T.Y.B.C.A. SCIENCE : SEMESTER-V

NEW SYLLABUS
CBCS PATTERN

DATA MINING AND DATA SCIENCE

Dr. DIPALI MEHER
Dr. PALLAWI BULAKH

NIRALI PRAKASHAN
ADVANCEMENT OF KNOWLEDGE

SPPU New Syllabus

A Book Of

Data Mining and Data Science

For B.C.A. (Science) : Semester - V

[Course Code BCA - 352 : Credits-04]

CBCS Pattern

As Per New Syllabus, Effective from June 2021

Dr. Dipali Meher

MCS, M.Phil, NET, Ph.D.
Assistant Professor
Department of Computer Science,
Modern College of Arts, Science and Commerce,
Ganeshkhind, Pune 16

Dr. Pallawi Bulakh

M.Sc. M.Phil., Ph.D., NET with JRF
Assistant Professor
Department of Computer Science,
Modern College of Arts, Science and Commerce,
Ganeshkhind, Pune-16

Member, BOS, Computer Application,
Savitribai Phule Pune University

Price 180.00

NIRALI PRAKASHAN
ADVANCEMENT OF KNOWLEDGE

N5872

DATA MINING AND DATA SCIENCE **ISBN 978-93-5451-258-2**

First Edition : August 2021
© : Authors

The text of this publication, or any part thereof, should not be reproduced or transmitted in any form or stored in any computer storage system or device for distribution including photocopy, recording, taping or information retrieval system or reproduced on any disc, tape, perforated media or other information storage device etc., without the written permission of Authors with whom the rights are reserved. Breach of this condition is liable for legal action.

Every effort has been made to avoid errors or omissions in this publication. In spite of this, errors may have crept in. Any mistake, error or discrepancy so noted and shall be brought to our notice shall be taken care of in the next edition. It is notified that neither the publisher nor the authors or seller shall be responsible for any damage or loss of action to any one, of any kind, in any manner, therefrom. The reader must cross check all the facts and contents with original Government notification or publications.

Published By :
NIRALI PRAKASHAN
Abhyudaya Pragati, 1312, Shivaji Nagar,
Off J.M. Road, Pune – 411005
Tel - (020) 25512336/37/39
Email : niralipune@pragationline.com

Polyplate

Printed By :
YOGIRAJ PRINTERS AND BINDERS
Survey No. 10/1A, Ghule Industrial Estate
Nanded Gaon Road
Nanded, Pune - 411041

DISTRIBUTION CENTRES

PUNE

Nirali Prakashan
(For orders outside Pune)
S. No. 28/27, Dhayari Narhe Road, Near Asian College
Pune 411041, Maharashtra
Tel : (020) 24690204; Mobile : 9657703143
Email : bookorder@pragationline.com

Nirali Prakashan
(For orders within Pune)
119, Budhwar Peth, Jogeshwari Mandir Lane
Pune 411002, Maharashtra
Tel : (020) 2445 2044; Mobile : 9657703145
Email : niralilocal@pragationline.com

MUMBAI

Nirali Prakashan
Rasdhara Co-op. Hsg. Society Ltd., 'D' Wing Ground Floor, 385 S.V.P. Road
Girgaum, Mumbai 400004, Maharashtra
Mobile : 7045821020, Tel : (022) 2385 6339 / 2386 9976
Email : niralimumbai@pragationline.com

DISTRIBUTION BRANCHES

DELHI

Nirali Prakashan
Room No. 2 Ground Floor
4575/15 Omkar Tower, Agarwal
Road Darya Ganj, New Delhi
110002
Mobile : 9555778814/9818561840
Email : delhi@niralibooks.com

BENGALURU

Nirali Prakashan
Maitri Ground Floor, Jaya
Apartments, No. 99, 6th Cross, 6th
Main, Malleswaram, Bengaluru
560003 Karnataka; Mob :
9686821074
Email : bengaluru@niralibooks.com

NAGPUR

Nirali Prakashan
Above Maratha Mandir, Shop No. 3,
First Floor, Rani Jhanshi Square,
Sitabuldi Nagpur 440012 (MAH)
Tel : (0712) 254 7129
Email : nagpur@niralibooks.com

KOLHAPUR

Nirali Prakashan
New Mahadvar Road, Kedar Plaza,
1st Floor Opp. IDBI Bank
Kolhapur 416 012 Maharashtra
Mob : 9850046155
Email : kolhapur@niralibooks.com

JALGAON

Nirali Prakashan
34, V. V. Golani Market, Navi Peth,
Jalgaon 425001, Maharashtra
Tel : (0257) 222 0395
Mob : 94234 91860
Email : jalgaon@niralibooks.com

SOLAPUR

Nirali Prakashan
R-158/2, Avanti Nagar, Near Golden
Gate, Pune Naka Chowk
Solapur 413001, Maharashtra
Mobile 9890918687
Email : solapur@niralibooks.com

marketing@pragationline.com | www.pragationline.com
Also find us on www.facebook.com/niralibooks

Preface ...

We take an opportunity to present this book entitled as **"Data Mining and Data Science"** to the students of **B.C.A. (Science), Semester - V** as per the New Syllabus, (CBCS Pattern 2019).

The book covers theory of Introduction to Data Mining, Data Warehousing, Classification, Clustering and Association Rule Mining, Introduction to Data Science, EDA and Data Visualization.

A special words of thank to Shri. Dineshbhai Furia, Mr. Jignesh Furia for showing full faith in us to write this text book.

We also thank Mr. Ilyas, Ms. Chaitali Takale, Mr. Ravindra Walodare, Mr. Sachin Shinde, Mr. Ashok Bodke, Mr. Moshin Sayyed and Mr. Nitin Thorat of M/s Nirali Prakashan for their excellent co-operation.

Although every care has been taken to check mistakes and misprints, any errors, omission and suggestions from teachers and students for the improvement of this text book shall be most welcome.

Authors

✍ ✍ ✍

Syllabus ...

Unit I Introduction to Data Mining **(10 Hrs)**

- 1.1 Definition Data mining
- 1.2 Data Mining issues
- 1.3 Stages of the Data Mining Process (KDD)
- 1.4 Data Mining Techniques/Tasks
- 1.5 Knowledge Representation Methods
- 1.6 Applications of Data mining
- 1.7 Data Pre-processing
  - 1.7.1 Data Cleaning
  - 1.7.2 Data Transformation
  - 1.7.3 Data Reduction
  - 1.7.4 Data Discretization

Unit II Data Warehousing **(08 Hrs)**

- 2.1 Introduction to Data Warehouse
- 2.2 Data Warehouse Architecture and its components
- 2.3 Data Modeling with OLAP
  - 2.3.1 Introduction
  - 2.3.2 Difference between OLTP and OLAP
  - 2.3.3 Data Mart
  - 2.3.4 Fact Table, Dimension Table, OLAP cube
  - 2.3.5 Different OLAP Operations
- 2.4 Schema Design
  - 2.4.1 Introduction
  - 2.4.2 Star and snow-Flake Schema
- 2.5 Introduction to Machine Learning
- 2.6 Introduction to Pattern Matching
- 2.7 Case study based on Schema Design

Unit III Classification **(12 Hrs)**

- 3.1 Introduction, Definition
- 3.2 Decision Tree
  - 3.2.1 Introduction
  - 3.2.2 Construction Principle
  - 3.2.3 Attribute Selection Measures
  - 3.2.4 Tree Pruning
- 3.3 Rule-Based Classification
  - 3.3.1 Using IF-THEN Rules for Classification
  - 3.3.2 Rule Extraction from a Decision Tree
- 3.4 Bayes Classification Methods
  - 3.4.1 Bayes' Theorem
  - 3.4.2 Naive Bayesian Classification
- 3.5 Bayesian Networks
- 3.6 Parameter and structure learning
- 3.7 Linear classifier

3.8 Perceptron
3.9 k-Nearest-Neighbor Classifiers
3.10 SVM classifiers
  3.10.1 Introduction
  3.10.2 Types of SVM
  3.10.3 Working of SVM
3.11 Regression
  3.11.1 Linear Regression
  3.11.2 Non linear Regression
3.12 Introduction to Prediction

Unit IV Clustering and Association Rule Mining **(10 Hrs)**

4.1 Cluster Analysis
  4.1.1 Introduction
  4.1.2 Requirements for Cluster Analysis
4.2 Hierarchical Methods
  4.2.1 Agglomerative Hierarchical Clustering
  4.2.2 Divisive Hierarchical Clustering
4.3 Partitioning Methods
  4.3.1 k-Means: A Centroid-Based Technique
  4.3.2 k-Medoids: A Representative Object-Based Technique
4.4 Introduction to Association Rule Mining Market Basket Analysis, Items, Itemsets and Large Itemsets
4.5 Apriori Algorithm
4.6 Kinds of Association Rules Mining
Multilevel association rules
Constraint Based Association rules mining

Unit V Introduction to Data Science **(10 Hrs)**

5.1 Basics of Data
5.2 What is Data Science?
5.3 Data science process
5.4 Stages in a Data Science project
5.5 Applications of Data Science in various fields
5.6 Basics of Data Analytics
5.7 Types of Analytics – Descriptive, Predictive, Prescriptive
5.8 Statistical Inference - Populations and samples - Statistical modeling - Probability Distributions

Unit VI EDA and Data Visualization **(10 Hrs)**

6.1 What is Exploratory Data Analysis?
6.2 Steps in EDA
6.3 Basic tools (plots, graphs and summary statistics) of EDA
6.4 Types of exploratory data analysis
6.5 Basic principles of data visualization
6.6 Benefits of Data Visualization
6.7 Data visualization techniques
6.8 Tools for data visualization

Contents ...

Bibliography

Books:

- Dunhan, M. (2003). Data Mining: Introductory and Advanced Topics. Prentice Hall. Engineering
- Agarwal, S.(2014). Data mining: Data mining concepts and techniques. In Proceedings – 2013. International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013. https://doi.org/10.1109/ICMIRA.2013.45
- Introduction to Data Preprocessing in Data Mining | by Sajeevan Wickramarathna | Tech x Talent | | Jun, 2021 | Medium

Websites:

- https://www.tutorialandexample.com/regression-in-data-mining/
- https://en.wikipedia.org/wiki/Time_series
- https://www.jigsawacademy.com
- http://deepai.org/
- https://www.javatpoint.com
- https://www. guru99.com
- https://www .dezyre.com
- https://www .OmniSci.com
- https://www .mygreatlearning.com
- https://www.researchgate.net/figure/The-fundamental-steps-of-the-exploratory-data-analysis-process_fig3_329930775
- http://www.csun.edu/edaasic/roosta/ECE595_Chap1.pdf
- https://www.oreilly.com/library/view/practical-statistics-for/9781491952955/ch01.html
- https://www.itl.nist.gov/div898/handbook/toolaids/pff/eda.pdf
- https://www.hcbravo.org/IntroDataSci/bookdown-notes/exploratory-data-analysis-summary-statistics.html

✍ ✍ ✍

1... Introduction to Data Mining

Learning Objectives...

- To understand the concept and issues of Data Mining.
- To study the KDD process in detail.
- To know the Data Mining techniques and knowledge representation methods.
- To understand the real world applications of Data Mining.
- To get information of Data Pre-processing.

1.1 INTRODUCTION

- Today, we live in a world where millions and trillions of data is generated daily. With the evolution of WWW (World Wide Web), there has been tremendous growth in information content, information processing and information retrieval. Information retrieval is also a major need as it is the basis for future prediction, analysis and decision making.
- Mining refers to the extraction of valuable things. Data mining, in turn, refers to the study of collecting, cleaning, processing and analysing the data and to retrieve meaningful information from huge data.

1.2 DEFINITION OF DATA MINING

- It is analysis of data and use of software techniques and statistical methods to find patterns in data.
- Data Mining deals with discovery of hidden knowledge, unexpected patterns and new rules from large data sets.
- Data mining is the use of algorithms to extract the information and patterns derived by the KDD (Knowledge Discovery in Databases) process.
- It is also known as the process of extracting hidden information from a large dataset. i.e. mining knowledge from data.
- Difference between Data Mining and Query Processing is as follows:

Table 1.1: Difference between Data Mining and Query Processing

Data Mining	Query Processing
Data Mining deals with the discovery of hidden Knowledge, unexpected pattern and new rules from large data sets.	Queries are the primary mechanism for retrieving information from a database and consist of questions presented to the database in a predefined format.
Data mining is analysis of data.	Query gives answer of a question fired to database.
Data mining uses algorithms for analysis of data.	In normal query no any algorithm is used.
In data mining, what is looking for is usually to be found.	In normal query, what is looking for exists in tables in database
Data mining uses data from various heterogeneous data sources.	Normal query can be executed on operational or traditional databases.
Data from various sources need to be integrated, pre-processed before data mining can be done.	In normal query data need not be integrated and pre-processed.
Query in data mining does not have any format.	Query is written in specific syntax such as, select *attributes* from *table names* where *conditions*
Output is analyzed data and presented as pattern or knowledge.	Output is subset of database.
For representation of output of data mining different visualization techniques such as graphical, icon based, hierarchical, geometric, pixel based and hybrid are used.	Output is displayed in terms of columns or sometime in text.
In data mining we are drowning in data, but starving for knowledge.	In normal query we are only drowning for data.
Data mining is actually one of the steps from KDD (Knowledge Discovery in Databases) process.	Normal query is used in data mining for extracting patterns.
There is no any fix languages used for data mining.	There are different languages used for writing query such as SQL, PL/SQL, MySQL
Different tools are used for data mining.	There is no tool used for normal query.
Data mining used in all databases.	Normal query used in relational databases.
Example 1: Develop a general profile of credit card customers. Differentiate poor credit risk customers from good credit card customers.	**Example 1:** List customers who use credit card to purchase more than ` 1000 worth groceries.
Example 2: Determine patients whose lifestyle is prone to getting a heart attack in near future.	**Example 2:** List patients who had at least one heart attack.

1.3 DATA MINING ISSUES

- Following issues were faced while doing data mining. These issues should be always considered by all data miners and data mining algorithms before going for data mining.

1. **Human interaction:** When a data mining task is to be undertaken, the goal is not clear. Users as well as the technical expert are unaware of the results. There is a need for a proper interface between the domain expert and users. The queries are formed by the experts based on the user's demand.
2. **Overfitting:** Overfitting is a statistical error. When a model is generated for a particular data set, it is supposed that the same model should accommodate future data sets as well. But overfitting occurs when the generated model is well suited for the training data set and it is not suited for the test data set or future data set. Overfitting can be reduced by increasing the training data set and by reducing model complexity.
3. **Outliers:** When a model is derived, there are some values of data that do not fit in the model. These values significantly different from the normal values, or they do not fit in any cluster. These values are called outliers. They can also be called as exceptions in the model derived. If a model includes outliers, it may not behave well for other significant values.
4. **Interpretation of the results:** Interpretation of the results obtained by data mining is a very crucial task. This interpretation is beyond only explanation of the results. This task requires expert analysis and interpretation. Hence, interpretation of the results is an issue in data mining.
5. **Visualization of the results:** Visualization of the results is useful to understand and quickly view the output of the different database algorithms.
6. **Large data sets:** Data mining models are generally designed to test the small data sets. But, when these models are applied to large data sets i.e. datasets with larger size then these models either fail or they wobble. There are many such models that work very well for the normal data sets but are inefficient in handling large data sets. The large data set issue can be handled with sampling and parallelization.
7. **High Dimensionality:** Dimensionality of the database refers to the different attributes that are present in the database. High dimensionality in a database leads to more number of attributes leading to confusion of choosing the attributes for the particular task. An increase in the number of attributes increases the complexity and efficacy of the algorithm. The solution to High Dimensionality is to reduce the number of attributes.
8. **Multimedia Data:** Many users demand the mining tasks for graphical, video or audio data. The multimedia data can be an issue in data mining as traditionally data mining tasks are designed for numeric or alphanumeric data.
9. **Missing data:** Sometimes the data is incomplete or missing. During the KDD process, this data may be filled with nearest estimates. These estimates may give false or invalid results creating problems.
10. **Irrelevant data:** Some data used in the mining task may not be relevant to the actual mining tasks. This data may either lead to invalid results or may vary the results.

11. **Noisy data:** The data which has no meaning is called noisy data. These values need to be corrected or replaced with meaningful data.
12. **Changing data:** The databases used for the mining tasks are subject to change. The algorithms run on the database at a particular time may show different results if the database is dynamic. This issue can be solved by running the data mining algorithm of the changed database completely.
13. **Integration:** The normal database querying results in the output as a traditional data processing system whereas KDD process gives the results which are unknown to the user. There is no union in KDD process and the traditional Query processing. The integration of these two will certainly give more profitable results.
14. **Application:** The output of the Data mining process results in unknown facts about the data. It is a big challenge to use this data for the purpose of decision making. This task is more crucial than developing or applying algorithms to database. Proper application or use of mining tasks will produce better results.

1.4 STAGES OF THE DATA MINING PROCESS (KDD)

- The KDD process (Knowledge discovery from database process) is the process of digging or finding the truth laid in databases which is not known yet and the things which are previously not discovered. The patterns, the stuff that has not been detected yet can be found with the KDD process. This extraction of unknown stuff through the KDD process is useful in automating summarization, pattern recognition and finding out the truth from facts and figures.
- The KDD process is divided into following steps:

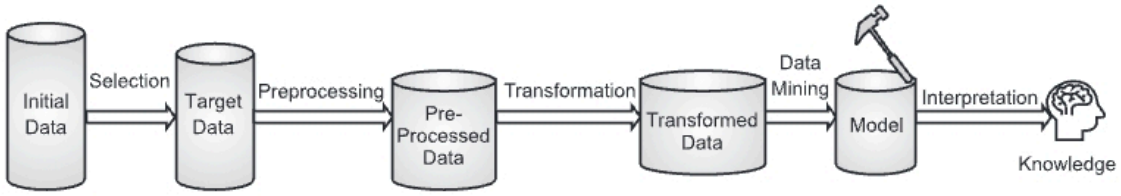


Fig. 1.1: KDD process

1. **Selection:** The data which is to be mined may not be necessarily from a single source. The data may have many heterogeneous origins. This data needs to be obtained from various data sources and files. The data selection is based on your mining goal. Data relevant to the mining task is selected from various sources.
2. **Pre-processing:** Pre-processing involves cleaning of the data and integration of the data. The data selected for mining purposes may have some incorrect, irrelevant values which leads to unwanted results. Some values may be missing or erroneous. Also, when data is collected from heterogeneous sources, it may involve varying data types and metrics. So, this data needs to be cleaned and integrated for noise elimination and inconsistency.
3. **Transformation:** Data transformation is the process of converting the data into the format which is suitable for processing. Here, data is molded in the form which is required by the data mining process.

4. **Mining:** The Mining process leads towards using methods, techniques to extract the pattern present in the data. The process involves transformation of relevant data records into patterns using classification. This step involves application of various data mining algorithms to the transformed data. Mining process generates the desired results for which the whole KDD process is undertaken.
5. **Visualization/Interpretation:** This is the last step in the KDD process. In this step, the data is presented to the user in the form of reports, tables or graphs. The presentation of the data to the users directly affects the usefulness of the results.

1.5 DATA MINING TECHNIQUES/TASKS

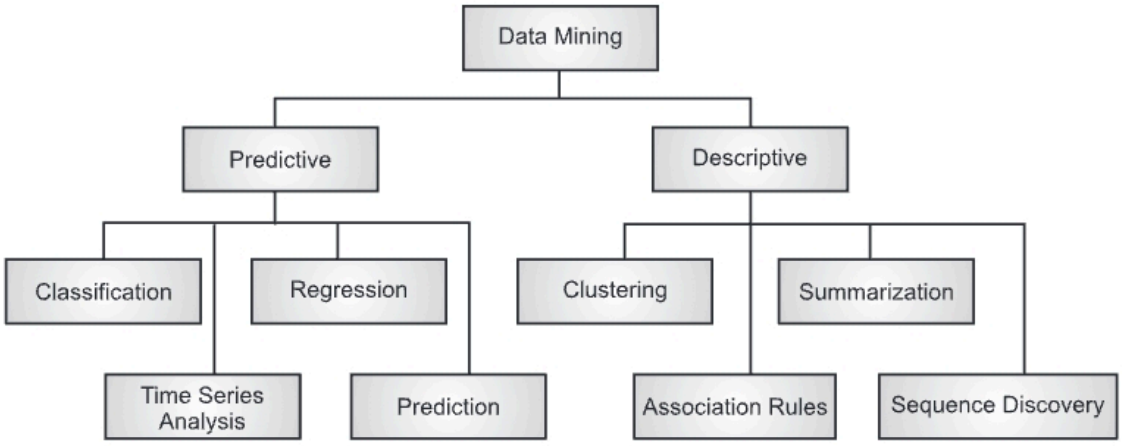


Fig. 1.2: Data Mining Tasks

- Data mining tasks can be categorized into two main types: Predictive and Descriptive.

A. Predictive Data mining:

- Predictive data mining tasks include the prediction based on the available data set in hand. These tasks give the model based on data and predict the future trends related to that data or unknown values that may be of interest for the future.
- The example of predictive tasks includes the prediction of future value of gold according to the current market trend. Also, prediction of high or low value of a share in the share market based on its previous growth is also a predictive data mining task.
- Predictive data mining includes Classification, Regression, Prediction and Time Series Analysis. Let's see the details of these tasks.

1. **Classification:** The dictionary meaning of classification is to classify, i.e. to categories or create groups of the data items according to particular criteria. In data mining, classification can be defined as arrangement of data items or making groups of data items based on the data points or observed values. The output of classification is a method that will decide the class of an object based on its attributes.

Examples of classification are:

(a) To find potential customers for a new product.

(b) To find the probable list of customers who are likely to apply for the credit card, based on previous data.

2. **Regression:** Regression can be defined as a data mining technique that is generally used for the purpose of predicting a range of continuous values (which can also be called "numeric values") in a specific dataset.
   - It is used to map data items to a real valued variable. Regression is very frequently used in business and market analysis. The main application involves financial prediction or forecasting, Environmental modeling and analysing trends and patterns.
   - There are two types of regression. One is Linear regression and another is Multiple regression.
     - In Linear regression, the relationship between two variables is established using a linear equation to observe the data. The output is a straight line which has only line dependent variables.
     - In Multiple regression, the relationship between two or more variables is established to predict the output and a single continuous dependent variable.
3. **Time Series Analysis:** Time series analysis is the process of recording of the data point at specific time intervals. This data is then used to predict the future values based on the data points recorded. Time series analysis can produce very important information for a business if used efficiently. The example of time series includes weather record, economic indicators, stock market analysis, workload projections etc.
4. **Prediction:** Prediction is a classification task. Prediction discovers the relationship between dependent variables and relationship between independent variables. It can also be viewed as estimation. The prediction is based on the data in hand and predictions or future trends of a phenomenon can be predicted using some predictive algorithms. The best example of prediction is the profit that could be gained out of sale. Predication is the technique of identifying the unavailable numerical data for a new process. Prediction applications include flooding, speech recognition, machine learning, and pattern recognition.

B. Descriptive Data Mining:

- Descriptive data mining tasks include the analysis of available data patterns or models to find out new interesting and significant information based on available data set.
- The example of descriptive data mining tasks includes the interchange in places of the super market according to the purchase pattern of the customers.
- Descriptive data mining includes Clustering, Summarization, Association Rules and Sequence Discovery. Let's discuss these tasks one by one.

1. **Clustering:** Clustering or cluster analysis is the method where the data points are grouped together according to their characteristics. The data points in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). Clustering can also be referred to as unsupervised learning.

Clustering can be used to find out the specific class of customer in the market. In life sciences, Clustering of similar character genes can give a new insight. Clustering can be used in outlier detection.

2. **Summarization:** Summarization is the process of finding the gist of the generated data. The process of Summarization divides the data into subsets with descriptions. Summarization is also called **Characterization or Generalization**. It extracts or derives representative information about the database. This may be accomplished by actually retrieving portions of the data.
3. **Association Rules:** Association rules find out the correlation among the data. Association rules find out a specific type of association between the data item. These associations are used to identify the frequency occurrence in the pattern and accordingly the strategies for business are changed or modified.

   An example of association rule is that in super market, generally, bread and butter are kept side by side or milk and bread are kept near to each other so that the person who buys milk will surely have the association of buying the bread as well. This phenomenon is also known as Market Basket Analysis.
4. **Sequence Discovery:** Sequence discovery, or sequential pattern mining, is a data mining technique that discovers statistically relevant patterns in sequential data. This mining program evaluates certain criteria, such as occurrence frequency, duration, or values in a set of sequences to find interesting hidden patterns. For example, most people who purchase CD players may be found to purchase CDs within one week.

   Other typical examples include customer shopping sequences (First buy computer, then CD-ROM, and then digital camera, within 3 months), Web clickstreams, bio-logical sequences, sequences of events in science and engineering, and in natural and social developments.

1.6 KNOWLEDGE REPRESENTATION METHODS

- When the data mining task is completed, the next part is visualization of the data. The mined data is visualized or represented using various visualization tools so as to understand the knowledge that has been devised out of the mining process. This method of visualization of data helps the user to understand the complex results of the mining process in an easy manner with the visualization tools.
- Visualization techniques include:
  - **Graphical:** This is a traditional graph structure including bar charts, pie charts, histograms, and line graphs may be used.
  - **Geometric:** Geometric techniques include the box plot and scatter diagram techniques.
  - **Icon-based:** This technique using figures, colors, or other icons can improve the presentation of the results.
  - **Pixel-based:** With these techniques, each data value is shown as a uniquely colored pixel.
  - **Hierarchical:** These techniques hierarchically divide the display area (screen) into regions based on data values.
  - **Hybrid:** The preceding approaches can be combined into one display.

1.7 APPLICATIONS OF DATA MINING

- Data mining is used by many organizations to improve the customer base. They focus on customer behavioral patterns, market analysis, profit areas and product improvement. The essential areas where data mining is used are as follows:

  (a) Education: Educational data mining deals with developing the methods to discover the knowledge from the education field. It is used to find out / project students' areas of interest, future learning capacities and other aspects. Educational institutions can apply different data mining techniques and take appropriate/ accurate decisions based on the outcome of the mining process. Also, the analysis of slow and fast learners and accordingly their teaching pattern can be determined.

  (b) Health and Medicine: Data mining can effectively be used in health care systems. During Covid 19 pandemic, the predictions of the covid 19 waves and the volume of patients was done using data mining. In Genetics also, data mining helps in determining the sequence of the genes and future trends.

  (c) Market Analysis: Market analysis is based on a particular pattern of purchase followed by customers. These patterns help the shop owner to understand the buying pattern of customers and accordingly useful decisions can be implemented so as to increase the profit of the store. Also, the market analysis helps to find out the different methodologies to retain the existing customers and gain new ones.

  (d) Fraud Detection: A fraud detection system helps in finding out the pattern of fraud, its potential attackers/ criminal detection and possible solutions using different data mining algorithms. These data mining methods provide timely and efficient solutions for detection and prevention of the frauds. Intrusion and lie detection can also be addressed by these mechanisms.

1.8 DATA PRE-PROCESSING

- We know that a tremendous amount of data is generated daily in today's world of web. The huge size of the data makes it vulnerable to changes. The resultant data may be incomplete, noisy. This weakens the quality of the data which is used for data mining purposes leads to invalid results. The process of cleaning the data and making it useful for the process of mining is called Data Pre-processing.
- The pre-processing process consists of many steps. Pre-processing can be performed manually or automatically. Let's discuss the steps involved in Data Pre-processing.

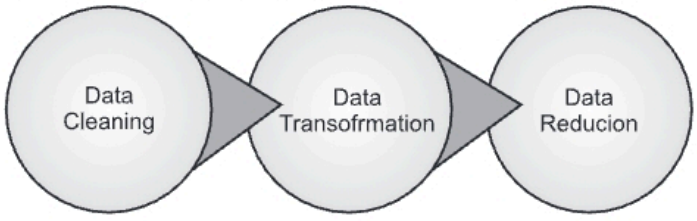


Fig. 1.3: Steps of Data Pre-processing

1.8.1 Data Cleaning

- The first step in data preprocessing is data cleaning. Data cleaning includes handling missing data and noisy data.

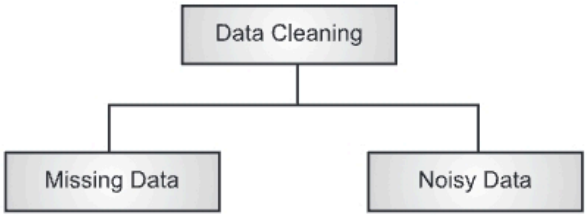


Fig. 1.4: Data Cleaning

(a) Missing data: Missing data is the case wherein some of the attributes or attribute data is missing or the data is not normalised. This situation can be handled by either ignoring the values or filling the missing value.

(b) Noisy data: This is data with error or data which has no meaning at all. This type of data can either lead to invalid results or can create the problem to the process of mining itself. The problem of noisy data can be solved with binning method, regression and clustering.

1.8.2 Data Transformation

- Data used for data mining is the data which comes from various heterogeneous platforms. This unstructured and structured data needs to be combined for smooth processing of data mining. This homogeneous data is then analysed to find out the patterns. The benefits of data transformation include the improved quality of the data. The user can get maximum value out of the available data. Transformation also helps in improving the performance of the queries. There are specialised tools for data transformation.

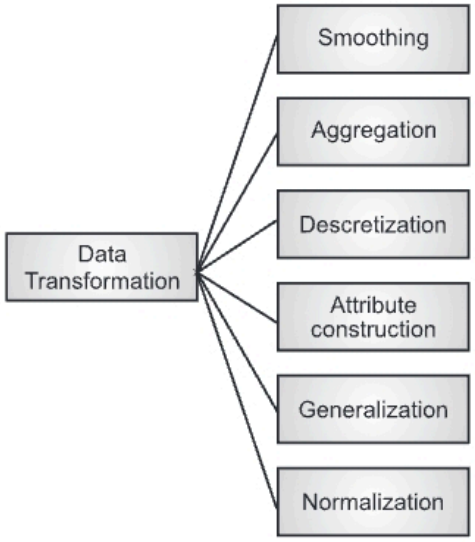


Fig 1.5: Data Transformation

- The various data transformation methods include:
  - **(a) Smoothing:** This is the process of removing the unnecessary data and cleaning the data so as to improve the functionality of the data.
  - **(b) Aggregation:** This is the process of collecting the data from heterogeneous platforms and converting it to a uniform format. This improves the quality of the data.
  - **(c) Discretization:** Large data sets are complex to handle. Discretization is the process of breaking up the data in small intervals. These chunks are continuous chunks and these are supported by all the existing frameworks.
  - **(d) Attribute construction:** To improve the efficiency in the mining process, some new attributes are generated from existing data sets.
  - **(e) Generalization:** This is the process of converting low level attributes to high level attributes using hierarchy.
  - **(f) Normalization:** In the process of Normalization, attributes are scaled within a specified range.

1.8.3 Data Reduction

- Data reduction is a process that reduced the volume of original data and represents it in a much smaller volume.

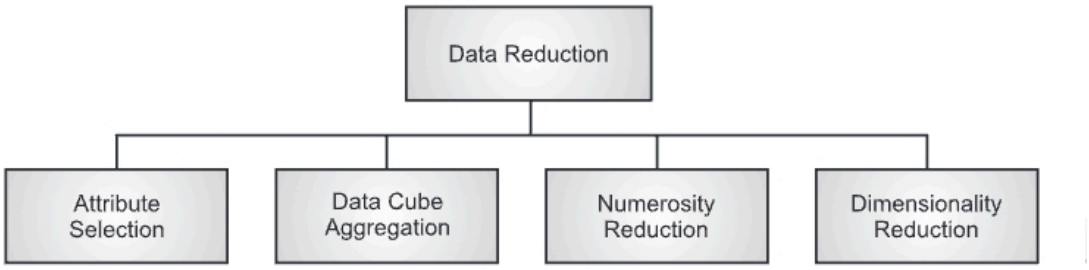


Fig. 1.6: Data Reduction Methods

- The various data reduction methods include:
  - **(a) Attribute Selection:** When data is collected from various sources, it may contain duplicate attributes. Some of the attributes are irrelevant. The Attribute Selection method is used to remove such redundant and unnecessary attributes from the data set. This process results in an improved data set.
  - **(b) Data Cube Aggregation:** In this reduction method, aggregation property is applied on selected data sets so as to get the data in a much simpler format.
  - **(c) Numerosity Reduction:** In this reduction method, actual data is substituted with a mathematical model of the data.
  - **(d) Dimensionality Reduction:** In this reduction method, duplicate attributes are removed to reduce the data size.

1.8.4 Data Discretization

Discretization:

- Large data sets are complex to handle. Discretization is the process of breaking up the data in small intervals. Here, the data size is reduced. But the data which is divided into intervals is kept continuous having some sequence. Every interval has its own name and later these intervals can be replaced with actual data. These chunks are continuous chunks and these are supported by all the existing frameworks.

Data Discretization techniques:

1. **Top-down Discretization:** If the process starts by first finding one or a few points (called split points or cut points) to split the entire attribute range, and then repeats this recursively on the resulting intervals, then it is called top-down discretization or splitting.
2. **Bottom-up Discretization:** If the process starts by considering all of the continuous values as potential split-points, removes some by merging neighborhood values to form intervals, then it is called bottom-up discretization or merging.

Concept Hierarchies:

- Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as numerical values for the attribute age) with higher-level concepts (such as youth, middle-aged, or senior).
- In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. This organization provides users with the flexibility to view data from different perspectives.

Data Discretization Methods:

Following are some Data discretization methods for numeric data:

1. **Binning:** This is a top-down unsupervised splitting technique based on a specified number of bins. In this method, values found for an attribute are grouped into a number of equal-width or equal-frequency bins. Then the values are smoothened using bin mean or bin median in each bean. Using this method recursively you can generate concept hierarchy.
2. **Histogram Analysis:** The histogram distributes an attribute's observed value into a disjoint subset, often called buckets or bins.
3. **Cluster Analysis:** Cluster analysis is a common form of data discretization. In this technique, a clustering algorithm can be applied to discretize a numerical attribute by partitioning the values of that attribute into clusters or groups.

Summary

- Data Mining deals with discovery of hidden knowledge, unexpected patterns and new rules from large data sets.
- Following issues were faced while doing data mining: Human Interaction, Overfitting, Outliers, Interpretation of result, Visualization of result, Large datasets, High Dimensionality, Multimedia Data, Missing data, Irrelevant data, Noisy data, Changing Data, Integration, Application.

- Knowledge discovery in databases (KDD) is the process of finding useful information and patterns in data.
- The KDD process is divided into following steps: Selection, Preprocessing, Transformation, Data Mining, Measuring , Interpretation/Visualization.
- Data mining tasks are Predictive and Descriptive.
- Predictive tasks include classification, regression, time series analysis and prediction.
- Descriptive tasks include clustering, summarisation, association rules and sequence discovery.
- Knowledge representation methods are graphical, geometric, icon based, pixel based and hierarchical and hybrid.
- Applications of data mining include education, health, finance and fraud detection.
- Data pre-processing is the process of cleaning the data and making it useful for the process of mining.
- The steps in data pre-processing are Data cleaning, Data transformation and Data reduction.
- Data reduction is a process that reduced the volume of original data and represents it in a much smaller volume.
- Discretization is the process of breaking up the data in small intervals.

Check Your Understanding

1. Data mining is also called ___.
   (a) Data Processing (b) Data Discovery
   (c) Knowledge discovery (d) Knowledge Processing.
2. Which of the following is an essential process in which the intelligent methods are applied to extract data patterns?
   (a) Warehousing (b) Data Mining
   (c) Text Mining (d) Data Selection
3. What are the functions of Data Mining?
   (a) Association and correctional analysis classification
   (b) Prediction and characterization
   (c) Cluster analysis and Evolution analysis
   (d) All of the above
4. Which of the following statements is correct about data mining?
   (a) It can be referred to as the procedure of mining knowledge from data
   (b) Data mining can be defined as the procedure of extracting information from a set of the data
   (c) The procedure of data mining also involves several other processes like data cleaning, data transformation, and data integration
   (d) All of the above

5. _____ is not data mining functionality?
   (a) Clustering and Analysis (b) Selection and Interpretation
   (c) Classification and Regression (d) Characterization and Discrimination
6. _____ is the output of KDD.
   (a) Query (b) Useful Information
   (c) Data (d) Information
7. The analysis performed to uncover interesting statistical correlations between associated-attribute-value pairs is called _____.
   (a) Mining of Association (b) Mining of Clusters
   (c) Mining of Correlation (d) None of the above
8. To remove noise and inconsistent data _____ is needed.
   (a) Data Cleaning (b) Data Transformation
   (c) Data Reduction (d) Data Integration
9. Which of the following is not a data pre-processing methods?
   (a) Data Visualization (b) Data Discretization
   (c) Data Cleaning (d) Data Reduction
10. Which of the following is the example of sequence data?
   (a) Weather forecast (b) Data matrix
   (c) Market basket data (d) Genome and DNA data of an organism

Answers

1. (c)	2. (b)	3. (d)	4. (d)	5. (b)	6. (b)	7. (c)	8. (a)	9. (a)	10. (d)

Practice Questions

Q.I Answer the following questions in short.

1. Define Data Mining.
2. List out steps of KDD process.
3. What are the types of data?
4. Compare descriptive and predictive data mining.
5. What is prediction?
6. Why we need to Pre-process the data?
7. What is Data integration?
8. List out the major issues in data mining.
9. What is data discretization?

Q.II Answer the following questions.

1. What is Data Cleaning? Describe various methods of Data Cleaning.
2. Discuss about the major issues in Data Mining.
3. Explain various steps in data pre-processing.
4. What is Data Mining?
5. Differentiate between Query processing and Data mining.

Q.III Define the terms.

1. Clustering
2. Classification
3. Regression
4. Time Series Analysis.
5. Data Cleaning
6. Data Preprocessing.

2... Data Warehousing

Learning Objectives...

- To introduce concept of Data Warehouse.
- To learn Data Warehouse Architecture and its components.
- To study Data Modelling with OLAP.
- To get information about Schema Design.
- To know about Machine Learning and Pattern Matching.

2.1 INTRODUCTION TO DATA WAREHOUSE

- As we know that there is tremendous growth in generation of data. Traditional database systems have a drawback of handling homogeneous data. Moreover, large organisations need a reliable solution to handle this huge data that too with heterogeneous format. Also, the system handling data should be powerful enough to manipulate, access and analyse this huge data with an effective system. The solution to this was provided in 1980 by William Inmon by using the term data warehouse. Data warehouse is a concept which supports decision support systems where a large amount of data is merged together.
- A Data Warehouse (DW) is a repository which is at the top of multiple databases. It can be defined as a process for collecting and managing data from varied sources to provide meaningful business insights.
- A data warehouse is a subject-oriented (it can be used to analyse a particular area or domain), integrated (it integrates multiple data sources in a single identification), time-variant (historical data is stored in a data warehouse, whereas in a transaction processing system recent data is stored) and non-volatile (once the data is moved to a warehouse, it doesn't change) collection of data in support of management's decision-making process.
- A data warehouse is used to analyse the data, to generate reports based on that analysis. Here the point to be noted is that, the data used for analysis or reporting is heterogeneous data collected from varied / multiple sources. The main purpose of data warehousing is to combine such data for the purpose of analysing and reporting. The data is then used for the strategic planning and decision making for the organizations. Data warehousing is different from normal data processing and querying.

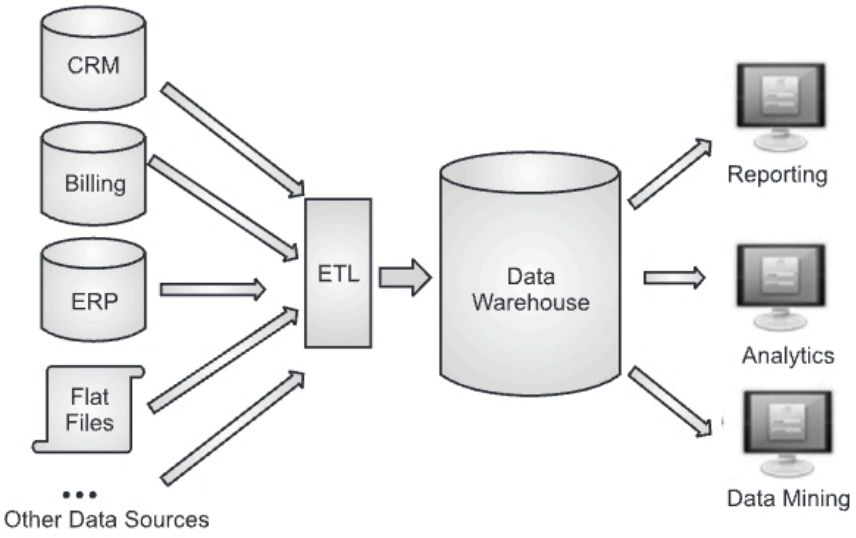


Fig. 2.1: Data Warehouse

Table 2.1: Difference between Data Mining and Data Warehousing

Data Mining	Data Warehousing
It is the process of discovering patterns from large data sets.	It is the process of extracting, transforming, aggregating and loading data from multiple data sources.
This is used for analysis of data.	This is used for periodical storage of data.
It analyses simple data.	It is used to store historical data and use it.
It improves knowledge extraction.	It improves performance of data.

Table 2.2: Difference between Database and Data Warehousing

Database	Data Warehousing
It is collection of interrelated and homogeneous data stored in rows and columns.	It is central storage of data from multiple data sources.
It is used for manipulation of data.	It is used for analysis and reporting from the data.
It is more of application oriented.	It is subject-oriented.
It uses online transaction processing system.	It uses an online analytical processing system.
It is based on the relational model.	It is based on data modeling techniques.

