.

3. **High Computational Power:**

o With the growth of **Big Data**, machine learning tasks require powerful computers with specialized hardware like **GPUs** or **TPUs** to handle the high computational load. The increasing complexity of tasks has made high-performance computing essential.

4. **Complexity of Algorithms:**

o Choosing the right machine learning algorithm, explaining how it works, applying it correctly, and comparing different algorithms are now critical skills for data scientists. This makes the selection and evaluation of algorithms a significant challenge.

5. **Bias-Variance Trade-off:**

o **Overfitting:** When a model performs well on training data but fails on test data, it's called overfitting. This means the model has learned the training data too well but lacks generalization to new data.

o **Underfitting:** When a model fails to perform well on both training and test data, it's called underfitting. The model is too simple to capture the patterns in the data.

o Balancing between overfitting and underfitting is a major challenge for machine learning algorithms.

1.6 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 machinelearning. CRISP-DM stands for Cross Industry Standard Process – Data Mining. This process involves six steps. The steps are listed below in Figure 1.11.

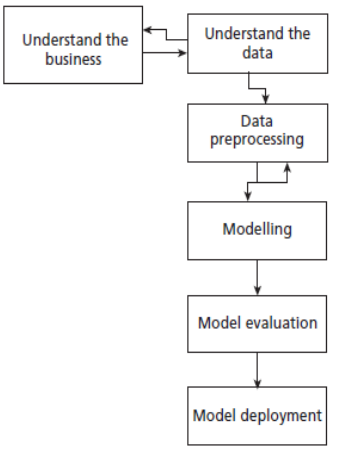


Figure 1.11: A Machine Learning/Data Mining Process

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 produceinaccurate results. This is a perennial problem for the classification models. Hence, suitablestrategies should be adopted to handle the missing data.

4. Modelling – This step plays a role in the application of data mining algorithm for the datato 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.

1.7 MACHINE LEARNING APPLICATIONS

Machine Learning technologies are used widely now in different domains. Machine learning appli cations 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 arecaptured by emoticons effectively. For movie reviews or product reviews, five stars or onestar 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 peoplewho 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.

The machine learning applications are enormous. The following Table 1.4 summarizes some ofthe machine learning applications.

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.	Telecommuni-cation	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

Key Terms:

- **Machine Learning** – A branch of AI that concerns about machines to learn automatically withoutbeing explicitly programmed.
- **Data** – A raw fact.
- **Model** – An explicit description of patterns in a data.
- **Experience** – A collection of knowledge and heuristics in humans and historical training data in case of machines.
- **Predictive Modelling** – A technique of developing models and making a prediction of unseen data.
- **Deep Learning** – A branch of machine learning that deals with constructing models using neural networks.
- **Data Science** – A field of study that encompasses capturing of data to its analysis covering all stagesof data management.
- **Data Analytics** – A field of study that deals with analysis of data.
- **Big Data** – A study of data that has characteristics of volume, variety, and velocity.
- **Statistics** – A branch of mathematics that deals with learning from data using statistical methods.
- **Hypothesis** – An initial assumption of an experiment.
- **Learning** – Adapting to the environment that happens because of interaction of an agent with the environment.
- **Label** – A target attribute.
- **Labelled Data** – A data that is associated with a label.

- **Unlabelled Data** – A data without labels.
- **Supervised Learning** – A type of machine learning that uses labelled data and learns with the help of a supervisor or teacher component.
- **Classification Program** – A supervisory learning method that takes an unknown input and assigns a label for it. In simple words, finds the category of class of the input attributes.
- **Regression Analysis** – A supervisory method that predicts the continuous variables based on the input variables.
- **Unsupervised Learning** – A type of machine leaning that uses unlabelled data and groups the attributes to clusters using a trial and error approach.
- **Cluster Analysis** – A type of unsupervised approach that groups the objects based on attributesso that similar objects or data points form a cluster.
- **Semi-supervised Learning** – A type of machine learning that uses limited labelled and largeunlabelled data. It first labels unlabelled data using labelled data and combines it for learning purposes.
- **Reinforcement Learning** – A type of machine learning that uses agents and environment interactionfor creating labelled data for learning.
- **Well-posed Problem** – A problem that has well-defined specifications. Otherwise, the problem is called ill-posed.
- **Bias/Variance** – The inability of the machine learning algorithm to predict correctly due to lackof generalization is called bias. Variance is the error of the model for training data. This leads to problems called overfitting and underfitting.
- **Model Deployment** – A method of deploying machine learning algorithms to improve the existing business processes for a new situation.

2.1 WHAT IS DATA?

- **Data** refers to raw facts that can be numbers, text, images, audio, or video.
- In computer systems, these facts are encoded in **bits**, allowing machines to process and store them.
- **Directly interpretable data**: Numbers or text, like "John is 25 years old."
- **Diffused data**: Data like images or videos that require computers to interpret, like identifying objects in a photo.

Types of Data Sources

1.**Flat files**: Simple files like CSV or text files.

2.**Databases**: Systems that store structured data.

3.**Data warehouses**: Centralized repositories for large volumes of data.

Operational vs. Non-operational Data

- **Operational data**: Data generated during regular business processes. Example: daily sales figures.
- **Non-operational data**: Data used for strategic decision-making, such as past sales data to predict future trends.

Data vs. Information

- **Data** alone is meaningless until it is processed to create **information**.
  - Example: A list of numbers is just data, but labeling it as "heights of students" gives it context, turning it into information.
- **Information** shows patterns, relationships, and associations.
  - Example: Analyzing sales data can reveal which products sell the most.

Big Data: Elements and Characteristics

What is Big Data?

- **Small data** can be processed using regular computers.
- **Big Data** is data that exceeds the capacity of standard computers and requires specialized tools.

The 6 Vs of Big Data

1. **Volume**:
   - Refers to the **size** of data.
   - Big Data is often measured in **petabytes (PB)** or **exabytes (EB)**, much larger than the gigabytes or terabytes of traditional data.
2. **Velocity**:
   - The **speed** at which data is generated and processed.
   - Thanks to IoT devices and the Internet, data arrives rapidly, often in real-time.
3. **Variety**:
   - The **diversity** of data formats:
     - **Form**: Data comes as text, audio, video, graphs, etc.
     - **Function**: Data from sources like conversations, transactions, or archives.
     - **Source**: Data can come from public sources, social media, or multimodal sources (combining different types).
4. **Veracity**:
   - Refers to the **accuracy** and **trustworthiness** of data.
   - Errors like technical glitches or human mistakes can affect the reliability of data, making veracity crucial.
5. **Validity**:
   - The **relevance** of data for a particular purpose, ensuring it is accurate and fit for decision-making.
6. **Value**:

- The **usefulness** of data based on the insights and information it provides, helping organizations make better decisions.

The data quality of the numeric attributes is determined by factors like precision, bias, and accuracy.

- Precision is defined as the closeness of repeated measurements. Often, standard deviation is used to measure the precision.
- Bias is a systematic result due to erroneous assumptions of the algorithms or procedures.
- Accuracy is the degree of measurement of errors that refers to the closeness of measurements to the true value of the quantity. Normally, the significant digits used to store and manipulate indicate the accuracy of the measurement.

2.1.1 Categories of Big Data

Big Data, data can be categorized into three types: **structured data**, **unstructured data**, and **semi-structured data**. Each type has its own characteristics, formats, and storage methods.

1. Structured Data

Definition: Structured data is organized and stored in a predefined format, such as a table in a database. This data is easy to search, retrieve, and analyze using tools like SQL.

Types of Structured Data:

- **Record Data**:
  - A dataset consists of a collection of measurements.
  - **Rows** (entities, cases, or records) represent objects.
  - **Columns** (attributes, features, or fields) represent measurements for each object.
  - A **label** refers to individual observations in the dataset.
- **Data Matrix**:

- A type of record data where all attributes are numeric.
- Data is represented as points in a multidimensional space where each attribute represents a dimension.
- Matrix operations can be applied to analyze this data.

- **Graph Data**:
  - Represents relationships between objects.
  - Example: In a web graph, **nodes** are web pages, and **edges** (hyperlinks) connect them.

- **Ordered Data**:
  - Objects have attributes with an implicit order.
  - Examples of ordered data:
    - **Temporal data**: Attributes associated with time, e.g., customer purchase patterns during festivals.
    - **Sequence data**: Sequence of elements without timestamps, e.g., DNA sequences (A, T, G, C).
    - **Spatial data**: Related to locations or positions, e.g., maps where points relate to geographical locations.

2. Unstructured Data

Unstructured data does not have a predefined organizational format. This type of data includes multimedia (video, image, audio) as well as text documents, blogs, and social media data.

- Examples of unstructured data include:
  - Videos on platforms like YouTube.
  - Images and photos.
  - Audio recordings, such as podcasts or songs.
  - Text documents, blogs, and posts from social media.

Key Point: It is estimated that around **80% of all data** is unstructured, making it a large and significant portion of Big Data.

3. Semi-Structured Data

Semi-structured data falls between structured and unstructured data. While it does not conform to a strict structure, it contains tags or markers that make it easier to organize.

- Examples of semi-structured data include:

o **XML/JSON files**: Contain data with embedded tags or fields.

o **RSS feeds**: Often follow a hierarchical structure, but not as rigid as a database.

o **Hierarchical data**: Data that follows a parent-child relationship, like in a directory tree.

2.1.2 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. There are different approaches to organize and manage data in storage files and systems from flat file to data warehouses. Some of them are listed below:

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 System It normally consists of database files and a database management system (DBMS). Database files contain original data and metadata. DBMS aims to manage data and improve operator performance by including various tools like database administrator, query processing, and transaction manager. A relational database consists of sets of tables. The tables have rows and columns. The columns represent the attributes and rows represent tuples. A tuple corresponds to either an object or a relationship between objects. A user can access and manipulate the data in the database using SQL.

Different types of databases are listed below:

1 A 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, which may have links to other tables. Normally, transaction databases are created for performing associational analysis that indicates the correlation among the items.

2. Time-series database stores time related information like log files where data is associated with a time stamp. This data represents the sequences of data, which represent values or events obtained over a period (for example, hourly, weekly or yearly) or repeated time span. Observing sales of product continuously may yield a time-series data.
3. Spatial databases contain spatial information in a raster or vector format. Raster formats are either bitmaps or pixel maps. For example, images can be stored as a raster data. On the other hand, the vector format can be used to store maps as maps use basic geometric primitives like points, lines, polygons and so forth.

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.