2.2 DATA WAREHOUSE ARCHITECTURE AND ITS COMPONENTS

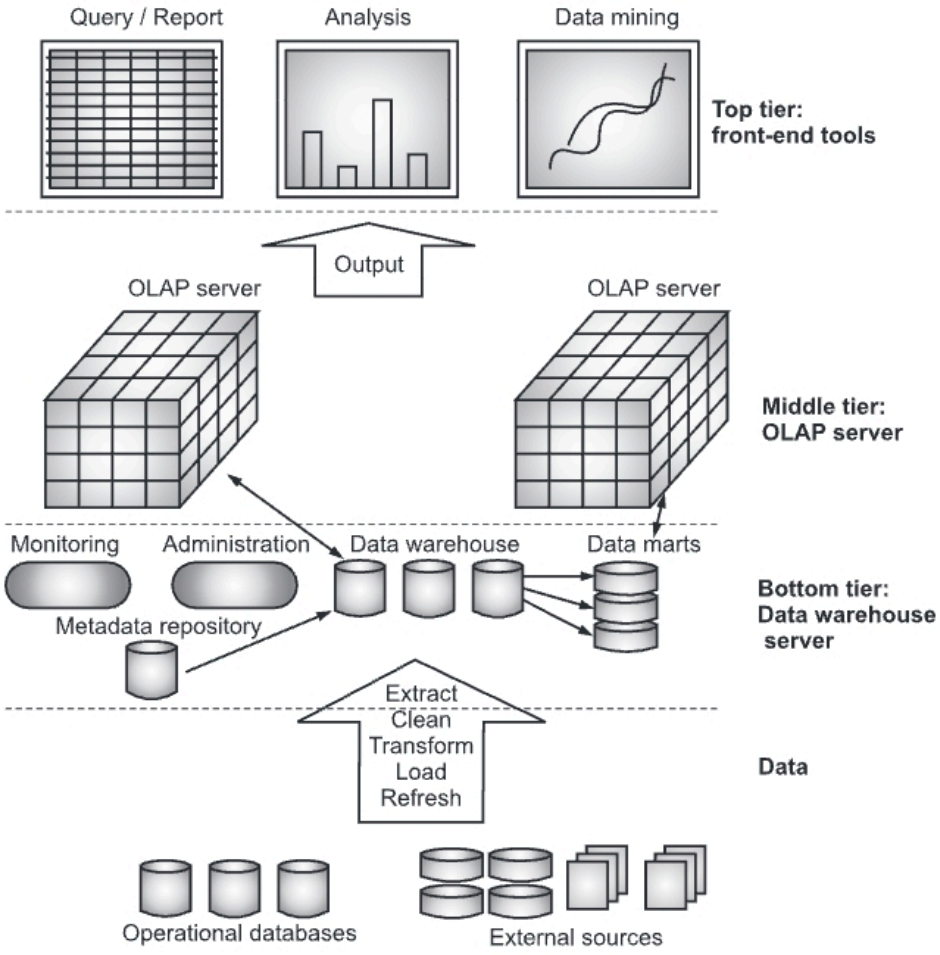


Fig. 2.2: Architecture of Data Warehouse

- The architecture illustrates gathering of data, the storage of data, and the querying and data-analysis support. The basic components of data warehousing includes:
  (a) Data migration
  (b) The warehouse
  (c) Access tools
- The data is extracted from the operational system. This extracted data must be reformatted, cleansed, integrated and summarized before loading into a warehouse. This process is done with the various tools available for transformation, cleaning, summarization and loading. These tools are called as **ETL tools** (Extract, Transform and Load).

- The main responsibilities of ETL tools include:
  - To remove the identity specifications of the data according to the rules.
  - To remove unwanted data from the operational database before loading it to the data warehouse.
  - To find and replace common names and definitions as the data is arriving from multiple destinations.
  - To summarize the data.
  - To recover the data with default values if there are missing values.
  - To remove the redundancy in data.
- This migration process is similar to that needed for data mining applications, but that data mining applications are not necessarily performed on summarized or business-wide data.
- Typical data warehouse architecture is three-tier architecture as follows:

1. **Bottom Tier:** Bottom tier consists of an actual data warehouse server. Summary data is replication of detailed information which stores a summary of the data present in the data warehouse. Metadata stores all the metadata definitions used by all processes within the data warehouse. **Metadata** is data about data. It is the card index describing how information is structured within the data warehouse. Metadata is used to map data sources to the common view of information within the data warehouse and used to automate the production of summary tables and also used to direct a query to the appropriate data source.
2. **Middle Tier:** The middle tier is an OLAP server. It is implemented using a Relational OLAP model or Multidimensional OLAP.
3. **Top Tier:** Top tier consists of front end tools that are used for querying the database which is present in the data warehouse to get the information for analysis.

2.3 DATA MODELING WITH OLAP

2.3.1 Introduction

- All we know is that an enormous amount of data is collected and stored on a daily basis. This data is mined for further growth of the organization. But data in its original form doesn't always give the expected results or output. This data needs to be structured because structured raw data will lead to making more specific decisions. The process of structuring your raw data is called **Data Modeling.** This is where OLAP comes in.
- OLAP (Online Analytical Processing) is a multidimensional database system which is able to analyse many data records quickly. This analysis will help the organizations in decision making and strategic planning.

The OLAP Process:

- Let us see how data is prepared for Online Analytical Processing (OLAP).

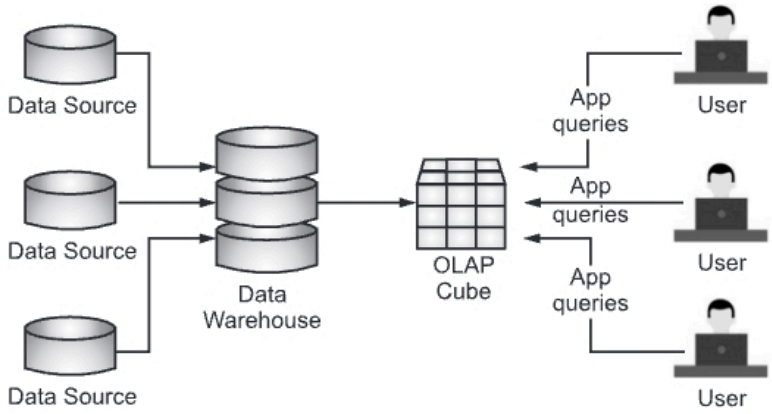


Fig. 2.3: Process of OLAP

- Following are steps of OLAP Process:

  Step 1: Data is first extracted from various data sources and formats, like text files and spreadsheets. This data is then stored in the Data Warehouse.

  Step 2: The data is cleaned, transformed, and stored in OLAP Cubes.

  Step 3: Once in the OLAP cubes, information is then pre-calculated and pre-aggregated in advance for further analysis.

  Step 4: Finally, the user gets the data from the OLAP cubes by running queries against them.

2.3.1.1 Terminologies associated with OLAP

- There are some important terminologies associated with OLAP. These are:
  1. **Cubes:** The multidimensional database structure used by OLAP is called a Cube. The data cube or cube allows fast analysis of the data according to the multiple dimensions of the data.
  2. **Dimensions:** Dimensions are the basis for the data structure of an OLAP data cube. These are the list of related data items and are used to organize the data in a similar category.
  3. **Measures:** An OLAP measure is a numeric value that aggregates the dimensions. It provides the details about the quantities of the user's interest.
  4. **Hierarchies:** Hierarchies are the subcategories of the dimensions. They are multilevel and they allow you to drill up or drill down the data.

2.3.1.2 Data Warehouse Models

There are three data warehouse models.

1. **Enterprise Warehouse:** This type of warehouse model collects all the information about the subjects across the organisation. It is cross functional in scope. It contains data along with its summary. An enterprise model of warehousing is an extensive business model and its design and implementation is a lengthy process.
2. **Data Mart:** A data mart is a subset of a data warehouse and it contains the data of a specific subject or a specific domain of an organization. For example, the marketing data mart of an organization may contain data related to marketing or a production data mart will contain the information and data related to production and so on.
3. **Virtual Data Warehouse:** A virtual data warehouse is a group of distinct databases which can be queried together. So a user can effectively and efficiently access all the data as if it was stored in one data warehouse.

2.3.2 Difference between OLTP and OLAP

Table 2.3: Difference between OLTP and OLAP

OLTP	OLAP
Online Transaction Processing system.	Online Analytical processing system.
It manages transaction oriented applications.	It manages the reports to multi-dimensional analytical queries.
It is an online database modifying system.	It is an online query answering system.
Its basic focus is on manipulating the database.	Its main focus is to analyse and extract the data for strategic decision making.
The queries are short and simple.	The queries are long and complex.
The modeling of OLTP is industry oriented.	The design of OLAP is subject or domain specific.
The main purpose is to control day to day transactions in the database.	Its main purpose is to find the hidden data and support decision making.

2.3.3 Data Mart

- A data mart is a subset of a data warehouse which is useful to a specific group of users. It is limited to a specific domain. For example, production data mart will have its scope to production related data like raw material, quantity, quality and so on. Marketing data mart will be related to marketing related attributes. The data contained in data marts tend to be summarized. Data marts are servers based on Linux or Windows.

Categorisation of Data Mart:

- The data marts can be categorized into independent data mart, dependent data mart or hybrid data mart. The categorisation of data mart is based on the source of data.
  1. **Dependent:** Dependent data marts are created by drawing data directly from operational, external or both sources.

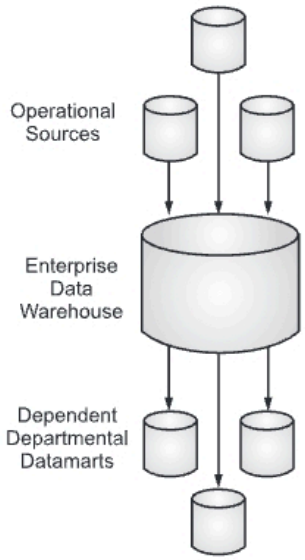


Fig. 2.4 (a): Dependent Data Mart

2. **Independent:** Independent data mart is created without the use of a central data warehouse.

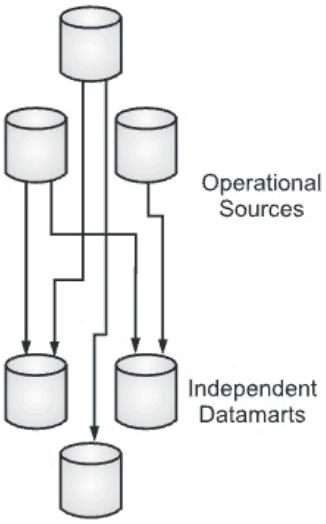


Fig. 2.4 (b): Independent Data Mart

3. **Hybrid:** This type of data marts can take data from data warehouses or operational systems.

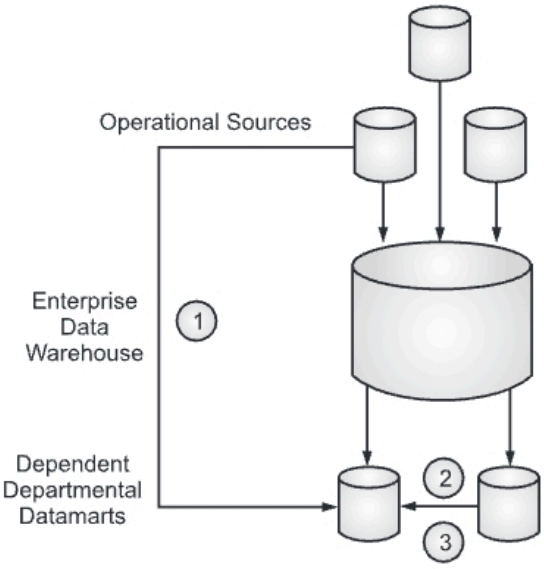


Fig. 2.4 (c): Hybrid Data Mart

Need of Data Mart:

- Since data mart is related to a specific domain, the time of retrieval of information is less with improved efficiency.
- It provides easy access to frequently requested data.
- They are easy to implement and the cost of implementation is less as compared to a data warehouse.
- A data mart is agile. In case of change in model, data mart can be built quicker due to a smaller size.
- A Data mart is defined by a single Subject Matter Expert.
- Data can be segmented and stored on different hardware/software platforms.

Advantages and Disadvantages of Data Mart:

Advantages:

- Data marts are domain specific; hence it is valuable to a specific group of users.
- It is cost effective and easy to implement.
- Data mart allows faster access of data.
- Data mart is easy to use as it is specifically designed for the needs of its users.
- A data mart can accelerate business processes.
- It is easy to implement and efficient to use.
- It contains historical data which enables the analyst to determine data trends.

Disadvantages:

- Many subsets of corporate data warehouse may create unnecessary burden.
- It is very hard to maintain the data mart if they are created with unrelated data.
- Data mart cannot provide company-wide data analysis as their data set is limited.

2.3.4 Fact Table, Dimension Table, OLAP Cube

Fact Table:

- A fact table is the central table in a star schema of a data warehouse. A fact table stores quantitative information for analysis and is often denormalized. This table is found at the center of a star schema or snowflake schema surrounded by dimension tables.
- A fact table consists of facts of a particular business process e.g., sales revenue by month by product. Facts are also known as measurements or metrics. A fact table record captures a measurement or a metric.

Dimension Table:

- A dimension table is a table in a star schema of a data warehouse. A dimension table stores attributes, or dimensions, that describe the objects in a fact table. For example, a Time dimension table stores the various aspects of time such as year, quarter, month, and day.

- The main characteristic of dimension tables is their multitude of attributes. Attributes are the columns that we summarize, filter, or aggregate.
- Dimensional tables have one primary key based on the underlying business key.

OLAP Cube:

- A data warehouse is based on a multidimensional data model which views data in the form of a data cube. A OLAP cube such as Sales, allows data to be modeled and viewed in multiple dimensions. Dimensions such as item (item_name, brand, type), time (day, week, month, quarter, year), location (street, city, state, country) and supplier (no, name).
- In data warehousing literature, an n-D base cube is called a **base cuboid**. The top most 0-D cuboid, which holds the highest-level of summarization, is called the **apex cuboid**. The lattice of cuboids forms a data cube.
- An OLAP cube optimized for data warehouse and Online Analytical Processing (OLAP) applications.
- An OLAP cube overcomes the limitations of relational databases by providing fast analysis of data. Cubes can display and sum large amounts of data. It is also provides searchable access to any data points to users.

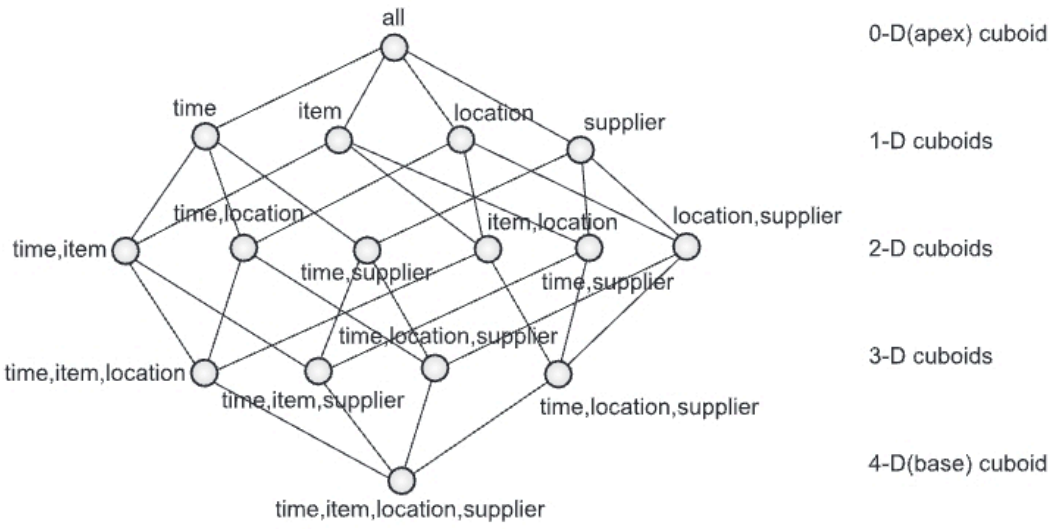


Fig. 2.5: A Lattice of Cuboids

2.3.5 Different OLAP Operations

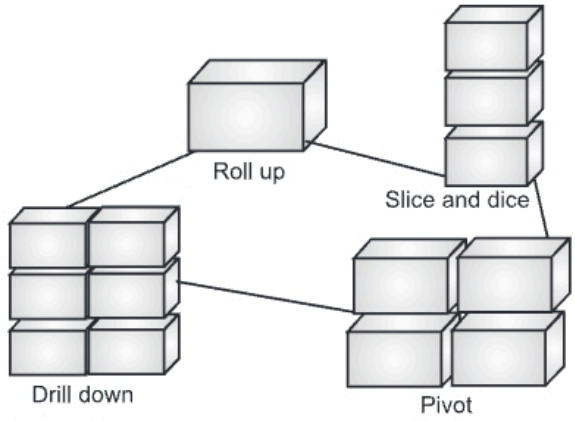


Fig. 2.6: Operations in OLAP

- OLAP operations are done on multidimensional data. This multidimensional data is organized in various dimensions. Every dimension includes multiple levels of abstraction. So, there are various OLAP operations to demonstrate these views.
- OLAP operations are based on a multidimensional view of data. Here is the list of OLAP operations:
  - Roll-up
  - Drill-down
  - Slice and dice
  - Pivot (rotate)

1. Roll-up:

- Roll-up performs aggregation on a data cube by either climbing up a hierarchy for a dimension or by reducing the dimensions. When roll-up is performed, some dimensions are reduced from the data.

Example 1: Consider the following example,

Location	Medal
Delhi	5
New York	2
Patiala	3
Los Angeles	5

Delhi, New York, Patiala and Los Angeles won 5, 2, 3 and 5 medals respectively. So in this example, we roll upon Location from cities to countries.

Roll-up operation:

Location	Medal
India	8
America	7

Example 2:

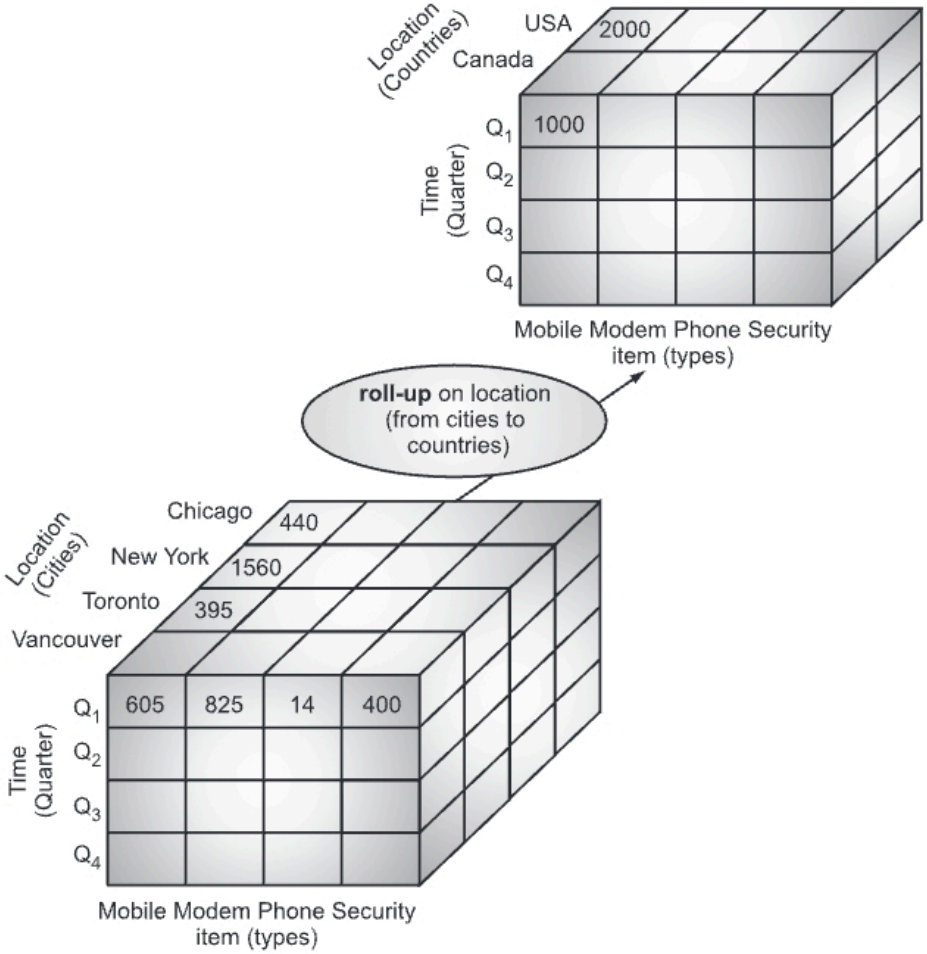


Fig 2.7 (a): Roll -up operation

- Roll-up is performed by climbing up a concept hierarchy for the dimension location.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up, the data is aggregated by ascending the location hierarchy from the level of the city to the level of the country.
- The data is grouped into cities rather than countries.
- When roll-up is performed, one or more dimensions from the data cube are removed.

2. Drill-down:

- The drill-down operation (also called roll-down) is the reverse operation of **roll-up**. It is performed by either of the following ways:
  - By stepping down a concept hierarchy for a dimension.
  - By introducing a new dimension.

Example 3:

Location	Medal
India	8
America	7

Location	Medal
Delhi	5
New York	2
Patiala	3
Los Angeles	5

Example 4:

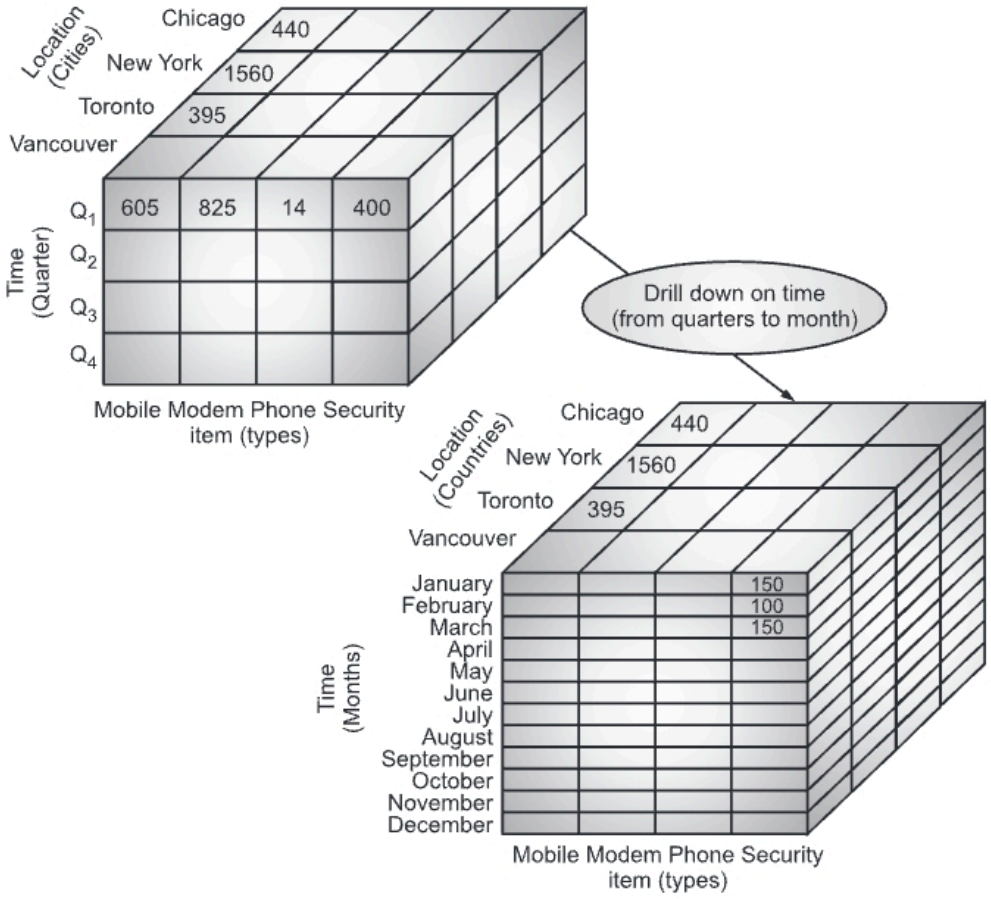


Fig 2.7(b): Drill-down operation

- Drill-down is performed by stepping down a concept hierarchy for the dimension time.
- Initially the concept hierarchy was "day < month < quarter < year."

- On drilling down, the time dimension is descended from the level of quarter to the level of month.
- When drill-down is performed, one or more dimensions from the data cube are added.
- It navigates the data from less detailed data to highly detailed data.

3. Slice and Dice:

- The slice operation performs a selection on one dimension of the given cube, resulting in a sub cube. It reduces the dimensionality of the cubes.

Example 5:

- For example, if we want to make a selection where Medal = 5.

Location	Medal
Delhi	5
Los Angeles	5

- The dice operation defines a sub-cube by performing a selection on two or more dimensions. For example, if we want to make a selection where Medal = 3 or Location = New York.

Location	Medal
Patiala	3
New York	2

Example 6 : The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Consider the following diagram that shows how slice works.

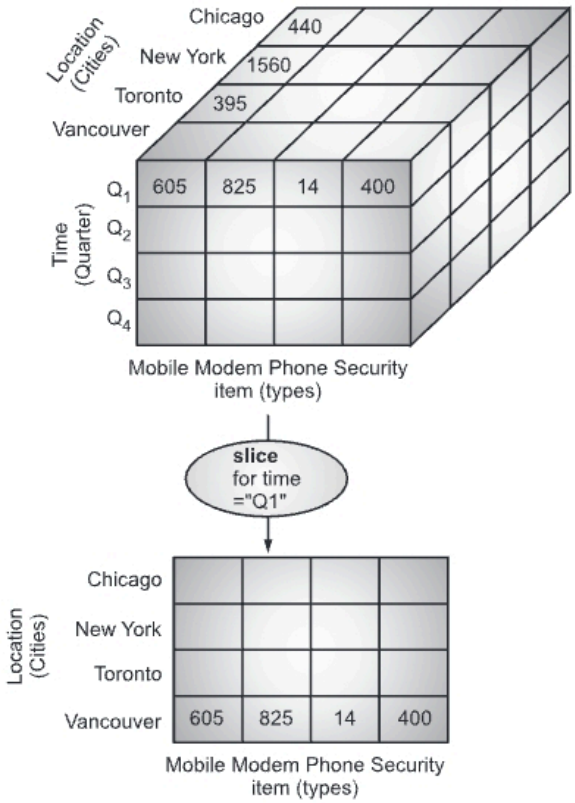


Fig. 2.7 (c): Slice Operation

- Dice selects two or more dimensions from a given cube and provides a new sub-cube. Consider the following diagram that shows the dice operation.

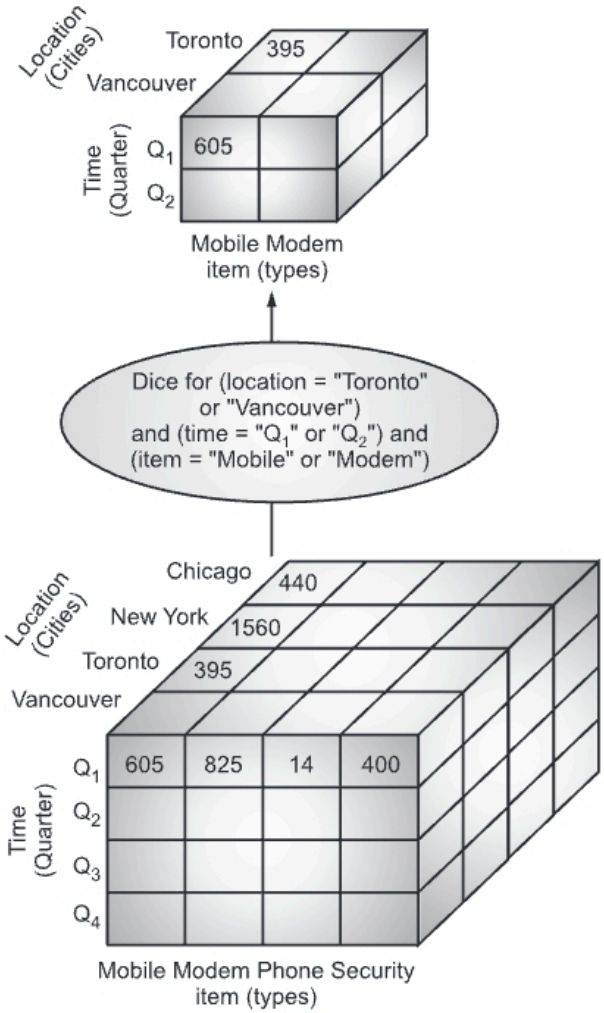


Fig. 2.7 (d): Dice Operation

- The dice operation on the cube based on the following selection criteria involves three dimensions:
  - (location = "Toronto" or "Vancouver")
  - (time = "Q1" or "Q2")
  - (item =" Mobile" or "Modem")

4. Pivot (Rotate):

- The pivot operation is also known as rotation. It rotates the data axis in view in order to provide an alternative presentation of data. Consider the following diagram that shows the pivot operation.

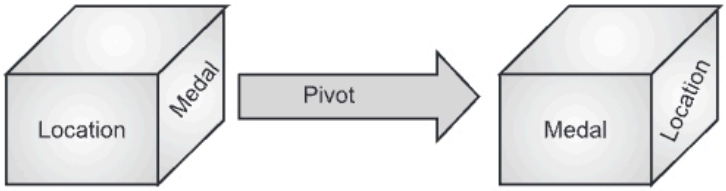


Fig. 2.7 (e): Pivot Operation

2.4 SCHEMA DESIGN

2.4.1 Introduction

- A database schema is the logical structure of the database. This structure contains the information and entails about the name and other details of all the records associated with that database. It also includes the relationships and association of the data amongst themselves. In case of a traditional database system, database schema describes the details where as in the case of data warehouse, need to maintain the schema as well. The database uses a relational model whereas in case of data warehouse, it uses star and snowflakes schema.
- The schema design in the data warehouse is based on fact tables and dimension tables. The fact tables contain data corresponding to any business process. It stores quantitative information for analysis. A dimension table stores data about how the data in fact table is being analyzed. They facilitate the fact table in gathering different dimensions on the measures which are to be taken.
- Let's discuss the two types of schema design in a data warehouse.

2.4.2 Star Schema

- Star schema is the most common schema in data warehouses. This is widely used to design the data warehouse. The basic architecture of star schema includes one fact table and many dimension tables. The advantage of star schema is that it is very efficient in handling the queries.

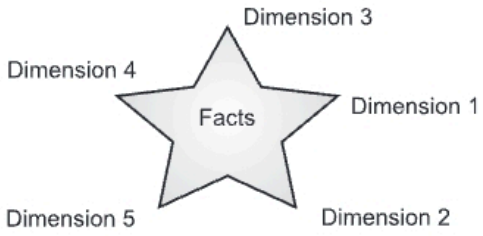


Fig. 2.8 (a): Star Schema

- Star schema contains one fact table associated with many dimension tables.
- The fact table contains primary information of the data warehouse.
- Dimension tables have details of the surrounding tables.
- The primary key which is present in each dimension is related to a foreign key which is present in the fact table.
- The fact table has two types of columns having foreign keys to dimension tables and measures which contain numeric facts.

- The fact tables are in 3NF form and the dimension tables are in denormalized form. Every dimension in the star schema should be represented by the only one-dimensional table. The dimension table should be joined to a fact table. The fact table should have a key and measure.

Example of Star Schema:

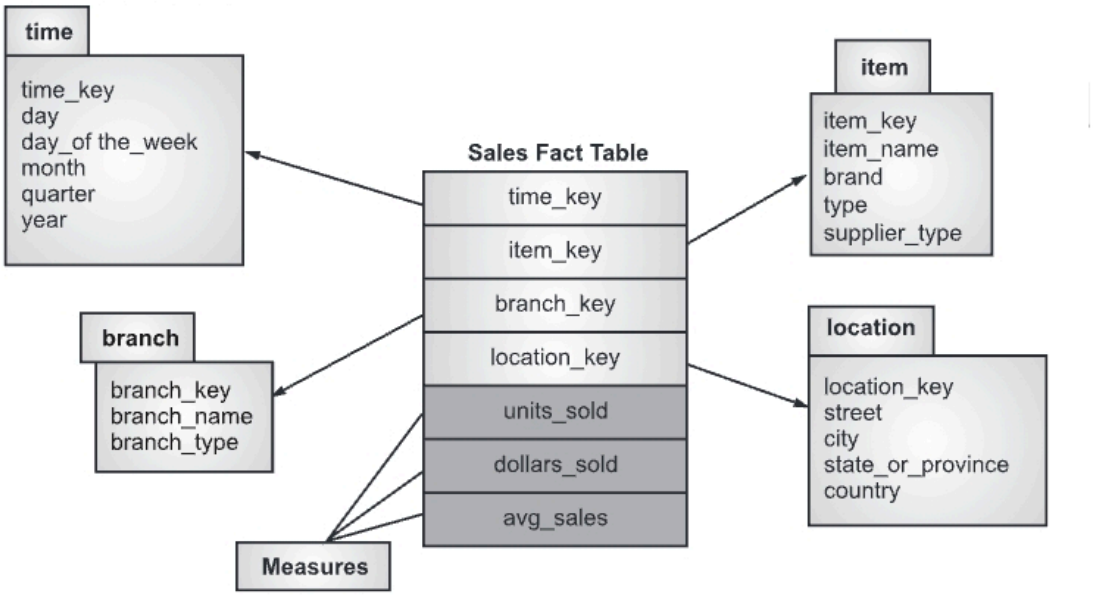


Fig. 2.8 (b): Example of Star Schema

- We are creating a schema which includes the sales data of a company. Sales are intended along following dimensions: time, item, branch, and location.
- The schema contains a central fact table for sales that includes keys to each of the four dimensions, along with two measures: dollar-sold and units-sold. The capacity of the fact table is reduced by the generation of dimension identifiers such as time_key and item_key via the system.
- Only a single table reproduces each dimension, and each table contains a group of attributes as it is shown in the star schema. The location dimension table includes the attribute set {location_key, street, city, state and country}. This restriction may introduce some redundancy. For example, two cities can be of same state and country, so entries for such cities in the location dimension table will create redundancy among the state and country attributes.

Advantages of Star schema:

- Its performance is good because simple queries are used.
- It contains single dimension tables.
- In star schema, both Dimension and Fact Tables are in De-Normalized form.
- It has less number of foreign keys and hence shorter query execution time.

Disadvantages of Star Schema:

- It has redundant data and hence difficult to maintain/change
- There are data integrity issues.
- Many-to-Many relationships are not supported.

2.4.3 Snowflake Schema

- Snowflake schema is a refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake. 'Snow flaking' or the normalization of the dimension tables can be done in many different ways.
- Snowflake schema is an arrangement of tables in a multidimensional database system.

Advantages to the Snowflake Schema:

- Data is structured.
- Data integrity is maintained.
- Less disk space is utilized.

Disadvantages of Snowflake Schema:

- It requires more complex queries.
- Complex queries decrease the performance.

Example of Snowflake Schema:

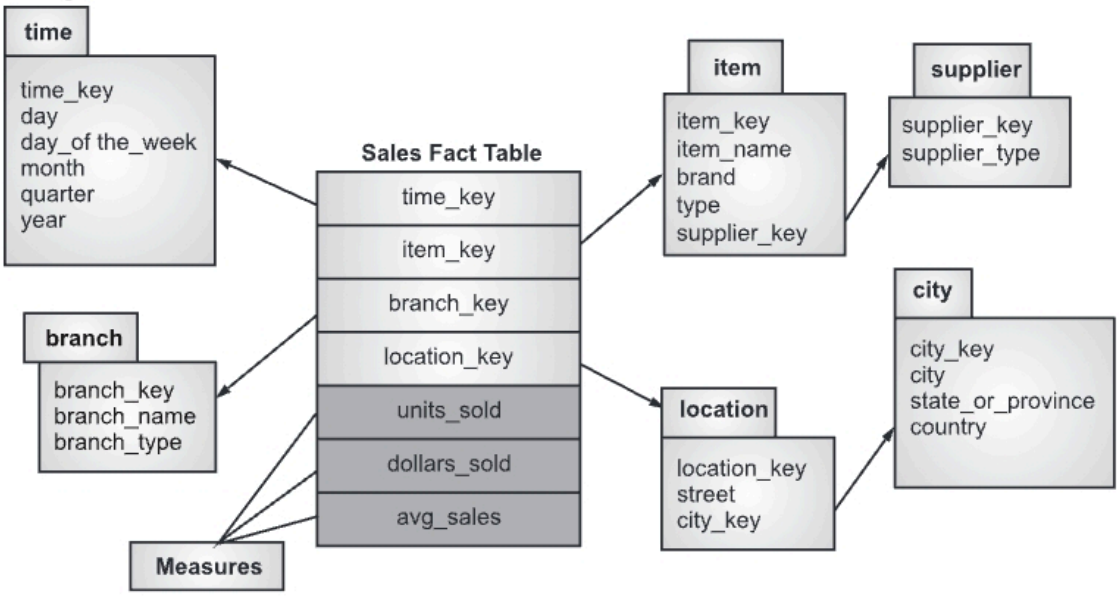


Fig. 2.9 : Example of Snowflake Schema

- In this example, the sales fact table is identical to that of the star schema, but the main difference is in the definition of dimension tables.
- The single dimension table for the item in the star schema is normalized in the snowflake schema, results in creation of new item and supplier tables.
- For instance, the item dimension table consists of the attributes item_key, item_name, brand, type, and supplier_key, where supplier_key is connected to the supplier dimension table, which holds supplier_key and supplier_type information.
- Similarly, the location dimension table involves the attributes location_key, street, and city_key, and city_key is linked to city dimension table containing the city, state and country attribute.

Difference between Star schema and Snowflakes schema.

Table 2.4: Star schema Vs Snowflakes schema

Star Schema	Snowflake Schema
Star schema is relational schema which follows the concept of facts and dimensions.	A snowflake schema is an extension of the star schema.
Data redundancy is high.	Low data redundancy.
Response time is fast.	It is less fast than star schema.
Tables in database are not normalised.	Data is normalised.
Single dimensional data is used.	Multidimensional data is used.
Top-down approach is used.	Bottom-up approach is used.
Simple in design.	Complex in design.

2.5 INTRODUCTION TO MACHINE LEARNING

- Machine learning is a growing technology which enables computers to learn automatically from past data.
- Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information. Currently, it is being used for various tasks such as image recognition, speech recognition, email filtering, Facebook auto-tagging, recommender system, and many more.

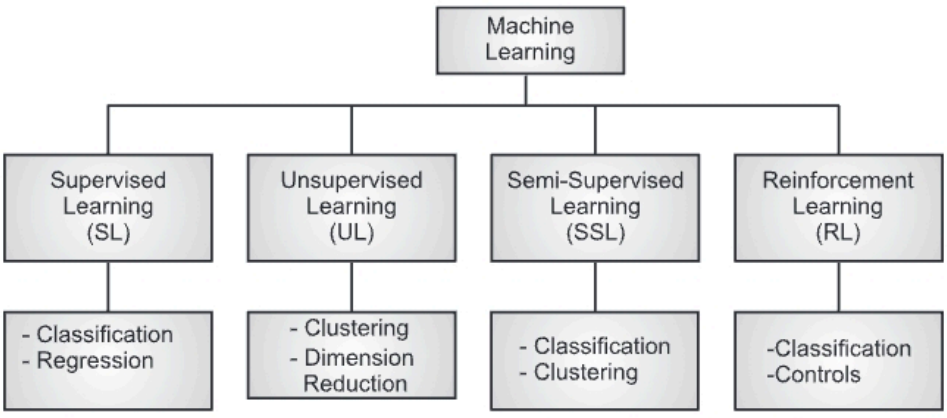


Fig 2.10: Classification of Machine Learning Technology

- Machine learning is the area of AI that examines how to write programs that can learn. In data mining, machine learning is often used for prediction or classification.

- With machine learning, the computer makes a prediction and then, based on feedback as to whether it is correct, "learns" from this feedback. It learns through examples, domain knowledge, and feedback. When a similar situation arises in the future, this feedback is used to make the same prediction or to make a completely different prediction.
- Statistics are very important in machine learning programs because the results of the predictions must be statistically significant and must perform better than a naive prediction.
- Applications that typically use machine learning techniques include speech recognition, training moving robots, classification of astronomical structures, and game playing.
- When machine learning is applied to data mining tasks, a model is used to represent the data (such as a graphical structure like a neural network or a decision tree). During the learning process, a sample of the database is used to train the system to properly perform the desired task. Then the system is applied to the general database to actually perform the task. This predictive modeling approach is divided into two phases.
    1. During the **training** phase, historical or sampled data are used to create a model that represents those data. It is assumed that this model is representative not only for this sample data, but also for the database as a whole and for future data as well.
    2. The **testing** phase then applies this model to the remaining and future data.

Types of ML:

- There are various types of Machine Learning: Supervised learning, Semi-supervised learning, Unsupervised learning and Reinforcement learning.
    1. **Supervised Learning:** A supervised approach learns by example. Given a training set of data plus correct answers, the computational model successively applies each entry in the training set. Based on its ability to correctly handle each of these entries, the model is changed to ensure that it works better with this entry if it were applied again. Given enough input values, the model will learn the correct behavior for any potential entry. Examples of supervised learning are classification and regression.

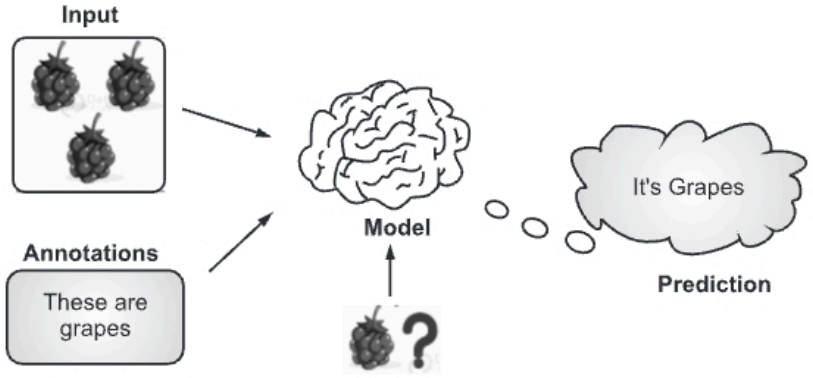


Fig. 2.11 (a): Supervised Learning

    2. **Unsupervised Learning:** With unsupervised data, data exists but there is no knowledge of the correct answer of applying the model to the data.

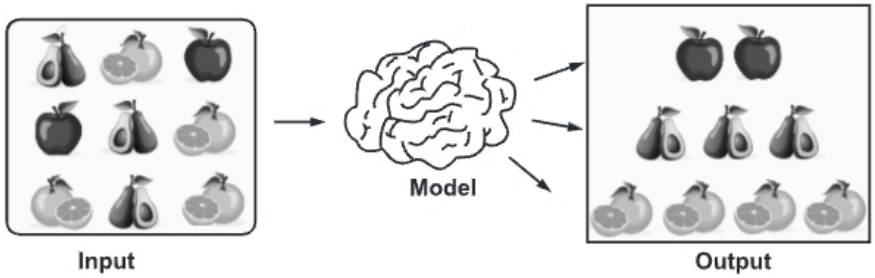


Fig. 2.11 (b): Unsupervised Learning

3. **Semi-supervised Learning:** Semi-supervised learning lies between supervised and unsupervised learning. During training, it uses a smaller labeled data set to guide classification and feature extraction from a larger, unlabeled data set. Semi-supervised learning can solve the problem of having not enough labeled data (or not being able to afford to label enough data) to train a supervised learning algorithm.