2.2 BIG DATA ANALYTICS AND TYPES OF ANALYTICS

The primary aim of data analysis is to assist business organizations to take decisions. For example, a business organization may want to know which is the fastest selling product, in order for them to market activities. Data analysis is an activity that takes the data and generates useful information and insights for assisting the organizations.

Data analysis and data analytics are terms that are used interchangeably to refer to the same concept. However, there is a subtle difference. Data analytics is a general term and data analysis is a part of it. Data analytics refers to the process of data collection, preprocessing and analysis. It deals with the complete cycle of data management. Data analysis is just analysis and is a part of data analytics. It takes historical data and does the analysis.

Data analytics, instead, concentrates more on future and helps in prediction.

There are four types of data analytics:

1. Descriptive analytics

2. Diagnostic analytics
3. Predictive analytics
4. Prescriptive analytics

Descriptive Analytics

Descriptive analytics is about summarizing the main features of the data you've collected. It tells you **what has happened** by using historical data and statistical techniques. The goal is to organize, describe, and present this data in an understandable way. Think of it as a report that explains "what is" without drawing any conclusions about why it happened.

Example: Imagine a store that collects data on monthly sales. Descriptive analytics would summarize this data by calculating the average sales, total revenue, or the most popular product in a given month.

Key Point: It doesn't explain why the sales were high or low—just tells you what the data shows.

Diagnostic Analytics

Diagnostic analytics answers the question **"Why did this happen?"** It's about understanding the root cause of an event. By examining the data closely, we look for patterns, trends, and relationships that explain the cause of an outcome.

Example: If the store's sales drop one month, diagnostic analytics would investigate why the drop happened. Maybe it's due to bad weather, a competitor's sale, or a new product that didn't perform well. The analysis focuses on finding and explaining the reasons behind the drop.

Key Point: It's all about cause and effect—identifying the reasons behind the data patterns.

Predictive Analytics

Predictive analytics looks into the future and answers the question **"What will happen?"** Using historical data and advanced algorithms (like machine learning), it predicts future trends and outcomes.

Example: The store uses data from previous years to predict what the sales will be in the upcoming holiday season. Algorithms analyze patterns like past holiday sales, customer behavior, and current market trends to make predictions.

Key Point: It focuses on forecasting future events based on current and past data.

Prescriptive Analytics

Prescriptive analytics goes a step further and asks **"What should we do?"** It not only predicts the future but also recommends actions to take. This type of analytics provides decision-making support by suggesting the best course of action to achieve desired outcomes.

Example: After predicting that sales will be low in the next quarter, prescriptive analytics suggests specific actions the store can take, such as launching a promotion, adjusting prices, or stocking more popular products. This helps businesses make better decisions and minimize risks.

Key Point: It's all about decision-making—helping businesses choose the best possible actions based on data.

2.3.1 Data Collection

The first task of gathering datasets are the collection of data. It is often estimated that most of the time is spent for collection of good quality data. A good quality data yields a better result. It is often difficult to characterize a 'Good data'. 'Good data' is one that has the following properties:

1. Timeliness – The data should be relevant and not stale or obsolete data.
2. Relevancy – The data should be relevant and ready for the machine learning or data mining algorithms. All the necessary

information should be available and there should be no bias in the data.

3. Knowledge about the data – The data should be understandable and interpretable, and should be self-sufficient for the required application as desired by the domain knowledge engineer.

Broadly, the data source can be classified as open/public data, social media data and multimodal data.

1.**Open or public data source** – It is a data source that does not have any stringent copyright

rules or restrictions. Its data can be primarily used for many purposes. Government census data are good examples of open data:

• Digital libraries that have huge amount of text data as well as document images

• Scientific domains with a huge collection of experimental data like genomic data and biological data

• Healthcare systems that use extensive databases like patient databases, health insurance data, doctors' information, and bioinformatics information

2.**Social media** – It is the data that is generated by various social media platforms like Twitter,

Facebook, YouTube, and Instagram. An enormous amount of data is generated by these platforms.

3.**Multimodal data** – It includes data that involves many modes such as text, video, audio and mixed types. Some of them are listed below:

• Image archives contain larger image databases along with numeric and text data

• The World Wide Web (WWW) has huge amount of data that is distributed on the Internet.

2.3.2 Data Pre-processing

Data Cleaning is the process of detecting and correcting (or removing) errors and inconsistencies in data to improve its quality before applying machine learning or data mining techniques. In the real world, raw data is often **'dirty'**, meaning it contains errors, missing information, or inconsistencies that can affect the results of the analysis.

Common Problems with Dirty Data:

1. **Incomplete Data**: When certain values are missing from the dataset.

2. **Inaccurate Data**: Data that has incorrect values or errors.

3. **Outliers**: Data points that are significantly different from the rest of the data, often due to errors or unusual circumstances.

4. **Missing Values**: Data entries where certain attributes are not provided.

5. **Inconsistent Values**: Mismatched or incorrectly formatted data values.

6. **Duplicate Data**: When the same data appears multiple times, unnecessarily.

Example of Dirty Data:

Let's refer to the table of patient data (from the image) to explain common data issues:

Table 2.1: Illustration of 'Bad' Data

Patient ID	Name	Age	Date of Birth (DoB)	Fever	Salary
1.	John	21		Low	–1500
2.	Andre	36		High	Yes
3.	David	5	10/10/1980	Low	" "
4.	Raju	136		High	Yes

Identifying the Problems:

1. **Incomplete Data**:

o For patients John, Andre, and Raju, the **Date of Birth (DoB)** is missing. This is an example of **missing values**.

2. **Inaccurate Data**:

o David's **age is recorded as 5**, but his **DoB** is **10/10/1980**, which makes his real age much older than 5. This is **inconsistent data**.

o Raju's **age is recorded as 136**, which is not realistic. This might be a **typographical error** or an **outlier**.

3. **Outliers**:

o Raju's age of 136 is an **outlier**, as it is an unrealistic value when compared to normal human lifespans. Outliers are often caused by data entry errors.

4. **Noisy Data**:

o John's **salary** is recorded as **-1500**, which is not possible. Salary cannot be negative, making this an example of **noisy data**.

o The entry for David's **salary** is simply blank (**" "**), which is another instance of missing data.

5. **Inconsistent Values**:

o In the **salary** column, Andre and Raju both have **'Yes'** recorded, which doesn't make sense in the context of salary data. A salary should be a numeric value, not a text response.

How to Address These Issues?

1. **Missing Data**:

o **Ignore the Tuple**: If a lot of values are missing in a row, you may choose to ignore or remove that row from the dataset.

o **Fill Values Manually**: Domain experts can manually fill missing values, but this is time-consuming.

o **Use Global Constants**: Fill missing values with a placeholder like 'Unknown' or '0'.

o **Use Average/Mean Values**: Replace missing numeric values (like salary) with the average value of that column.

- **Prediction Techniques**: Machine learning algorithms can predict missing values based on patterns in other data (e.g., using decision trees).

2. **Inaccurate Data**:

- Correct entries by referring to other reliable data sources or consult domain experts. For example, David's age should be corrected based on his actual DoB.

3. **Handling Outliers**:

- Investigate outliers to determine if they are errors or legitimate data. Raju's age of 136 may be a typo and can be corrected if the correct age is known.

4. **Noisy Data**:

- Noisy data, like John's negative salary (-1500), can be corrected by setting a minimum limit (e.g., salary cannot be below 0). In this case, either correct or remove the invalid entry.

5. **Inconsistent Values**:

- Standardize the format for fields like salary. For Andre and Raju, change the text entries ('Yes') to numeric values or fill in missing data using estimation techniques.

Methods for Handling Missing Data:

1. **Ignoring the Tuple**: Discard rows with missing data (not ideal when a lot of data is missing).

2. **Filling Manually**: Domain experts analyze and fill the missing values.

3. **Global Constant**: Fill missing values with 'Unknown' or 'None'.

4. **Attribute Mean**: Replace missing numerical values with the average for that attribute.

5. **Class-based Mean**: Use the mean value of the same class or group to fill missing data.

6. **Most Probable Value**: Predict missing values using machine learning algorithms like decision trees.

Removal of Noisy or Outlier Data

In data analysis, **noise** refers to random errors or variations in the data that can distort the results of analysis. Noise can affect data accuracy and, if not removed, can lead to misleading conclusions. Therefore, it's important to clean noisy data before applying any analysis or machine learning algorithms.

What is Noise?

- Noise is random error or variance in measured values.
- It can appear as outliers, missing values, or inconsistent data.
- Noise reduction is an essential step in **data cleaning** to improve the quality of analysis.

Techniques for Removing Noise:

One common method to remove noisy data is **binning**, which organizes data into groups (bins) and then applies smoothing techniques to remove noise. Binning methods can also be used for **data discretization**, which reduces the number of values for easier analysis.

Binning Method:

- **Step 1**: Sort the data in increasing order.
- **Step 2**: Divide the sorted data into equal-frequency bins (also called **buckets**).
- **Step 3**: Apply smoothing techniques within each bin to reduce noise.

Smoothing Techniques for Binning:

1. **Smoothing by Means**:

- Replace all values in the bin with the mean (average) of the bin values.

Example:

- Given data: **S = {12, 14, 19, 22, 24, 26, 28, 31, 34}**
- First, divide into bins of size 3:
  - **Bin 1**: {12, 14, 19}
  - **Bin 2**: {22, 24, 26}
  - **Bin 3**: {28, 31, 34}
- Now apply smoothing by means (replace all values with the bin's mean):
  - **Bin 1** (mean = 15): {15, 15, 15}
  - **Bin 2** (mean = 24): {24, 24, 24}
  - **Bin 3** (mean ≈ 31): {31, 31, 31}
- **Explanation**: Each value in the bin is replaced by the mean of the bin to smooth the data.

2. **Smoothing by Medians**:

- Replace all values in the bin with the **median** of the bin values (the middle value when the data is sorted).

Example:

- Given the same data and bins:
  - **Bin 1** (median = 14): {14, 14, 14}
  - **Bin 2** (median = 24): {24, 24, 24}
  - **Bin 3** (median = 31): {31, 31, 31}
- **Explanation**: Each value in the bin is replaced by the median, which reduces the effect of outliers or extreme values.

3. **Smoothing by Bin Boundaries**:

- Replace each value in the bin with the closest **boundary value** (minimum or maximum value in the bin).

Example:

- Given the same data and bins:
  - **Bin 1** (boundary values: 12 and 19): {12, 12, 19}
  - **Bin 2** (boundary values: 22 and 26): {22, 22, 26}
  - **Bin 3** (boundary values: 28 and 34): {28, 34, 34}
- **Explanation**: For each bin, values are replaced by the closest boundary value (either the minimum or maximum of that bin).
- **Example**: In Bin 1, the original data was {12, 14, 19}. The boundaries are 12 and 19, so the value 14 is closer to 12, and it's replaced by 12.