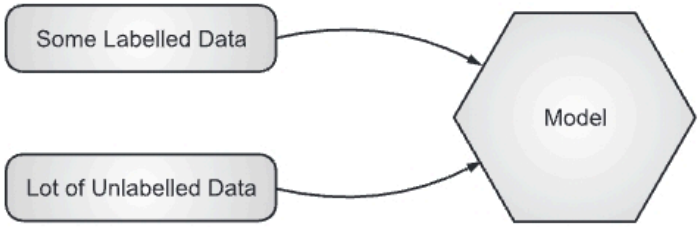


Fig. 2.11 (c): Semi-supervised Learning

4. **Reinforcement Machine Learning:** Reinforcement machine learning is a behavioral machine learning model that is similar to supervised learning, but the algorithm is not trained using sample data. This model learns as it goes by using trial and error. A sequence of successful outcomes will be reinforced to develop the best recommendation or policy for a given problem.

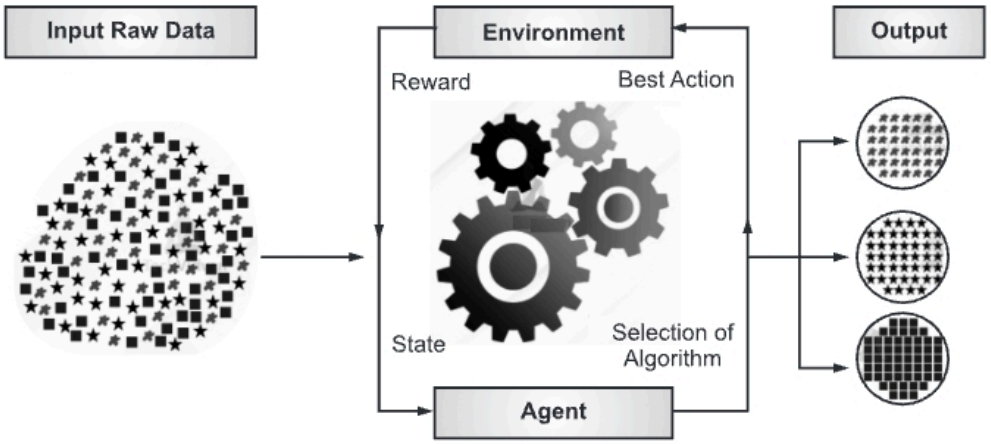


Fig. 2.11 (d): Reinforcement Learning in ML

2.6 INTRODUCTION TO PATTERN MATCHING

- As we know that data mining is all about finding the hidden truth behind the data. It is very interesting that in the large amount of data, some of the data patterns get repeated.
- Pattern matching or pattern recognition is the process to find occurrences of a predefined pattern in data. Pattern matching also finds the number of occurrences of a particular pattern in data.
- Pattern matching is a kind of supervised learning and it identifies the suitable match from different groups of data and enhances the number of events to improve the performance and efficiency of all events. Pattern matching works using the regular expressions used in a programming language.

Applications of Pattern Matching:

- There are many applications of pattern matching. Some of the applications are listed below:
  - In text editing applications, pattern matching is useful in finding occurrences of a string in the text being edited.
  - Web search engines use pattern matching for information retrieval from various data servers.
  - Pattern matching is used in Time series analysis to study the patterns of behavior in data obtained from two different time series to find the similarity.
  - Pattern matching can be treated as classification where the predefined patterns are the classes under consideration. The classification of the data is based on similarity between them.

2.7 CASE STUDY BASED ON SCHEMA DESIGN

Case 1: Autolife Auto Insurance company has a huge generation of data resources from business operations, however. So, to handle this data and to get ease of access for every information; the Company has decided to make major steps for informed decision making and important data analysis.

Objective:

Autolife Company has an OLTP database which keeps records on motor vehicle insurance information. This database contains detailed information in respect of drivers, vehicle and claim information. Current database model has been designed for fast data entry and is sufficient for individual client's specific information as well as fast transaction processing. Company critically needs to make comprehensive analysis like identification of contracts with a high loss ratio and low overall customer value. Appropriate actions needed to be taken with high risk customers, such as premium adjustment, loss prevention measures and in some cases contract cancellations and reduced gross claim expenditures. By making evidenced based management and informed decisions, companies will focus on profitable customers by lowering their premiums and overcoming competitive pressure. It will help companies make better risk management and overall profitability for the company. Autolife Company is in urgent need to utilize the existing data resources efficiently for better risk management and obtain competitive advantage in the Auto Insurance Industry.

Solution:

Autolife Company has decided to implement a Data Warehouse to leverage its data resources. Autolife Company needs to reorganize the existing process of information delivery and to establish one single, unified and integrated data warehouse. A data warehouse is an integrated subject oriented, time-variant, non-volatile database that provides support for decision making.

In order to support decision making Autolife Company decided to reorganize the data into Star Schema in the Data warehouse. In effect, the star schema creates a near equivalent multidimensional database schema from the existing OLTP relational database. It will help in advance data analysis for Risk management and overcoming competitive pressure.

Structure of Star Schema:

Star schema yields an easily implemented model for multidimensional data analysis while still preserving the relational structure on which the operation database is built. The basic star schema has four components: facts, dimensions, attributes and attributes hierarchies. The Star schema would most likely be a read-only database due to the widespread redundancy introduced into the model.

Fact Table:

Autolife Company has factual data in Claim Information such as date, location, type of accident, cause of accident, liability, recovery cost. Fact tables contain the quantitative data or factual data about a business. This information is numerical, additive measurements and can consist of many columns and millions or billions of rows.

Dimensions:

Claim Information facts can be analysed by dimensions such as Driver, Location, Time, and Automotive. Dimension tables are usually smaller and hold descriptive data that reflects the dimensions.

Attributes:

For example, Driver name, Driver ID, gender, age group, race, and other attributes. Some of these attributes might relate to each other hierarchically.

Attribute Hierarchies:

Provide top-down data Aggregation, Drill down or rol-up data analysis. For example, in time dimension there are Attribute hierarchies such as day, week, month, quarter, and year. When the decision maker wants to see company yearly claim information then they are using year hierarchy level. They can further drilldown to quarter level sales quantity, as per their needs. Same as in Location dimension, data can be analysed by Country, Region, Province City and Town.

Benefits of Data warehouse to Autolife Company:

By organizing the Autolife Company data around star scheme companies can analyse information like what customers are high risk and what group of customers is profitable. What cities have more accidents ratios and what time of the year accidents happen? What habit of drivers may be considered high risk? What vehicles are considered low risk ? and so on.

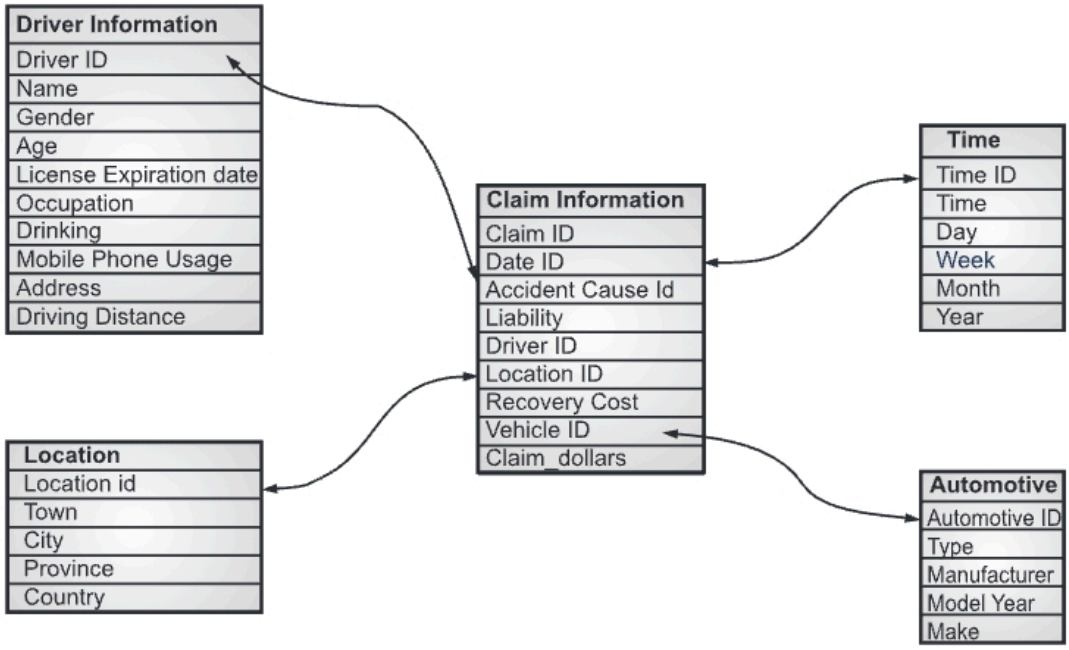


Fig. 2.12: Schema design for the Autolife Company

Summary

- A data warehouse is subject-oriented (can be used to analyse a particular area or domain), integrated (It integrates multiple data sources in a single, time-variant (Historical data is stored in a data warehouse).
- The basic components of data warehousing includes data migration, the warehouse and access tools.
- Typical data warehouse architecture is three-tier architecture: bottom tier, middle tier and top tier.
- The process of structuring your raw data is called data modeling.
- OLAP (Online Analytical Processing) is a multidimensional database system which is able to analyse many data records quickly.
- A data mart is subset of a data warehouse which is useful to a specific group of users. It is confined to a specific domain.
- The categories of data mart are dependent, independent and hybrid.
- A fact table is the central table in a star schema of a data warehouse. A fact table stores quantitative information for analysis and is often denormalized.
- A dimension table is a table in a star schema of a data warehouse. A dimension table stores attributes, or dimensions, that describe the objects in a fact table.
- A data warehouse is based on a multidimensional data model which views data in the form of a data cube.
- OLAP Operations are done on multidimensional data Roll-up these operations are Roll-up, Drill-down, Slice and dice and Pivot (rotate).

- A database schema is the logical structure of the database. The database uses a relational model whereas in case of data warehouse, it uses star and snowflakes schema.
- There are two types of schemas in data warehouse: Star schema and snowflake schema.
- Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information
- Pattern matching is a kind of supervised learning and it identifies the suitable match from different groups of data and enhances the number of events to improve the performance and efficiency of all events.

Check Your Understanding

1. OLAP stands for _____.
   (a) Online Analytical Processing (b) Online Analysis Processing
   (c) Online Transaction Processing (d) Online Aggregate Processing
2. Data that can be modeled as dimension attributes and measure attributes are called _____ data.
   (a) Multidimensional (b) Single dimensional
   (c) Measured (d) Dimensional
3. What do data warehouses support?
   (a) OLAP (b) OLTP
   (c) OLAP and OLTP (d) Operational databases
4. _____ is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management decisions.
   (a) Data Mining (b) Data Warehousing
   (c) Web Mining (d) Text Mining
5. The data Warehouse is _____ .
   (a) read only (b) write only
   (c) read write only (d) None of the mentioned
6. The important aspect of the data warehouse environment is that data found within the data warehouse is _____ .
   (a) subject-oriented (b) time-variant
   (c) integrated (d) All of the above
7. _____ describes the data contained in the data warehouse.
   (a) Relational data (b) Operational data
   (c) Metadata (d) Informational data

8. What is Machine learning?
   (a) The autonomous acquisition of knowledge through the use of computer programs.
   (b) The autonomous acquisition of knowledge through the use of manual programs.
   (c) The selective acquisition of knowledge through the use of computer programs.
   (d) The selective acquisition of knowledge through the use of manual programs.
9. What is true about Machine Learning?
   (a) Machine Learning (ML) is the field of computer science.
   (b) ML is a type of artificial intelligence that extracts patterns out of raw data by using an algorithm or method.
   (c) The main focus of ML is to allow computer systems to learn from experience without being explicitly programmed or human intervention.
   (d) All of the above.
10. _____ is a good alternative to the star schema.
   (a) Star schema (b) Snowflake schema
   (c) Fact constellation (d) Star-snowflake schema
11. Which of the following is not a kind of data warehouse application?
   (a) Information processing (b) Analytical processing
   (c) Data mining (d) Transaction processing

Answers

1. (a)	2. (a)	3. (a)	4. (b)	5. (a)	6. (d)	7. (c)	8. (a)	9. (d)	10. (c)
11. (d)									

Practice Questions

Q.I Answer the following questions in short.

1. What is the need of a data warehouse?
2. What are OLTP and OLAP database systems?
3. List the major steps involved in the ETL process.
4. What is the need for a separate database for decision makers?
5. What is a data warehouse?
6. What are the benefits of building an enterprise data warehouse?
7. List major components of any data warehouse system.
8. List the characteristics of OLAP systems.

Q.II Answer the following questions.

1. What is the major difference between the star schema and the snowflake schema?
2. List some differences between an OLTP system and a data warehouse system.
3. Describe the features of a data warehouse.
4. What are the major differences between OLTP and a data warehouse system?
5. Explain the star scheme technique of modeling a data warehouse.
6. Explain a multidimensional view and a data cube.
7. Describe in detail the concept of Machine Learning.

Q.III Define the terms.

1. Star schema
2. Snowflake schema
3. OLAP Cube
4. Supervised learning
5. Unsupervised learning
6. Pattern Matching

3... Classification

Learning Objectives…

- To know the concept of classification.
- To learn about Decision tree.
- To get information of Rule-based Classification.
- To study Bayes Classification Methods and Bayesian Networks.
- To know about Parameter and structure learning, Linear classifier, Perception, k-Nearest-Neighbour Classifiers.
- To get knowledge of SVM classifiers.
- To learn about Regression and Prediction.

3.1 INTRODUCTION

- There are two forms of data analysis that can be used for extracting models. One is describing important classes or to predict future data trends. These two forms are as follows:
  - Classification
  - Prediction

Definition of Classification:

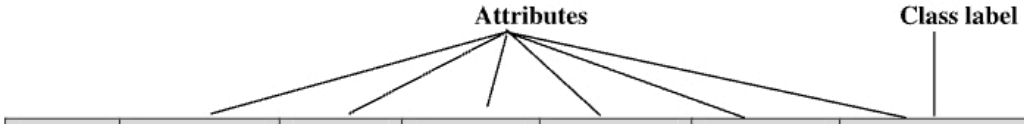
- As the name suggests, classification is the process of classifying the data. It is a data mining technique which is done for analysis of the data. It is the process of finding the model that defines the classes and their concepts. It identifies and categorizes the sub population of the data.
- The main goal of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data.
- The classification can be defined as,

  "Given a database $D = \{t_1, t_2, \ldots, t_n\}$ of tuples (items, records) and a set of classes $C = \{C_1, \ldots, C_m\}$, the classification problem is to define a mapping $f : D \rightarrow C$ where, each t_i is assigned to one class. A class, C_j, contains precisely those tuples mapped to it; that is,

  $C_j = \{t_i \mid f(t_i) = C_j, 1 \Leftarrow i \Leftarrow n, \text{ and } t_i \in D\}$."
- As defined, classification is a mapping from the database to the set of classes. These classes are predefined. Each data tuple is assigned to exactly one class. The classification can be implemented in two steps:

1. The first step is to create a model by examining the training data. The model is created by giving the training data as input data. These models are treated as classification rules, decision trees.
2. In the second step, this model is applied for the purpose of classification of the unknown data objects (also called as test data set). The results of unknown data sets are compared with the model created in the first step with the training data set. The training data set is always different from the test data set.

Example of Classification - Training Data Set:



Patient Id	Sore throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1.	Yes	Yes	Yes	Yes	Yes	Strep throat
2.	No	No	No	Yes	Yes	Allergy
3.	Yes	Yes	No	Yes	No	Cold
4.	Yes	No	No	No	No	Strep throat
5.	No	Yes	No	Yes	No	Cold
6.	No	No	No	Yes	No	Allergy
7.	No	No	Yes	No	No	Strep throat
8.	Yes	No	No	Yes	Yes	Allergy
9.	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

Supervised Learning (Classification):

- Since class label is provided it is known as Supervised Learning.
- Typically the model is represented in the form of classification rules, decision trees or mathematical formulae.

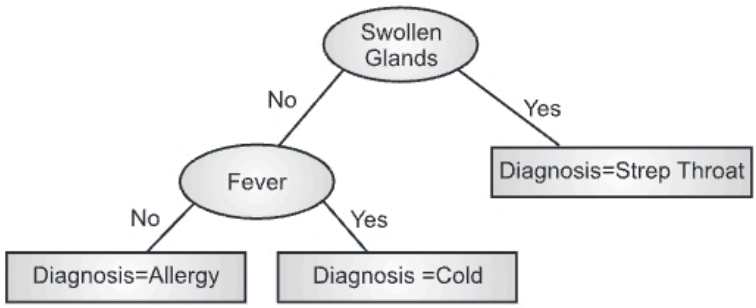
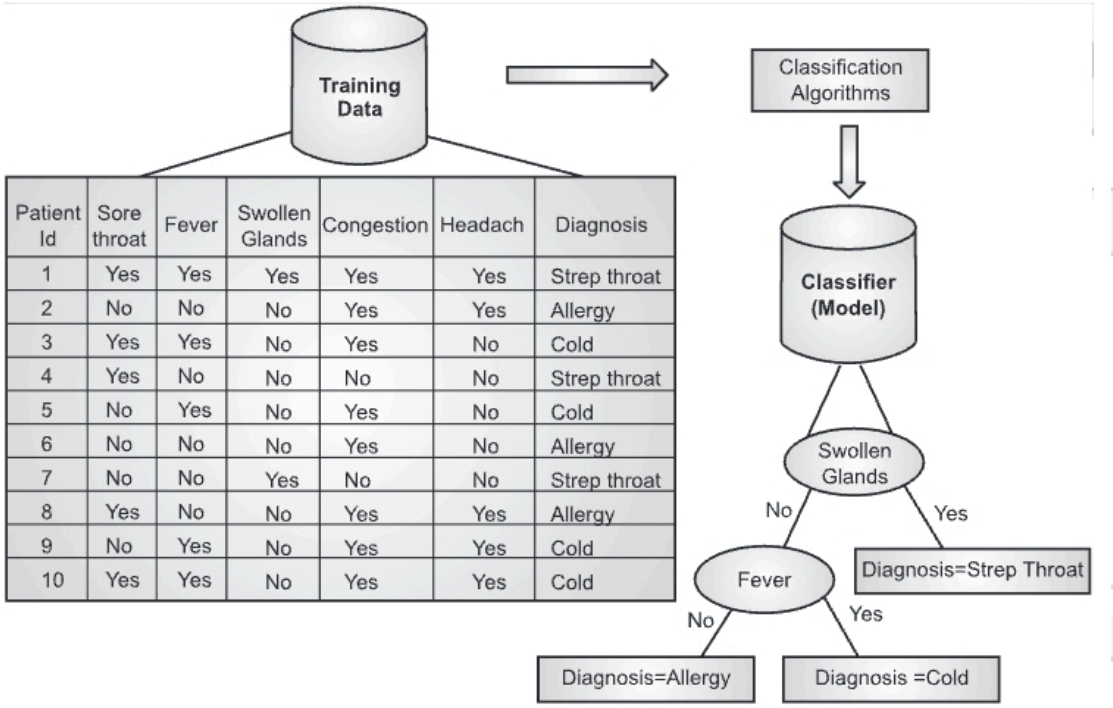


Fig. 3.1 (a): Classification Model

- In the second step, the model is used for classification. First it is used on a test data to check its accuracy then it can be used to classify future data tuples whose class label values are not known.



Patient Id	Sore throat	Fever	Swollen Glands	Congestion	Headach	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	No	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

Fig. 3.1 (b): Implementation of Classification

3.1.1 Methods to Solve Classification Problems

- There are three basic ways to solve a classification problem.
  1. Specifying boundaries.
  2. Using probability distribution.
  3. Using posterior probabilities.

1. **Specifying boundaries:** Here classification is performed by dividing the input space of potential database tuples into regions where each region is associated with one class.
2. **Using probability distributions:** For any given class, C_j, $P(t_i \mid C_j)$ is the PDF for the class evaluated at one point, t_i. If a probability of occurrence for each class $P(C_j)$ is known (perhaps determined by a domain expert), then $P(C_j)\ P(t_i \mid C_j)$ is used to estimate the probability that t_i is in class C_j.
3. **Using posterior probabilities:** Given a data value t_i, we would like to determine the probability that t_i is in a class C_j. This is denoted by $P(C_j \mid t_i)$ and is called the posterior probability. One classification approach would be to determine the posterior probability for each class and then assign t_i to the class with the highest probability.

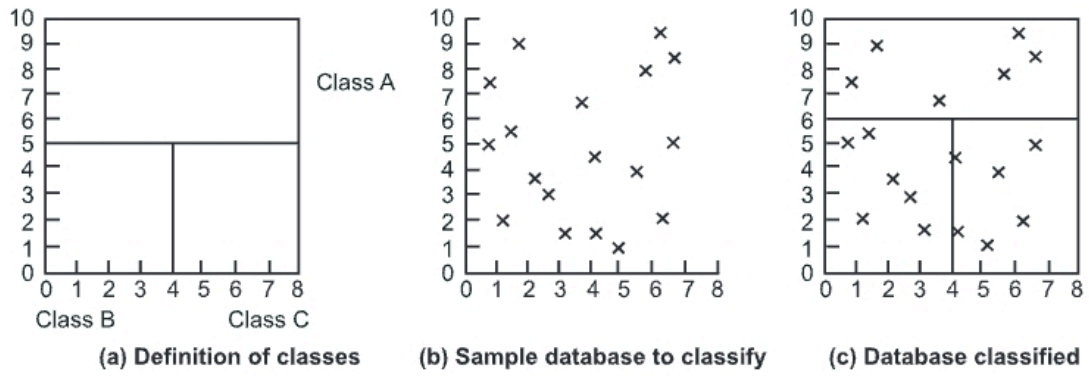


Fig. 3.2: Methods to solve a Classification Problem

3.1.2 Applications of Classification

The classification process has many applications. Some of them are listed below:

- **Credit approval:** Applicant as good or poor credit risk.
- **Target marketing:** Profile of a good customer.
- **Medical diagnosis:** Develop a profile of stroke victims. Cancer tumor cell identification.
- **Fraud detection:** Determine a credit card purchase is fraudulent.
- **Email spam classification:** Filter the spam e-mail automatically.
- **Banking:** Bank customer's loan pay willingness prediction.

3.1.3 Classifier

The algorithm which implements the classification on a dataset is known as a Classifier. There are two types of Classifier:

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

   Examples: SPAM or NOT SPAM, TRUE or FALSE, YES or NO, MALE or FEMALE, etc.

2. **Multi-class Classifier:** If a classification problem has more than two outcomes then it is called as Multi-class Classifier.

   Example: Classifications of types of flowers, Classification of types of movies.

3.1.4 Classification Algorithms

- The main aim of the Classification algorithm is to identify the category of a given dataset, and these algorithms are mainly used to predict the output for the categorical data.
- There are many classification algorithms based on categorisation. These are as shown in the Fig. 3.3.

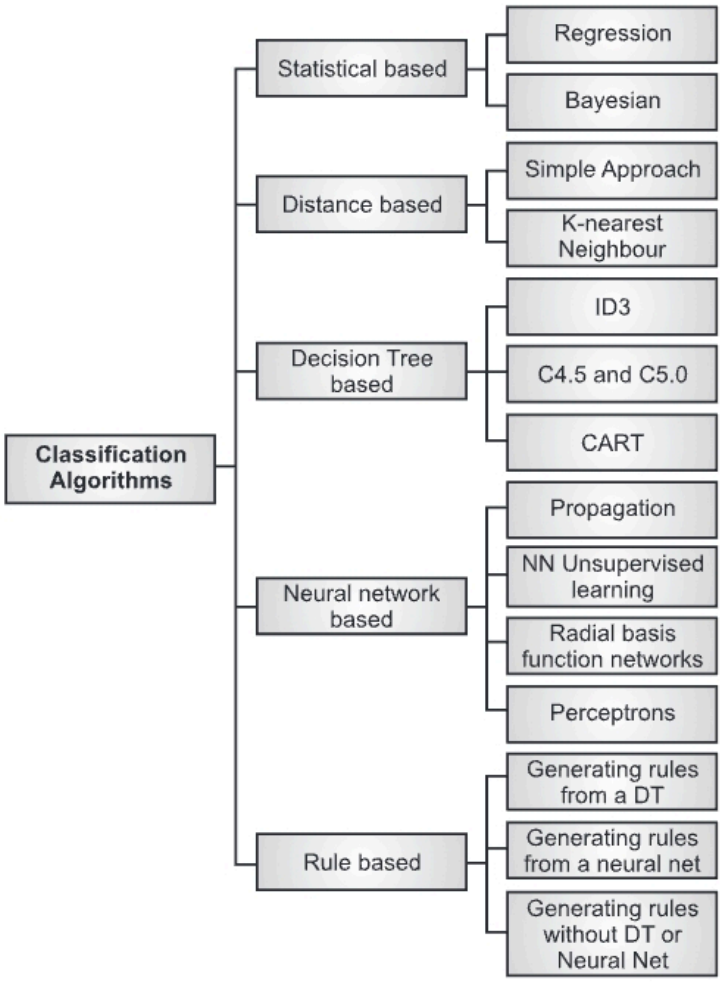


Fig. 3.3: Types of Classification Algorithms

1. **Statistical based classification:** Statistical classification is the broad supervised learning approach that trains a program to categorize new, unlabeled information based upon its relevance to known, labeled data.
2. **Distance based classification:** Distance based algorithms are nonparametric methods that can be used for classification. These algorithms classify objects by the difference between them as measured by distance functions.
3. **Decision tree based classification:** Decision tree builds classification in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. It can handle both categorical and numerical data.
4. **Neural network based classification:** Artificial neural networks are relatively basic electronic networks of neurons based on the neural structure of the brain. They process records one at a time, and learn by comparing their classification of the record (i.e., largely arbitrary) with the known actual classification of the record.

5. **Rule based classification:** This classification scheme make use of IF-THEN rules for class prediction.

3.2 DECISION TREE

3.2.1 Introduction

- A decision tree is the approach, where a classification process is modelled as a tree. Once a tree is constructed, it is applied to each and every tuple in the database. The result is a classification. Decision trees are the most useful approach in the classification problem.
- There are two basic steps involved in the decision tree classification process.
  1. Creation of the tree.
  2. Application of the tree to the database.
- Decision tree is hierarchical structure consisting of nodes and directed edges. The tree has three types of nodes:
  1. **Root node:** The topmost node in the tree is the root node. It has no incoming edges and zero or more outgoing edges.
  2. **Internal Node:** It denotes a test on an attribute. This node has exactly one incoming and two or more outgoing edges.
  3. **Leaf node(terminal node):** It holds a class label. This node has exactly one incoming and no outgoing edges.
- Each branch/edge denotes the outcome of a tree.
- Creating or building a tree is a complex process as compared to applying the tree to the database.

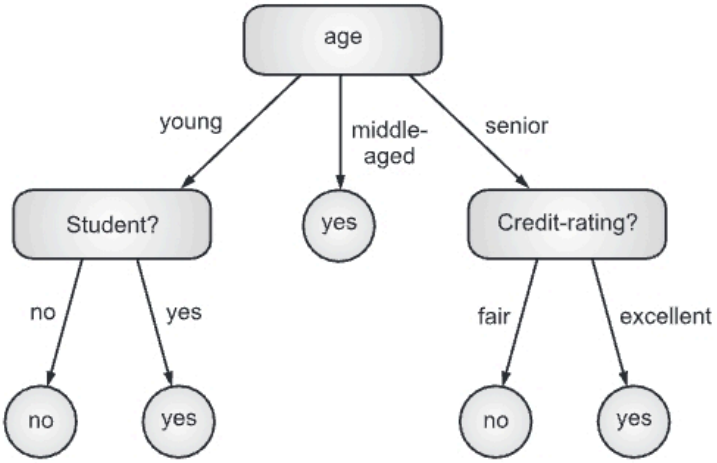


Fig. 3.4: Structure of decision tree

Definition: Decision Tree (DT):

Given a database $D = \{t_1, \ldots, t_n\}$ where $t = (t_{i1}, \ldots, t_{ih})$ and the database schema contains the following attributes $\{A_1, A_2, \ldots, A_h\}$. Also given is a set of classes $C = \{C_j, \ldots, C_m\}$.

- A decision tree (DT) or classification tree is a tree associated with D that has the following properties:
  - Each internal node is labeled with an attribute, Ai.

- Each arc is labeled with a predicate that can be applied to the attribute associated with the parents.
- Each leaf node is labeled with a class, C_j.

Advantages of Decision Tree:

1. DTs are easy to use and efficient.
2. Rules can be generated that are easy to interpret and understand.
3. They can handle nonlinear parameters easily.
4. The scalability of DTs is good as the tree size is independent of the database size.
5. Trees can be constructed for many attributes in the tree.
6. It is not necessary to normalise the data.
7. It requires less time.

Disadvantages of Decision Tree:

1. They do not easily handle continuous data. These attribute domains must be divided into categories to be handled.
2. The mathematical calculation of the decision tree mostly requires more memory.
3. The mathematical calculation of the decision tree is time consuming.
4. The space and time complexity of the decision tree model is relatively higher.
5. Decision tree model training time is more as complexity is high.

3.2.2 Construction Principle

- A decision tree is a supervised learning where we train the data. Here, the algorithm divides the dataset into subsets according to the most significant attribute in the database. It is labelled as root node. And from this root node the dataset is splitted into subsets. This splitting is known as decision nodes. If there is no possibility to divide the decision node, it is known as the leaf node.

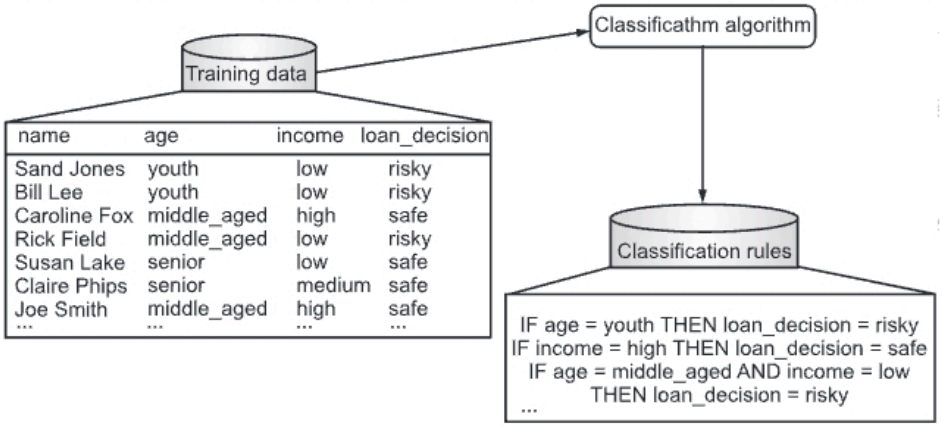
Algorithm for Decision Tree building:

```
Input : D //Training data
Output : T //Decision tree
DTBuild algorithm://simple algorithm to illustrate naive approach to building DT
1. T = 0 ;
2. Determine best splitting criterion ;
3. T = Create root node and label with splitting attribute ;
4. T = Add arc to root node f o r each split predicate and label ;
5. for each arc do
6. D = Database created by applying splitting predicate to D ;
7. if stopping point reached for this path, then
8. T = Create leaf node and label with appropriate class ;
9. else 1
10. T = DTBuild(D) ;
11. T = Add T to arc ;
```

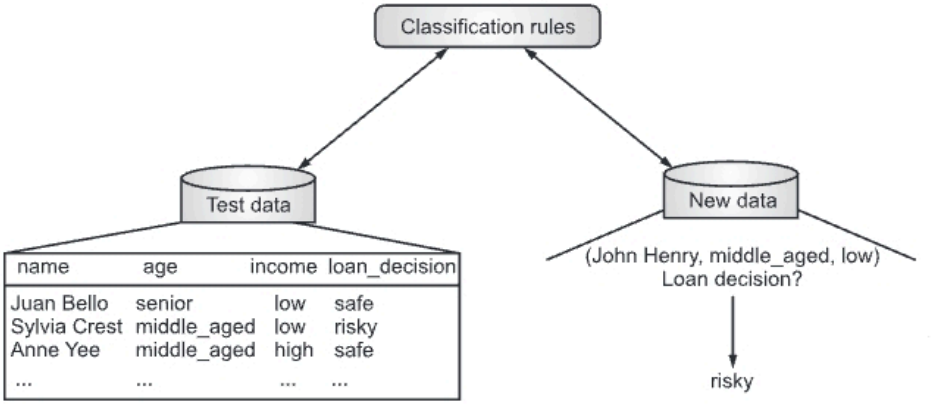
- The algorithm designed for the decision, decides how to identify the attribute and how the splitting should be done. The process is a top-down approach. The top region presents all the observations at a single place and then it is splitted into two or more branches which may further split.
- In this approach, it considers the current node and future nodes. The decision tree algorithms will continue running until stop criteria such as the minimum number of observations etc. is reached.
- When a decision tree is constructed, many nodes may contain outliers or noisy data. To remove this noisy data or outliers, a tree pruning method is applied. Thus, removing unwanted data improves the accuracy of the tree.
- Some of the decision tree algorithms include Hunt's Algorithm, ID3, CD4.5, and CART.

Two-step process to construct classification model:

1. **Learning Step:** The training data is fed into the system to be analyzed by a classification algorithm. In following example, the class label is the attribute i.e. "loan decision". The model built from this training data is represented in the form of decision rules.
2. **Classification:** Test dataset are fed to the model to check the accuracy of the classification rule. If the model gives acceptable results then it is applied to a new dataset with unknown class variables.



(a) Learning step



(b) Classification

Fig. 3.5

Predictors				Target
Outlook	**Temp.**	**Humidity**	**Windy**	**Play Golf**
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mlid	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

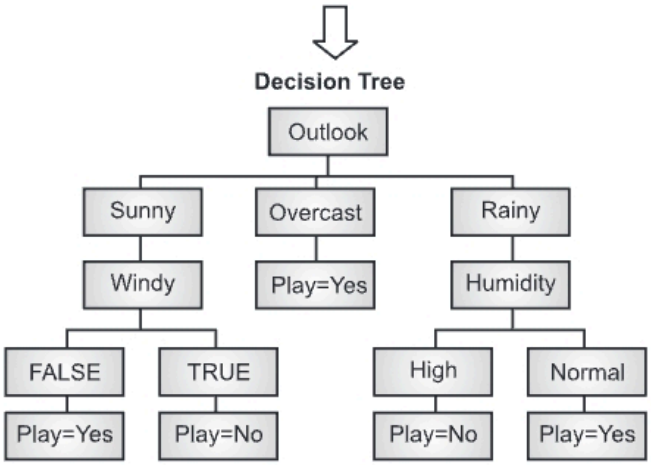


Fig. 3.6: Example of decision tree with dataset.

3.2.3 Attribute Selection Measures

- An attribute selection measure is a heuristic for selecting the splitting criterion that "best" separates a given data partition, D, of class-labeled training tuples into individual classes.
- If we were to split D into smaller partitions according to the outcomes of the splitting criterion, ideally each partition would be pure (i.e., all the tuples that fall into a given partition would belong to the same class). Conceptually, the "best" splitting criterion is the one that most closely results in such a scenario. Attribute selection measures are also known as splitting rules because they determine how the tuples at a given node are to be split.
- The attribute selection measures provide the ranking for every attribute. The attribute which is having the best rank will be selected as splitting attribute. If the splitting attribute is continuous-valued or if we are restricted to binary trees, then, respectively, either a split point or a splitting subset must also be determined as part of the splitting criterion.

- The tree node created for partition D is labelled with the splitting criterion, branches are grown for each outcome of the criterion, and the tuples are partitioned accordingly.

Measures of Attribute selection:

- Following are Attribute selection measures:

(a) Entropy: Entropy is the measurement of impurities or randomness in the data points. If a dataset contains homogeneous subsets of observations, then no impurity or randomness is there in the dataset, and if all the observations belong to one class, the entropy of that dataset becomes zero.

(b) Information gain: It is an attribute selection measure used by ID3. Here, the attribute with maximum information gain is selected. This attribute minimizes the information needed to classify the tuples in the resulting partitions and reflects the least randomness or "impurity" in these partitions. Such an approach minimizes the expected number of tests needed to classify a given tuple and guarantees that a simple (but not necessarily the simplest) tree is found.

Information Gain = Entropy before splitting – Entropy after splitting

Given a probability distribution such that,

$$P = (p_1, p_2, \ldots, p_n)$$

Where, p_i is the probability of a data point in the subset of D_i of a dataset D,

Therefore, it is defined as,

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

(c) Gini index: The Gini coefficient (Gini index or Gini ratio) is a statistical measure. It computes the degree of probability of a specific variable that is wrongly being classified when chosen randomly and a variation of gini coefficient. It works on categorical variables, provides outcomes either be "successful" or "failure" and hence conducts binary splitting only.

In 1984, it was proposed by Leo Breiman as an impurity measure for decision tree learning. It is given by the following equation/formula;

$$\text{Gini}(P) = \sum_{i=1}^{n} p_i(1 - p_i) = 1 - \sum_{i=1}^{n} (p_i)^2$$

Where $P = (p_1, p_2, \ldots, p_n)$, and pi is the probability of an object that is being classified to a particular class.

- The degree of Gini index varies from 0 to 1,
  - Where, value as 0 represents that all the elements be allied to a certain class, or only one class exists there.
  - Value as 1 indicates that all the elements are randomly distributed across various classes.
  - A value of 0.5 denotes the elements are uniformly distributed into some classes.

(d) Gain ratio:

- In decision tree learning, Gain Ratio or Uncertainty Coefficient is used to normalize the information gain of an attribute against how much entropy that attribute has.
- It is proposed by John Ross Quinlan.
- Formula of Gain ratio is given by,

Gain Ratio = Information Gain/Entropy

- From the above formula, it can be stated that if entropy is very small, then the gain ratio will be high and vice versa.
- Be selected as splitting criterion, Quinlan proposed following procedure,
  1. First, determine the information gain of all the attributes, and then compute the average information gain.
  2. Second, calculate the gain ratio of all the attributes whose calculated information gain is larger or equal to the computed average information gain, and then pick the attribute of higher gain ratio to split.

3.2.4 Tree Pruning

- The pruning of tree pruning is a process where the anomalies in the training data are removed due to outliers or noise in the data.
- The pruned trees are beneficial as the pruned trees are smaller and less complex.
- Also, decision trees that are trained on any training data do have the feat of overfitting the training data. So, pruning will remove the overfitting problem as well.
- The pruning process cuts off or crops the lower ends of the tree so as to make it simple and less complex.

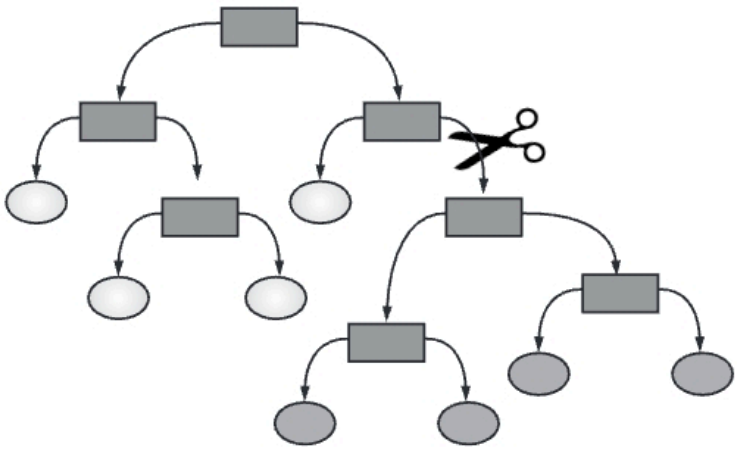


Fig. 3.7 (a): Pruning Process

- The example is given in the figure below:

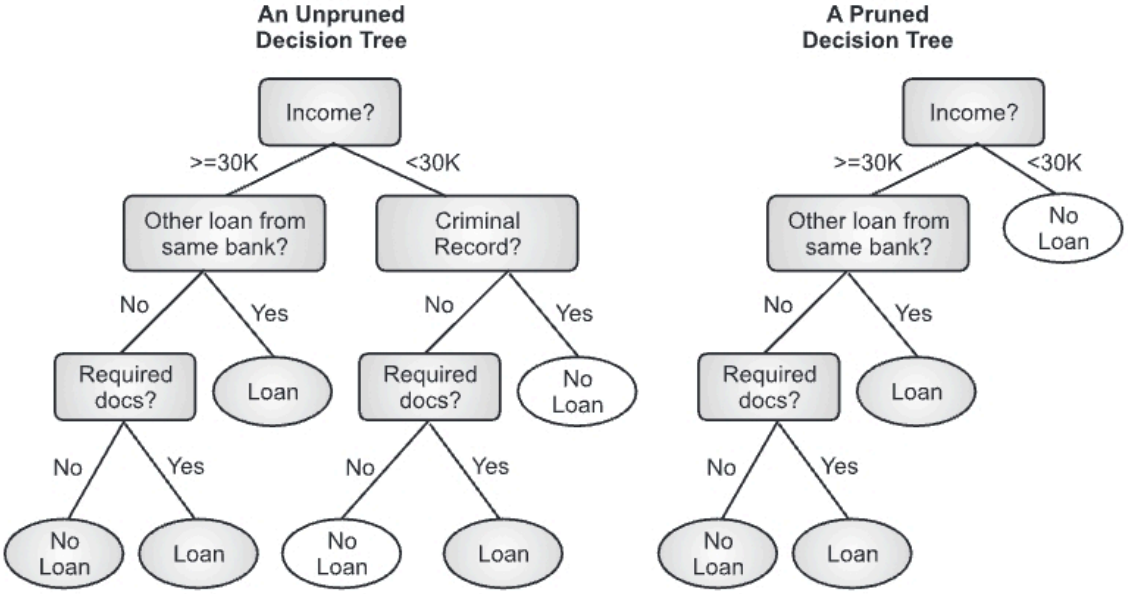


Fig. 3.7(b): Unpruned and Pruned Decision Tree

- There are two methods of decision tree pruning:
  - **(a) Pre-pruning:** The tree is pruned by halting its construction early.
  - **(b) Post-pruning:** This approach removes a sub-tree from a fully grown tree.

3.3 RULE-BASED CLASSIFICATION

- The very simple approach to perform classification is to apply if-then rules that cover all the possible cases. Here the learned model is represented as if- then rules. Here, at first, rules that are used for classification are examined. Then, the method with which these rules are generated are studied. These rules are either generated from a DT or directly from training data.