Why Use Binning to Remove Noise?

- **Smoothing by Means**: Reduces random noise by averaging the values within each bin.
- **Smoothing by Medians**: More robust against outliers than using means since medians are less sensitive to extreme values.
- **Smoothing by Bin Boundaries**: Eliminates noise by forcing all values within the bin to adhere to the boundaries, creating a more consistent dataset.

Data Integration and Data Transformations

Data integration involves routines that merge data from multiple sources into a single data source.

So, this may lead to redundant data. The main goal of data integration is to detect and remove redundancies that arise from integration. Data transformation routines perform operations like normalization to improve the performance of the data mining algorithms. It is necessary to transform data so that it can be processed. This can be considered as a preliminary stage of data conditioning. Normalization is one such technique. In normalization, the attribute values are scaled

to fit in a range (say 0-1) to improve the performance of the data mining algorithm. Often, in neural networks, these techniques are used. Some of the normalization procedures used are:

1. Min-Max
2. z-Score

Min-Max Procedure It is a normalization technique where each variable V is normalized by its difference with the minimum value divided by the range to a new range, say 0–1. Often, neural networks require this kind of normalization. The formula to implement this normalization is given as:

$$min\text{-}max = \frac{V - min}{max - min} \times (new\ max - new\ min) + new\ min \qquad (2.1)$$

Here max-min is the range. Min and max are the minimum and maximum of the given data, new max and new min are the minimum and maximum of the target range, say 0 and 1.

Example 2.2: Consider the set: V = {88, 90, 92, 94}. Apply Min-Max procedure and map the marks to a new range 0–1.

Solution: The minimum of the list V is 88 and maximum is 94. The new min and new max are 0 and 1, respectively. The mapping can be done using Eq. (2.1) as:

For marks 88,

$$min\text{-}max = \frac{88-88}{94-88} \times (1-0) + 0 = 0$$

Similarly, other marks can be computed as follows:

For marks 90,

$$min\text{-}max = \frac{90-88}{94-88} \times (1-0) + 0 = 0.33$$

For marks 92,

$$min\text{-}max = \frac{92-88}{94-88} \times (1-0) + 0 = \frac{4}{6} = 0.66$$

For marks 94,

$$min\text{-}max = \frac{94-88}{94-88} \times (1-0) + 0 = \frac{6}{6} = 1$$

So, it can be observed that the marks {88, 90, 92, 94} are mapped to the new range {0, 0.33, 0.66, 1}. Thus, the Min-Max normalization range is between 0 and 1.

z-Score Normalization This procedure works by taking the difference between the field value

and mean value, and by scaling this difference by standard deviation of the attribute.

$$V* = V - \mu/\sigma \qquad (2.2)$$

Here, s is the standard deviation of the list V and m is the mean of the list V.

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

Solution: The mean and Sample Standard deviation (s) 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.

Data Reduction

Data reduction reduces data size but produces the same results. There are different ways in which data reduction can be carried out such as data aggregation, feature selection, and dimensionality reduction.

2.4 DESCRIPTIVE STATISTICS

Descriptive statistics is a branch of statistics that does dataset summarization. It is used to summarize and describe data. Descriptive statistics are just descriptive and do not go beyond that.

In other words, descriptive statistics do not bother too much about machine learning algorithms and its functioning.

Let us discuss descriptive statistics with the fundamental concepts of datatypes.

Dataset and Data Types

A dataset can be assumed to be a collection of data objects. The data objects may be records, points, vectors, patterns, events, cases, samples or observations. These records contain many attributes. An attribute can be defined as the property or characteristics of an object.

For example, consider the following database shown in sample Table 2.2.

Table 2.2: Sample Patient Table

Patient ID	Name	Age	Blood Test	Fever	Disease
1.	John	21	Negative	Low	No
2.	Andre	36	Positive	High	Yes

Every attribute should be associated with a value. This process is called measurement. The type of attribute determines the data types, often referred to as measurement scale types.

The data types are shown in Figure 2.1.

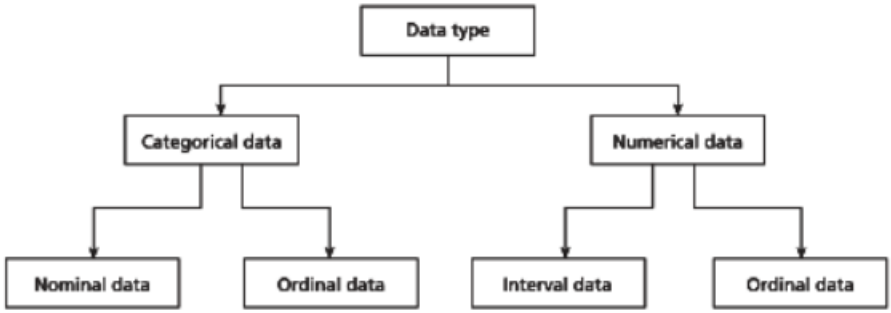


Figure 2.1: Types of Data

Broadly, data can be classified into two types:

1. Categorical or qualitative data

2. Numerical or quantitative data

Categorical or Qualitative Data The categorical data can be divided into two types. They are nominal type and ordinal type.

•Nominal Data – In Table 2.2, patient ID is nominal data. Nominal data are symbols and cannot be processed like a number. For example, the average of a patient ID does not make any statistical sense. Nominal data type provides only information but has no ordering among data. Only operations like (=, ≠) are meaningful for these data. For example, the patient ID can be checked for equality and nothing else.

•Ordinal Data – It provides enough information and has natural order. For example, Fever = {Low, Medium, High} is an ordinal data. Certainly, low is less than medium and medium is less than high, irrespective of the value. Any transformation can be applied to these data to get a new value.

Numeric or Qualitative Data It can be divided into two categories. They are interval type and ratio type.

•Interval Data – Interval data is a numeric data for which the differences between values are meaningful. For example, there is a difference between 30 degrees and 40 degrees. Only the permissible operations are + and -.

•Ratio Data – For ratio data, both differences and ratio are meaningful. The difference between the ratio and interval data is the position of zero in the scale. For example, take the Centigrade-Fahrenheit conversion. The zeroes of both scales do not match.

Hence, these are interval data.

Another way of classifying the data is to classify it as:

1.Discrete value data

2.Continuous data

Discrete Data This kind of data is recorded as integers. For example, the responses of the survey can be discrete data. Employee identification number such as 10001 is discrete data.

Continuous Data It can be fitted into a range and includes decimal point. For example, age is a continuous data. Though age appears to be discrete data, one may be 12.5 years old and it makes sense. Patient height and weight are all continuous data.

Third way of classifying the data is based on the number of variables used in the dataset. Based on that, the data can be classified as univariate data, bivariate data, and multivariate data. This is shown in Figure 2.2.

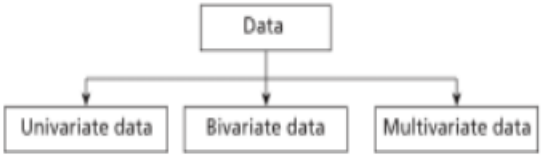


Figure 2.2: Types of Data Based on Variables

2.5 UNIVARIATE DATA ANALYSIS AND VISUALIZATION

Univariate analysis is the simplest form of statistical analysis. As the name indicates, the dataset has only one variable. A variable can be called as a category. Univariate does not deal with cause or relationships. The aim of univariate analysis is to describe data and find patterns. Univariate data description involves finding the frequency distributions, central tendency measures, dispersion or variation, and shape of the data.

2.5.1 Data Visualization

Let us consider some forms of graphs

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.

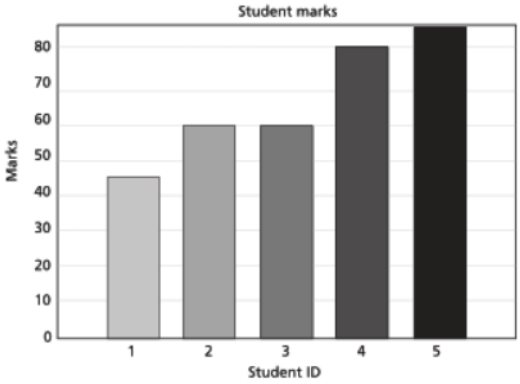


Figure 2.3: Bar Chart

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.

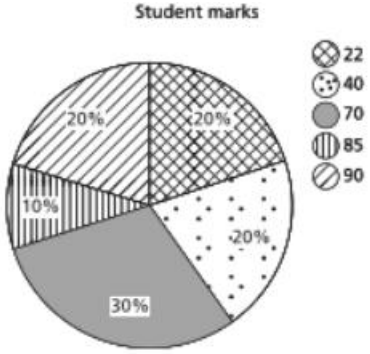


Figure 2.4: Pie Chart

It can be observed that the number of students with 22 marks are 2. The total number of students are 10. So, 2/10 × 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.

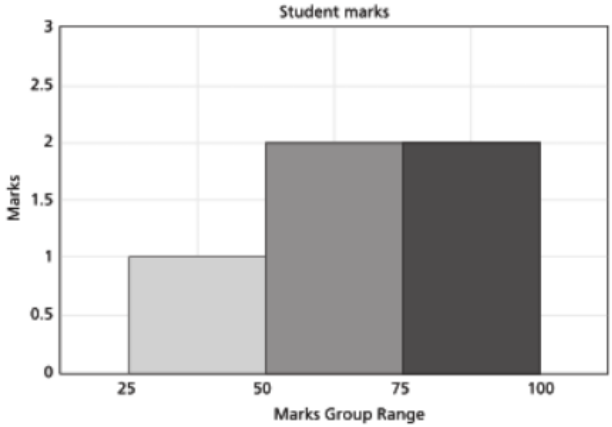


Figure 2.5: Sample Histogram of English Marks

Histogram conveys useful information like nature of data and its mode. Mode indicates the peak of dataset. In other words, histograms can be used as charts to show frequency, skewness present in the data, and shape.

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.