3.3.1 Using IF-THEN Rules for Classification

- If- then approach is a very simple approach. Likewise we do the classification of results based on the percentage of the student. If percentage > 90 then its Outstanding, percentage > 80 then its Distinction, if percentage > 60 then result is First class and so on…
- A if - then approach based classification rule r consists of **if** (predicate) part and **then** part i.e. the consequences of predicate.

```
r = (a, c)
If (condition) then (conclusion)
```

- Here r is the classification rule, `a` is predicate and `c` is consequence of the predicate. The predicate is evaluated as true or false against each tuple in the dataset.
- The `if` part i.e. predicate is known as **antecedent** or precondition and the conclusion part or consequence part is known as rule **consequent**. In the rule part, there might be one or more than one conditions that can be logically ANDed.

- These rules correspond to the Decision Tree that can be created. DT can always be used to generate rules, but they are not equivalent.

3.3.2 Rule Extraction from a Decision Tree

- The process to generate a rule from a DT is straightforward and is outlined in Algorithm. This algorithm will generate a rule for each leaf node in the decision tree. All rules with the same consequent could be combined together by ORing the antecedents of the simpler rules.

Algorithm:

```
Input :
T // Decision tree
Output :
R // Rules
Gen algorithm:
/ Illustrate simple approach to generating classification rules from a DT
   1. R = 0
   2. for each path from root to leaf in T do
   3. a = True
   4. for each non - leaf node do
   5. a = a ∧ (label of node combined with label of incident outgoing arc)
   6. c = label of leaf node
   7. R = R U r = (a, c)
```

- Using this algorithm, the following rules are generated for the DT in Fig. 3.8:

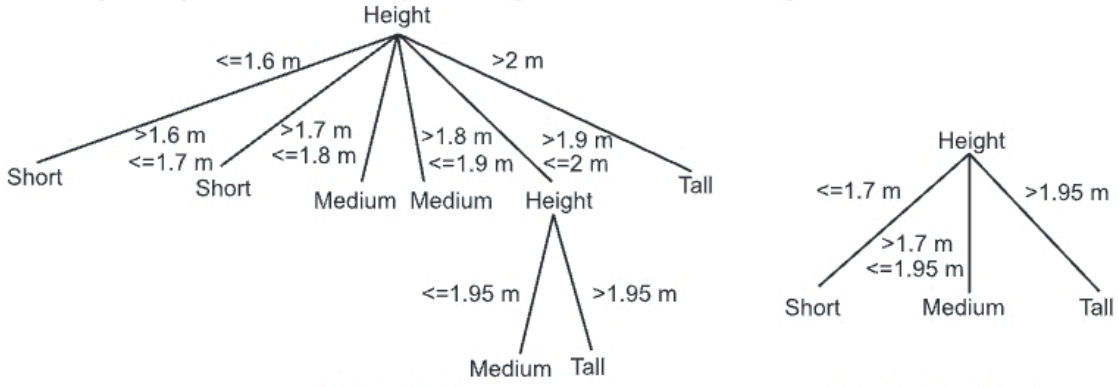


Fig. 3.8: Rule-extraction for DT

```
{ ( (Height ≤ 1.6 m), Short)
( ((Height > 1.6 m) ∧ (Height ≤ 1.7 m)) , Short)
(((Height > 1.7 m) ∧ (Height ≤ 1.8 m)), Medium)
(((Height > 1.8 m) ∧ (Height ≤ 1.9 m)), Medium)
( ((Height > 1.9 m) ∧ (Height ≤ 2 m) ∧ (Height ≤ 1.95 m)), Medium)
```

(((Height > 1.9 m) ∧ (Height ≤ 2 m) ∧ (Height > 1.95 m)), Tall)
((Height > 2 m), Tall) }

- An optimized version of these rules is then:

{ ((Height ≤ 1 .7 m), Short)
(((Height > 1 .7 m) ∧ (Height ≤ 1 .95 m)), Medium)
((Height > 1 .95 m), Tall) }

3.4 BAYES CLASSIFICATION METHODS

- **Bayesian classification** is a probabilistic approach to learning and inference based on a different view of what it means to learn from data, in which probability is used to represent uncertainty about the relationship being learnt.
- Bayesian classification is based on Bayes' Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.
- **Naive Bayes** is a type of **classifier** which uses the **Bayes'** Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class. The class with the highest probability is considered as the most likely class.

3.4.1 Bayes' Theorem

- Bayes' Theorem is named after Thomas Bayes. There are two types of probabilities:
  1. Posterior Probability [P(H/X)]
  2. Prior Probability [P(H)]

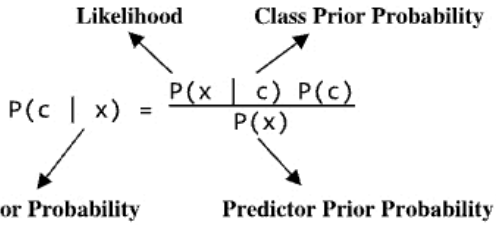
  Where, X is a data tuple and H is some hypothesis.
- According to Bayes' Theorem,

$$P(H/X) = \frac{P(X/H)P(H)}{P(X)}$$

3.4.2 Naïve Bayesian Classification

- Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.
- It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.
- For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.
- Naive Bayes' model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

- Bayes' theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



$$P(c \mid x) = P(x_1 \mid c) \times P(x_2 \mid c) \times \ldots \times P(x_n \mid c) \times P(c)$$

Where,

- $P(c|x)$ is the posterior probability of class (c, target) given predictor (x, attributes).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of predictor given class.
- $P(x)$ is the prior probability of *predictor*.

3.4.2.1 How does the Naive Bayes algorithm work?

- Let us understand it using an example given below.

Example: We have a training data set of weather and corresponding target variable 'Play' (suggesting possibilities of playing). Now, we need to classify whether players will play or not based on weather condition.

Solution: Let's follow the below steps to perform it.

Step 1: Convert the data set into a frequency table.

Step 2: Create Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.

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

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

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

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

3.4.2.2 Applications of Naive Bayes Algorithms

- **Real time Prediction:** Naive Bayes is an eager learning classifier and it is sure fast. Thus, it could be used for making predictions in real time.
- **Multi class Prediction:** This algorithm is also well known for multi class prediction feature. Here we can predict the probability of multiple classes of target variable.
- **Text classification/ Spam Filtering/ Sentiment Analysis:** Naive Bayes classifiers mostly used in text classification (due to better result in multi class problems and independence rule) have higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments).
- **Recommendation System:** Naive Bayes Classifier and Collaborative Filtering together builds a Recommendation System that uses machine learning and data mining techniques to filter unseen information and predict whether a user would like a given resource or not.

3.5 BAYESIAN NETWORKS

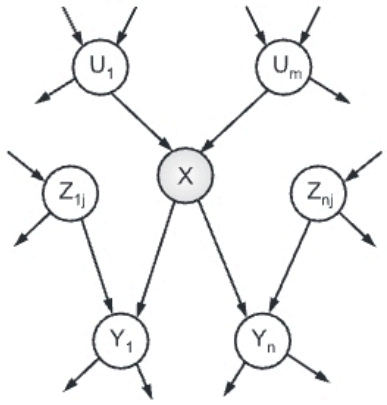


Fig. 3.9: Bayesian Network

- **Bayesian networks** are a type of probabilistic graphical model that uses Bayesian inference for probability computations. Bayesian networks aim to model conditional dependence, and therefore causation, by representing conditional dependence by edges in a directed graph. Through these relationships, one can efficiently conduct inference on the random variables in the graph through the use of factors.
- Using the relationships specified by our Bayesian network, we can obtain a compact, factorized representation of the joint probability distribution by taking advantage of conditional independence.

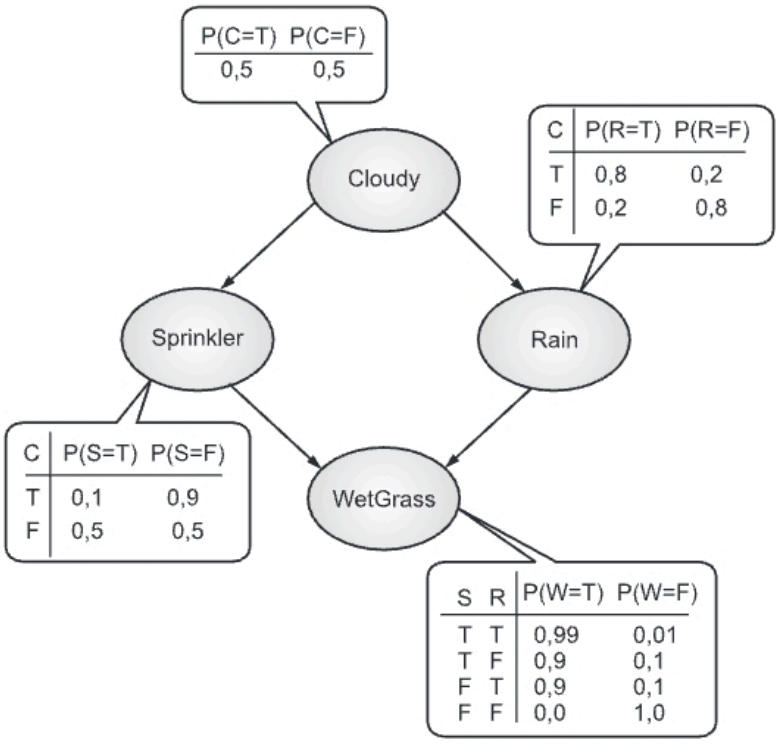


Fig. 3.10: Representation of Probability Distribution

- A Bayesian network is a **directed acyclic graph** in which each edge corresponds to a conditional dependency, and each node corresponds to a unique random variable. Formally, if an `edge (A, B)` exists in the graph connecting random variables A and B, it means that `P(B|A)` is a **factor** in the joint probability distribution, so we must know `P(B|A)` for all values of B and A in order to conduct inference.
- In the above example, since Rain has an edge going into `WetGrass`, it means that `P(WetGrass|Rain)` will be a factor, whose probability values are specified next to the `WetGrass` node in a conditional probability table.

3.6 PARAMETER AND STRUCTURE LEARNING

Parameter Learning:

- Parameter learning is the process of using data to learn the distributions of a Bayesian network or Dynamic Bayesian network. Generally, a Bayesian network is illustrated by an expert and is then used to perform inference.
- In order to fully specify the Bayesian network and thus fully represent the joint probability distribution, it is necessary to specify for each node X the probability distribution for X conditional upon X's parents. The distribution of X conditional upon its parents may have any form. It is common to work with discrete or Gaussian distributions since that simplifies calculations.

- Sometimes, only constraints on a distribution are known then one can use the principle of maximum entropy to determine a single distribution, the one with the greatest entropy given the constraints. (Analogously, in the specific context of a dynamic Bayesian network, the conditional distribution for the hidden state's temporal evolution is commonly specified to maximize the entropy rate of the implied stochastic process.)
- Often these conditional distributions include parameters that are unknown and must be estimated from data, e.g., via the maximum likelihood approach. Direct maximization of the likelihood (or of the posterior probability) is often complex given unobserved variables.
- A classical approach to this problem is the expectation-maximization algorithm. This algorithm alternates computing expected values of the unobserved variables conditional on observed data, with maximizing the complete likelihood (or posterior) assuming that previously computed expected values are correct. Under mild regularity conditions, this process converges on maximum likelihood (or maximum posterior) values for parameters.
- A fully Bayesian approach to parameters is to treat them as additional unobserved variables and to compute a full posterior distribution over all nodes conditional upon observed data, then to integrate out the parameters. This approach can be expensive and lead to large dimension models, making classical parameter-setting approaches more tractable.

Structure learning:

- The task of structure learning for Bayesian networks refers to learn the structure of the Directed Acyclic Graph (DAG) from data. There are two methods for the structure learning: Score-based approach and Constraint-based approach.
  1. **Score-based approach:** The score-based approach first defines a criterion to evaluate how well the Bayesian network fits the data, then searches over the space of DAGs for a structure with maximal score. The score-based approach is a search problem and consists of the definition of score metric and the search algorithm.
  2. **Constraint-based approach:** The constraint-based case employs the independence test to identify a set of edge constraints for the graph and then finds the best DAG that satisfies the constraint. This approach works well with some other prior (expert) knowledge of structure but requires lots of data samples to guarantee testing power. So it is less reliable when the number of samples is small.

3.7 LINEAR CLASSIFIER

- Linear classifiers classify data into labels based on a linear combination of input features. Therefore, these classifiers separate data using a line or plane or a hyperplane (a plane in more than 2 dimensions).
- They can only be used to classify data that is linearly separable. They can be modified to classify non-linearly separable data.
- There are two major algorithms for linear binary classification:
  (a) Perceptron
  (b) SVM

3.8 PERCEPTRON

- The simplest Neural Network is called a Perceptron. A Perceptron is a single neuron with multiple inputs and one output.
- A simple perceptron can be used to classify into two classes. It takes an input, aggregates it (weighted sum) and returns 1 only if the aggregated sum is more than some threshold, else it returns 0. In the context of supervised learning and classification, this can be used to predict the class of a sample. So, a perceptron can be used to solve a classification problem.

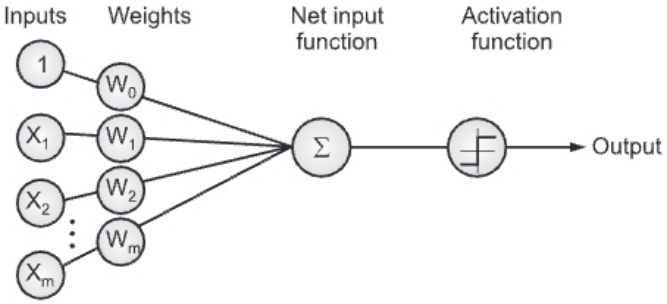


Fig. 3.11: Process of Perceptron

- A Perceptron is an algorithm for supervised learning of binary classifiers. This algorithm enables neurons to learn and process elements in the training set one at a time.
- There are two types of Perceptron: Single layer and Multilayer.
  1. **Single layer:** Single layer perceptron can learn only linearly separable patterns.
  2. **Multilayer:** Multilayer perceptron or feed forward neural networks with two or more layers have the greater processing power.
- The Perceptron algorithm learns the weights for the input signals in order to draw a linear decision boundary.

3.9 K-NEAREST-NEIGHBOR CLASSIFIERS

- K-Nearest-Neighbor (KNN) is a supervised learning algorithm which is used for classification and regression. The major application of KNN is in classification of predictive problems. K-nearest-neighbours stores all available cases and classifies new cases based on a similarity between the data items. KNN does not make assumptions on underlying data (non-parametric algorithm).
- Example:

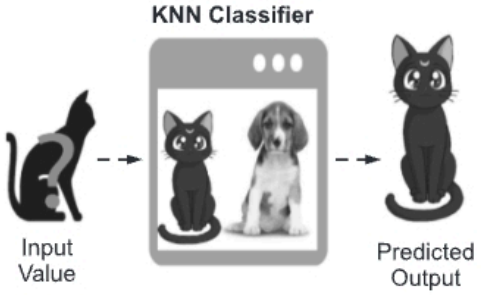


Fig. 3.12: Classification using KNN

- Suppose the image creature is present and it looks the same as a cat and dog but you want to know either it is cat or dog. For this we use the KNN algorithm and works on a similarity measure. The model will find same feature of new data set to cats and dogs images which are based on features it put to see it is dog or cat.

Working of KNN Algorithm:

Step 1: Select the number K of the neighbours.

Step 2: Calculate the Euclidean distance of **K number of neighbours.**

Step 3: Take the K nearest neighbours as per the calculated Euclidean distance.

Step 4: Among these k neighbours, count the number of the data points in each category.

Step 5: Assign the new data points to that category for which the number of the neighbour is maximum.

Step 6: Our model is ready.

For example, consider the following data points.

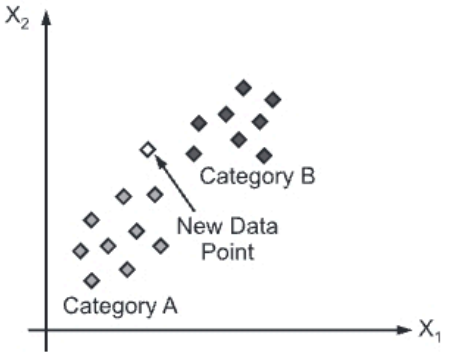


Fig. 3.13: Data Points

- And there is a new data point which is to be classified with KNN.
- The first step is to choose the number of neighbours. So, let's take k = 5.
- The next step is to calculate the Euclidean distance between the data points. It can be calculated as:

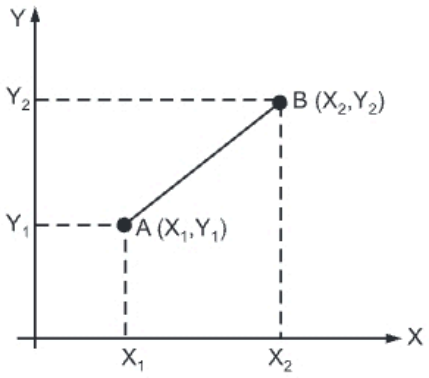


Fig. 3.14: Calculation of the Euclidean Distance

- The Euclidean distance formula:

  $d = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$

  where,

  (x_1, y_1) : the coordinates of one point.

  (x_2, y_2) : the coordinates of the other point.

  d : the distance between (x_1, y_1) and (x_2, y_2).
- With the Euclidean distance we get the nearest neighbours, as three nearest neighbours in class A and two nearest neighbours in class B. Therefore, the new data point must belong to Class A.

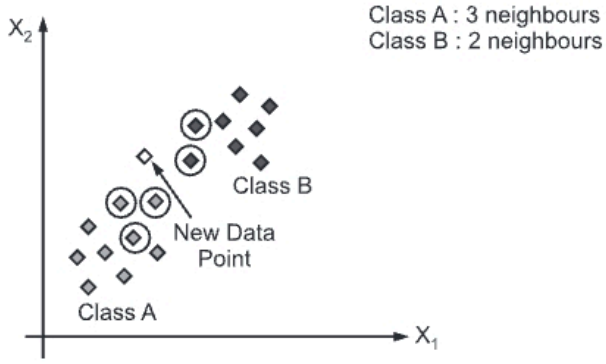


Fig. 3.15: K-Nearest-Neighbors

Advantages of KNN Algorithm:

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

Disadvantages of KNN Algorithm:

- Always needs to determine the value of K.
- It is computationally expensive due to high storage requirements.

3.10 SVM CLASSIFIERS

3.10.1 Introduction

- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. Support Vector Machine is a method for the classification of both linear and nonlinear data. SVM is an algorithm that uses a nonlinear mapping to transform the original training data into a higher dimension.
- The main objective of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a **Hyperplane**.
- The following figure explains the concept.
  - **Support Vectors:** Data Points that are closest to the hyperplane are called support vectors. Separating line will be defined with the help of these data points.
  - **Hyperplane:** It is a decision plane or space which is divided between a set of objects having different classes.

- **Margin:** It may be defined as the gap between two lines on the closet data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

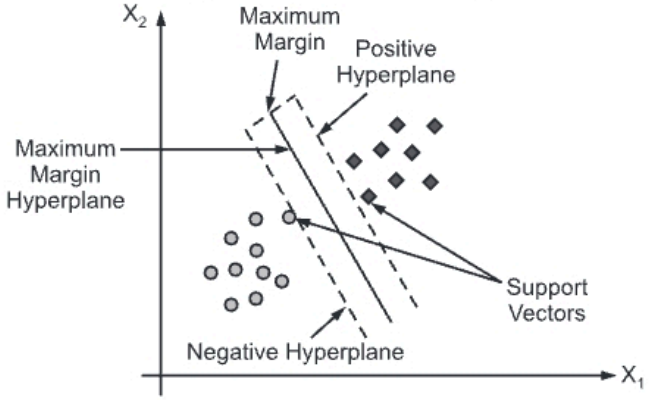


Fig. 3.16: Support Vector Machine

- SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.
- SVM algorithms can be used for Face detection, image classification, text categorization, etc.

3.10.2 Types of SVM

There are two types of SVM:

1. **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
2. **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

3.10.3 Working of SVM

Linear SVM:

- The working of the SVM algorithm is as follows:
- Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x_1 and x_2. We want a classifier that can classify the pair (x_1, x_2) of coordinates in either green or blue. Consider the below figure:

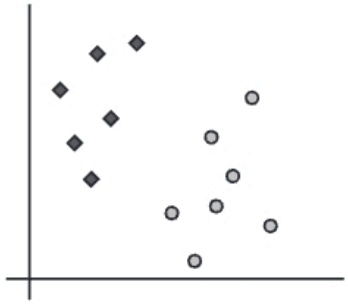


Fig. 3.17: Dataset with two tags

- Since it is a two dimensional space, by drawing a straight line, we can easily separate the green and blue data sets into two classes. But there can be multiple lines that can separate these classes. Consider the below figure.

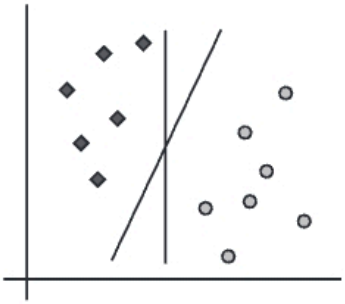


Fig. 3.18: Hyperplane

- The SVM algorithm gives the optimal / best line or decision boundary that separates the data sets. This region is called a hyperplane.
- SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called the margin. And the goal of SVM is to maximize this margin. The hyperplane with maximum margin is called the **Optimal Hyperplane**.

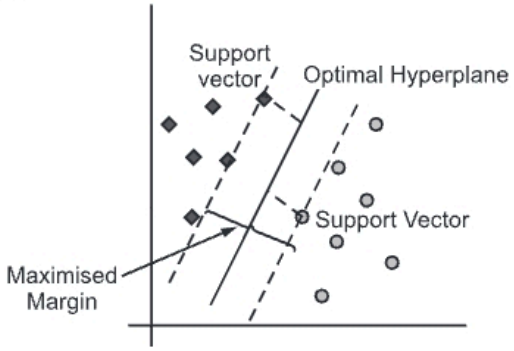


Fig. 3.19: Optimal Hyperplane

Non-Linear SVM:

- The linear SVM is applicable to the data which can be separated by a straight line. But, if the data is not linear, it is not possible to draw a single straight line. One such example is given in the below figure.

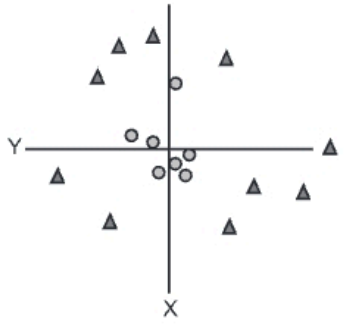


Fig. 3.20: Non-Linear Dataset

- To solve such a problem, we add one more dimension i.e. Z dimension along with the X dimension and y dimension to separate the data points. It can be calculated as,

$$Z = X^2 + Y^2$$

By adding the third dimension, the sample space will become as below image:

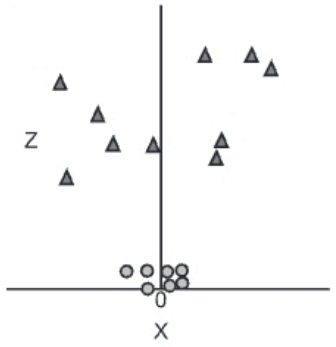


Fig. 3.21: Addition of third dimension Z

- Non-linear SVM will divide the data points into classes in the following way. Consider the below image:

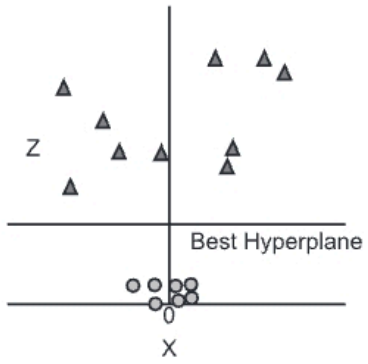


Fig. 3.22: Non-linear SVM

- Since we are in 3D Space, hence it is looking like a plane parallel to the X-axis. If we convert it in 2D space with Z = 1, then it will become as:

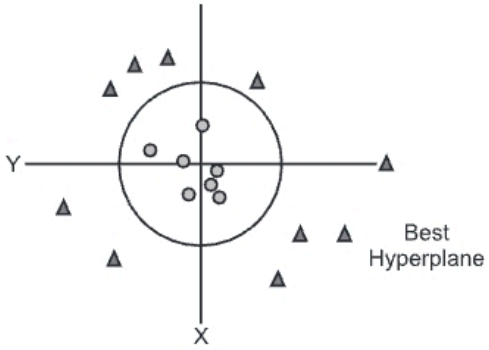


Fig. 3.23: Best Hyperplane

- Hence we get a circumference of radius 1 in case of non-linear data.

3.11 REGRESSION

- Regression is a data mining technique that is used for predicting a range of continuous values in a specific dataset. Regression is a supervised learning technique which predicts any continuous valued attribute. It analyses the relationship between a target variable (dependent) and its predictor variable (independent).
- Regression is used across multiple industries for business and marketing planning, financial forecasting, environmental modeling and analysis of trends.

Table 3.1: Difference between Regression and Classification

Regression	Classification
In Regression, the output variable must be of continuous nature or real value.	In Classification, the output variable must be a discrete value.
The task of the regression algorithm is to map the input value (x) with the continuous output variable (y).	The task of the classification algorithm is to map the input value(x) with the discrete output variable (y).
Regression Algorithms are used with continuous data.	Classification Algorithms are used with discrete data.
In Regression, we try to find the best fit line, which can predict the output more accurately.	In Classification, we try to find the decision boundary, which can divide the dataset into different classes.
Regression algorithms can be used to solve the regression problems such as Weather Prediction, House price prediction, etc.	Classification Algorithms can be used to solve classification problems such as Identification of spam emails, Speech Recognition, Identification of cancer cells, etc.
They are divided into Linear and Non-linear Regression.	The Classification algorithms can be divided into Binary Classifiers and Multi-class Classifiers.

3.11.1 Linear Regression

- Linear regression is the method where regression models the relationship between two variables by fitting a linear equation to observe the data. It is a simple type of regression.
- Linear regression attempts to find the mathematical relationship between variables.
- If the outcome is a straight line then it is considered a linear model and if it is a curved line, then it is a non-linear model.
- The relationship between dependent variables is given by a straight line and it has only one independent variable.

$$Y = \alpha + \beta X$$

Where, Model 'Y' is a linear function of 'X'.

- The value of 'Y' increases or decreases in a linear manner according to which the value of 'X' also changes.

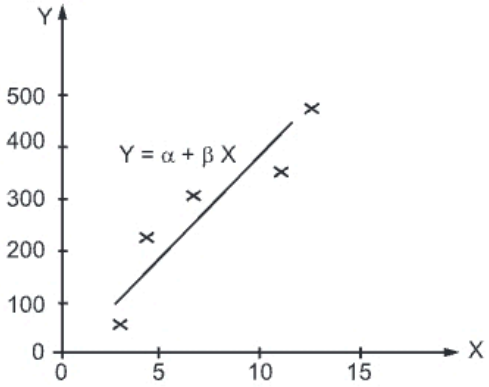


Fig. 3.24: Linear Regression

3.11.2 Non-linear Regression

- Non-linear regression is a form of regression analysis in which data is fit to a model and then expressed as a mathematical function.
- Non-linear regression is a curved function of an X variable (or variables) that is used to predict a Y variable.
- Non-linear regression can show a prediction of population growth over time.
- Non-linear models can be modeled by a polynomial function. For example,

$$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

Convertible to linear with new variables.

$$x_2 = x^2, \quad x_3 = x^3$$

$$y = w_0 + w_1 x + w_2 + x_2 + w_3 x_3$$

- Other functions can also be transformed to linear models.

3.12 INTRODUCTION TO PREDICTION

- Prediction is one of the data mining processes. It is used to find a numerical output. Here, the training dataset contains the inputs and corresponding numerical output values. According to the training dataset, the algorithm derives the model or a predictor. When the new data is given, the model should find a numerical output. This method does not have the class label. The model predicts a continuous-valued function or ordered value.
- Predictive data mining is data mining that is used for the purpose of using business intelligence or other data to forecast or predict trends. The prediction can help the decision makers to make better decisions.
- Prediction is the method of recognizing the missing or not available numerical data for a new process of observing.
- In prediction, the authenticity depends on how well a given predictor can guess the value of a predicted attribute for new data.
- In prediction, the model can be known as the predictor.
- XLMiner (XLMiner is a comprehensive data mining add-in for Excel) supports the use of four prediction methods: Multiple linear regression, K-nearest- neighbours, Regression tree and Neural network.

Summary

- Classification is the process of classifying the data. It is a data mining technique which is done for analysis of the data. It is the process of finding the model that defines the classes and their concepts. It identifies and categorizes the sub population of the data.
- A decision tree is the approach, where a classification process is modeled as a tree. Once a tree is constructed, it is applied to each and every tuple in the database. The result is a classification. Decision trees are the most useful approach in the classification problem.
- A decision tree is a supervised learning where we train the data. Here, the algorithm divides the dataset into subsets according to the most significant attribute in the database.
- An attribute selection measure is a heuristic for selecting the splitting criterion that "best" separates a given data partition.
- The pruning of decision tree is a process where the anomalies in the training data are removed due to outliers or noise in the data.
- Approach to perform classification is to apply if-then rules that cover all the possible cases. Here the learned model is represented as if- then rules. Here, at first, rules that are used for classification are examined.
- A if - then approach based classification rule r consists of *if* (predicate) part and *then* part i.e. the consequences of predicate.

$$r = (a, c)$$

If (condition) then (conclusion)

- Bayesian classification is a probabilistic approach to learning and inference based on a different view of what it means to learn from data, in which probability is used to represent uncertainty about the relationship being learnt.
- Bayes' Theorem is named after Thomas Bayes. There are two types of probabilities:
  - Posterior Probability [P(H/X)]
  - Prior Probability [P(H)]

  Where, X is a data tuple and H is some hypothesis.

  According to Bayes' Theorem,

$$P(H/X) = \frac{P(X/H)P(H)}{P(X)}$$

- Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem.
- Bayesian networks are a type of probabilistic graphical model that uses Bayesian Inference for probability computations.
- A Bayesian network is a directed acyclic graph in which each edge corresponds to a conditional dependency, and each node corresponds to a unique random variable.
- Parameter learning is the process of using data to learn the distributions of a Bayesian network or Dynamic Bayesian network
- The task of structure learning for Bayesian networks refers to learn the structure of the directed acyclic graph (DAG) from data.

- K-Nearest-Neighbour is a supervised learning algorithm which is used for classification and regression.
- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.
- Support Vector Machines (SVMs) is a method for the classification of both linear and nonlinear data.
- Regression is a data mining technique that is used for predicting a range of continuous values in a specific dataset.
- Regression is a supervised learning technique which predicts any continuous valued attribute.
- Linear regression is the method where regression models the relationship between two variables by fitting a linear equation to observe the data.
- Non-linear regression is a form of regression analysis in which data is fit to a model and then expressed as a mathematical function.
- Prediction is a data mining task that discovers relationships between independent variables and relationships between dependent variables.

Check Your Understanding

1. Which classification scheme makes use of IF-THEN rules for class prediction?
   (a) Rule-based classification (b) Statistical based classification
   (c) Distance based classification (d) Decision tree based classification
2. What is a Decision Tree?
   (a) Flow-Chart
   (b) Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label.
   (c) Flow-Chart & Structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label.
   (d) None of the mentioned
3. Decision Nodes are represented by ____ .
   (a) Disks (b) Squares
   (c) Circles (d) Triangles
4. ____ is the process of finding a model that describes and distinguishes data classes or concepts.
   (a) Data Characterization (b) Data Classification
   (c) Data Discrimination (d) Data Selection
5. ____ is used to map a data item to a real valued prediction variable.
   (a) Classification (b) Regression
   (c) Clustering (d) Summarization
6. ____ involves the learning of the function that does mapping.
   (a) Classification (b) Clustering
   (c) Regression (d) Summarization

7. _____ assumes that the target data fit into some known type of function.
   (a) Classification (b) Clustering
   (c) Regression (d) Summarization
8. What is the objective of the support vector machine algorithm?
   (a) To find an optimal hyperplane in an N-dimensional space that distinctly classifies the data points where N is the number of features.
   (b) To find an optimal hyperplane in an N-dimensional space that distinctly classifies the data points where N is the number of samples.
   (c) To find an optimal hyperplane in an N-dimensional space that distinctly classifies the data points where N is the number of target variables.
   (d) None of these
9. Support Vector Machine (SVM) can be used for _____.
   (a) classification only (b) regression only
   (c) classification and regression both (d) None of these
10. In SVM, the dimension of the hyperplane depends upon which one?
   (a) The number of features (b) The number of samples
   (c) The number of target variables (d) All of the above
11. What is Reinforcement learning?
   (a) All data is unlabelled and the algorithms learn to inherent structure from the input data.
   (b) All data is labelled and the algorithms learn to predict the output from the input data.
   (c) It is a framework for learning where an agent interacts with an environment and receives a reward for each interaction.
   (d) Some data is labelled but most of it is unlabelled and a mixture of supervised and unsupervised techniques can be used.
12. A nearest neighbour approach is best used _____.
   (a) with large-sized datasets.
   (b) when irrelevant attributes have been removed from the data.
   (c) when a generalized model of the data is desirable.
   (d) when an explanation of what has been found is of primary importance
13. Which of the following option is true about k-NN algorithm?
   (a) It can be used for classification.
   (b) It can be used for regression.
   (c) It can be used in both classification and regression.
   (d) None of the above

14. Which of the following statement is true about KNN algorithm?
    1. KNN performs much better if all of the data have the same scale.
    2. KNN works well with a small number of input variables (p), but struggles when the number of inputs is very large.
    3. KNN makes no assumptions about the functional form of the problem being solved.

    (a) 1 and 2 (b) 1 and 3

    (c) Only 1 (d) All of the above

Answers

1. (a)	2. (c)	3. (b)	4. (b)	5. (b)	6. (c)	7. (c)	8. (d)	9. (a)	10. (a)
11. (c)	12. (b)	13. (c)	14. (d)						

Practice Questions

Q.I Answer the following questions in short.

1. What is classification?
2. What are the advantages of using decision trees?
3. Write any two disadvantages of decision trees.
4. List the classification categorization algorithms.
5. What are SVM classifiers?
6. What is Perceptron?
7. What is decision tree pruning?

Q.II Answer the following questions.

1. How prediction is different from classification?
2. Compare and contrast classification methods.
3. Describe the K-nearest neighbour classifiers and case –based reasoning.
4. Describe Bayesian classification.
5. Describe about Prediction.
6. Explain about basic decision tree induction algorithm.
7. Explain the working of SVM classifier.
8. Write a short note on linear and non-linear regression.

Q.III Define the terms.

1. Regression
2. Information Gain
3. Gini index
4. Gain ratio
5. Classifier

4...
Clustering and Association Rule Mining

Learning Objectives...

- To describe a case where clustering is appropriate, and what insight it might extract from the data.
- To explain the K-means and k-medoid clustering algorithm.
- To interpret the output of a K-means and k-medoid algorithm analysis.
- To identify when it is necessary to scale variables before clustering.
- To describe the advantages, limitations, and assumptions of the K-means and K-medoids clustering algorithm.
- To figure out which objects are placed together for market basket analysis to increase sale.
- To describe the advantages, limitations, and assumptions of the Apriori Association rule mining algorithm.
- To interpret the output of an Apriori algorithm.

4.1 CLUSTER ANALYSIS

- Cluster is known as a group of similar things or objects/data points positioned or occurring closely together. Clusters are created with some similar characteristics found in data.
- There is high intra cluster similarity and low inter cluster similarity i.e., objects in the same cluster are similar and different clusters are dissimilar. Class label is not available in clustering, so it is always called un-supervised learning.
- The goal of clustering is to decide the internal grouping in a set of unlabelled data.
- Clustering analysis is used in market research, pattern recognition, data analysis, and image processing.

4.1.1 Introduction

- Clustering is unsupervised learning. Clustering methods are used to find similarities and relationship patterns in data then clustering will be done on that data having the same features. Such data will be place in groups. These groups are called as clusters.
- Clustering determines intrinsic grouping among the present labelled data.

- Some assumptions were made to find similar features in data. Every assumption will construct different types of valid clusters.

  Many definitions for clusters have been proposed:

  - Clustering is set of like elements. Elements from different clusters are not alike.
  - If clustering is done on distance measure the distance between points in a cluster is less than the distance between a point in the cluster and any point outside it.
- Clustering is similar to **database segmentation**, in which similar tuples (records) in a database are grouped together (partition/segment).
- The following Fig. 4.1 will give clear ideas about Clustering.

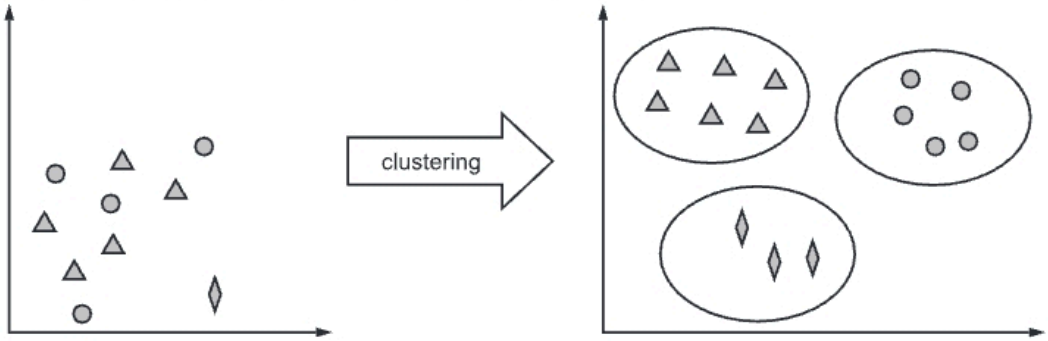


Fig. 4.1: Clustering

- It is the process of making a group of abstract objects into classes of similar objects.
- The advantage of clustering over classifications is that it is adaptable to changes and ready to find a single common feature among data points.

Definition:

Given a database $D = \{t_1, t_2, \ldots, t_n\}$ of tuples and an integer value k, the clustering problem is to define a mapping $f: D \rightarrow \{1, \ldots, k\}$ where each t_i is assigned to one cluster K_j, $1 \leq j \leq k$. A cluster, K_j contains precisely those tuples mapped to it; that is, $K_j = \{t_i \mid f(t_i) = K_j, 1 \leq i \leq n, \text{ and } t_i \in D\}$.

Here k = number of clusters to be created is an input value.

Applications of Clustering:

- Clustering in Data mining has following useful applications:
  - **Data summarization and compression:** Clustering will be useful in the fields like image processing and vector quantization which requires data summarization, compression and reduction.
  - **Collaborative systems and customer segmentation:** Clusters can be created based on similarity measure so it can find similar products, similar users. Clusters can be used in the area of collaborative systems and customer segmentation where similarity measure is important.
  - **Serve as a key intermediate step for other data mining tasks:** Cluster algorithms generate a compact summary of data used for classification, testing, hypothesis generation. So, clustering will be worked as a key intermediate step for other data mining tasks/algorithms.

- **Trend detection in dynamic data:** Clustering can also be applied for trend detection in dynamic data sets as clusters of similar trends can be created.
- **Social network analysis:** In social network analysis, like Facebook/Twitter clustering will be sued for generating sequences in images, videos or audios.
- **Biological data analysis:** In biological data as example to detect cancer, clustering can be used for making cluster images, videos.
- **Marketing:** Finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records.
- **Biology:** Data Mining helps in the classification of animals and plants are done using similar functions or genes in the field of biology.
- **Libraries:** Book ordering.
- **Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds.
- **City-planning:** Identifying groups of houses according to their house type, value and geographical location
- **Earthquake Studies:** Clustering observed earthquake epicenters to identify dangerous zones.
- **WWW:** Document classification; clustering weblog data to discover groups of similar access patterns.
- **Medicine:** In the field of medicine, Clustering diseases, cures for diseases, or symptoms of diseases can lead to very useful taxonomies.
- **Psychiatry:** The correct diagnosis of clusters of symptoms such as paranoia, schizophrenia, etc. is essential for successful therapy.
- **Archeology:** Researchers have attempted to establish taxonomies of stone tools, funeral objects, etc. by applying cluster analytic techniques.

Basic features of Clustering:

- Some basic features of clustering (as opposed to classification) as follows:
  1. The (best) number of clusters is unknown.
  2. There may not be any a priori knowledge concerning the clusters.
  3. Cluster results are dynamic.

Problems occur while Clustering:

- When clustering is applied to real world database many problems can occur.
  - **Outlier handling is difficult:** Outliers are the elements do not generally fall into any cluster. They are treated as solitary (lonely) clusters. If clustering algorithms attempt to increase size of clusters the outliers will be forced to be placed in some clusters. Such process may create poor clusters.
  - Dynamic data in the database implies that the cluster membership may change over time.
  - Interpreting the semantic meaning of each cluster is difficult. In classification we already know the class labels but when clustering process finishes cluster creation, the exact meaning of each cluster may not be obvious. In this situation help of domain experts is needed for assigning label to each cluster and interpret the meaning of each cluster.

- There is no any correct answer to clustering problem. Many solutions can be found. It is not easy to determine exact number of clusters. In this situation again, domain experts are required.
- The most important issue is what data is used for clustering. In case of classification, prior knowledge about attributes is there so it is supervised learning but in clustering unsupervised learning aims no prior knowledge about data to cluster.

4.1.2 Requirements for Cluster Analysis

- The following points will clear idea about importance of clustering in data mining:
  - **Scalability:** Highly scalable clustering algorithms are needed to deal with large databases such as big data.
  - **Ability to deal with different kinds of attributes:** Algorithms should be applicable to any kind of data sets such as numerical data, categorical, and binary data.
  - **Discovery of clusters with attribute shape:** Clustering algorithms determines clusters of random shape. Different types of distance measures are used to find out circular shape clusters of small sizes.
  - **High dimensionality:** The desired clustering algorithm should be able to handle low as well as high dimensional data sets.
  - **Ability to deal with noisy data:** Algorithms should be designed in such a way that they should able to handle noisy data. Some algorithms do not handle noisy data so poor quality of clusters as designed.
  - **Interpretability:** Clustering algorithms should produce interpretable, comprehensible, and usable results.

4.1.3 Types of Clustering

- Clustering algorithms are classifying among themselves as hierarchical or partitional.