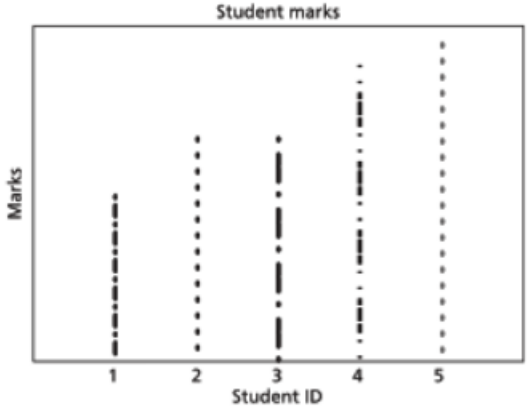


Figure 2.6: Dot Plots

2.5.2 Central Tendency

Therefore, a condensation or summary of the data is necessary. This makes the data analysis easy and simple. One such summary is called central tendency. Thus, central tendency can explain the characteristics of data and that further helps in comparison. Mass data have tendency to concentrate at certain values, normally in the central location. It is called measure of central tendency (or averages). Popular measures are mean, median and mode.

1. Mean – Arithmetic average (or mean) is a measure of central tendency that represents the 'center' of the dataset. Mathematically, the average of all the values in the sample (population) is denoted as x. Let x1, x2, … , xN be a set of 'N' values or observations, then the arithmetic mean is given as:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (2.3)$$

For example, the mean of the three numbers 10, 20, and 30 is 20

•**Weighted mean** – Unlike arithmetic mean that gives the weightage of all items equally, weighted mean gives different importance to all items as the item importance varies.

Hence, different weightage can be given to items. In case of frequency distribution, mid values of the range are taken for computation. This is illustrated in the following computation. In weighted mean, the mean is computed by adding the product of proportion and group mean. It is mostly used when the sample sizes are unequal.

•**Geometric mean** – Let x1, x2, … , xN be a set of 'N' values or observations. Geometric mean

is the Nth root of the product of N items. The formula for computing geometric mean is given as follows:

$$\text{Geometric mean} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{N}} = \sqrt[N]{x_1 \times x_2 \times \cdots \times x_N} \qquad (2.4)$$

Here, n is the number of items and xi are values. For example, if the values are 6 and 8, the geometric mean is given as In larger cases, computing geometric mean is difficult. Hence, it is usually calculated as:

$$\text{Anti-log of } \frac{\log(x_1) + \log(x_2) + \cdots + \log(x_N)}{N} \qquad (2.5)$$

$$= \text{anti-log } \frac{\sum_{i=1}^{n} \log(x_i)}{N} \qquad (2.6)$$

The problem of mean is its extreme sensitiveness 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.

2. 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. A median class is that class where (N/2)th item is present.

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

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

Median class is that class where N/2th item is present. Here, i is the class interval of the median class and L1 is the lower limit of median class, f is the frequency of the median class, and cf is the cumulative frequency of all classes preceding median.

3. Mode – Mode is the value that occurs more frequently in the dataset. In other words, the value that has the highest frequency is called mode.

2.5.3 Dispersion

The spreadout of a set of data around the central tendency (mean, median or mode) is called dispersion. Dispersion is represented by various ways such as range, variance, standard deviation, and standard error. These are second order measures. The most common measures of the dispersion data are listed below:

Range Range is the difference between the maximum and minimum of values of the given list of data.

Standard Deviation The mean does not convey much more than a middle point. For example, the following datasets {10, 20, 30} and {10, 50, 0} both have a mean of 20. The difference between these two sets is the spread of data. Standard deviation is the average distance from the mean of the dataset to each point.

The formula for sample standard deviation is given by:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}} \qquad (2.8)$$

Here, N is the size of the population, xi is observation or value from the population and m is the population mean. Often, N – 1 is used instead of N in the denominator of Eq. (2.8).

Quartiles and Inter Quartile Range It is sometimes convenient to subdivide the dataset using coordinates. Percentiles are about data that are less than the coordinates by some percentage of the total value. kth percentile is the property that the k% of the data lies at or below Xi. For example, median is 50th percentile and can be denoted as Q0.50. The 25th percentile is called first quartile (Q1) and the 75th percentile is called third quartile (Q3). Another measure that is useful to measure dispersion is Inter Quartile Range (IQR). The IQR is the difference between Q3 and Q1.

Interquartile percentile = Q3 – Q1 (2.9)

Outliers are normally the values falling apart at least by the amount 1.5 × IQR above the third quartile or below the first quartile.

Interquartile is defined by Q0.75 – Q0.25. (2.10)

Example 2.4: For patients' age list {12, 14, 19, 22, 24, 26, 28, 31, 34}, find the IQR.

Solution: The median is in the fifth position. In this case, 24 is the median. The first quartile is median of the scores below the mean i.e., {12, 14, 19, 22}. Hence, it's the median of the list below 24. In this case, the median is the average of the second and third values, that is, Q0.25 = 16.5. Similarly, the third quartile is the median of the values above the median, that is {26, 28, 31, 34}. So, Q0.75 is the average of the seventh and eighth score. In this case, it is 28 + 31/2 = 59/2 = 29.5.

Hence, the IQR using Eq. (2.10) is:

= Q0.75 – Q0.25

= 29.5-16.5 = 13

Five-point Summary and Box Plots The median, quartiles Q1 and Q3, and minimum and maximum written in the order < Minimum, Q1, Median, Q3, Maximum > is known as five-point summary. Example 2.5: Find the 5-point summary of the list {13, 11, 2, 3, 4, 8, 9}.

Solution: The minimum is 2 and the maximum is 13. The Q1, Q2 and Q3 are 3, 8 and 11, respectively. Hence, 5-point summary is {2, 3, 8, 11,

13}, that is, {minimum, Q1, median, Q3, maximum}. Box plots are useful for describing 5-point summary. The Box plot for the set is given in Figure 2.7.

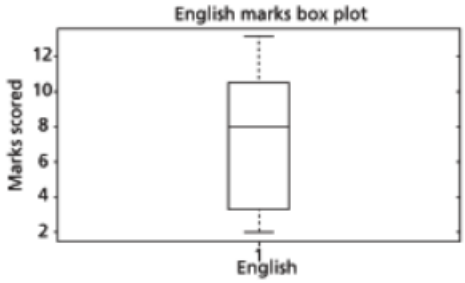


Figure 2.7: Box Plot for English Marks

2.5.4 Shape

Skewness and Kurtosis (called moments) indicate the symmetry/asymmetry and peak location of the dataset.

Skewness

The measures of direction and degree of symmetry are called measures of third order. Ideally, skewness should be zero as in ideal normal distribution. More often, the given dataset may not have perfect symmetry (consider the following Figure 2.8).

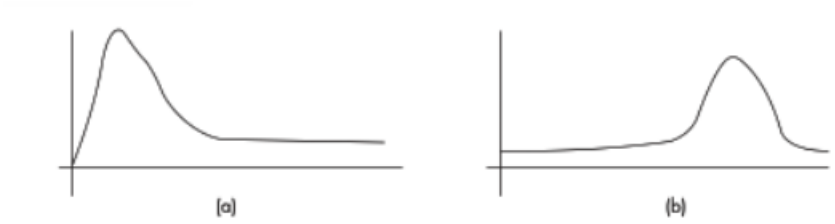


Figure 2.8: (a) Positive Skewed and (b) Negative Skewed Data

Generally, for negatively skewed distribution, the median is more than the mean. The relationship between skew and the relative size of the mean and median can be summarized by a convenient numerical skew

index known as Pearson 2 skewness coefficient.

$$\frac{3 \times (\mu - median)}{\sigma} \quad (2.12)$$

Also, the following measure is more commonly used to measure skewness. Let X1, X2, …, XN be a set of 'N' values or observations then the skewness can be given as:

$$\frac{1}{N} \times \sum_{i=1}^{N} \frac{(x_i - \mu)^3}{\sigma^3} \quad (2.13)$$

Here, m is the population mean and s is the population standard deviation of the univariate data. Sometimes, for bias correction instead of N, N - 1 is used.

Kurtosis

Kurtosis also indicates the peaks of data. If the data is high peak, then it indicates higher kurtosis and vice versa. Kurtosis is measured using the formula given below:

$$\frac{\sum_{i=1}^{N} (x_i - \bar{x})^4 / N}{\sigma^4} \quad (2.14)$$

It can be observed that N - 1 is used instead of N in the numerator of Eq. (2.14) for bias correction. Here, x and s are the mean and standard deviation of the univariate data, respectively.

Some of the other useful measures for finding the shape of the univariate dataset are mean absolute deviation (MAD) and coefficient of variation (CV).

Mean Absolute Deviation (MAD)

MAD is another dispersion measure and is robust to outliers. Normally, the outlier point is detected by computing the deviation from median and by dividing it by MAD. Here, the absolute deviation between the data and mean is taken. Thus, the absolute deviation is

given as:

$$|x - \mu| \quad (2.15)$$

The sum of the absolute deviations is given as $\Sigma|x - \mu|$

Therefore, the mean absolute deviation is given as: $\frac{\Sigma|x - \mu|}{N}$ (2.16)

Coefficient of Variation (CV)

Coefficient of variation is used to compare datasets with different units. CV is the ratio of standard deviation and mean, and %CV is the percentage of coefficient of variations.

2.5.5 Special Univariate Plots

The ideal way to check the shape of the dataset is a stem and leaf plot. A stem and leaf plot are a display that help us to know the shape and distribution of the data. In this method, each value is

split into a 'stem' and a 'leaf'. The last digit is usually the leaf and digits to the left of the leaf mostly form the stem. For example, marks 45 are divided into stem 4 and leaf 5 in Figure 2.9. The stem and leaf plot for the English subject marks, say, {45, 60, 60, 80, 85} is given in Figure 2.9.

Stem	Leaf
4	5
5	
6	0 0
7	
8	0 5

Figure 2.9: Stem and Leaf Plot for English Marks

It can be seen from Figure 2.9 that the first column is stem and the second column is leaf. For the given English marks, two students with 60 marks are shown in stem and leaf plot as stem-6 with 2 leaves with 0. The normal Q-Q plot for marks x = [13 11 2 3 4 8 9] is given below n Figure 2.10.

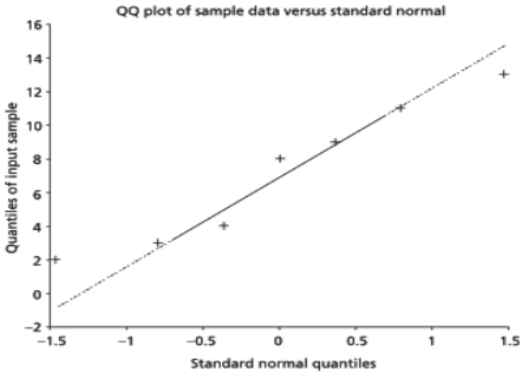


Figure 2.10: Normal *Q-Q* Plot