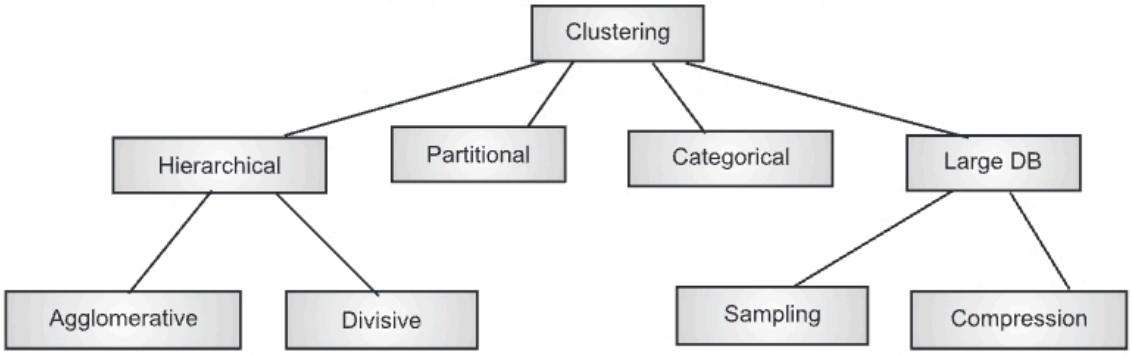


Fig. 4.2: Types of Clustering Algorithms

1. **Hierarchical Clustering:**
   - In this type of clustering method, a nested set of clusters is created.
   - Every level in hierarchy has distinct set of clusters.
   - At lowest level, every item is its own cluster. These clusters are unique.
   - As level in the hierarchy increases, items are grouped together in cluster.

- At the highest level, all items belong to one cluster.
- The desired numbers of clusters are not defined.
- These algorithms can be classified into Agglomerative or Divisive.
  - **Agglomerative:** In this type, clusters are created in bottom-up fashion.
  - **Divisive:** In this type, clusters are created in top-down fashion.

2. **Partitional Clustering:**
   - In this type of clustering method, only one set of clusters is created.
   - The desired numbers of clusters are defined.
3. **Categorical Clustering:**
   - These algorithms work on categorical databases. In Categorical databases, values describe some characteristic or category. For example, what is your favourite colour?
4. **Large DB Clustering:**
   - These algorithms work on Large databases. i.e. Big data.
   - These algorithms familiarize to memory constraints either by sampling or compression technique.
   - In case of sampling, data structures are used. These data structures can be compressed or pruned to fit into memory irrespective of size of database.
   - Clustering algorithms can be categorized based on use of mathematical formula such as graph theoretic or matrix algebra for running that algorithm. Adjacency matrix with distance measures is given as input to clustering algorithms.

4.1.4 Similarity or Distance Measure used in Clustering

- A similarity measure can be defined as the distance between various data points. It is simply the ordinary distance between two points. Euclidean distance is extensively used in clustering problems.
- The default distance measure used with the K-means algorithm is also the Euclidean distance.
- The Euclidean distance determines the root of square differences between the coordinates of a pair of objects as shown in following equation.

$$Dist_{XY} = max_k \ |X_{ik} - X_{jk}|$$

4.1.5 Outliers

- Outliers are isolated data points from given data set. They reside far from rest of the data. They generally do not fall into any cluster.
- Outlier values are much different from remaining points in the dataset. They always represent error in the database. For example, Average height of person is 5 to 6 feet. A person having 6.5 feet height from same dataset is the outlier.
- Due to presence of outliers clustering algorithms do not scale well. For example, kindly refer Fig. 4.3 consider three clusters with solid line. In this situation, outlier will fall in the cluster by itself. Now consider two clusters with dashed line then different data points will be which are placed in two solid line cluster will be placed in one cluster. This problem may occur because number of clusters to be formed should be given as input to the clustering algorithms in advance.

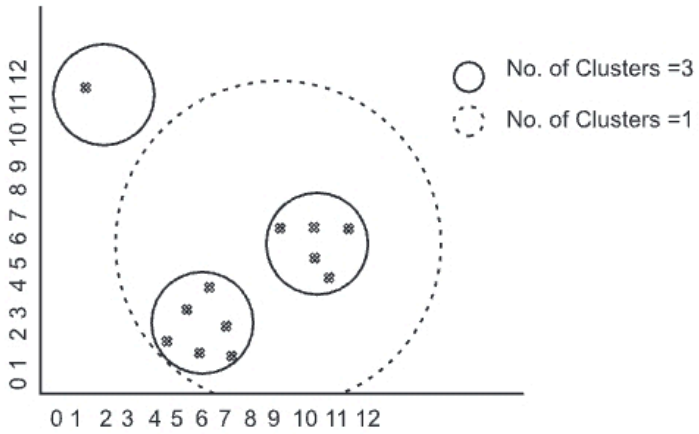


Fig. 4.3: Outliers in Clustering

- Clustering algorithms should find the outliers and remove them then only they can perform well.
- Every time while removing outliers care must be taken otherwise algorithm may get distracted. For example, consider a problem is to predicate flooding. In case of normal water flow values are seems to be normal. In case of flooding extremely high-water level values will be occur that is outlier. If this value is removed by data mining clustering algorithm then it shows flooding never occurred. So clustering algorithm will be designed by considering flooding value.
- Outlier detection is also known as **Outlier Mining**. It is the process of identification of outliers from data set.
- Always removing outlier is not the option so, clustering algorithms must be designed to treat different with outliers.
- Some Outlier detection techniques are based on statistical techniques also.
- **Discordancy tests** are used to detect outliers. These tests are not very realistic for real world data sets as these datasets may not follow normal data distributions.
- Distance measures techniques also can be used to detect outliers. For example, Salary of CEO among the salaries of other employees.

4.2 HIERARCHICAL METHODS

- As mentioned earlier, hierarchical clustering algorithms actually creates sets of clusters.

Data structure for hierarchical clustering:

- **Dendrogram:** It is a tree data structure used to illustrate the hierarchical clustering technique with set of different clusters. Elements of Dendrogram as follows:
  - **Root:** It contains one cluster where all the data elements are together.
  - **Leaves:** It contains single element cluster.
  - **Internal Node:** They formed by merging the clusters in the lower level of the tree(children of the same parent).
  - **Level of tree:** It is associated with distance measures used for merging the clusters.
- All clusters in dendrogram at particular level are combined because children(low level) clusters had a distance between them less than the distance value associated with this level in tree.

Example 1:

- Consider 6 elements {A, B, C, D, E, F}. There are 6 parts of one diagram. Following figure shows 5 different clusters.
  - **Part a:** There are six clusters. Here is one cluster for each element.
  - **Part b:** There are four clusters. (i.e. Two sets of two elements (two elements are closer to each other))
  - **Part c:** There are three clusters. Second cluster is updated by adding nearest elements.
  - **Part d:** There are two clusters. Two clusters are merged here. (Two cluster elements are closer to each other than the remote element cluster.
  - **Part e:** There is one cluster. All the six elements are merged.

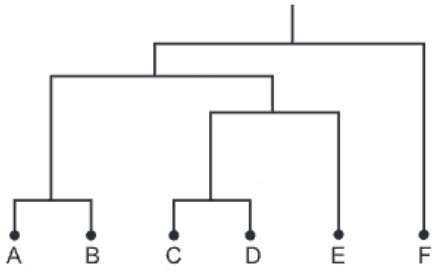


Fig. 4.4: Dendrogram for Example 1

- **Space complexity** of hierarchical clustering algorithms is $O(n^2)$ (Space required for adjacency matrix) where there are n items to cluster. Space required for dendrogram id = O(kn).
- **Time complexity** of hierarchical clustering algorithms is $O(kn^2)$ (one iteration for each level in the dendrogram).
- Algorithms can merge the clusters which are near to each other at low level or new clusters can be created at each level with progressively larger distances.
- Application of hierarchical algorithms in biology- plant and animal taxonomies can be viewed as hierarchy of clusters.

4.2.1 Agglomerative Hierarchical Clustering

- The first step of these algorithms starts with each individual item in its own cluster and iteratively merges clusters until all items belong in one cluster.
- Agglomerative algorithms differ in how clusters are merged together at each level.
- Typical agglomerative algorithm is as follows:

Input:

```
D = {t1, t2 , …, tn} // Set of elements
A //Adjacency matrix showing distance between elements
```

Output:

```
DE // Dendrogram represented as a set of ordered triples
```

Agglomerative algorithm:

```
1. d = 0;
2. k = n;
3. K = {{ t1}, ..., {tn}};
4. DE = { (d, k, K) }; // Initially dendrogram contains each element
5. in its own cluster.
6. repeat
7. old k = k;
8. d = d + 1;
9. Ad = Vertex adjacency matrix for graph with threshold
10. distance of d;
11. (k, K) = New Clusters (Ad, D) ;
12. if old k ≠ k then
13. DE = DE U (d, k, K); // New set of clusters added to dendrogram.
14. until k = 1
```

- This algorithm takes input as set of elements and distance between them in terms of n × n adjacency matrix(n is number of vertices). Matrix contains distance value between two points. `A[i, j] = dist(ti, tj).`
- The above algorithm gives output as dendrogram DE, which is represented as set of ordered triplets as {d, k, K} where d: threshold distance, k: number of clusters, K: set of clusters. So, the dendrogram in the example 1 is represented as following:

  {(0, 5), {{A}, {B}, {C}, {D}, {E}}),
  (1, 3, {{A, B}, {C, D}, {E}})
  (2, 2, {{A, B, C, D}, {E}}),
  (3, 1, {{A, B, C, D, E}})}
- Drawing of dendrogram creates many sets of clusters. Depending on threshold value user can determine which set of clusters will be used.
- Procedure called NewClusters() is used to determine how to create next level of clusters from the previous level. This step shows how different agglomerative algorithms differ in their working. This procedure merges two clusters / multiple clusters from prior level. Algorithm also determines which clusters are merged together (when they have same distance values). Agglomerative algorithms also differ in their working using this point.

Techniques of Hierarchical Clustering:

- Different techniques are used to determining the distance criteria based on graph theory concepts. Following are the three criteria:
  - Single Link
  - Complete Link
  - Average Link

1. Single Link:

- This idea is based on finding maximum connected components in a graph.
- A connected component in a graph means path should exist between two vertices/components/ nodes.
- In this technique, two clusters are merged together if there is at least one edge that connects the two clusters; that is, `min dist(point1, point2)` ⇐ `threshold distance`. Due to this reason this technique is called the **Nearest Neighbour Clustering Technique.**

Example 2: Consider following Table 4.1 It is adjacency matrix shows distance between two vertices/points.

Table 4.1: Adjacency Matrix for single link algorithm

Item	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

- Based on above table, the graph will display as follows (Fig 4.5).
- This graph shows all edges with all connections with vertices.

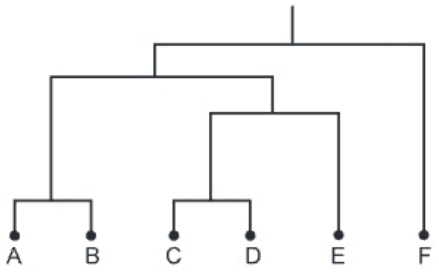
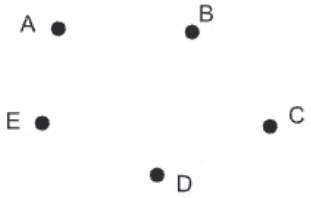
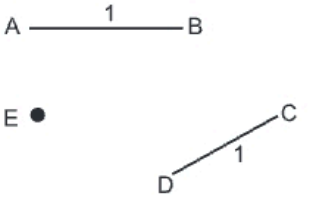
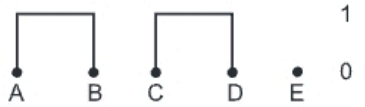
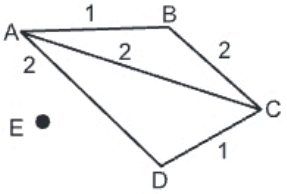
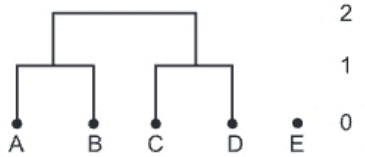
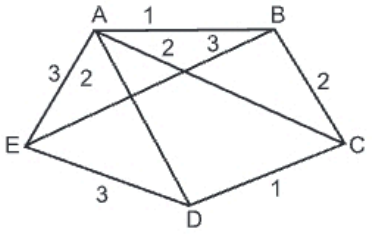
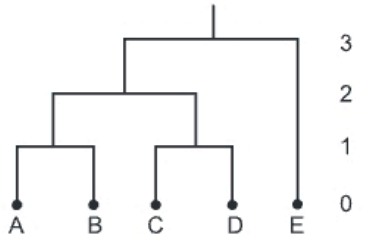


Fig. 4.5 : Graph with all distances for Table 4.1

- Following Table 4.2 shows graph with edges with threshold (distance) 1, 2, 3 constructions with single link algorithm. Right hand side number shows the threshold distance which is used to merge the clusters at each level.
- Steps for construction is as follows:

Table: 4.2: Stepwise Dendrogram construction

Steps	Graphs	Dendrogram
Step 0		
Step 1: It will combine the connected clusters. With threshold distance 1 or less Three clusters are formed as : {A, B}, {C, D}, (E)		
Step 2: It will combine the connected clusters. With threshold distance 2 or less Two clusters are formed. So two clusters from previous level {A, B}, {C, D} are merged together and form one cluster. So, Now the clusters are: {A, B, C, D}, {E}		
Step 3: It will combine the connected clusters. With threshold distance 3 or less. Here the graph is connected, so the two clusters from the last level are merged into one large cluster that contains all elements as: {A, B, C, D, {E}}		

- The single link algorithm replaces New Cluster procedure in the agglomerative algorithm with a procedure to find connected components of a graph.
- Procedure to find connected components of a graph takes input as adjacency matrix and produces output as a set of connected components defined by a number and an array containing membership of each component.

Problem arises in Single Link Algorithm:

- Inefficient as time complexity of connected components procedure is $O(n^2)$ for each iteration.
- More variations of single link algorithm are possible. One of them is:
  - **MST (Minimum Spanning Tree):** A procedure called MST produces a minimum spanning tree taking an adjacency matrix as input. Clusters are merged in increasing order of distance. After merging of cluster, the distance of then in MST becomes infinity.
  - **Time complexity of MST** : $O(n^2)$ (as once minimum spanning tree is created for n-1 nodes then loop is repeated for n-1 times).

2. **Complete Link:** This algorithm is same as single link and it searches for groups rather than connected components. This group is known as **Clique**. A clique is a maximal graph in which there is an edge between any two vertices.

- A procedure maximum distance between two clusters is used to find maximum distance between two clusters. Clusters are merged if `max dist ⇐ threshold dist`. This is also known as clique procedure. This procedure finds all cliques in the graph.
- **Time complexity** is $O(n^2)$.
- More compact clusters are found. Following Fig. 4.6 shows dendrogram with complete link algorithm.

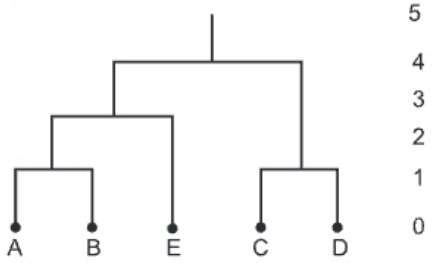


Fig. 4.6: Dendrogram for Complete Link Algorithm

3. Average Link: This technique merges two clusters if the `average distance (point1, point2) < distance threshold`. The complete graph is examined not a threshold graph link single link or average link algorithm at each step. Following Fig. 4.7 shows dendrogram with complete link algorithm.

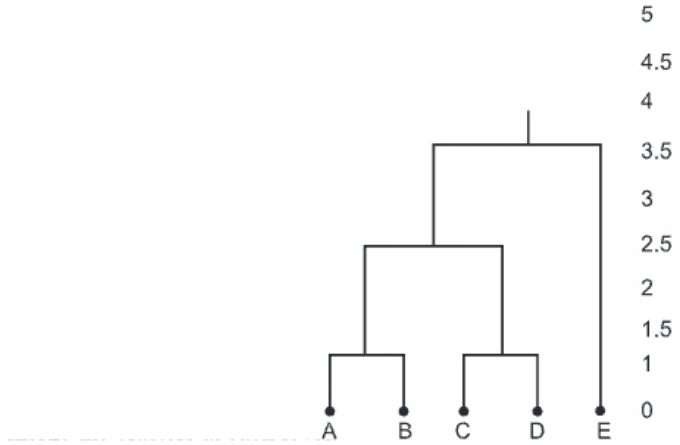


Fig. 4.7: Dendrogram for Average Link Algorithm

- **Space complexity** of agglomerative algorithm: $O(n^2)$ (where, n : number of items to cluster).

Disadvantages of Agglomerative algorithms:

- In case of large databases, the time and space complexity of `NewClusters()` procedure is high.
- This algorithm is not incremental. As and when new clusters are added or old clusters are removed the whole algorithm must be rerun.
- These algorithms do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects.
- The space required for the adjacency matrix is $O(n^2)$.
- Because of the iterative nature of the algorithm, the matrix must be accessed multiple times.
- Agglomerative approach is not incremental, when elements are added or removed, the entire algorithm has to rerun.
- These algorithms can never undo what was done previously.

4.2.2 Divisive Hierarchical Clustering

- This algorithm works in reverse way of agglomerative algorithm.
- In this type of clustering technique, all items are placed in one cluster then clusters are split. Splitting of cluster process continues till all items are placed in their own cluster.
- Splitting process is done for all those items which are not sufficiently close to other elements.
- An example to explain above algorithm is based on the MST version of the single link algorithm.
- In this technique, edges will be cut out from the largest to the smallest in MST.

Example 3:

- Consider a cluster containing all items: {A, B, C, D, E}.

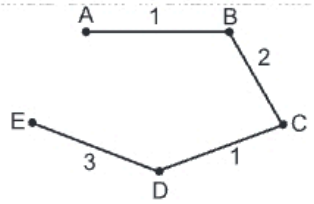


Fig 4.8 (a): MST for Example 3.

- At MST type, there is largest edge between D and E. So, this edge can be cut and we then split the one cluster into two: {E} and {A, B, C, D}.
- Then we remove the edge between B and C. So, one large cluster can be divided into two clusters: {A, B} and {C, D}.These clusters will then be split at the next step.
- The order depends on how a specific implementation would treat identical values. Looking at the dendrogram in Figure 4.4, we see that we have created the same set of clusters as with the agglomerative approach, but in **reverse order.**

- Following Fig. 4.8 (b) will clear idea about both the algorithms.

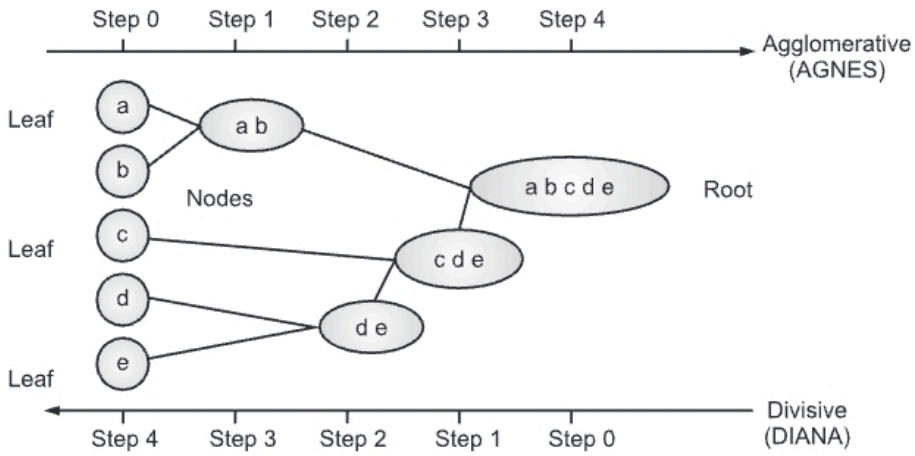


Fig. 4.8 (b): Agglomerative versus Division Clustering

4.3 PARTITIONING METHODS

- These algorithms do not use hierarchical structures. This algorithm used iterative process.
- The agglomerative and divisive clustering algorithms create clusters in several steps whereas Partitional algorithms create clusters in one step only.
- So, in this type of algorithms only one set of clusters is created.
- As there is only one set of output containing clusters, user must give initially value of k. i.e. number of clusters to be created.
- Some criterion or metric function is used to determine good set of clusters. Such criteria may be average distance between clusters.
- One common criteria or measure used in this type of algorithms is a squared error metric. This measures the squared distance from each point to the centroid for the associated cluster.

$$\sum_{m=1}^{k} \sum_{t_{mi} \in K_m} dis(C_m, t_{mi})^2$$

Problem with Partitional algorithms:

- They suffer from a combinatorial explosion due to the number of possible solutions. Searching for all possible clustering alternatives will not be possible. For example n (number of items to cluster) = 19 k, (number of clusters) = 4. So, the possible combinations of cluster creations will be S(n, k):

$$S(n, k) = \frac{1}{k!} \sum_{i=1}^{k} (-1)^{k-1} \binom{k}{i} (i)^n$$

- There will be 11,259,666,000 different ways to cluster 19 items in 4 clusters.

4.3.1 K-MEANS: A Centroid-Based Technique

K-Means Clustering:

- This is iterative clustering algorithm. In this, items are moved among set of clusters until desired set is reached.
- It is one type of squared error algorithm and some convergence criteria should be defined to obtain the final result.

Input :

```
D = {t1,  t2, ..., tn) // Set of elements
k // Number of desired clusters
```

Output :

```
K // Set of clusters
```

K-means algorithm:

```
1. assign initial values for means m1, m2, ..., mk;
2. repeat
3. assign each item ti to the cluster which has the closest mean;
4. calculate new mean for each cluster;
5. until convergence criteria is met;
```

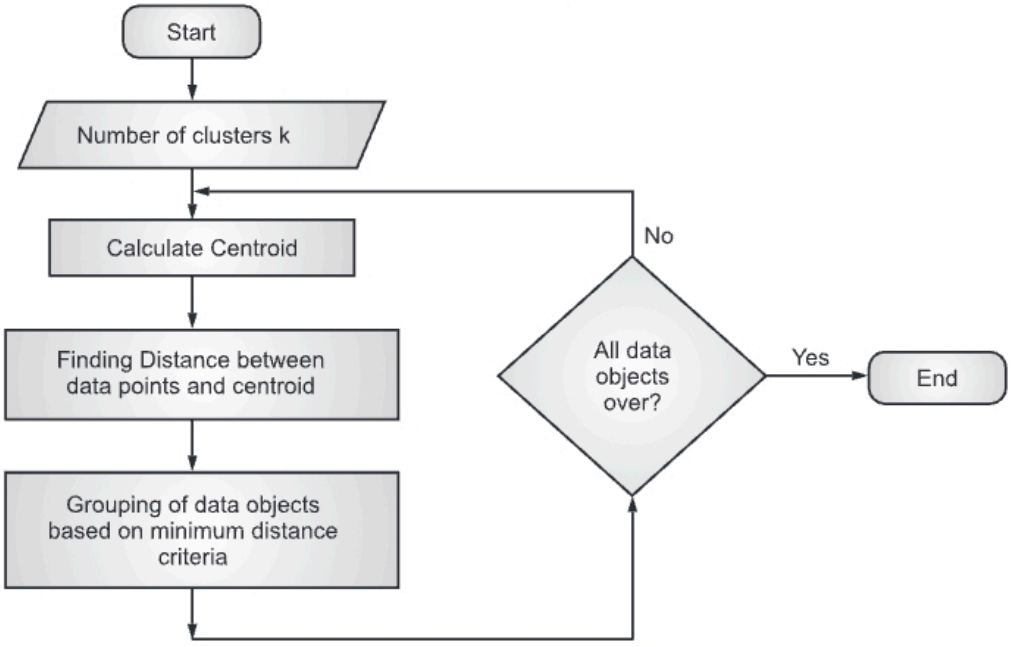


Fig. 4.9: K-Means clustering flowchart

Example 4:

- Suppose we have following items to cluster {2, 4, 10, 12, 3, 20, 30, 11, 25}

  k = 2(number of clusters to be created).
- There are two means for two clusters. m_1, m_2; Assume $m_1 = 2$, $m_2 = 3$

- Using Euclidian distance formula clusters are:

  K_1 = {2-2, 4-2, 10-2, 12-2, 3-2, 20-2, 30-2, 11-2, 25-2}

  Answer of first cluster K_1 = {0, 2, 8, 10, 1, 18, 28, 9, 23}

  K_2 = {2-3, 4-3, 10-3, 12-3, 3-3, 20-3, 30-3, 11-3, 25-3}

  Answer of second cluster K_2 = {1, 1, 7, 9, 0, 17, 27, 8, 22}

  Dataset = {2, 4, 10, 12, 3, 20, 30, 11, 25}

Table 4.3: Example 4 Clustering solution

Steps	m_1	m_2	K_1	K_2	Compare distances and place the data points in the respective cluster
Step I	2	3	{2}	{3, 4, 10, 11, 12, 20, 30, 25}	{0, 2, 8, 10, 1, 18, 28, 9, 23} ↕ ↕ ↕ ↕ ↕ ↕ ↕ ↕ ↕ {1, 1, 7, 9, 0, 17, 27, 8, 22}
Step II (recalculate the mean) for cluster 1 $m_1 = \frac{(2)}{1} = 2$(average method from statistics) cluster 2 $m_2 = \frac{(3+4+10+11+12+20+30+25)}{8} = \frac{115}{8} = 14.375$ approx 14.36					
	2	14.36	{2, 3, 4}	{10, 11, 12, 20, 30, 25}	{ 0, 2, 8, 10, 1, 18, 28, 9, 23} ↕ ↕ ↕ ↕ ↕ ↕ ↕ ↕ {12.36,10.36,4.36,2.36,11.36,5.64,15.64,3.36,1.64}
Step III (recalculate the mean) for cluster 1 $m_1 = \frac{(2+3+4)}{3} = \frac{9}{3} = 3$ (average method from statistics) cluster 2 $m_2 = \frac{(10+11+12+20+30+25)}{6} = \frac{108}{6} = 18$					
	3	18	{2, 3, 4, 10}	{11, 12, 20, 30, 25}	{1, 1, 7, 9, 0, 17, 27, 8, 22} ↕ ↕ ↕ ↕ ↕ ↕ ↕ ↕ ↕ {16, 14, 8, 6, 15, 2, 12, 7, 7}
Step IV (recalculate the mean) for cluster 1 $m_1 = \frac{(2+3+4+10)}{4} = \frac{19}{4} = 4.75$ (average method from statistics) cluster 2 $m_2 = \frac{(11+12+20+30+25)}{5} = \frac{98}{5} = 19.6$					
	4.75	19.6	{2, 3, 4, 10, 11}	{12, 20, 30, 25}	{2.75,0.75,5.25,7.25,1.75,15.25,25.25,6.25,20.25} ↕ ↕ ↕ ↕ ↕ ↕ ↕ ↕ ↕ {17.6, 15.6, 9.6, 7.6, 16.6, 0.4, 10.4, 8.6, 5.4}

... (Contd.)

Step V (recalculate the mean) for

cluster 1 $m_1 = \frac{(2+3+4+10+11)}{5} = \frac{30}{5} = 6$ (average method from statistics)

cluster 2 $m_2 = \frac{(12+20+30+15)}{4} = \frac{87}{4} = 21.75$

	6	21.75	{2, 3, 4, 10, 11, 12}	{20, 30, 25}	{4, 2, 4, 6, 3, 14, 24, 5, 19} ↕ ↕ ↕ ↕ ↕ ↕ ↕ ↕ ↕ {19.75,17.75,11.75,9.75,18.75,1.75,8.25,10.75,3.25}

Step VI (recalculate the mean) for

cluster 1 $m_1 = \frac{(2+3+4+10+11+12)}{6} = \frac{42}{6} = 7$ (average method from statistics)

cluster 2 $m_2 = \frac{(20+30+25)}{3} = \frac{75}{3} = 25$

	7	25	{2, 3, 4, 10, 11, 12}	{20, 30, 25}	{5, 3, 3, 5, 4, 13, 23, 4, 18} ↕ ↕ ↕ ↕ ↕ ↕ ↕ ↕ ↕ {23, 21, 15, 13, 22, 5, 5, 14, 0}

Step VII (recalculate mean) for

cluster 1 $m_1 = \frac{(2+3+4+10+11+12)}{6} = \frac{42}{6} = 7$

cluster 2 $m_2 = \frac{(20+30+25)}{3} = 25$

Now mean does not change from step VI so criterian function converges. Hence final clusters are obtainedas follows

$C_1 = \{2, 3, 4, 10, 11, 12\}$ with $m_1 = 7$

$C_2 = \{20,30,25\}$ with $m_2 = 25$

Example 5:

Suppose that the data mining task is to cluster the following eight points (with (x; y) representing location) into three clusters.

A1(2; 10); A2(2; 5); A3(8; 4); A4(5; 8); A5(7; 5); A6(6; 4); A7(1; 2); A8(4; 9):

The distance function is Euclidean distance. Suppose initially we assign A_1, A_4 and A_7 as the center of each cluster, respectively. Use the k-means algorithm to show only:

(a) The three cluster centers after the first round of execution.

(b) The final three clusters.

The distance function between two points $a = (x_1, y_1)$ and $b = (x_2, y_2)$ is defined as:

$$\rho(a, b) = |x_2 - x_1| + |y_2 - y_1|.$$

Solution:

Table 4.4 (a): Solution of K-Means Example 5

		(2, 10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2, 10)				
A2	(2, 5)				
A3	(8, 4)				
A4	(5, 8)				
A5	(7, 5)				
A6	(6, 4)				
A7	(1, 2)				
A8	(4, 9)				

- First, we list all points in the first column of the table above. The initial cluster centers – means, are (2, 10), (5, 8) and (1, 2) - chosen randomly.
- Next, we will calculate the distance from the first point (2, 10) to each of the three means, by using the distance function:

point **mean 1**

x_1, y_1 x_2, y_2

(2, 10) (2, 10)

$\rho(a, b) = |x_2 - x_1| + |y_2 - y_1|$

$\rho(\text{point, mean 1}) = |x_2 - x_1| + |y_2 - y_1|$

$= |2 - 2| + |10 - 10|$

$= 0 + 0$

$= 0$

point **mean 2**

x_1, y_1 x_2, y_2

(2, 10) (5, 8)

$\rho(a, b) = |x_2 - x_1| + |y_2 - y_1|$

$\rho(\text{point, mean 2}) = |x_2 - x_1| + |y_2 - y_1|$

$= |5 - 2| + |8 - 10|$

$= 3 + 2$

$= 5$

point **mean 3**

x_1, y_1 x_2, y_2

(2, 10) (1, 2)

$$\rho(a, b) = |x_2 - x_1| + |y_2 - y_1|$$
$$\rho(\text{point, mean 2}) = |x_2 - x_1| + |y_2 - y_1|$$
$$= |1 - 2| + |2 - 10|$$
$$= 1 + 8$$
$$= 9$$

So, we fill in these values in the Table 4.4 (b).

Table 4.4 (b): Solution of K-Means Example 5

		(2, 10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2, 10)	0	5	9	1
A2	(2, 5)				
A3	(8, 4)				
A4	(5, 8)				
A5	(7, 5)				
A6	(6, 4)				
A7	(1, 2)				
A8	(4, 9)				

- So, which cluster should the point (2, 10) be placed in? The one, where the point has the shortest distance to the mean – that is mean 1 (cluster 1), since the distance is 0.

Cluster 1 **Cluster 2** **Cluster 3**

(2, 10)

- After Iteration 1, Table 4. 4(b) values are shown in Table 4.4 (c).

Table 4.4 (c): Solution of K-Means Example 5

		(2, 10)	(5, 8)	(1, 2)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2, 10)	0	5	9	1
A2	(2, 5)	5	6	4	3
A3	(8, 4)	12	7	9	2
A4	(5, 8)	5	0	10	2
A5	(7, 5)	10	5	9	2
A6	(6, 4)	10	5	7	2
A7	(1, 2)	9	10	0	3
A8	(4, 9)	3	2	10	2

- Therefore, after Iteration I New clusters are as follows:

Cluster 1	Cluster 2	Cluster 3
(2, 10) → A1	(8, 4) → 3	(2, 5) → A2
	(5, 8) → A4	(1, 2) → A7
	(7, 5) → A5	
	(6, 4) → A6	
	(4, 9) → A8	

- Next, we need to re-compute the new cluster centers (means). We do so, by taking the mean of all points in each cluster.

  For Cluster 1, we only have one point A1(2, 10), which was the old mean, so the cluster center remains the same.

  For Cluster 2, we have $\left(\frac{(8+5+7+6+4)}{5}, \frac{(4+8+5+4+9)}{5}\right) = (6, 6)$

  For Cluster 3, we have $\left(\frac{(2+1)}{2}, \frac{(5+2)}{2}\right) = (1.5, 3.5)$

  Formula = $|x_2 - x_1| + |y_2 - y_1|$

Table 4.4 (d): Solution of K-Means Example 5

	(x_1, y_1) (x_2, y_2) →	(2, 10)	(6, 6)	(1.5, 3.5)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2, 10)	0	4 + |–4| = 8	7	1
A2	(2, 5)	5	4 + 1 = 5	2	3
A3	(8, 4)	12	|–2| + 2 = 4	7	2
A4	(5, 8)	5	1 + |–2| = 3	8	2
A5	(7, 5)	10	|–1| + 1 = 2	7	2
A6	(6, 4)	10	0 + 2 = 2	5	2
A7	(1, 2)	9	5 + 4 = 9	2	3
A8	(4, 9)	3	2 + |–3| = 5	8	1

- Therefore, New clusters after iteration II areas follows:

Cluster 1	Cluster 2	Cluster 3
(2, 10) → A1	(8, 4) →A3	(2, 5) → A2
(4, 9) → A8	(5, 8) → A4	(1, 2) → A7
	(7, 5) → A5	
	(6, 4) → A6	

- Next, we need to re-compute the new cluster centers (means). We do so, by taking the mean of all points in each cluster.

  For Cluster 1, we have $\left(\frac{(2+4)}{2}, \frac{(9+10)}{2}\right) = (3, 9.5)$

  For Cluster 2, we have $\left(\frac{(8+5+7+6)}{4}, \frac{(4+8+5+4)}{4}\right)$

  For Cluster 3, we have $\left(\frac{(2+1)}{2}, \frac{(5+2)}{2}\right) = (1.5, 3.5)$

Table 4.4 (e) : Solution of K-Means Example 5

	(x_1, y_1) (x_2, y_2) →	(3, 9.5)	(6.5, 5.25)	(1.5, 3.5)	
	Point	**Dist Mean 1**	**Dist Mean 2**	**Dist Mean 3**	**Cluster**
A1	(2, 10)	1 + |–.5| = 1.5	4.5 + |–4.5| = 9.0	|–.5| + 6.5 = 7	1
A2	(2, 5)	1 + 4.5 = 5.5	4.5 + 0.25 = 4.75	.5 + |–1.5| = 2	3
A3	(8, 4)	|–5| + 5.5 = 10.5	|–2.5| + 1.25 = 3.75	|–6.5| + |–.5| = 7	2
A4	(5, 8)	|–2| + 1.5 = 3.5	1.5 + |–2.75| = 4.25	|–3.5| + |–4.5| = 8	1
A5	(7, 5)	|–4| + 4 = 8	|–1.5| + 0.25 = 1.75	|–5.5| + |–1.5| = 7	2
A6	(6, 4)	|–3| + 5.5 = 8.5	0 + 1.25 = 1.25	|–4.5| + |–.5| = 5	2
A7	(1, 2)	2 + 7.5 = 9.5	5.5 + 3.25 = 8.75	0.5 + 1.5 = 2	3
A8	(4, 9)	|–1| +.5 = 1.5	2.5 + |–3.75| = 9.25	|–2.5| + |–5.5| = 8	1

- Therefore, New clusters after iteration III are as follows:

Cluster 1	**Cluster 2**	**Cluster 3**
(2, 10) → A1	(8, 4) → A3	(2, 5) → A2
(5, 8) → A4	(7, 5) → A5	(1, 2) → A7
(4, 9) → A8	(6, 4) → A6	

- Next, we need to re-compute the new cluster centers (means). We do so, by taking the mean of all points in each cluster.

  For Cluster 1, we have $\left(\frac{(2+5+4)}{3}, \frac{(10+8+9)}{3}\right) = (3.66, 9)$

  For Cluster 2, we have $\left(\frac{(8+7+6)}{3}, \frac{(4+5+4)}{3}\right) = (7, 4.33)$

  For Cluster 3, we have $\left(\frac{(2+1)}{2}, \frac{(5+2)}{2}\right) = (1.5, 3.5)$

Cluster 1	Cluster 2	Cluster 3
(2, 10) → A1	(8, 4) → 3	(2, 5) → A2
(5, 8) → A4	(7, 5) → A5	(1, 2) → A7
(4, 9) → A8	(6, 4) → A6	

Time complexity of K-Means algorithm:

O(tkn)

where, t: number of iterations, n: number of items to be clustered, k: number of clusters

Advantage of K-Means Algorithm:

- K-Means algorithm uses Euclidian distance measure for efficient solutions.

Disadvantages of K-Means Algorithm:

- This algorithm doesn't work with categorical data. As means must be defined with numerical data.
- Only convex(curved) shapes clusters are created.
- This algorithm doesn't handle noisy data and outliers well. A small number of such data can influence mean value.
- Typical value for k(number of clusters to be created) is range from 2 to 10.
- It is not time efficient.
- This algorithm doesn't scale well for big data.
- It does not save distance calculation information. If it will save this information then in the next step of iteration, number of distance calculations must be reduced.

- One variation of K-Mean is k-modes which handles categorical data. It uses modes instead of means.

4.3.2 k-Medoids: A Representative Object-based Technique

Concept of PAM:

- k-Medoids clustering is called **Partitioning Around Medoids** (PAM).
- Leonard Kaufman and Peter J. Rousseeuw given the name k-Medoid with their PAM (Partitioning Around Medoids) algorithm.
- A problem with K-Means algorithm is clustering algorithm is that the final centroids(means) are not interpretable. i.e. centroids are not the actual point but the mean of points present in that cluster. So, the coordinates of centroids that never looks like real points from the dataset.
- This algorithm chooses actual data points as centres (medoids or exemplars).
- This algorithm can be used with random or arbitrary distance measures.
- This algorithm is working as more robust than k-means algorithm for handling outliers.
- This is a traditional partitioning technique used for clustering.
- *n* objects must be placed into k clusters. Value of k should be known in advance.
- The basic idea in this algorithm is to make final centroids as actual data points. This allows the centroids to be interpretable. So, this algorithm is called as a **representative object-based technique**.

Medoid:

- The medoid of a cluster is defined as the object in the cluster whose average dissimilarity to all the objects in the cluster is minimal, that is, it is a most centrally located point in the cluster.
- Clusters are represented using medoids.

PAM Algorithm:

Simple steps of algorithm are as follows:

- Use real objects to represent the clusters (called medoids).

  Step 1: Select k representative objects arbitrarily.

  Step 2: For each pair of selected object(i) and non-selected object(h),

  Calculate the Total swapping Cost (TC_{ih}).

  Step 3: For each pair of i and h,

  If $TC_{ih} < 0$, i is replaced by h.

  Then assign each non-selected object to the most similar representative object.

  Step 4: Repeat Steps 2-3 until there is no change in the medoids or in TC_{ih}.

- **Total swapping cost**: $TC_{ih} = \sum_j C_{jih}$

  Where, C_{jih} is the cost of swapping i with h for all non-medoid objects j.

  C_{jih} will vary depending upon different cases.

Algorithm:

Input:

D = {t_1, t_2, …, t_n} //Set of elements

A //Adjacency matrix showing distance between element s

k//Number of desired clusters

Output :

K//Set of clusters

PAM Algorithm:

```
1. arbitrarily select k medoids from D;
2. repeat
3. for each t_h not a medoid do
4. for each medoid t_i do
5. calculate TC_ih ;
6. find i, h where TC_ih is the smallest;
7. if TC_ih < 0, then
8. replace medoid t_i with t_h;
9. until TC_ih > = 0;
10. for each t_i ∈ D do
11. assign t_i to K_j , where dis(t_i, t_j) is the smallest over all
    medoids;
```

Explanation:

- At first step, a random set of k items are considered as medoids.
- At every step of algorithm every point in the cluster will be examined that can it be medoids for that cluster?
- If such data items are found then replace the existing medoids of that cluster with new medoids.
- All pairs of medoids with non-medoid objects (data points) were formed to check whether it improves the quality of the algorithm and clustering will be done at its best, replaces the medoids.
- Quality of algorithm is determined by checking sum of all distances between non medoid objects and medoids within the clusters.
- A data item assigned to the cluster represented by medoid to which it is closest (i.e. having minimum distance between medoid and data point).
- Assume K_i is the cluster represented by medoid t_i. Now, suppose that t_i is a current medoid and user want to determine whether it should be replaced with a non-medoid t_h. This swapping can be done only if the overall impact to the cost (sum of the distances to cluster medoids) emphasizes an improvement.
- **The cost of exchange or swapping of medoids** with non-medoid objects must be determined. i.e. the cost is the change to the sum of all distances from items to their cluster medoids.

  Cost: C_{jih} , medoid: t_i, Non-medoid: t_h
- Following are the four situations must be considered while calculating this cost.
  - $t_j \in K_i$; but $\exists$ another medoid t_m where $dis(t_j, t_m) \Leftarrow dis(t_j, t_h)$
  - $t_j \in K_i$; but $dis(t_j,t_h) \Leftarrow dis(t_j, t_m)$ $\forall$ other medoids t_m
  - $l_j \in K_m$, $\notin K_i$, and $dis(t_j, t_m) \Leftarrow dis(t_j, t_h)$
  - $f_j \in K_m$, $\notin K_i$, but $dis(t_j, t_h)$, $\Leftarrow dis(t_j, t_m)$
- **Total effect to quality by changing of medoid** is given by following equation.

$$TC_{ih} = \sum_{j=1}^{n} C_{jih}$$

Difference between Centroid and Medoid:

- Medoid is always selected from dataset but Centroid may or may not select from data set.
- Centroid does not belong to data set but medoid always belongs to data set.

Examples of Centroid and Medoid:

1. To calculate medoid or centroid, consider data set as {1, 2, 3}.
   a. Centroid is $= \frac{(1+2+3)}{3} = \frac{6}{3} = 2$ (addition of all data objects/number of objects)
   b. Medoid (central value from data) is 2

2. To calculate medoid or centroid data set as {1, 3, 7, 10}

   c. Centroid is = $\frac{(1 + 3 + 7 + 10)}{4} = \frac{21}{4} = 5.25$

   d. Medoid is = 3 or 7

3. To calculate medoid or centroid data set as {51, 13, 22, 65, 92}

   e. Centroid is = $\frac{(51 + 13 + 22 + 65 + 92)}{5} = \frac{243}{5} = 48.6$

   f. Medoid is = 22

Examples of PAM:

Example 6:

Dataset = {1, 2, 3, 20, 21, 22} and K = Number of clusters = 2.

Initially assume that medoids are 1 and 2 which are represented by M1(1) M2(2).

Absolute distance measure technique is used to solve this example.

Step 1:

Table 4.5 (a): Step 1 for K-medoid Example 6

Dataset	M1(1)	M2(2)	Minimum of column M1(1) and M2(2)
1	(1 – 1) = 0	(2 – 1) = 1	0 < 1 = 0
2	(2 – 1) = 1	(2 – 2) = 0	1 > 0 = 0
3	(3 – 1) = 2	(3 – 2) = 1	2 > 1 = 1
20	(20 – 1) = 19	(20 – 2) = 18	19 > 18 = 18
21	(21 – 1) = 20	(21 – 2) = 19	20 > 19 = 19
22	(22 – 1) = 21	(22 – 2) = 20	21 > 20 = 20
			Σ i.e. (0 + 0 + 1 + 18 + 19 + 20) = 58

Step 2:

- Now we cannot change medoids simultaneously. So other non-medoid objects in the data set who will become medoids are 3, 20, 21, 22. We can change one medoid at a time. Currently we keep M2(2) i.e. medoid 2 as it is. So, we can change medoid M1(1)and there are four choices 3, 20, 21, 22. We have to find the best choice.

  (keep in mind absolute so take modulus of answer)

- Now we have to find Total Cost.

  $TC_{1,3}$ (Total Cost of changing medoid from 1 to 3)

  $TC_{1,20}$ (Total Cost of changing medoid from 1 to 20)

  $TC_{1,21}$ (Total Cost of changing medoid from 1 to 21)

  $TC_{1,22}$ (Total Cost of changing medoid from 1 to 22)

Table 4.5 (b): Step 2 for K-medoid Example 6

M2(2) keep this medoid

M1(1) needs to change so choices are

Dataset	M(2)	M(3)		M(20)		M(21)		M(22)	
			Min with M(2)		Min with M(2)		Min with M(2)		Min with M(2)
1	1	(1 – 3) = 2	min(1, 2) = 1	(1 – 20) = 19	min(1, 19) = 1	(1 – 21) = 20	min(1, 20) = 1	(1 – 22) = 21	min(1, 21) = 1
2	0	(2 – 3) = 1	min(0, 1) = 0	(2 – 20) = 18	min(0, 18) = 0	(2 – 21) = 19	min(0, 19) = 0	(2 – 22) = 20	min(0, 20) = 0
3	1	(3 – 3) = 0	min(1, 0) = 0	(3 – 20) = 17	min(1, 17) = 1	(3 – 21) = 18	min(1, 18) = 1	(3 – 22) = 19	min(1, 19) = 1
20	18	(20 – 3) = 17	min(18, 17) = 17	(20 – 20) = 0	min(18, 0) = 0	(20 – 21) = 1	min(18, 1) = 1	(20 – 22) = 2	min(18, 2) = 2
21	19	(21 – 3) = 18	min(19, 18) = 18	(21 – 20) = 1	min(19, 1) = 1	(21 – 21) = 0	min(19, 0) = 0	(21 – 22) = 1	min(19, 0) = 0
22	20	(22 – 3) = 19	min(20, 19) = 19	(22 – 20) = 2	min(20, 2) = 2	(22 – 21) = 1	min(20, 1) = 1	(22 – 22) = 0	min(20, 1) = 1
New Cost		TC1, 3	= 55	TC1, 20	= 5	TC1, 21	= 4	TC1, 22	= 5
New Cost – Old Cost			**55 – 58 = –3**		**5 – 58 = –53**		**4 – 58 = –54**		**5 – 58 = –53**

$TC_{1,3} = 1 + 0 + 0 + 17 + 18 + 19 = 55$

$TC_{1,20} = 1 + 0 + 1 + 0 + 1 + 2 = 5$

$TC_{1,21} = 1 + 0 + 1 + 1 + 0 + 1 = 4$

$TC_{1,22} = 1 + 0 + 1 + 2 + 0 + 1 = 5$

(Kindly note: don't take absolute value i.e. modulus for new cost-old cost)

- Comparing above Total Costs we get –54 is minimum cost. As TC1,21

  So, we will keep 21 as medoid. Now we have M2(2) and M1(21) as medoid.

Step 3:

- So, medoid 1 is as it is M1(1), instead of 2 as medoid we can find best choice from non-medoid objects from data set i.e. 3, 20, 21, 22.

  $TC_{2,3}$ (Total Cost of changing medoid from 2 to 3)

  $TC_{2,20}$ (Total Cost of changing medoid from 2 to 20)

  $TC_{2,21}$ (Total Cost of changing medoid from 2to 21)

  $TC_{2,22}$ (Total Cost of changing medoid from 2 to 22)

Table 4.5 (c): Step 3 for K-medoid Example 6

Dataset	M(1)	M(3)		M(20)		M(21)		M(22)	
			Min with M(1)		Min with M(1)		Min with M(1)		Min with M(1)
1	0	(1− 3) = 2	min(0, 2) = 0	(1-20) = 19	min(0, 19) = 0	(1 − 21) = 20	min(0, 20) = 0	(1 − 22) = 21	min(0, 21) = 0
2	1	(2 − 3) = 1	min(1, 1) = 1	(2 − 20) = 18	min(1, 18) = 1	(2 − 21) = 19	min(1, 19) = 1	(2 − 22) = 20	min(1, 20) = 1
3	2	(3 − 3) = 0	min(2, 0) = 0	(3 − 20) = 17	min(2, 17) = 2	(3 − 21) = 18	min(2, 18) = 2	(3 − 22) = 19	min(2, 19) = 2
20	19	(20 − 3) = 17	min(19, 17) = 17	(20 − 20) = 0	min(19, 0) = 0	(20 − 21) = 1	min(19, 1) = 1	(20 − 22) = 2	min(19, 2) = 2
21	20	(21 − 3) = 18	min(20, 18) = 18	(21 − 20) = 1	min(20, 1) = 1	(21 − 21) = 0	min(20, 0) = 0	(21 − 22) = 1	min(20, 1) = 1
22	21	(22 − 3) = 19	min(21, 19) = 19	(22 − 20) = 2	min(21, 2) = 2	(22 − 21) = 1	min(21, 1) = 1	(22 − 22) = 0	min(21, 0) = 0
New Cost		**TC2, 3**	**= 55**	**TC2, 20**	**= 6**	**TC2, 21**	**= 5**	**TC2, 22**	**= 6**
New Cost – Old Cost			**55 − 58 = −3**		**6 − 58 = −52**		**5 − 58 = −53**		**6 − 58 = −52**

$TC_{2,3} = 0 + 1 + 0 + 17 + 18 + 19 = 55$

$TC_{2,20} = 0 + 1 + 2 + 0 + 1 + 2 = 6$

$TC_{2,21} = 0 + 1 + 2 + 1 + 0 + 1 = 5$

$TC_{2,22} = 0 + 1 + 2 + 2 + 1 + 0 = 6$

$TC_{2,21} = -53$ which is minimum cost.

- Now, $TC_{1,21} = -54$ and $TC_{2,21} = -53$ comparing these two values $TC_{2,21}$ is minimum value.

  So, we are swapping 1 with 21.

  Therefore, new medoids are (2,21).

Example 7:

K-Medoid for 2-D Data:

{(8, 7), (3, 7), (4, 9), (9, 6), (8, 5), (5, 8), (7, 3), (8, 4),(7, 5),(4, 5)}

It is improved version of k-means.

D = |x – x1| + |y – y1| + |z – z1|… for any dimension use this formula.

K = 2: Number of clusters to form: Select two points as initial medoids as (8, 5), (4, 5) for two clusters.

Table 4.6 (a): Step 1 for Example 7

x	y	For cluster C1 (8, 5) D = |x – 8| + |y – 5|	For cluster C2 (4, 5) D = |x – 4| + |y – 5|	Min (column 1, column 2)	Final clusters
8	7	|8 – 8| + |7 – 5| = 0 + 2 = 2	|8 – 4| + |7 – 5| = 4 + 2 = 6	Min(2, 6) = 2	C1
3	7	|3 – 8| + |7 – 5| = 5 + 2 = 7	|3 – 4| + |7 – 5| = 1 + 2 = 3	Min(7, 3) = 3	C2
4	9	|4 – 8| + |9 – 5| = 4 + 4 = 8	|4 – 4| + |9 – 5| = 0 + 4 = 4	Min(8, 4) = 4	C2
9	6	|9 – 8| + |6 – 5| = 1 + 1 = 2	|9 – 4| + |6 – 5| = 5 + 1 = 6	Min(2, 6) = 2	C1
8	5	|8 – 8| + |5 – 5| = 0 + 0 = 0	|8 – 4| + |5 – 5| = 4 + 0 = 4	Min(0, 4) = 0	C1
5	8	|5 – 8| + |8 – 5| = 3 + 3 = 6	|5 – 4| + |8 – 5| = 1 + 3 = 4	Min(6, 4) = 4	C2
7	3	|7 – 8| + |3 – 5| = 1 + 2 = 3	|7 – 4| + |3 – 5| = 3 + 2 = 5	Min(3, 5) = 3	C1
8	4	|8 – 8| + |4 – 5| = 0 + 1 = 1	|8 – 4| + |4 – 5| = 4 + 1 = 5	Min(1, 5) = 1	C1
7	5	|7 – 8| + |5 – 5| = 1 + 0 = 1	|7 – 4| + |5 – 5| = 3 + 0 = 3	Min(1, 3) = 1	C1
4	5	|4 – 8| + |5 – 5| = 4 + 0 = 4	|4 – 4| + |5 – 5| = 0 + 0 = 0	Min(4, 0) = 0	C2

- First set K1 = {(8, 7), (9, 6), (8, 5), (5, 8), (7, 3), (8, 4), (7, 5)}
- Second set K2 = {(3, 7), (4, 9), (5, 8), (4, 5)}
- Find Total Cost:

  Cost = {(8, 5), (8, 7)} + {(8, 5), (9, 6)} + {(8, 5), (7, 3)} + {(8, 5), (8, 5)} + {(8, 5), (8, 4)} + {(8, 5), (7, 5)}

  Cost = 2 + 2 + 0 + 3 + 1 + 1 = 9

  Cost = {(4, 5), (3, 7)} + {(4, 5), (4, 9)} + {(4, 5), (5, 8)} + {(4, 5), (4, 5)}

  Cost = 3 + 4 + 4 + 0 = 11

  Total Cost (8, 5), (4, 5) = 9 + 11 = 20
- Now change the medoids with the nearest value and check for the cost change the medoid first(8, 4).

Table 4.6 (b) : Step 2 for Example 7

x	y	For cluster C1 (8, 4) D = |x – 8| + |y – 5|	For cluster C2 (4, 5) D = |x – 4| + |y – 5|	Min(column1, column2)	Final clusters
8	7	|8 – 8| + |7 – 4| = 0 + 3 = 3	|8 – 4| + |7 – 5| = 4 + 2 = 6	Min(3, 6) = 3	C1
3	7	|3 – 8| + |7 – 4| = 5 + 3 = 8	|3 – 4| + |7 – 5| = 1 + 2 = 3	Min(8, 3) = 3	C2
4	9	|4 – 8| + |9 – 4| = 4 + 5 = 9	|4 – 4| + |9 – 5| = 0 + 4 = 4	Min(9, 4) = 4	C2
9	6	|9 – 8| + |6 – 4| = 1 + 2 = 3	|9 – 4| + |6 – 5| = 5 + 1 = 6	Min(3, 6) = 3	C1
8	5	|8 – 8| + |5 – 4| = 0 + 1 = 1	|8 – 4| + |5 – 5| = 4 + 0 = 4	Min(1, 4) = 1	C1
5	8	|5 – 8| + |8 – 4| = 3 + 4 = 7	|5 – 4| + |8 – 5| = 1 + 3 = 4	Min(7, 4) = 4	C2
7	3	|7 – 8| + |3 – 4| = 1 + 1 = 2	|7 – 4| + |3 – 5| = 3 + 2 = 5	Min(2, 5) = 2	C1
8	4	|8 – 8| + |4 – 4| = 0 + 0 = 0	|8 – 4| + |4 – 5| = 4 + 1 = 5	Min(0, 5) = 0	C1
7	5	|7 – 8| + |5 – 4| = 1 + 1 = 2	|7 – 4| + |5 – 5| = 3 + 0 = 3	Min(2, 3) = 2	C1
4	5	|4 – 8| + |5 – 4| = 4 + 1 = 5	|4 – 4| + |5 – 5| = 0 + 0 = 0	Min(5, 0) = 0	C2

- First set K1 = {(8, 7), (9, 6), (8, 5), (7, 3), (8, 4), (7, 5)}
- Second set K2 = {(3, 7), (4, 9), (5, 8), (4, 5)}
- Calculate total cost:

  Cost = {(8, 4),(8, 7)} + {(8, 4), (9, 6)} + {(8, 4), (8, 5)} + {(8, 4), (7, 3)} + {(8, 4), (8, 4)} + {(8, 4), (7, 5)}

  Cost = 3 + 3 + 1 + 2 + 0 + 2 = 11

  Cost = {(4, 5), (3, 7)} + {(4, 5), (4, 9)} + {(4, 5), (5, 8)} + {(4, 5), (4, 5)}

  Cost = 3 + 4 + 4 + 0 = 11

  Therefore, **Total cost (8, 4), (4, 5) = 22.**
- Now comparing two cost if (Cost2 - Cost1) > 0 then terminate

  In this situation (22 – 20) > 0 i.e. 2 > 0
- So final answer is :

  C1 {(8, 7), (9, 6), (8, 5), (7, 3), (8, 4), (7,5)}

  C2 {(3, 7), (4, 9), (5, 8), (4, 5)}
- To display clusters graphically:

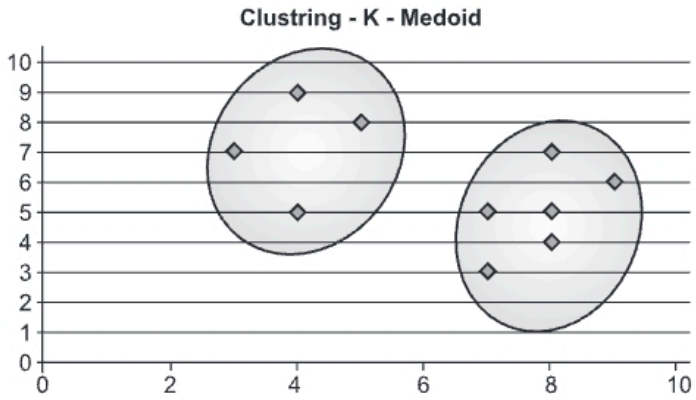


Fig. 4.10: Display of clusters in graphical mode for Example 7

Time complexity of PAM is:

- $O(k(n-k)^2)$ for each iteration where n is number of objects and k is number of clusters.

Advantages of PAM:

- This algorithm simple to understand and easy to implement.
- K-Medoid Algorithm is fast and converges in a fixed number of steps.
- PAM is less sensitive to outliers than other partitioning algorithms.
- PAM works efficiently for small data sets but does not scale well for large data sets.

Disadvantages of PAM:

- This algorithm not suitable for clustering non-spherical (arbitrary/random shaped) groups of objects. This is because it relies on minimizing the distances between the non-medoid objects and the medoid (the cluster centre) – briefly, it uses compactness as clustering criteria instead of connectivity.
- It may obtain different results for different runs on the same dataset because the first k medoids are chosen randomly.

4.4 INTRODUCTION TO ASSOCIATION RULE MINING

- Use of data mining is to extract useful patterns from data irrespective of type of data. Finding patterns is noting but finding relationships among data. These relationships help users to create groups of data items.
- The kind of purchasing items in group denotes some kind of relationships among them. This relationship in data mining is called as association rule.
- These association rules are often used in marketing, advertising, inventory applications and in retail business.
- It is also used in fault prediction in telecommunication networks. Association rules are used to show the relationships between data items. Association rules are used to detect common usage in purchasing items.
- The existence of item in a transaction is best captured by a likehood measure or a probability.
- For example: Consider patients medical data set which consists of patient symptoms and illness that a patient suffered from. When association rule is applied to this data set doctor can find out any correlation among the symptoms and illness.

- Following example will clear idea about association rule mining.
- **Example:** A Grocery store keeps record of items purchased by customer at bill counter. The manager receives transaction report summary. This summary contains different types of items and their quantity. Such transaction report summary will be generated periodically. Manager observed that number of times Butter is purchased at the same number of times bread is purchased. The percentage of Butter and bread is 100%. It is also found that 33.3% of the time Butter is purchased, Jelly is also purchased. However, Butter exists in only about 50% of the overall transactions.
- The database should be in the form of tuples where association rule is to be found. Each tuple is the list of items purchased at one time.
- The support of an item (or set of items) is the percentage of transactions in which that item (or items) occurs.

Example 8:

Consider Table 4.7 which shows sample transaction database of a Grocery shop:

Table 4.7: Sample Database for Example 8

Transaction	Items
t_1	Bread, Jelly, Butter
t_2	Bread, Butter
t_3	Bread, Milk, Butter
t_4	Biscuit, Bread
t_5	Biscuit, Milk

- For above Example 8, Support will be calculated as follows:

Table 4.7 (a): Count of Support for Example 8

Set	Support	Set	Support
Biscuit	40	Biscuit, Bread, Milk	0
Bread	80	Biscuit, Bread, Butter	0
Jelly	20	Biscuit, Jelly, Milk	0
Milk	40	Biscuit, Jelly, Butter	0
Butter	60	Biscuit, Milk, Butter	0
Biscuit, Bread	20	Bread, Jelly, Milk	0
Biscuit, Jelly	0	Bread, Jelly, Butter	20
Biscuit, Milk	20	Bread, Milk, Butter	20
Biscuit, Butter	0	Jelly, Milk, Butter	0
Bread, Jelly	20	Biscuit, Bread, Jelly, Milk	0
Bread, Milk	20	Biscuit, Bread, Jelly, Butter	0
Bread, Butter	60	Biscuit, Bread, Milk, Butter	0
Jelly, Milk	0	Beer, Jelly, Milk, Butter	0
Jelly, Butter	20	Bread, Jelly, Milk, Butter	0
Milk, Butter	20	Biscuit, Bread, Jelly, Milk, Butter	0
Biscuit, Bread, Jelly	0		

- **Definition:** Given a set of items $I = \{I_1, I_2, \ldots, I_m\}$ and a database of transactions $D = \{t_1, t_2, \ldots, t_n\}$ where, $t_i = \{I_{i1}, I_{i2}, \ldots, I_{ik}\}$ and $I_{ij} \in I$, an association rule is an implication of the form $X \Rightarrow Y$ where $X, Y \subset I$ are sets of items called item sets and $X \cap Y = \phi$.
- In simple words , if $\{i_1, i_2, \ldots, i_k\} \rightarrow j$ means: if a basket contains all of $i_1, \ldots, i_k$ then it is likely to contain j.
- The support(s) for an association rule $X \Rightarrow Y$ is the percentage of transactions in the database that contain $X \cup Y$.
- The confidence or strength(α) for an association rule $X \Rightarrow Y$ is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that contain X.

Example 9:

Example of rule, support, confidence and lift.

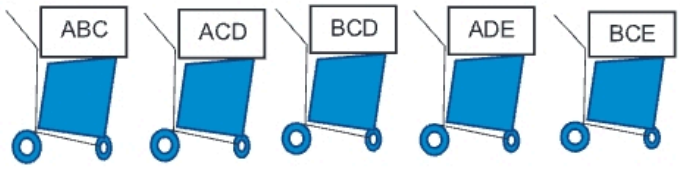


Fig. 4.11: Market basket analysis for Example 9

Table 4.8: Rule, Support and Confidence for Example 9

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B\&C \Rightarrow D$	1/5	1/3	5/6

Example10:

- Basket contains the items as follows:

  $B_1 = \{a, b, c\}$ $B_2 = \{a, p, j\}$

  $B_3 = \{a, b\}$ $B_4 = \{c, j\}$

  $B_5 = \{a, b, c\}$ $B_6 = \{a, b, c, j\}$

  $B_7 = \{c, b, j\}$ $B_8 = \{b, c\}$

  Using association rule we can say $\{a, b\} \rightarrow c$

Confidence of rule: $Conf(X \rightarrow Y) = supp(X \cup Y)/supp(X)$.

For above example, Confidence = 2/4 = 50%.

- Selection of association rules is based on support and confidence. Confidence measures strength of the rule. Support measure frequency of its occurrence in database. Generally large confidence values and smaller support is used. For example, look at table 4.7 No Bread $\Rightarrow$ Butter it has α = 75% (this rule holds 75% of time i.e. 3/4 times Bread occurs as Butter)

- Jelly ⇒ Milk (this rule does not hold because they were not purchased together). So, an advertising agency will advertise for bread and butter by keeping some discount on any one of them to increase the sale of that particular product.
- Efficiency of association rule algorithms depends on number of times database is scanned and maximum number counted itemsets.

4.4.1 Market Basket Analysis, Items, Itemsets and Large Itemsets

Market Basket Analysis :

- Market basket analysis is a data mining technique used by retailers to increase sales of items by well understanding customer purchasing patterns.
- Analysis of large data sets for example to find out purchase history, to find out product groupings, to find out products that are likely to be purchased together can be done using market basket analysis.
- Combinations of items that occur frequently in the transactions will be examined. Such examinations of transactions allow retailers to identify relationships between the items that people frequently purchase together.
- For example, people who buy bread and peanut butter also buy jelly. People who buy sugar may buy milk.
- Association rules are used in market basket analysis to predict the occurrence of an items based on the presence of other items in a transaction. It is applied to uncover customer purchase patterns and drive personalized marketing decisions.
- Market basket analysis is used to determine which products are brought together. This information is used to design super market layout so that to design promotional campaigns such that products' purchase can be improved.
- Implementation of market basket analysis requires a background in statistics and data science, as well as some algorithmic computer programming skills.

Itemsets:

- A collection of items purchased by a customer is an itemset.

Large Itemsets :

- In general association rule mining is a two-step process:
  1. **Finding all frequent itemsets:** Each of found itemsets will occur as frequently as a predetermined minimum support count, min sup.
  2. **Generation of strong association rules from the frequent itemsets:** These generated rules must satisfy minimum support and minimum confidence.
- **Definition:** A large (frequent) itemset is an itemset whose number of occurrences is above a threshold, s. Use the notation L to indicate the complete set of large itemsets and *l* to indicate a specific large itemset.
- When large itemset is found then association rule is also found i.e. $X \Rightarrow Y$ .This must have $X \cup Y$ in the set of frequent itemsets.
- Kindly note that subset of any large itemset is also large.

Large Itemset Algorithm:

Input:

```
D // Database of transactions
I // Items
L  // Large itemsets
S // Support
α // Confidence
```

Output :

```
R // Association Rules satisfying s and α
```

ARGen algorithm:

```
1. R = ϕ ;
2. for each l ∈ L do
3. for each x ⊂ 1 such that x ≠ ϕ do
4. if  support(l)  >= α then
5. support(x)
6. R=R U { x ==> (1 - x)} ;
```

Example 10:

Example of Large Itemset (ARGen Algorithm).

L = {{Biscuit}, {Bread}, {Milk}, {Butter}, {Bread, Butter}}

Consider l = {Bread, Peanut Butter}

There are two non-empty subsets of l: {Bread} and {Butter}

With first non-empty subset,

$$\frac{\text{Support } (\{\text{Bread, Butter}\})}{\text{Support } (\{\text{Bread}\})} = \frac{60}{80} = 0.75$$

The above example state that confidence of association rule is `Bread=>Butter` is 75% which is greater than α so it is valid association rule added to R.

Applications of Market Basket Analysis are as follows:

- **Retailer's technique:** This technique used by large retailers like Amazon, Flipkart to uncover associations between items.
- **Manufacturing:** Predictive analysis of equipment failure can be done.
- **Pharmaceutical:** Discovery of co-occurrence relationships among diagnosis in case of disease in patient.
- **Bioinformatics:** Pharmaceutical active ingredients prescribed to different patient groups according to their disease.
- **Financial/Criminology:** Fraud detection based on credit card usage data.
- **Customer Behavior:** Associating purchases with demographic and socio-economic data of customers.

4.5 APRIORI ALGORITHM

- This is one of the well-known best algorithms for generating association rules.
- It is powerful algorithm for mining frequent itemsets for Boolean association rules.
- Name of the algorithm is based on the priority it uses. i.e. Apriori.
  - Apriori property based on fact that it uses prior knowledge of frequent itemset properties.
  - This property uses iterative approach as level wise search.
  - At any level k itemsets are used to explore (k + 1) itemsets.
  - At first step, whole database is scanned and count of each individual item is found. Assume minimum support.
  - Consider those items which satisfy minimum support. Set of such frequent itemset is found.
  - The resulting set is denoted as L_1 (Level 1).
  - L_1 is used to find L_2 (L_2 is the frequent itemsets 2).
  - L_2 is used to find L_3 and the process continues till no more frequent k-itemsets can be found.
  - Every time database has to be scanned for find L_k frequent itemset.

Apriori Property(Large Itemset Property):

- Any subset of a large itemset must be large.

  OR

- All nonempty subsets of a frequent itemset must also be frequent.

- Apriori employs an iterative approach known as level-wise search, where k-itemsets are used to explore k + 1-itemsets.
  - Initially, scan DB once to get frequent 1-itemset.
  - Generate length (k + 1) candidate itemsets from length k frequent itemsets.
  - Test the candidates against DB.
  - Terminate when no frequent or candidate set can be generated.
- **Apriori Pruning Principle:** If there is any itemset which is infrequent, its superset should not be generated/tested.

  Method:

  L_k denotes the set of frequent k-itemsets- Large itemset.

  C_k is the superset of L_k – Candidate for Large itemset.

  - Apriori Algorithm is a Two-step process is followed consisting of **join** and **prune** actions to generate L_k from L_{k-1}.
    - **Join Step:** Apriori assumes that items within a transaction or itemset are sorted in lexicographic order.

      The Candidate set C_k is generated by taking the join $L_{k-1} \times L_{k-1}$, where members of L_{k-1} are joinable if their first k–2 items are in common. This ensures that no duplicates are generated.

- **Prune step:** To reduce the size of Ck, Apriori property is used as follows:
  - Any (k−1)-itemset that is not frequent cannot be a subset of a frequent k-itemset. Hence if any (k−1)-subset of a candidate k-itemset is not in L_{k-1}, the candidate cannot be frequent and can be removed from C_k.
  - The count of each candidate in C_k is used to determine L_k (minimum support count).

Algorithm Apriori_generate(Lk):

```
1. for each itemset l1 in Lk
2. for each itemset l2 in Lk
3. If  k – 1 elements in l1 and l2 are equal
   // If l1[1] = l2[1] and l1[2] = l2[2] and … l1[k – 1] = l2[k – 1] and
   //l1[k] < l2[k]
4. C = l1 × l2
5. add C to  Ck + 1
6. for each k subset s of c
7. if  s does not belong to Lk then
8. delete c
9. break
```

The Apriori Algorithm:

C_k: Candidate itemset of size k

L_k: frequent itemset of size k

Algorithm Apriori:

```
1. L1 = {frequent items};
2. for(k = 1; Lk ! = ϕ; k++) do
3. begin
4. Ck+! = Apriori_generate(Lk)
   // candidates generated from Lk;
5. for each transaction t in database do
6. increment the count of all candidates in Ck+!  that are contained in t
7. Lk + 1 = candidates in Ck+1 with min_support
8. end
9. return ∪k Lk;
```

Example 11:

Consider database shown in Table 4.9 where supmin=2. Apply Apriori algorithm and find frequent itemsets.

Table 4.9: Sample Database for Example 11

Tid	Items
10	A, B, E
20	B, E
30	B, C
40	A, B, D
50	A, C
60	B, C
70	A, C
80	A, B, C, E
90	A, B, C

Solution:

Tid	Items
20	B, E
30	B, C
40	A, B, D
50	A, C
60	B, C

C_1 — 1st scan →

{A}	6
{B}	7
{C}	6
{D}	1
{E}	3

L_1 →

Itemset	sup
{A}	6
{B}	7
{C}	6
{E}	3

L_1

Itemset	sup
{A}	6
{B}	7
{C}	6
{E}	3

$L_1 \times L_1$ →

C_2

Itemset	sup
{A, B}	4
{A, C}	4
{A, E}	2
{B, C}	4
{B, E}	3
{C, E}	1

2nd scan →

L_2

Itemset	sup
{A, B}	4
{A, C}	4
{A, E}	2
{B, C}	4
{B, E}	3

L_2

Itemset	Sup
{A, B}	4
{A, C}	4
{A, E}	2
{B, C}	4
{B, E}	3

$L_2 \times L_2$ →

C3 = {{A, B, C}, {A, B, E}, {A, C, E}, {B, C, E}}

The 2 item subsets of {A, B, C} are {A, B}, {B, C}, {A, C} which are all in L2

The 2 item subsets of {A, B, E} are {A, B}, {B, E}, {A, E} which are all in L2

The 2 item subsets of {A, C, E} are {A, C}, {C, E} and {A, E}. {C, E} is not in L2

Remove {A, C, E}

The 2 item subsets of {B, C, E} are {B, C}, {C, E} and {B, E} . {C, E} is not in L2

Remove {B, C, E}

C_3

Itemset
{A, B, C}
{A, B, E}
{A, C, E}
{B, C, E}

3rd scan →

L_3

Itemset	sup
{A, B, C}	2
{A, B, E}	2

Itemset
{A, B, C}
{A, B, E}

$L_3 \times L_3$ →

C_4 = {{A, B, C, E}}

The 3 item subsets of {A, B, C, E} are {A, B, C}, {B, C, E}, {A, C, E} and {A, B, E},

{B, C, E} and {A, C, E} are not in L_3

Remove {A, B, C, E}

Thus C_4 is empty and algorithm terminates.

Time complexity of algorithm:

$O(2^d)$ where, d: Total number of unique items in the transaction dataset.

Advantages of Apriori Algorithm:

- This algorithm uses breadth-first search.
- Easy to implement.
- Uses large itemset (Apriori) property.

Disadvantages of Apriori Algorithm:

- This algorithm scans the database multiple times for generating candidate sets.
- Apriori algorithm requires huge memory space as they deal with large number of candidates itemset generation. So, space complexity is high.
- Apriori algorithm execution time is more wasted in producing candidates every time.

There are different variations of Apriori algorithm to remove its drawbacks. Few of them are Partition algorithm, Sampling algorithm, Dynamic Itemset Counting (DIC) and Direct hashing and pruning (DHP).

4.6 KINDS OF ASSOCIATION RULES

- Association rules can be categorized as follows:
  1. Types of values handled
     - Boolean association rules
     - Quantitative association rules
  2. Levels of abstraction involves
     - Single level association rules
     - Multilevel association rules
  3. Dimensions of data involved
     - Single dimensional association rules
     - Multidimensional association rules

4.6.1 Mining Multilevel Association Rules

- It is the variation of generalized rules.
- Due to sparsity (i.e. Not having enough something) characteristics in multidimensional database, it is difficult to find strong associations between data items at lower level.
- Strong associations can be found at higher level. Data mining applications should provide ability to find strong association between data items at multiple levels.
- Association rules generated from mining data at multiple levels of abstraction are called multiple-level or multilevel association rules.
- In this method, itemsets may occur from any level in the hierarchy.
- Using different variations of Apriori algorithm the concept hierarchy is traversed in a top-down manner and large itemsets are generated.
- If large itemset found at level i then large itemsets are also generated at level i+1. Large k-itemsets at one level in the concept hierarchy are used as candidates to generate large k-itemsets for children at the next level.
- There is more support for itemsets occurring at higher levels in the concept hierarchy.
- Minimum support is always checking for each level in the concept hierarchy.
- Frequency of itemsets at higher levels is more then frequency of itemsets at lower levels. Following are the rules that can be applied:
  - The minimum support for all nodes in the hierarchy at the same level is identical.
  - If α is the minimum support for level i in the hierarchy and αi-1 is the minimum support for level i -1, then $\alpha_i - 1 > \alpha_i$;
- There are many variations for this approach. These variations are related with support threshold. Few of them are described as follows:
  - **Uniform Support:** Same value of minimum support is used for every level of abstraction.

- **Reduced Support:** At every level of abstraction, there is its own minimum support threshold. So minimum support at lower levels reduces.
- **Group-based Support:** Users and domain experts always think that group mechanism can be comfortable for mining specific items. For example, a user could set up the minimum support thresholds based on product price, in order to pay particular attention to the association patterns containing items in these categories.

4.6.2 Constraint Based Association Rules Mining

- Association rules are used in numerous studies which are focusing on performance and functionality issues on the rules. Studies have been carried out firstly for efficient computation of association rules then computation of needed rules. In efficient computation- Apriori algorithm, Lattice of itemsets techniques were used.
- Most of these studies basically considered the data mining exercise in isolation. These studies did not explore how data mining can interact with the human user, which is a key component in the broader picture of knowledge discovery in databases. Hence, studies provide little or no support for user focus. Therefore, the user usually needs to wait for a lengthy period of time to get numerous association rules, out of which only a small fraction may be interesting to the user. In other words, the user often incurs a high computational cost that is disproportionate to what he wants to get. This calls for constraint-based association rule mining.
- It is might possible that data mining does not cover all the rules of given data set. Some rules may ne unrelated or show uninteresting for data miners/users. This is known as constraint-based association rule mining.
- It aims to develop a systematic method by which the user can find important association among items in a database of transactions. By doing so, the user can then figure out how the presence of some interesting items (i.e., items that are interesting to the user) implies the presence of other interesting items in a transaction.

Table 4.10 : Association Rule Notations

Term	Description
D	Database of transactions
ti	Transaction in D
s	Support
X,Y	Item sets
$X \Rightarrow Y$	Association rule
L	Set of large itemsets
l	Large itemset in L
C	Set of candidate itemsets
p	Number of partitions
α	Confidence

Summary

- Clustering is unsupervised learning. Clustering methods are used to find similarities and relationship patterns in data.
- Clustering is like *database segmentation*, in which similar tuples (records) in a database are grouped together(partition/segment).
- There are various issues while applying clustering in data mining like scalability, ability to deal with different kinds of attributes, discovery of clusters with attribute shape, high dimensionality, ability to deal with noisy data and interpretability etc.
- Hierarchical and Partitional are two basic types of clustering.
- In hierarchical clustering method, a nested set of clusters is created.
- In divisive type of clustering, clusters are created in top-down fashion.
- Outliers are isolated data points from given data set. They reside far from rest of the data. They generally do not fall into any cluster.
- Dendrogram, a tree data structure is used to illustrate the hierarchical clustering technique with set of different clusters.
- K-means and K-medoids are two algorithms of Partitional clustering.
- Association rules are used to find patterns among data sets.
- Apriori algorithm is the well-known algorithm used for association rule mining.
- Market basket analysis is a data mining technique used by retailers to increase sales of items by well understanding customer purchasing patterns.
  - Apriori algorithm works in two steps i.e. join and prune step.
  - There are various kinds of association rules generated by applying different algorithms on data set.

Check Your Understanding

1. The first phase of Apriori algorithm is ____.
   (a) Candidate generation (b) Itemset generation
   (c) Pruning (d) Partitioning
2. The apriori frequent itemset discovery algorithm moves _____ in the lattice.
   (a) upward (b) downward
   (c) breadthwise (d) lengthwise
3. After the pruning of a priori algorithm, _____ will remain.
   (a) only candidate set (b) no candidate set
   (c) only subset (d) only itemset
4. The _____ item sets find all sets of items (items sets) whose support is greater than the user-specified minimum support.
   (a) non- frequent (b) frequent
   (c) item group (d) item set

5. A collection of one or more items is called as _____.
   (a) Itemset (b) Confidence
   (c) Support (d) Support Count
6. Frequency of occurrence of an itemset is called as _____.
   (a) Support (b) Confidence
   (c) Support Count (d) Itemset
7. The number of iterations in apriori _____.
   (a) increases with the size of the data
   (b) decreases with the increase in size of the data
   (c) increases with the size of the maximum frequent set
   (d) decreases with increase in size of the maximum frequent set
8. Frequent item sets is ____.
   (a) Superset of only closed frequent item sets
   (b) Superset of only maximal frequent item sets
   (c) Subset of maximal frequent item sets
   (d) Superset of both closed frequent item sets and maximal frequent item sets
9. _____ clustering techniques starts with all records in one cluster and then tries to split that cluster into small pieces.
   (a) Agglomerative (b) Partition
   (c) Divisive (d) Numeric
10. Which of the following clustering requires merging approach?
    (a) Partitional (b) Hierarchical
    (c) Naive Bayes (d) Divisive
11. _____ clustering technique start with as many clusters as there are records, with each cluster having only one record.
    (a) Agglomerative (b) Divisive
    (c) Partition (d) Numeric
12. Which of the following clustering techniques start with as many clusters as there are records or observations with each cluster having only one observation at the starting?
    (a) Agglomerative clustering (b) Fuzzy clustering
    (c) Divisive clustering (d) Model-based clustering
13. Which statement is true about the K-Means algorithm?
    (a) The output attribute must be cateogrical.
    (b) All attribute values must be categorical.
    (c) All attributes must be numeric.
    (d) Attribute values may be either categorical or numeric.

14. Which of the following statements about the K-means algorithm are correct?
    (a) The K-means algorithm is sensitive to outliers.
    (b) For different initializations, the K-means algorithm will definitely give the same clustering results.
    (c) The K-means algorithm is non sensitive to outliers.
    (d) The K-means algorithm can detect non-convex clusters.

Answers

1. (a)	2. (a)	3. (b)	4. (b)	5. (a)	6. (c)	7. (d)	8. (d)	9. (c)	10. (b)
11. (a)	12. (a)	13. (c)	14. (a)						

Practice Questions

Q.I Answer the following questions in short.

1. What is large itemset?
2. What is Apriori property?
3. What is large itemset property?
4. What is time complexity of Apriori algorithm?
5. What is time complexity of k-Means algorithm?
6. What is time complexity of k-Medoid algorithm?

Q.II Answer the following questions.

1. What are different applications of clustering?
2. What are different applications of market basket analysis.
3. Explain k-Means algorithm in brief.
4. Explain k-Medoid algorithm in brief.
5. Explain Apriori algorithm in brief.
6. Explain different requirements of clustering.
7. What are different types of clustering?
8. What are different types of association rules?
9. What are advantages and disadvantages of K-means algorithm?
10. What are advantages and disadvantages of K-Medoid algorithm?
11. What are pros and cons of Apriori algorithm?
12. Differentiate between Agglomerative and Divisive clustering method.

13. Solve the following Problems:

(i) Consider the following transaction table and generate candidate itemsets and frequent itemsets, where the minimum support count = 2.

TID	List of Items
T1	I1, I2, I3
T2	I2, I4
T3	I2, I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I 3
T8	I1, I2, I3, I5
T9	I1, I2, I5

Apply Apriori algorithm to find the candidate itemset and frequent itemset.

(ii) Consider the following transactions table and generate the candidate itemsets and frequent itemsets where the minimum support count is 2.

TID	List of Items
1	A, B, C
2	B, D
3	B, E
4	A, B, D
5	A, E
6	B, E
7	A, E
8	A, B, E, C
9	A, B, C

Apply Apriori algorithm to find the candidate itemset and frequent itemset.

(iii) Consider the following transactions table and generate the candidate itemsets and frequent itemsets where the minimum support count is 2.

TID	List of Items
1	Milk, Tea, Cake
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

(iv) Use the K-Means algorithm and Euclidean distance to cluster the following 10 examples into 3 clusters.

Point	X	Y
A	3	3
B	8	5
C	4	4
D	2	4
E	7	7
F	5	8
G	3	5
H	4	8
I	6	9
J	9	6

Perform K-means algorithm and show how calculations performed at each iteration. Assume initial cluster centres A, E and H. Draw a 10 by 10 space with all the 10 points and show the clusters and the new centroids after the each iteration.

Q.III Define the terms.

1. Clustering
2. Outlier
3. Medoid
4. Mean
5. Centroid
6. Association rule
7. Support
8. Confidence
9. Large itemset
10. Apriori property

✍ ✍ ✍

5... Introduction to Data Science

Learning Objectives…

- To explore a wide variety of data for planning and decision-making purpose.
- To detect and diagnose common data issues, such as missing values, special values, outliers, inconsistencies, and localization.
- To identify and articulate trends and patterns in data gathered over time.
- To find out different industrial and academic applications of Data Science and Data Analytics.
- To walk through data science life cycle.
- To get tips and precautions to be taken for making data science and a part of organization.

5.1 INTRODUCTION

- Data science is the most demanding field in 21st century. Organizations are looking for candidates having knowledge of data science. We know that as any application to be developed must be in online mode. Huge amount of data generated in all online applications. This data is not a data science.
- Application takes input as data and generates more data itself. It cannot be application with data but it is a data product. Data Science enables creation of data products. For example, Google is the master in creation of data products. Spell checking was a difficult problem but by suggestion Google made it simpler.
- There are many more companies like Facebook, LinkedIn also knows how to use data. Many times, using social media user is not aware that what he is searching for spending more time online, and leaving a trail of data wherever they go. But the application traces user navigations using Artificial intelligence.
- This chapter gives ideas about data with its type and science needed to discover useful patterns from it with useful applications by using systematic process.

5.2 BASICS OF DATA

- Data is in any form in numbers, text, audio, video or any statistical measurement, words, observations, or any input given to application.

- There are two types of data: Qualitative and Quantitative.

1. **Qualitative Data**: These are the things you can observe but cannot be measured. It consists of words, pictures, symbols but not numbers. This data is also called as categorical data as information is stored by category. It gives answers of questions like, how the things have happened? Why the things have happened?
    - For Example, Colors of rainbow, Ethnicity such as American Indian, Asian, etc. These types of data are collected through interviews, open ended questionnaire, surveys. This data is in the form of words/text.
    - There are 2 general types of qualitative data: Nominal data and Ordinal data.
        (i) **Nominal Data:** It is used for just for labelling variables without any type of quantitative value. The word nominal comes from the Latin word "nomen" which means 'name'. The nominal data could just be called "labels". Examples of nominal data: Gender (Men, Women), Hair colour (Black, White, Golden, Brown), Marital status(Married, Single, Widowed). In this type of data, there is no intrinsic ordering to the variables.
        (ii) **Ordinal Data:** This data shows where a number is in order. This data is placed into some kind of order by their position on a scale. Ordinal data may indicate superiority. No arithmetic can be done with ordinal number as they only show sequence. Ordinal variables are considered as "in between" qualitative and quantitative variables. User can assign numbers to ordinal data to show their relative position. User cannot do arithmetic with that numbers. For example, "First", "Second", "Third" etc., Grades of students in result sheet like O, A+, A, B etc., Rating given to product, Economic values for a variable such as low, medium, high.
2. **Quantitative Data**: This type of data is easiest to explain. It answers the questions: How many, How often, How much. It is all about numbers. Stuffs that can be measured, counted. This data is easily open to statistical manipulation and can be represented by a wide variety of statistical types of graphs and charts such as Line Chart, Bar graph, Scatter plot etc.
    - For example, Count of colors in rainbow, Scores in exam like 85.56, Person's shoe size, Weight of a person, Temperature in room etc.
    - These are again data classified into **Discrete** and **Continuous**. This type of data is expressed in numbers (numeric). The values of data can be large or small. Numeric values are corresponding to specific category or label. For example, Number of Students in Class.

Table 5.1: Example of Discrete Data

Class	Number of students enrolled
FY	240
SY	238
TY	200
SC I	60
MSC II	65

(i) **Discrete:** Data that is countable. It involves only integers. They cannot be subdivided into parts. It has limit. For example, Count of colors in a rainbow, Number of students in a class (you cannot have 3.5 students), Number of employees working in the company, Days of month, Number of questions those student have answered correctly etc.

(ii) **Continuous:** Data can take any value in a range and would require infinite specificity to be accurate. It can be measured on a scale or continuum and can have almost any numeric value. For example, height (in meters, centimeters, feet, inches), time, temperature. Different measurement instruments can be used to measure continuous data. For example, Height of student, Time required to complete a project, Speed of car.

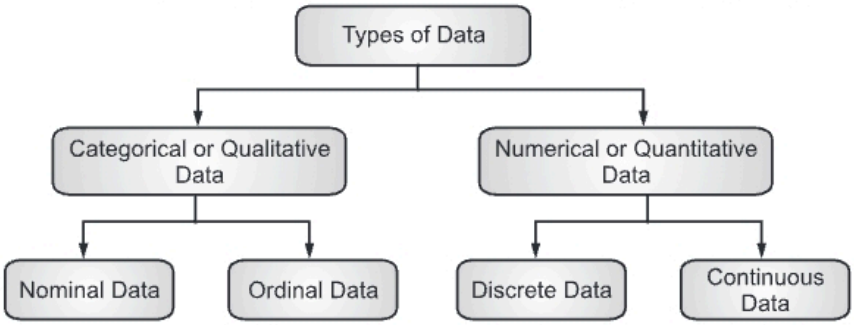


Fig. 5.1: Types of Data

- Data can be also categorized as:

(i) Structured Data:

- Data whose elements are addressable for effective analysis.
- It has been organized into a formatted repository that is known as database.
- It concerns all data which can be stored in database SQL in a table with rows and columns.
- They have relational keys and can easily be mapped into pre-designed fields.
- Such data is most processed in the development and simplest way to manage information.
- **Example:** Relational data.

(ii) Semi-structured Data:

- Semi-structured data is information that does not reside in a relational database but that has some organizational properties that make it easier to analyse.
- With some processes, user can store them in the relation database (it could be very hard for some kind of semi-structured data), but Semi-structured exist to ease space.
- **Example**: XML data.

(iii) Unstructured Data:

- Data which is not organized in a predefined manner or does not have a predefined data model; thus, it is not a good fit for a mainstream relational database.
- There are alternative platforms for storing and managing, it is increasingly widespread in IT systems and is used by organizations in a variety of business intelligence and analytics applications.
- **Example:** Word, PDF, Text, Media logs, NoSQL Data.

- Following table will clear idea about difference between Structured, Semi-structured and Unstructured data.

Table 5.2: Difference between Structured, Semi-structured and Unstructured data

Properties	Structured data	Semi-structured data	Unstructured data
Technology	It is based on Relational database table.	It is based on XML/RDF (Resource Description Framework).	It is based on character and binary data.
Transaction Management	Matured transaction and various concurrency techniques.	Transaction is adapted from DBMS not matured.	No transaction management and no concurrency.
Version Management	Versioning over tuples, row, tables.	Versioning over tuples or graph is possible.	Versioned as a whole
Flexibility	It is schema dependent and less flexible.	It is more flexible than structured data but less flexible than unstructured data.	It is more flexible and there is absence of schema.
Scalability	It is very difficult to scale DB schema.	Its scaling is simpler than structured data.	It is more scalable.
Robustness	Very robust	New technology, not very spread	—
Query performance	Structured query allow complex joining	Queries over anonymous nodes are possible	Only textual queries are possible

- Data with any type work together and help organizations and business to build successful data-driven decision-making process.

5.3 WHAT IS DATA SCIENCE?

- The term "Data Science" was coined at the beginning of the 21st Century. It is attributed to William S.
- Modern business companies and organizations uphold data science is an integral part of IT due to huge amount of data generated in any process of business.
- User may think data science is related to computer science, but it is separate field. The basic difference between data science and computer science is that computer science involves creation of algorithms and writing programs for processing and recording data while data science covers any type of data analysis which may or may not involve computers.
- Data science is related with Statistics and Mathematics which includes the collection, organization, analysis, and presentation of data.
- Artificial intelligence encapsulates the concepts of three fields: Statistics, Mathematics, and Programming language and acts as the machinery or brain of data science.

- Domain expertise is a helpful component in the verification of causal and logical relationships in models and conclusions.

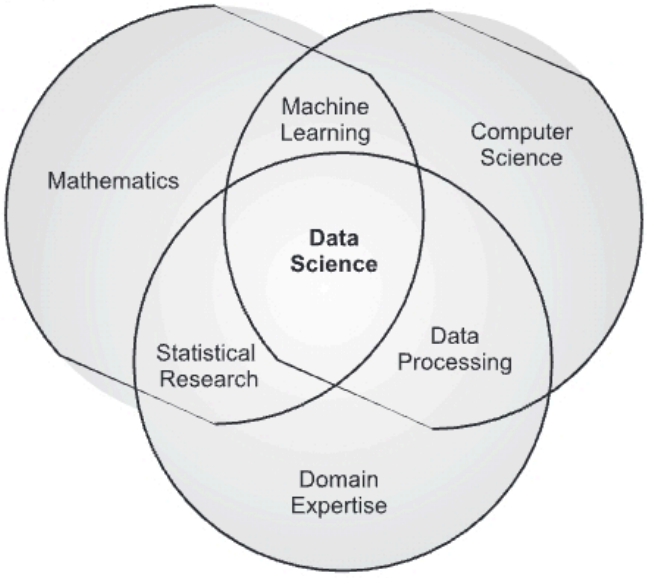


Fig. 5.2: Concept of Data Science

- Data science includes preparing data for investigation and analysis purpose and processing, performing advanced data analysis, and presenting the results to disclose patterns from data and enable stakeholders to draw informed conclusions.
- Data science deals with study of data. It also includes developing methods for recording, storing, preprocessing, analyzing data to find out useful patters/information from data. For example, A College has petabytes of student's data may use data science to develop effective ways to store, manage and analyzes the data. The company may use statistical techniques and scientific methods to perform tests and extract results to provide significant information to its users/ developers and all stake holders of business.
- In short, Data Science is all about:
  - Asking meaningful questions to analyse the data.
  - Modeling data using different tools and techniques.
  - Using different visualization tools to represent the data and its output.
  - Data understanding to get better decisions about data and analysis of data and reaching to final conclusions.

5.3.1 Data Scientist

- The central job of data scientist is to understand the data, extract useful information from it and how to apply these data for solving problems.

Skills required for Data Scientist:

- Programming in Python or R (either works).
- Fluency with popular packages and workflows for data science tasks in your language of choice.

- Knowledge of writing SQL queries.
- Statistics knowledge and methods.
- Basic Machine learning and Modeling skills.
- Ability to work with unstructured data from various sources like video and social media.

5.3.2 Need of Data Science

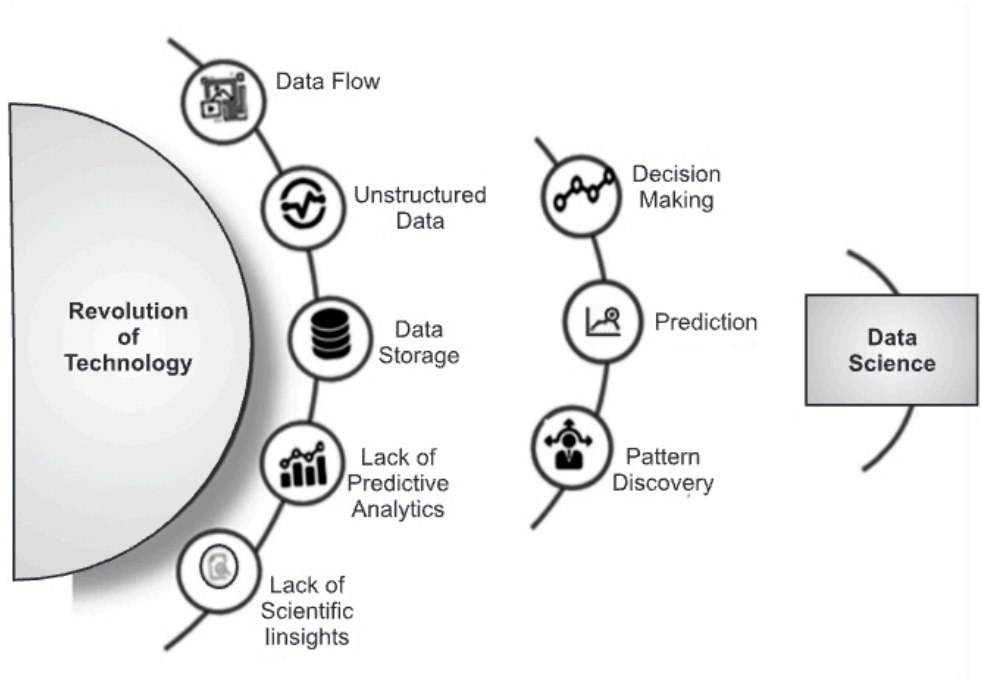


Fig. 5.3: Need of Data Science

- Around 25 years ago, data is available in structure form only. Such data can be easily stored in Excel and RDBMS software's which can be processed easily. Due to development of World Wide Web data is becoming so vast, unstructured and speed of its generation is very vast i.e. 2.5 quintals per day which is then known as Big data. Research estimation by 2020 states that 1.7 MB of data will be created by every second by single person on earth.
- Now in virtual mode, every company needs data to process, analyse and increase their sale. Handle such huge amount of data is challenging task and requires powerful, integral, efficient algorithms, tools, methods. This technology is known as Data Science.
- Following are some main reasons to use data science technology:
  (a) Data science technology is used to convert huge, unstructured data into structured data and gives meaningful insights to be used in decision making process.
  (b) To start any small or big business setup data science is the key point to be looked first. Companies like Amazon, Google uses data science algorithms for better customer experience.
  (c) Uses of data science in transportation like driverless (self-driving) car which is future of transportation.

(d) With the help of prediction algorithms data science technology will be used in Elections, Weather forecasting, Market basket analysis, Surveys, Ticket confirmations of flight, Movie, Train etc.

- As data science tries to find out useful insights from data needed for any business to grow up.

5.3.3 Difference between BI and Data Science

Table 5.3: Distance between BI and Data Science

Criterion	Business Intelligence	Data Science
Perception	Looking Backward	Looking Forward
Data Source	Business intelligence deals with structured data, e.g., data warehouse, Mostly SQL.	Data science deals with structured and unstructured data, e.g., weblogs, feedback, etc. Like logs, SQL, NoSQL or text.
Method	Analytical (historical data)	Scientific (goes deeper to know the reason for the data report)
Skills	Statistics and Visualization are the two skills required for business intelligence.	Statistics, Visualization, and Machine learning are the required skills for data science.
Focus	Business intelligence focuses on both Past and present data.	Data science focuses on past data, present data, and also future predictions. Analysis and Neuro-linguistic Programming.
Tools	Pentaho, Microsoft BI, QlikView	R, Python, TensorFlow.

5.3.4 Components of Data Science

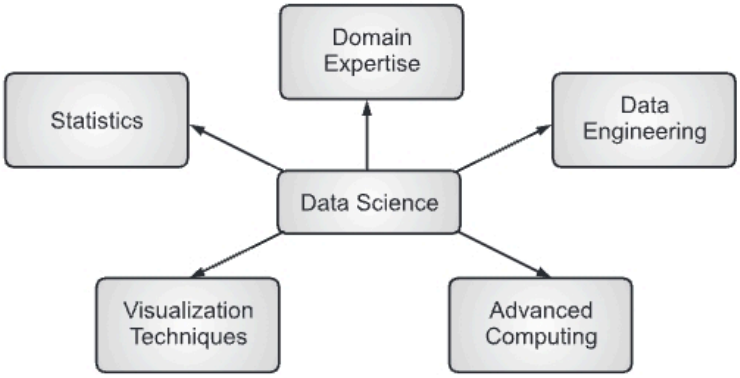


Fig. 5.4: Components of Data Science

- The components of Data Science are as following:

1. **Statistics:** Without statistics data science cannot be thought of. It is the basic and critical unit of data science. Huge amount of numerical data is given to algorithm, in that different statistical measures are applied and useful output is generated.
2. **Visualization:** Huge amount of data is represented in quick glance which is easy to understand by all business processes and its stakeholders and society.

3. **Machine Learning:** Different machine learning algorithms are used to make predictions about future/unpredicted data.
4. **Deep Learning:** It is a new machine learning research technique where an algorithm selects the analysis model to follow.
5. **Advanced Computing:** Data science is nothing but data computing. So, many advanced computer techniques tools are need to process data in data science.
6. **Data Engineering:** It is preparation of systems for collecting, validating, and preparing that high-quality data. Data engineers gather and prepare the data and data scientists use the data to promote better business decisions.
7. **Domain Expertise:** Domain experts have knowledge about data domains. The useful insight which is output of data science needed to be clearly understood by user. If user is unable to interpret results the domain experts are needed to interpret the results.

5.4 DATA SCIENCE PROCESS

- Before going for systematic process of data science a goal must be set. The purpose of goal is making sure that all stakeholders should understand what, how, and why of the question to be solved. Information and context should be cleared before going for problem solving.

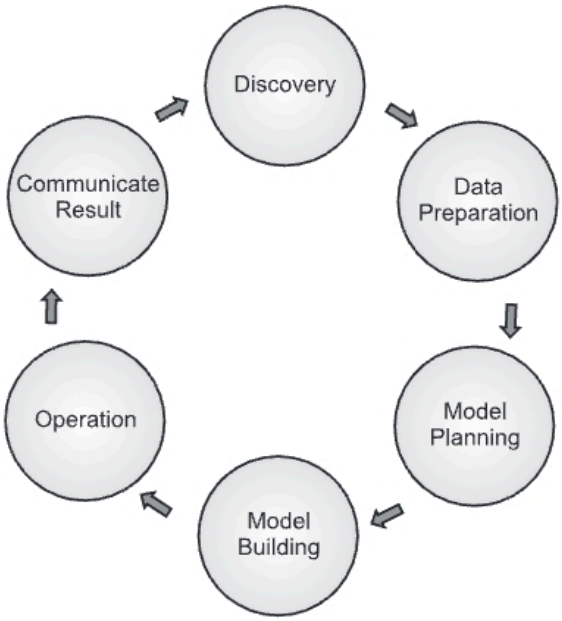


Fig. 5.5: Data Science Process

Step 1 - Discovery: This step involves obtaining data from all identified, recognized internal and external sources. This acquired data helps user to answer the business problem under study.

- Different kind of data can be obtained such as,
  - Logs from webserver
  - Social media data

- Census database
- Stream data
- Data collected from online repositories
- Data Collected from Customer Relationship Management database/ software.

Step 2 - Data Preparation: The collected data have many conflicts like missing value, blank columns, incorrect data format, which needs to be cleaned. Before using data for modelling it should be clean, transformed and in a single format. So, Data transformation is done at this step.

- Remember the 'Garbage in Garbage out' Philosophy". That means, if the data is unfiltered and irrelevant then the results of the analysis will not mean anything (wrong predictions were done and wrong results are obtained). Missing data can be predicted. Erroneous data can be corrected.
- Clean data gives better predictions. This step is also known as the **Data scrubbing** or **Exploratory Data Analysis (EDA)**. Data should be organized keep in orderly fashion and unwanted data is removed. Make one standard format for collected data.

Step 3 - Model Planning: In this step of data science process, model for training data set is finalized. At this stage, user needs to determine the method and technique to draw the relation between input variables and output variables. Planning for a model which is used different statistical formulas and visualization tools.

- For this purpose, different software and tools like SQL analysis services, R, and SAS/access are used. Data mining techniques such as regression, classification and machine learning algorithms are used at this stage. Models are planned to get useful insights that are essential for data-driven decision-making.

Step 4 - Model Building: Models must be building using training data set and their performance is evaluated using test data set. Actual model building process happens here. Data scientist distributes datasets for training and testing.

- Techniques like association, classification, and clustering are applied to the training data set. The model once prepared is tested against the "testing" dataset.

Step 5 - Operationalize: In this stage, user will deliver the final baselined model with reports, code, and technical documents. Model is deployed into a real-time production environment after thorough testing. At this step, preparation of visual insights to discover useful patterns in the data will be done.

Step 6 - Communicate Results: This is known as presenting your results and automating the analysis. In this stage, the key results are transferred to all stakeholders. This helps user to decide if the results of the project are a success or a failure based on the inputs from the model. Actionable insight is a key outcome that user show how data science can bring about predictive analytics and later prescriptive analytics.

- Here we can also learn that how to repeat a positive result or prevent a negative outcome. Different Visual techniques will be used. For example, what are the most important features that influence the class labels (**Y** variables)?

5.5 STAGES IN DATA SCIENCE PROJECT

- Data mining practitioners typically achieve timely, reliable results by following a structured, repeatable process that involves these stages:

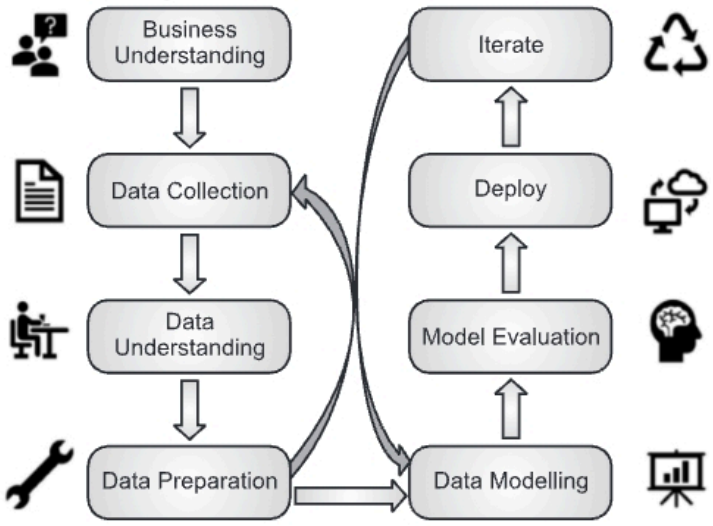


Fig. 5.6: Stages Data Science Project

1. **Business Understanding:** Success of an organization depends on the quality of questions asked and how effectively they are answered by that business. The computation power for data processing must be greater than the amount of data collected in any business.

   While using Google, instead of asking question like "What constitutes relevant search results" Google will try to find out "Which ads are relevant to user?" Questions are asked depending on kind of business. For example, Uber - What percentage of time do drivers actually drive? How steady is their income?

   Asking such type of questions is necessary to increase sale in any business. Such questions will get on the journey of data science. If correct questions are asked the correct data is collected for business for further processing.

2. **Data Collection:** Once user is having clear business understanding, data collection become easier. It means breaking a problem into smaller ones. Data scientist in business will decide what data is needed, what different sources of that data are and how that data should be processed to get desired output.

3. **Data Understanding:** Data understanding stage of data science project answers the question that whether collected data will solve the given business problem.

   For example, Uber wanted to understand whether people would choose to be drivers for them. For this, they collected the following data:

   (i) Their current income.

   (ii) The percentage cut cab companies charge them.

   (iii) The number of rides they fulfil.

(iv) The amount of time they sit idle.

(v) Fuel costs.

Steps (i) to (iv) will allow data scientists to understand about drivers but Step (v) doesn't do anything with driver. Actually, fuel cost will have impact on Uber model but the data collected for Step (v) in not needed.

4. **Data Preparation:** Now, we have collected data so prepare it for further analysis.

   Various statistical techniques can be used like Mean, Median, and Mode also graphical techniques such as Histograms, Bar Plots, and Charts will be used to understand the data. When user will have clear understanding about data then it will be prepared for further analysis.

   Data preparation includes following steps:

   (i) Handling missing data.

   (ii) Correcting invalid values.

   (iii) Removing duplicates.

   (iv) Structuring the data to be fed into an algorithm.

   (v) Feature engineering.

   Data collection, understanding, preparation step takes 70% to 90% of overall data science project time.

   At the end of this step, if user feels that data is not proper then user can go back to first step i.e. Data Collection.

5. **Data Modeling:** It is used to find patterns and behaviors in data. Patterns helps user in two ways.

   (a) Descriptive Modeling: It quantifies relationships in data in a way that is often used to classify customers or prospects into groups. For example, to categorize customers by their product preferences and life stages.

   (b) Predictive Modeling: Prediction about future models can be done. For example, user can predict stock exchange values. i.e. concept of linear regression.

   Machine learning will be used for Data modelling. There are three steps: Training, Validation and Testing. These stages depend of whether mode of learning is supervised or unsupervised. In case of supervised learning these steps are used.

6. **Model Evaluation:** After of data modelling stage, models are evaluated on some statistical measure such as:

   (a) Accuracy: How well the model performs i.e. it should describe the data accurately.

   (b) Relevance: Does the created model answer the original business question under study.

7. **Deploy:** At last, all data science project is deployed in real world. This deploy through Android app, Web app or any enterprise software. The prepared data model should be shown to real world. Once peoples use it, feedback data will be generated. Such feedback data will be used to decide life or depth of any data science project.

8. **Iterate:** A data science project is an iterative process. User should keep on repeating the various steps until he will be able to fine tune the methodology to specific case.

- Consequently, user will have most of the above stages going in parallel.

5.6 APPLICATIONS OF DATA SCIENCE IN VARIOUS FIELDS

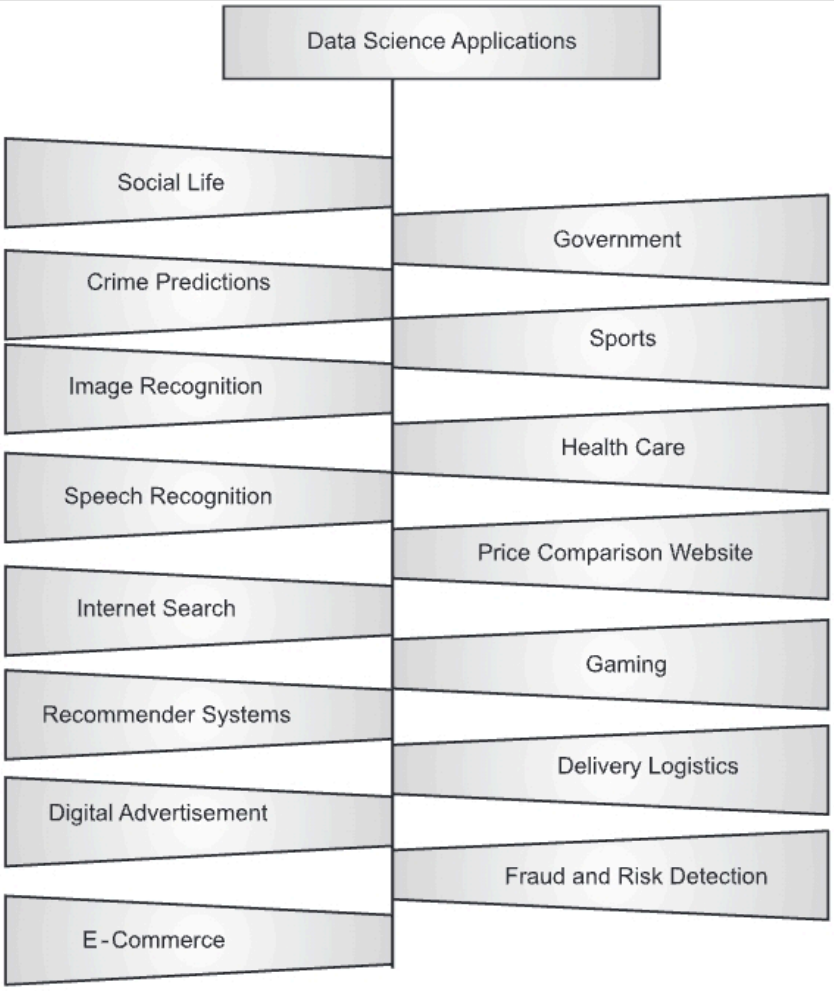


Fig. 5.7: Applications of Data Science

(a) **Image Recognition:** Scanning barcode from any device by using WhatsApp web feature from smart phone. Used for Facebook suggestions to tag friends for the post created by user. Google provides the option to search for images by uploading them. It uses image recognition and provides related search results. All these examples use data science technique in background and give required results to user.

(b) **Speech Recognition:** In this application, you can simply speak out the message and it will be converted to text. Alexa, Siri Google voice call, WhatsApp call uses speech recognition to recognize user commands. Using these facilities like typing in Google Doc by talking is luxury like peaking the content text to type, send. For example, Virtual assistant application program used to set an alarm, playing music on demand, weather forecasting etc.

(c) **Internet Search:** Search engines like Google, Yahoo, Bing uses data science to make search at fast and real time.

(d) **Digital Advertisement:** Depending on user search pattern advertisements are curated for users on website. Data science algorithms are used to display banners on different websites, digital billboards at the airports. Though internet surfing and the entire digital marketing spectrum are most significant applications of data science and machine learning.

(e) **Recommender Systems:** Amazon, YouTube always give suggestions of similar products, things according to user surfing behaviour, search history and wish list.

(f) **Price Comparison Websites:** Jungalee, PriceDekho, Trivago, Google and any other website which compares prices for same products across different platforms to get best deal for any product for user.

(g) **Gaming:** Opponent can analyses the player's moves and can increase difficulty level while playing with the help of machine learning and data science.

(h) **Delivery Logistics:** Freight giants like UPS, FedEx, and DHL use practices of data science to discover optimal routes, delivery times, and transport modes among many others. A plus with logistics is the data obtained from the GPS devices installed.

Data science is used in America for optimization of road routes. Through optimization of routes it helps to save thousands of gallons of gas when spread across hundreds of trips and vehicles among companies also that are not explicitly eco-focused.

(i) **Optimizing Package Routing:** Data science is used to optimize package transport from drop-off to delivery. Network Planning Tools (NPT) uses machine learning and AI to crack challenging logistics puzzles, such as how packages should be rerouted around bad weather or service bottlenecks.

(j) **HealthCare:** Data science is used in detecting tumors (for example, Google has developed tool LYNA for identifying Brest cancer tumours), artery stenosis, organ description employs various methods and frameworks like Map reduce to find ideal parameters for tasks such as lung texture sorting. It applies machine learning methodologies, support vector machines, content-based medical image indexing, and wavelet analysis for stable texture classification.

(k) **Sports:** Players, team managers, coaches and fans rely on sports analytics before making decisions or developing strategies to win games.

(l) **Government:** Government is one of the biggest data collectors. All the government agencies right from census data to National security intelligence rely heavily on trained data scientists to help them in making better decisions.

(m) **E-commerce:** Citizens of that same town can each shop in their own personalized digital mall, also known as the internet. Online retailers often automatically tailor their web storefronts based on viewers' data profiles. That can mean tweaking page layouts and customizing highlighted products among other things. Some stores may also adjust prices based on what consumers seem able to pay, a practice called personalized pricing. Even websites that sell nothing (not directly, anyway) features personalized advertisements such as Amazon.

- All above applications of data science uses algorithms or create algorithms to analyses/filters data and make it suitable for user specific purpose. So, Analysis of data is come in to picture to deal with huge amount of unstructured web data.

5.7 BASICS OF DATA ANALYTICS

- Data Analytics (DA) is the process of examining data sets to draw inferences about the information they contain, increasingly with the aid of specialized systems and software.
- Data analytics is used in many industries to allow companies and organization to make better business decisions and in the sciences to verify or disprove existing models or theories.
- Following points will clear ideas about Data Analytics:
  1. Define the question or goal behind the analysis: What are you (user) trying to discover/ find out?
  2. Right data collection to answer this question.
  3. Perform data cleaning/scrubbing to improve data quality, removal of unnecessary data and prepare it for analysis process. Right data, right format.
  4. Data Manipulation using any tools or techniques.
  5. Data Analysis and interpretation using statistical tools and techniques, finding correlations, patterns, trends, outliers in data.
  6. Data presentation in meaningful ways. Different graphs, charts, tables visualization techniques can be used so that useful insights in the data can be communicated to all stake holders and finding its applications.

5.7.1 Difference between Data Science and Data Analytics

- There is difference between Data Science and Data Analytics. Data Science is study of data includes developing methods to extract useful information (i.e. knowledge) from data.
- Following points will clear ideas about Data Science and Data Analytics.

Table 5.4: Difference between Data Science and Data Analytics

Factors	Data Science	Data Analytics
Meaning	Data Science is a universal field which includes data cleansing, preparation, and analysis.	Driving insights and fashions from the data. It is a combination of Business Intelligence and Business Analytics.
Skill set	• Predictive Data Modeling • Predictive Analytics • Advanced Statistics • Engineering /Programming	• Static Data Modeling • Business Intelligence Tools • Intermediate Statistics • Solid Programming Skills • Regular Expression (SQL)
Scope	Macro	Micro
Exploration	• Search Engine Exploration • Machine Learning • Artificial Intelligence • Big Data-often Unstructured	• Data Visualization Techniques • Designing Principles • Big Data- Mostly Structured
Goals	Discover new questions to drive innovation.	Use existing information to uncover Actionable Data.

5.7.2 Difference between Data Analysis and Data Analytics

Table 5.5: Difference between Data Analysis and Data Analytics

Data Analysis	Data Analytics
Data analysis is the process of studying a given data set (in close detail), dividing them into small components, and studying the subcomponents individually and their relationship with each another.	Data analytics is a more comprehensive term referring to a discipline that comprises the complete management of data, including collection, cleaning, organizing, storing, administering, and analysis of data with the help of specialized tools and techniques.
Data analysis, a subset of data analytics, refers to specific actions.	Data analytics is the broad field of using data and tools to make business decisions.
According to Merriam Webster, Data Analysis is the division of a whole into small components.	According to Merriam Webster, Data Analytics is the science of logical analysis.
Data analysis looks backward over time and works on the facts and figures of what has happened.	Data analytics work towards modelling the future or predicting a result.
Data analysis is a process or method.	Data analytics is an overarching discipline (science).
Data analysis is a process involving the collection, manipulation, and examination of data for getting a deep insight.	Data analytics is taking the analysed data and working on it in a meaningful and useful way to make well-versed business decisions.
Data analysis helps design a strong business plan for businesses, using its historical data that tell about what worked, what did not, and what was expected from a product or service.	Data analytics helps businesses in utilizing the potential of the past data and in turn identifying new opportunities that would help them plan future strategies. It helps in business growth by reducing risks, costs, and making the right decisions.
The analysis restructures existing available information or data.	Data Analytics uses this analyzed information to predict what may happen.
In an example of an apparel brand, the business/brand owner analyses last year's sales data to gain insight into profit trends and sales trends as per seasons, months, and weeks. This analysis of what has happened is basically an in-depth review of the past facts.	Analytics combines the results from the analysis of last year's sales data with logical reasoning to predict future sales pattern and design and plan accordingly.
In data analysis, experts explore past data, break down the macro elements into the micros with the help of statistical analysis, and draft a conclusion with deeper and significant insights.	Data analytics utilizes different variables and creates predictive and productive models to challenge in a competitive marketplace.

... (Contd.)

Data Analysis	Data Analytics
Tools used for data analysis are Open Refine, Rapid Miner, KNIME, Google Fusion Tables, Node XL, Wolfram Alpha, Tableau Public, etc.	Tools used in Data analytics are Python, Tableau Public, SAS, Apache Spark, Excel, etc.
The life cycle of data analytics also comprises data analysis as one of the significant steps.	Data analytics is more extensive in its scope and includes data analysis as a sub-component.
Data analysis is actually studying past data to understand 'what happened?'	Data analytics predicts 'what will happen next or what is going to be next?'

5.8 TYPES OF ANALYTICS

- Data analytics is a wide field. There are different types of data analytics: Descriptive, Diagnostic, Predictive, Prescriptive and Cognitive Analytics. Each type has a different goal and a different place in the data analytics process. These are also primary data analytics applications in business. Many organizations do not know where to begin, what kind of analytics can nurture business growth, and what these different types of the analytics mean.

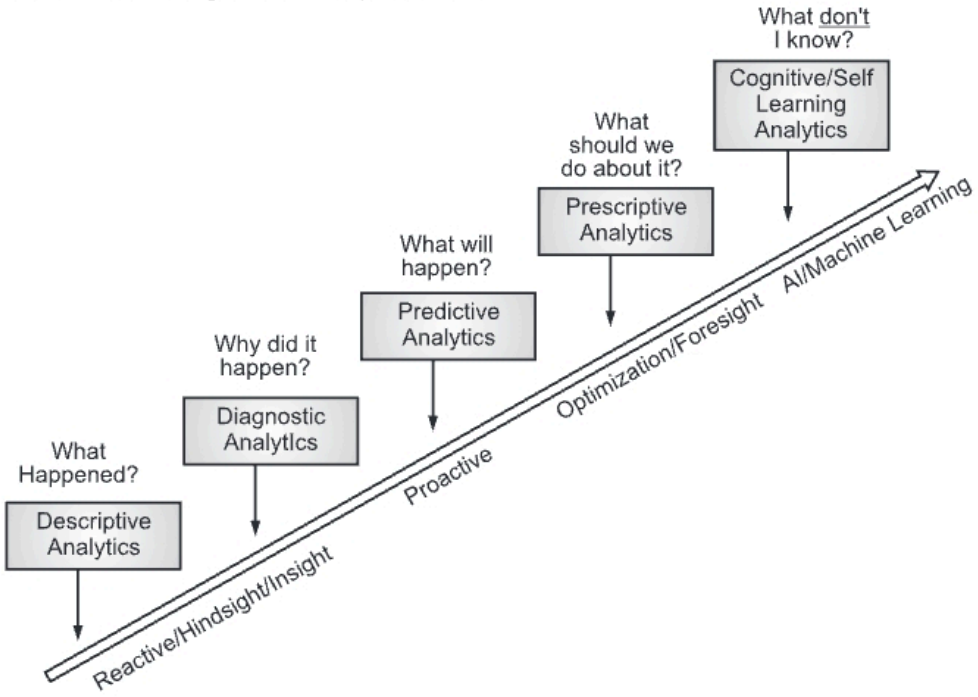


Fig 5.8: Analytics Maturity Model

(a) **Descriptive Analytics:**

- It is most basic form of analytics. This type of data analytics used by 90% of organizations.
- It answers the questions "What has happened?" .It describes main features of collection of data. It is commonly applied to large volume of data i.e. census data. It analyses the data coming in real time and historical data for insights for future study.

- The main objective of this is to find out the reasons behind precious success or failure in the past. Past refers to historical.
- A business learns from past behaviors to understand how they will impact future outcomes. It is used to find overall performance of a company at aggregate level with various aspects.
- It is based on standard aggregate function in databases. Most of the social analytics are descriptive analytics. They summarize certain groupings based on simple counts of some events. The numbers of followers, likes, posts, fans are event counters. These metrics are used for social analytics like the average response time, the average number of replies per post, % index, number of page views, etc. that are the outcome of basic arithmetic operations.
- The best example to explain descriptive analytics is the results that a business gets from the web server through Google Analytics tools. The outcomes help understand what actually happened in the past and validate if a promotional campaign was successful or not based on basic parameters like page views.
- The biggest use of descriptive analysis in business is to track Key Performance Indicators (KPIs). KPIs describe how a business is performing based on chosen benchmarks. Business applications of descriptive analysis include:
  (i) KPI dashboards
  (ii) Monthly revenue reports
  (iii) Sales leads overview

(b) Diagnostic Analytics:

- Analytics performed on the internal data to understand the "why" behind what happened is referred to as diagnostic analytics.
- This kind of analytics is used by businesses to get an in-depth insight into a given problem provided they have enough data at their disposal.
- Diagnostic analytics helps identify anomalies and determine casual relationships in data.
- For example, e-Commerce giants like Amazon can drill the sales and gross profit down to various product categories like Amazon Echo to find out why they missed on their overall profit margins.
- Diagnostic analytics also find applications in healthcare for identifying the influence of medications on a specific patient segment with other filters like diagnoses and prescribed medication.

(c) Predictive Analytics:

- Predictive analytics uses the understanding of the past to make "predictions" about the future. Analysing past data patterns and trends can accurately inform a business about what could happen in the future. This helps in setting realistic goals for the business, effective planning, and restraining expectations.
- Predictive analytics is used by businesses to study the data and ogle into the crystal ball to find answers to the question "What could happen in the future based on previous trends and patterns?"
- Predictive analytics provides better recommendations and more future-looking answers to questions that cannot be answered by BI.

- Predictive analytics helps predict the likelihood of a future outcome by using various statistical and machine learning algorithms but the accuracy of predictions is not 100%, as it is based on probabilities.
- To make predictions, algorithms take data and fill in the missing data with the best possible guesses. This data is pooled with historical data present in the CRM systems, ERP, and HR systems to look for data patterns and identify relationships among various variables in the dataset.
- Organizations like Walmart, Amazon, and other retailers leverage predictive analytics to identify trends in sales based on purchase patterns of customers, forecasting customer behaviour, forecasting inventory levels, predicting what products customers are likely to purchase together so that they can offer personalized recommendations, predicting the number of sales at the end of the quarter or year.
- Predictive analytics can be further categorized as:
  (i) **Predictive Modeling:** What will happen next, if ?
  (ii) **Root Cause Analysis:** Why this actually happened?
  (iii) **Data Mining:** Identifying correlated data
  (iv) **Forecasting:** What if the existing trends continue?
  (v) **Monte-Carlo Simulation:** What could happen?
  (vi) **Pattern Identification and Alerts:** When should action be invoked to correct a process.

(d) Prescriptive Analytics:

- Prescriptive analytics is the next step of predictive analytics.
- Prescriptive analytics advises on possible outcomes and results in actions that are likely to maximize key business metrics.
- It basically uses simulation and optimization to ask "What should a business do?"
- Prescriptive analytics is an advanced analytics concept based on:
  (i) Optimization that helps to achieve the best outcomes.
  (ii) Stochastic optimization helps understand how to achieve the best outcome and identify data uncertainties to make better decisions.
- Prescriptive analytics is a combination of data and various business rules. The data for prescriptive analytics can be both internal (within the organization) and external (like social media data). Business rules are preferences, best practices, boundaries, and other constraints. Mathematical models include natural language processing, machine learning, statistics, operations research, etc.
- Prescriptive analytics is comparatively complex and hence not used in day to day business activities.
- Prescriptive analytics can be used in healthcare to enhance drug development, finding the right patients for clinical trials, etc.

(e) Cognitive Analytics:

- This is the most advanced type of business analytics that applies human intelligence to certain tasks by combining many technologies such as artificial intelligence, semantics, machine, and deep learning algorithms.

- The goal is to understand and copy how a human brain makes a decision and comes with a system or computer that does the same.
- Some of the tasks that can be performed using cognitive analytics are chatbots, virtual assistants, recognizing objects in an image, and segmentation of those images.
- It works by searching the entire available "knowledge base" to locate real-time data.
- It collects and makes real-time data sources such as text, images, audio, and video available to these analytics tools for decision-making.
- This Involves machine learning and natural language processing.

5.9 STATISTICAL INFERENCE - POPULATION AND SAMPLES - STATISTICAL MODELING - PROBABILITY DISTRIBUTIONS

5.9.1 Role of Statistics in Data Science

- "Data Scientists" means a professional who uses scientific methods to liberate and create meaning from raw data.
- "Statistics" means the practice or science of collecting and analysing numerical data in large quantities.
- Statistics is foundational to Data Science; there is strong relationship between these two fields.
- Statistics is one of the most important disciplines to provide tools and methods to find structure in and to give deeper insights into data.
- Some of the more important Statistics concepts used in Data Science include probability distributions, statistical significance, hypothesis testing, and regression.
- Statistics is also used for summarizing the data fairly quickly.
- Visual representation helps identify outliers, specific trivial patterns, and certain metric summary such as Mean, Median, Variance, that helps in understanding the middlemost value, and how the Outlier affects the rest of the data.
- Machine Learning is rapidly growing field at the intersection of computer science and statistics concerned with finding patterns in data. It is responsible for various advancements in technology, product recommendations to Speech recognition to Autonomous driving.

5.9.2 Populations and Samples

- The first step of every statistical analysis you will perform is to determine whether the data you are dealing with is a population or a sample.

5.9.2.1 Population

- Population is an entire pool of data from where a statistical sample is extracted. It can be visualized as a complete data set of items that are similar in nature. The size of population may be either infinite or finite.
- **Example:** A list consisting of the name of all the employees in a company.

5.9.2.2 Sample

- Sample is a subset of the population, i.e. it is an integral part of the population that has been collected for analysis.

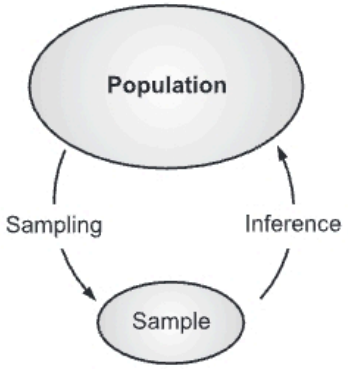


Fig 5.9: Population and Sample

- **Sampling:** It is a method that allows us to get information about the population based on the statistics from a subset of the population (sample), without having to investigate every individual.
- Every sampling type comes under two broad types:
  - **Probability sampling** : Random selection techniques are used to select the sample.
  - **Non-probability sampling:** Non-random selection techniques based on certain criteria are used to select the sample.
- **Example:** In the process of election, all the people of the country elect the candidate by polling in ballots or EVMs etc. In this process, all the people's votes are considered. Here, Population is the overall count of the people.
- Usually, before and after the election, NEWS channels conduct Opinion polls (i.e.) entry poll (before the election), and exit poll (after election) surveys. In these opinion polls, the poll samples of 5000–10000 people are taken. This sample represents the views of the people of the country.
- Best results depend on how well the sample represents the population. The sample must contain the characteristic of the population. It should represent the population.

5.9.2.3 Basic Terms used in Statistics

- **Variable:** A value whose characteristics such as quantity can be measured, it can also be addressed as a data point, or a data item.
- **Distribution:** The sample data that is spread over a specific range of values.
- **Parameter:** It is a value that is used to describe the attributes of a complete data set (also known as 'population'). Example: Average, Percentage.
- **Quantitative analysis:** It deals with specific characteristics of data- summarizing some part of data, such as its mean, variance, and so on.
- **Qualitative analysis:** This deals with generic information about the type of data, and how clean or structured it is.
- **Mean:** It is understood as the central most value when the data points are arranged in a descending or ascending order, or the most likely value.
- **Mode:** It can be understood as the data point that occurs the greatest number of times, i.e. the frequency of the value in the dataset would be very high.

- **Median:** It is a measure of central tendency of the data set. It is the middle number that can be found by sorting all the data points in a dataset and picking the middle-most element. If the number of data points in a dataset is odd, one single middle value is picked up, whereas two middle values are picked and their mean is calculated if the number of data points in a dataset is even.
- **Range:** It refers to the value that is calculated by finding the difference between the largest and the smallest value in a dataset.
- **Quartile:** Quartiles are values that divide the data points in a dataset into quarters. It is calculated by sorting the elements in order and then dividing the dataset into four equal parts.
- **Variance:** The average of the difference between every value and the mean of that specific distribution.
- **Standard deviation:** It can be understood as the measure that indicates the dispersion that occurs in the data points of the input data.
- **Support:** It is set of values that can be assumed with non-zero probability by the random variable.
- **Hypothesis:** Hypothesis is an assumption that is made on the basis of some evidence. This is the initial point of any investigation that translates the research questions into a prediction. It includes components like variables, population and the relation between the variables. A research hypothesis is a hypothesis that is used to test the relationship between two or more variables.

5.9.3 Statistical Inference

- It is the process of using data analysis to infer properties of an underlying distribution of probability.
- Statistical inference makes propositions about a population, using data drawn from the population with some form of sampling.
- Inferential statistics help us draw conclusions from the sample data to estimate the parameters of the population.
- The sample is very unlikely to be an absolute true representation of the population and as a result, we always have a level of uncertainty when drawing conclusions about the population.

5.9.4 Statistical Modeling

- Statistical Modeling refers to the process of applying statistical analysis to datasets. A statistical model is a mathematical relationship between one or more random variables and other non-random variables.
- The application of statistical Modeling to raw data helps data scientists approach data analysis in a strategic manner, providing intuitive visualizations that aid in identifying relationships between variables and making predictions.
- Common data sets used for statistical modelling are Internet of Things (IoT) sensors, census data, public health data, social media data, imagery data, and other public sector data that benefit from real-world predictions.

5.9.4.1 Statistical Modelling Techniques

- The first step in developing a statistical model is gathering data, which may be sourced from spreadsheets, databases, data lakes, or the cloud.
- The most common statistical Modeling methods for analysing this data are categorized as either **Supervised learning** or **Unsupervised learning**.

- Some popular statistical model examples include Logistic regression, Time-series, Clustering, and Decision trees.

  (a) Supervised learning: Supervised learning techniques include regression models and classification models:

  - **Regression model:** It is a type of predictive statistical model that analyses the relationship between a dependent and an independent variable. Common regression models include logistic, polynomial, and linear regression models. Use cases include forecasting, time series Modeling, and discovering the causal effect relationship between variables.
  - **Classification model:** It is a type of machine learning in which an algorithm analyses an existing, large and complex set of known data points as a means of understanding and then appropriately classifying the data; common models include models include decision trees, Naive Bayes, nearest neighbour, random forests, and neural networking models, which are typically used in Artificial Intelligence.

  (b) Unsupervised learning: Unsupervised learning techniques include clustering algorithms and association rules:

  - **K-means clustering**: This aggregates a specified number of data points into a specific number of groupings based on certain similarities.
  - **Reinforcement learning:** It is an area of deep learning that concerns models iterating over many attempts, rewarding moves that produce favourable outcomes and correcting steps that produce undesired outcomes, therefore training the algorithm to learn the optimal process.

5.9.4.2 Types of Statistical Models

- There are three main types of statistical models: Parametric, Non-parametric and Semi-parametric:

  1. **Parametric:** A family of probability distributions that has a finite number of parameters.
  2. **Non-parametric:** Models in which the number and nature of the parameters are flexible and not fixed in advance. The complexity of the model is unbounded and grows with the amount of data.
  3. **Semi-parametric:** It is a hybrid model. The parameter of this model has both a finite-dimensional component (parametric) and an infinite-dimensional component (nonparametric).Cox proportional hazard model is a popular example of semi-parametric assumptions.

5.9.5 Probability Distributions

- A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range.
- This range will be bounded between the minimum and maximum possible values, but precisely where the possible value is likely to be plotted on the probability distribution depends on a number of factors. These factors include the distribution's mean (average), standard deviation, skewness, and kurtosis.
- It is a mathematical description of a random phenomenon in terms of its sample space and the probabilities of events (subsets of the sample space).
- For instance, if X is used to denote the outcome of a coin toss ("the experiment"), then the probability distribution of X would take the value 0.5 (1 in 2 or 1/2) for X = heads, and 0.5 for X = tails (assuming that the coin is fair). Examples of random phenomena include the weather condition in a future date, the height of a randomly selected person, the fraction of male students in a school, the results of a survey to be conducted, etc.

- Some of them include the normal distribution, chi square distribution, binomial distribution, and Poisson distribution.

Types of Distributions:

1. **Bernoulli Distribution:** A **Bernoulli distribution** has only two possible outcomes, namely 1 (success) and 0 (failure), and a single trial. So, the random variable X which has a Bernoulli distribution can take value 1 with the probability of success, say p, and the value 0 with the probability of failure, say q or 1-p.
2. **Uniform Distribution:** When you roll a fair die, the outcomes are 1 to 6. The probabilities of getting these outcomes are equally likely and that is the basis of a uniform distribution. Unlike Bernoulli Distribution, all the n number of possible outcomes of a uniform distribution are equally likely.
3. **Binomial Distribution:** A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution.
4. **Normal Distribution:** It represents the behaviour of most of the situations in the universe. The large sum of (small) random variables often turns out to be normally distributed, contributing to its widespread application.
5. **Poisson Distribution:** Suppose you work at a call centre; approximately how many calls do you get in a day? It can be any number. Now, the entire number of calls at a call centre in a day is modelled by Poisson distribution.
   - Some more examples are:
     1. The number of emergency calls recorded at a hospital in a day.
     2. The number of thefts reported in an area on a day.
     3. The number of customers arriving at a salon in an hour.
     4. The number of suicides reported in a particular city.
     5. The number of printing errors at each page of the book.

   Assumptions for Poisson distribution:
   - A distribution is called **Poisson distribution** when the following assumptions are valid:
     1. Any successful event should not influence the outcome of another successful event.
     2. The probability of success over a short interval must equal the probability of success over a longer interval.
     3. The probability of success in an interval approaches zero as the interval becomes smaller.
6. **Exponential Distribution:** Let's consider the call centre example one more time. What about the interval of time between the calls? Here, exponential distribution comes to our rescue. Exponential distribution models the interval of time between the calls.

   Other examples are:
   1. Length of time between metro arrivals.
   2. Length of time between arrivals at a gas station.
   3. The life of an Air Conditioner.

   Exponential distribution is widely used for survival analysis. From the expected life of a machine to the expected life of a human, exponential distribution successfully delivers the result.

5.10 CHALLENGES OF DATA SCIENCE TECHNOLOGY

- Data science is broadening its branches all over the world. But there includes a lot of challenges which delays a data scientist while dealing with data.
- Let us see some of the major challenges faced by data scientists.
  - High variety of information & data is required for accurate analysis.
  - Not adequate data science talent pool available.
  - Management does not provide financial support for a data science team.
  - Unavailability of/difficult access to data.
  - Data Science results not effectively used by business decision makers.
  - Explaining data science to others is difficult.
  - Privacy issues.
  - Lack of significant domain expert.
  - If an organization is very small, they can't have a Data Science team.

Summary

- There are different types of data like qualitative and quantitative.
- Depending on nature it can be also categorised as structured, semi-structured and unstructured data.
- Due to shift of data from physical to virtual mode there is need of data science
- Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes.
- Statistics, Visualization, Deep Learning, Machine Learning, Artificial intelligence are important Data Science concepts.
- Data Science Process goes through Discovery, Data Preparation, Model Planning, Model Building, Operationalize, and Communicate Results.
- R, SQL, Python are essential Data science tools.
- The predictions of Business Intelligence are looking backward while for Data Science is looking forward.
- Important applications of Data science are Internet Search, Recommendation Systems, Image & Speech Recognition, Gaming world, Online Price Comparison, Recommender systems, Delivery Logistics, Digital Advertisement, E-commerce, Fraud and Risk Detection, Crime Predictions and Social life etc.
- High variety of information & data is the biggest challenge of Data Science technology.
- Driving insights and fashions from the data. It is a combination of Business Intelligence and Business Analytics.
- There are different types of data analytics like Descriptive analytics, Diagnostic analytics, Predictive analysis, Prescriptive analytics, Cognitive analytics etc.
- Statistics plays a big role in data science. With the help of many statistical measures models of data can be prepared to give meaningful insights used for decision making purpose in any business.

Check Your Understanding

1. Which of the following is the most important language for Data Science?

   (a) Java (b) R

   (c) Ruby (d) Basic

2. Which of the following approach should be used to ask Data Analysis question?
   (a) Find only one solution for particular problem.
   (b) Find out the question which is to be answered.
   (c) Find out answer from dataset without asking question.
   (d) Find many solutions for particular problem.
3. Which of the following is one of the key data science skills?
   (a) Statistics (b) Machine Learning
   (c) Data Visualization (d) All of the above
4. Which of the following is the common goal of statistical modelling?
   (a) Inference (b) Summarizing
   (c) Subsetting (d) subdividing
5. Which of the following focuses on the discovery of (previously) unknown properties of data?
   (a) Velocity (b) Variety
   (c) Volume (d) Vast
6. CNN is mostly used for which type of data?
   (a) Both Structured and Unstructured (b) Structured Data
   (c) Unstructured data (d) Distributed data
7. _____ type of data contains only numbers.
   (a) Quantitative (b) Unstructured
   (c) Structured (d) Qualitative
8. Which of the following is the measure of dispersion?
   (a) head (b) tail
   (c) support (d) standard deviation
9. In _____ type of data things cannot be measured.
   (a) Qualitative (b) Detailed
   (c) Quantitative (d) Summary
10. _____ type of data shows number in order.
    (a) Descriptive (b) Ordinal Data
    (c) Summary (d) Detailed
11. Number of students of a class is an example of _____ .
    (a) Discreate data (b) Ordinal data
    (c) Nominal data (d) Summary Data
12. Which of the following is type of structured data?
    (a) XML data (b) Relational data
    (c) Word file (d) PDF data
13. RDF stands for _____ .
    (a) Resource Detailed Framework (b) Resource Data Framework
    (c) Resource Description Facility (d) Resource Description Framework
14. _____ is future of Artificial Intelligence.
    (a) Data Digging (b) Data Science
    (c) Data (d) Knowledge
15. Which of the following is the first step of data science process?
    (a) Model building (b) Operation
    (c) Data preparation (d) Discovery

16. Which of the following is the first stage of data science project?
    (a) Data Collection (b) Data Understanding
    (c) Iterate (d) Business Understanding

Answers

1. (b)	2. (b)	3. (d)	4. (a)	5. (b)	6. (c)	7. (a)	8. (d)	9. (a)	10. (b)
11. (a)	12. (b)	13. (d)	14. (b)	15. (d)	16. (d)				

Practice Questions

Q.I Answer the following questions in short.

1. What is Qualitative type of data?
2. What is Quantitative type of data?
3. Explain types of Data with one example each.
4. Explain any one measure of dispersion with example.
5. Give example of discrete data.
6. List the different components of data science.
7. List different types of Data Analytics.
8. What is Cognitive Analytics?
9. List the different process of Data Science.
10. What are two different types of Data Modeling?
11. What are different statistical modelling techniques?

Q.II Answer the following questions.

1. Explain in detail different types of data?
2. What is the difference between Data Science and Data Analytics?
3. What is difference between BI and Data Science?
4. Differentiate between Structured data, Semi structured data and Unstructured data.
5. Explain various applications of Data Science.
6. Which skills required for data scientists need to succeed?
7. Explain need of Data Science.
8. Explain process of Data Science with neat diagram.
9. Describe different stages of Data Science Project?
10. Explain various steps of Data Preparation.
11. What are different categories of Predictive Analysis? Explain in detail.
12. Explain with example concept of hypothesis.
13. Explain continuous data with example.
14. Describe different types of probability distributions?
15. Describe different challenges of data science technology?

Q.III Define the terms.

1. Data Science
2. Data Analytics
3. Population
4. Sampling
5. Hypothesis
6. Probability Distribution
7. Standard Deviation

6...

EDA and Data Visualization

Learning Objectives…

- To examine the data for distribution, outliers, and anomalies to direct specific testing of designed hypothesis.
- To describe Exploratory Data Analysis and Visualization concepts.
- To study data analysis and visualization models and algorithms.
- To learn applicability of different data analysis and visualization model's techniques to solve real world problems.

6.1 INTRODUCTION

- Different types of data are an invaluable resource and data visualization is one of the most efficient tools for analysing and communicating interesting ideas and insights from the data. But badly conceived, incorrectly created or downright untruthful visualizations miss the whole point of visualizing data.
- To understand the process of getting useful insights from data, Exploratory Data Analysis method is used. In EDA method, with the help of different statistical measures data is tuned which gives knowledge based on decisions will be made. Different visualization techniques will clear the idea about data analysis and its use with its applications. User must follow some rules for data visualization and can use different readily available tools and techniques to get clear idea about data.
- In the previous chapter, we have seen what data and its types are. Before going for Exploratory data Analysis, let's us have a quick view on different dimensionality of Data Sets.

6.1.1 Dimensionality of Data Sets

It is the number of attributes that objects in the data set have. If data set contains high number of attributes, then such high dimensionality of data set also become a problem.

1. Univariate Data Set:

- Univariate data is a collection of information characterized by or depending on only one random variable.
- For example, how many months does it take for avocado plants to produce their fruit? Data is gathered for the purpose of answering the question (a research question/s.)

- Univariate data does not answer research questions about relationships between variables, but rather it is used to describe one characteristic or attribute that varies from observation to observation. So, univariate data means Measurement made on one variable per subject.

2. Bivariate Data Set:

- It deals with two variables that can change. These variables can be compared to find relationships between them. If one variable is influencing another variable, then it is said that data has dependent and independent variable.
- An independent variable is a condition or piece of data in an experiment that can be controlled or changed.
- A dependent variable is a condition or piece of data in an experiment that is controlled or influenced by an outside factor, most often the independent variable.
- For example, if Prakash is studying for college examinations then he might track study time and score of college examination. It is observed that more time spend on study better his examination score become. So, Prakash exam score is dependent on time spent for study. Following table 6.1 will clear this concept.

Table 6.1: Example of Bivariate Data Set

Number of Hours For Study	Examination Score
5	70
6	85
7	90
8	97

So, bivariate means measurement made on two variables per subject.

3. Multivariate Data Set:

- A data set consisting of two or more than two variables is referred to as multivariate dataset.
- For example, a dataset of height of students will be called univariate data ('height of students' being the only variable) whereas a dataset of height and weight of students in a class will be a bivariate/ multivariate dataset (since there are two variables, i.e. Height of students and weight of students). So, multivariate means measurement made on many variables per subject.

6.2 WHAT IS EXPLORATORY DATA ANALYSIS?

- EDA belongs to the critical process of performing preliminary investigations on data to discover patterns, to spot irregularities, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.
- EDA is an approach to analyse data sets and to find out main characteristics from it.
- Sometimes statistics approach/methods are not used at preliminary level for analysis but still data should tell us beyond hypothesis testing and formal modelling.
- EDA was coined by John Turkey to encourage statisticians to investigate the data, and possibly formulate hypotheses that could lead to new data collection and experiments.
- Encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.

Difference between EDA and IDA:

- **IDA** means Initial Data Analysis.
- IDA focuses on:
  - Checking assumptions required for model fitting.
  - Hypothesis testing.
  - Handling missing values.
  - Making transformations of variables as per need.
- IDA is part of EDA.
- **EDA** is a method/philosophy for data analysis that employs a variety of techniques (mostly graphical).
- The goal of EDA is to accomplish the following:
  - Maximize insight into a data set.
  - Uncover underlying structure.
  - Extract important variables.
  - Detect outliers and anomalies.
  - Test underlying assumptions,
  - Develop parsimonious models.
  - Determine optimal factor settings.
- It focuses on how data analysis is carried out. EDA is not a collection of techniques, but it is a philosophy that how we explore the data set i.e., what we look for in the data set? How we investigate data set? How we present the analysis to users.

6.2.1 How does EDA differ from Classical Data Analysis (CDA)?

Table 6.2: Difference between Exploratory Data Analysis approach and Classical Data Analysis approach

Factors	EDA	CDA
Model	• The Exploratory Data Analysis approach does not apply deterministic or probabilistic models on the data. • On the contrary, the EDA approach allows the data to suggest acceptable models that best fit to the data.	• The classical approach uses models (both deterministic and probabilistic) on the data. • Deterministic models include, for example, regression models and analysis of variance (ANOVA) models. • The most common probabilistic model assumes that the errors about the deterministic model are normally distributed. This assumption affects the validity of the ANOVA F tests.

... (Contd.)

Factors	EDA	CDA
Focus	It is focus on the data i.e. its structure, outliers, and models suggested by the data.	It is focus on the model by estimating parameters of the model and generating predicted values from the model.
Techniques	Generally graphical techniques such as histograms, scatter plots, character plots, box plots, interactive histograms, probability plots, residual plots, and mean plots.	Generally quantitative in nature. They include ANOVA, t tests, chi-squared tests, and F tests.
Rigor/Rigidity	Lack of rigor by being very suggestive, indicative, and insightful about what the appropriate model should be.	Classical techniques are rigorous, formal, and objective.
Data Treatment	It shows all the available data. So, no loss of information.	Classical estimation techniques have the characteristic of taking all the data and mapping the data into a few numbers ("estimates"). Due to mapping there is a loss of information.
Assumptions	Make little or no assumption that means they present and show the data with fewer burdening assumptions.	Make underlying assumptions (e.g., normality).

6.2.2 Reasons of using EDA

Following are the main reasons of using EDA:

- To catch the mistakes.
- For checking of assumptions.
- To see patterns in data.
- To generate hypothesis.
- Primary selection of appropriate models.
- Determining relationships among the explanatory variables.
- Assessing the direction and rough size of relationships between explanatory and outcome variables.

6.3 STEPS IN EDA

- **EDA process** usually follows six distinct steps. These are: (i) Distinguish Attributes, (ii) Univariate Data Analysis, (iii) Detect Interactions Among Attributes, (iv) Detect Missing & Aberrant Values, (v) Detect Outliers, and (vi) Feature Engineering.
- As displayed in Fig. 6.1, the analysis begins with identification of attributes in a dataset that gives a clear understanding of the data to be analysed. To understand individual attributes and their relationships with each other different analysis such as univariate, bivariate, and multivariate analyses

are performed. Then cleaning and data preparation tasks are performed, where missing, aberrant values and outliers are detected and imputed. Lastly, the process ends with feature engineering, where features are transformed or combined to generate new features.

- Before going for exploratory data analysis, the preliminary step is to load the database.

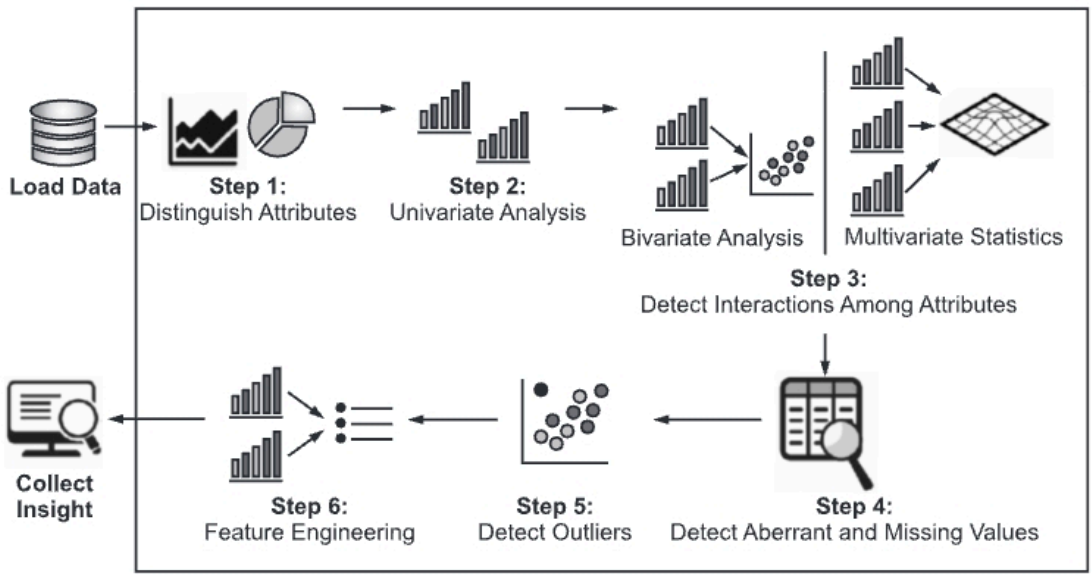


Fig. 6.1: Fundamental Steps of Exploratory Data Analysis

Step 1 - Distinguish Attributes:

- This step is also called as variable identification. Identification of input and output variable will be done in this step.
- Next to identify data type and category of the variables. This step assists users to prepare clear analysis goals.
- There are different types of attributes as we have seen in Fig 6.2. Databases commonly have quantitative (numerical) or qualitative (categorial) attributes.
- Statistical techniques can be applied to all types of datasets. It is therefore important to clearly distinguish and understand the meaning of every attribute in data set under study before analysing data.
- There are various online commercial tools are available to visualize all data set with their attribute in tabular format or graphical format with attribute values and data types. (For example, tools like Domino, Tabalu, Taggle, Taco, Micorsoft power BI, IBM Watson Analytics, DimScanner, ForeSight, Podiumetc.).
- Consider an example, of variable identification as follows: Suppose, user wants to predict, whether the students will play cricket or not (Refer below Table 6.3 for data set). Here, user needs to identify predictor variables, target variable, data type of variables and category of variables.

Table 6.3: Data Set for Variable Identification

Stud_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight (kgs)	Play Cricket
S1	M	65	178	61	1
S2	F	75	174	56	0
S3	M	45	163	62	1
S4	M	57	175	70	0
S5	F	59	162	67	0

- Following Figure 6.2 gives clear idea about variable identification.

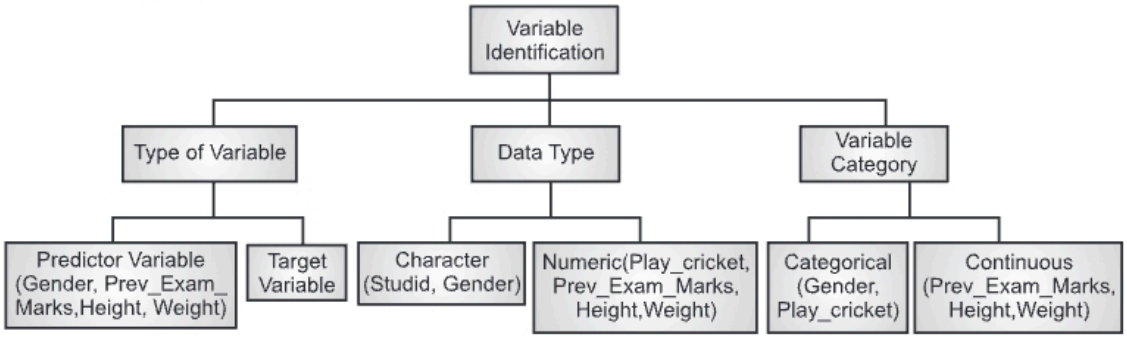


Fig. 6.2: Variable Identification

Step 2 - Univariate data Analysis:

- After the attributes in a data set are identified, it is necessary to perform univariate analysis for getting richer and deeper understanding of each attribute.
- It allows the determination of attribute combinations /patterns for successive analysis. Some details such as centrality (i.e., mean, median, and mode) and dispersion (i.e., range, variance, standard deviation, skewness, and kurtosis) of attributes in the data will be detected.
- The centrality measure helps user to determine approximate of average values.
- The dispersion measure helps user to identify the spread of the value between its lowest and highest bounds.
- This analysis helps user to identify missing values or outliers in a dataset and to discretize continuous variables.
- Tools such as interactive histograms, box plots, pie charts, line graphs are used here. So, in this step's variables are explored one by one.
- Method to perform univariate analysis will depend on whether the variable type is categorical or continuous.
- For example, just have a look on these methods and statistical measures for categorical and continuous variables individually:
  - **(i) Continuous Variables:** In continuous variables, users need to understand the central tendency and spread of the variable. These are measured using various statistical metrics visualization methods as shown below Table 6.4 and Fig. 6.3.

Table 6.4: Use of Statistical Measures for Continuous Variables

Central tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Bar Plot
Mode	IQR	Box Plot
Min	Variance	Line chart
Max	Standard deviation	
	Skewness and Kurtosis	

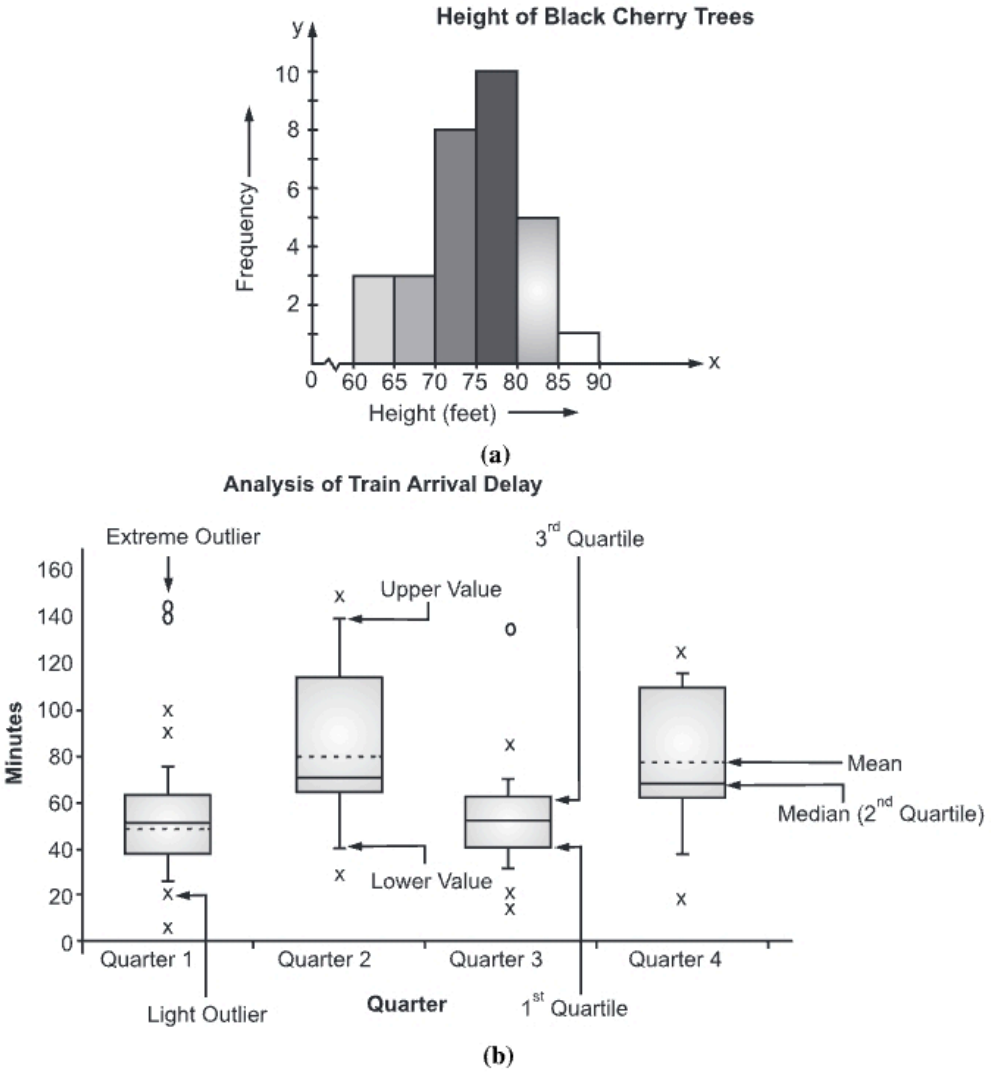


Fig. 6.3: Use of Statistical Visualization Methods for Continuous Variables

(ii) **Categorical Variables:** For this type of variables user will use frequency table to understand distribution of each category. It can be measure using percentage such as count and count% against each category. Bar chart can be used for visualization of categorical variables.

Step 3 - Detect Interactions Among Attributes:

- After the univariate analysis step relationships among different attributes in a dataset is to be found out. This helps to identify incompatibilities among attribute values, but it also enables analysts to generate optimal feature combinations) for subsequent analysis.
- There are two ways to perform analysis of attribute relationships: Bivariate and Multivariate statistics.

(1) Bivariate Statistics:

- Bivariate statistics only analyses the association of a chosen pair of attributes, the intersection of more than two variables are analysed using multivariate statistics.
- Bivariate analysis needs to be performed before multivariate analysis. So, user will get clear idea about attribute pair with one another, and more pairs can be formed for further analysis.
- Interactive filtering and aggregation of attributes are the two ways to perform bivariate analysis.
- Association and disassociation between two variables are found out at predefined significance level. This analysis can be performed for categorical and continuous variables.
- The combinations can be made such as:

(a) Categorical and Categorical: To find out relationship between these two variables following variables can be used.

(i) **Two-way Table:** Count and count% will be used here. The row represents the category of one variable and the columns represent the categories of the other variable. Count or count% of observations available in each combination of row and column categories can be shown. For example, two-way table of number of peoples playing in games gender wise as follows shown in Table 6.5.

Table 6.5: Two-way table of number of peoples playing in games

	Baseball	Basketball	Football	Total
Male	13	15	20	48
Female	23	16	13	52
Total	36	31	33	100

(ii) **Stacked Column Chart:** This method is used to display two-way table in visual form. Stacked charts display parts of a whole, only trends in the bottom series and in the total of all series can be accurately assessed. The stacked bar chart (aka stacked bar graph) extends the standard bar chart from looking at numeric values across one categorical variable to two. Each bar in a standard bar chart is divided into several sub-bars stacked end to end, each one corresponding to a level of the second categorical variable. For example, here is a stacked column chart showing the popularity of different colours among a group of children.

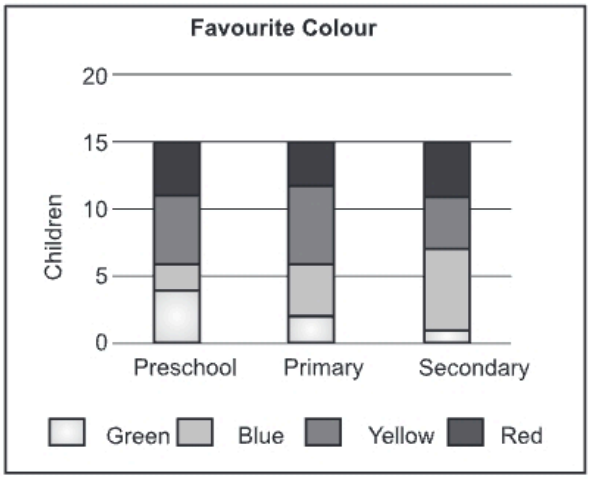


Fig. 6.4: Example of Stacked Column Chart

(iii) Chi-Square test: This is used to find out statistical significance of relationship between variables. It tests whether the evidence in the sample is strong enough to generalize that the relationship for a larger population as well. Chi-square is based on the difference between the expected and observed frequencies in one or more categories in the two-way table. It returns probability for the computed chi-square distribution with the degree of freedom.

(b) Continuous and Continuous: Scatter plot will be used for bivariable analysis between two continuous variables. It is clear way to find out the relationship between two variables. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear.

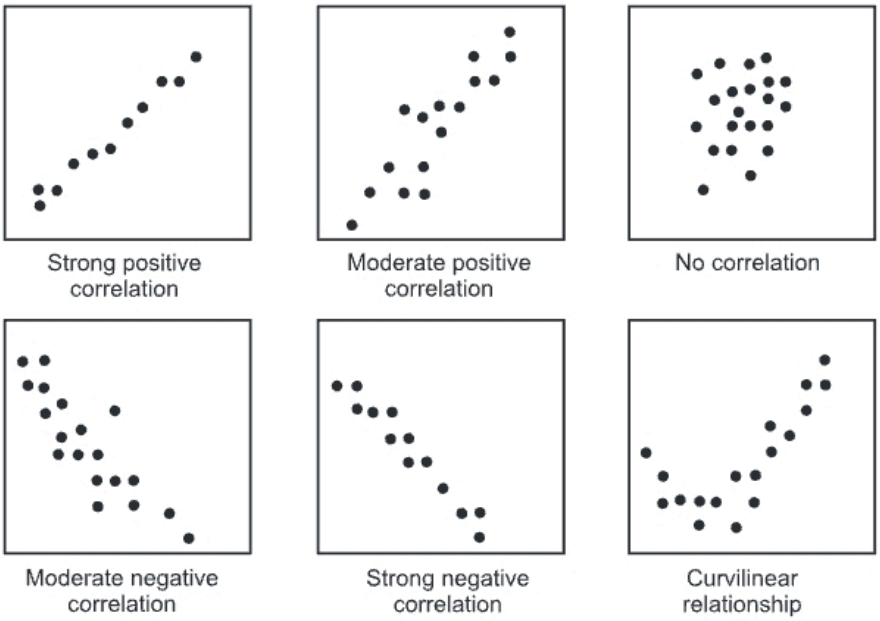


Fig. 6.5: Example of Scalar Plot

- Scatter plot shows the relationship between two variables but does not indicates the strength of relationship amongst them. To find the strength of the relationship, we use Correlation. Correlation varies between -1 and +1.
  - −1: Perfect negative linear correlation
  - +1: Perfect positive linear correlation
  - 0 : No Correlation

(c) Categorical and Continuous: When user is exploring relation between categorical and continuous variables, box plot can be drawn for each level of categorical variables. If levels are small in number, it will not show the statistical significance. To look at the statistical significance we can perform Z-test, T-test or ANOVA.

(i) Z-Test/ T-Test: Either test assess whether mean of two groups are statistically different from each other or not.

$$z = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(ii) ANOVA: It assesses whether the average of more than two groups is statistically different.

(2) Multivariate statistics:

- When pair of attributes is analysed next step is to perform analysis for multiple attributes. Clustering and dimensionality reduction techniques can be used for multivariate statistics.

Step 4 - Detect aberrant and missing values:

- Missing and unusual values may bias the result of analysis. Missing data in the training data set can slashes the power / fit of a model or can lead to a biased model because user have analysed the behaviour and relationship with other variables incorrectly. It can lead to wrong prediction or classification. For example, Consider Table 6.6

Table 6.6: Data set with Missing Values

Name	Weight	Gender	Play Cricket/ Not
A	58	M	Y
B	61	M	Y
C	58	F	N
D	55		Y
E	55	M	N
F	64	F	Y
G	57		Y
H	57	M	N

Table 6.7: Summary of Missing Values

Gender	#Students	# Play Cricket	#Play Cricket
Female	2	1	50%
Male	4	2	50%
Missing	2	2	100%

- In Table 6.6, missing values must be identified. The inference from this data set is that the chances of playing cricket by males are higher than females. On the other hand, if you look at the Table 6.8 which shows data after treatment of missing values (based on gender), we can see that females have higher chances of playing cricket compared to males.

Table 6.8: Data after Treatment of Missing Values

Name	Weight	Gender	Play Cricket/ Not
A	58	M	Y
B	61	M	Y
C	58	F	N
D	55	F	Y
E	55	M	N
F	64	F	Y
G	57	F	Y
H	57	M	N

Table 6.9: Summary of Missing Values

Gender	#Students	# Play Cricket	#Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

- Data may have missing values and can occur at two stages of EDA as follows:

(i) Data Extraction: Problems may occur at data extraction stage. You can do either double checking for the data which is to extract from database or hashing techniques can be used.

(ii) Data Collection: Errors occur at time of data collection are difficult to detect. There are four types as the following:

(a) Missing Completely at Random (MCAR): When the probability of missing variable is same for all observations.

(b) Missing at Random (MAR): When variable is missing at random and missing ratio varies for different values / level of other input variables.

(c) **Missing that depends on unobserved predictors:** When the missing values are not random and are related to the unobserved input variable.

(d) **Missing that depends on the missing value itself:** When the probability of missing value is directly correlated with missing value itself.

Methods used to treat missing values:

(i) **Deletion:** Listwise detection and Pairwise detection techniques can be used here.

(ii) **Mean/Mode/Median Imputation:** It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable. Generalized imputation and similar case imputation are the two techniques used here.

(iii) **Prediction Model:** Predictive method can be designed to predict missing values.

(iv) **KNN (K-nearest Neighbour) Computation:** The missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing. The similarity of two attributes is determined using a distance function.

Step 5 - Detect Outliers:

- This process happens after univariate, bivariate and multivariate analysis. Outlier is an observation that appears far away and diverges from an overall pattern in a sample. Outliers can add bias to the analysis leading to misinterpretation of attribute properties. Outliers are of three types:
  - (a) **Univariate:** They can be detected by calculation of the Inter-Quartile Range (IQR) of individual variables.
  - (b) **Bivariate:** It can be detected using combining two attributes and calculating their correlation coefficient.
  - (c) **Multivariate:** They can be detected by inspecting correlations among different attributes. Factor analysis techniques are used.
- Different visual techniques are used to display the outlier based on their types.
- Univariate and bivariate outliers can be detected using box plot, interactive histogram, and scatter plot.

Types of Outliers:

- There are different types of outliers:
  - (i) **Data entry Errors:** Human errors while doing data entry, data collection or recording can cause outliers in data.
  - (ii) **Measurement Error:** Errors detected while instrument used for measurement is faulty.
  - (iii) **Experimental Error:** Errors detected while doing actual experiment.
  - (iv) **Intentional Error:** Self-reported measures that involves sensitive data.
  - (v) **Data Processing Error:** While performing data mining, data can be extracted from multiple sources. It is possible that some manipulation or extraction errors may lead to outliers in the dataset.

(vi) **Sampling Error:** Different observations are collected in sample rather than specified.

(vii) **Natural Outlier:** When an outlier is not artificial i.e. not due to any above error then it is natural outlier.

Example:

- Following Table 6.10 is the simple example of outlier.

Table 6.10: Example of Outlier

Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 300
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation: 1.04	Standard Deviation = 85.03

- Outliers can be removed by many techniques such as Deleting observations, Transforming and binning values, Imputing and Treating them separately.

Step 6 - Feature Engineering:

- This is the final step of EDA. This step is performed after obtaining detailed insights about the dataset. This step is the core step of Exploratory Data Visualization. It is the science of extracting more information from existing data. Bringing out useful information from data is known as feature engineering. This step is further performed in two steps:

(1) **Variable Transformation**: Transformation is replacement of variable by function. It is the process of changing the distribution or relationship of a variable with others. For example, replacing a variable x by the square / cube root or logarithm x is a transformation. Variable transformations convert complex non-linear relationships into linear relationships; and standardize values to obtain a better understanding.

Reasons of using variable transformation:

(i) **Change of scale:** Change of scale can be done to standardise the values of variables. Such transformation must not change the shape of variable distribution.

(ii) **Transformation of non-linear relationship into linear relationship:** Transformation helps to convert non-linear relationship into linear relationship. Scatter plot can be used to find the relationship between two continuous variables. Logarithm method is commonly used for transformation.

(iii) **Symmetric distribution is preferred over skewed distribution:** Symmetric distributions are easier to interpret and generate inferences. When user is having skewed distribution, transformations are used to skewness. For right skewed distribution, Square / cube root or Logarithm of variable techniques can be used. For Left skewed distribution, Square / cube or Exponential of variables methods can be used.

(iv) **Implementation point view:** Binning of variables can be done here. For example, an experienced person having age more as compared to fresher and gives better performance. From an implementation standpoint, launching age based program might present implementation challenge. However, categorizing the sales agents in three age group buckets of < 30 years, 30-45 years and > 45 and then formulating three different strategies for each group is a judicious approach.

Methods used for Variable Transformation:

(i) **Logarithm:** Log of variable is to be found. It is very common method. It is used to change the shape of distribution of the variable on a distribution plot. It is used for reducing right skewness of variables. This method cannot be applied for zero or negative values.

(ii) **Square / Cube root:** This method is the sound effect on variable distribution. Not significant method as logarithm method. Cube root can be applied to negative values including zero. Square root can be applied to positive values including zero.

(iii) **Binning:** This method is used to categorize the variables. This method is carried out on original values, percentile, or frequency. Decision of categorization technique is based on business understanding.

(iv) **Normalization:** It helps to convert skewed distributions into more symmetric distributions.

(2) **Variable/Feature Creation:** This process is used to generate new variables/ features based on existing variables. For example, we have date format for birthdate as dd-mm-yyyy. Then we can separate out day, month, year, week, weekday, from this date. Following Table 6.11 shows simple example of variable creation.

Table 6.11: Simple Example of Variation Creation

Sr. No.	Sname	Sdate	New_day	New_Month	New_Year
S1	M	13-sept-2011	13	9	2011
S2	F	29-feb-2011	29	9	2011
S3	M	26-apr-2013	26	4	2013
S4	F	1-jan-2010	01	01	2010

Different techniques used for variable creation:

- There are various techniques used to create variables. Let's see some of the commonly used methods:

1. **Creating derived variables:** New variables are created from existing variables using set of functions. For example, to give name to a person what salutation we can user like Mr., Miss, Mrs, Master, Smt. For new variable. Methods such as taking log of variables, binning variables and other methods of variable transformation can also be used to create new variables.
2. **Creating dummy variables:** It is used to convert categorical variable into numerical variables. They are also called as indicator variables. It is useful to take categorical variable as a predictor in statistical models. Categorical variables can take value 0 and 1. For example, variable 'gender' has values: 1(Male), 0 (Female). Dummy variables can be created for more than two classes of categorical variables with n or (n-1) dummy variables.

- Following Table 6.12 gives clear ideas about example of creating dummy variable.

Table 6.12: Example of Dummy Variable

Sr. No.	Sname	Sgender	NewVar_Gender
S1	A	Male	1
S2	B	Female	0
S3	C	Female	0
S4	D	Male	1

- After performing all correct steps for EDA useful insights can be obtained based on which decisions can be made.

6.4 BASIC TOOLS (PLOTS, GRAPHS AND SUMMARY STATISTICS) OF EDA

- There are various tools and techniques used for Exploratory Data Analysis. Let's see some of the commonly used tools:

6.4.1 Design Capture Tools

- These tools capture a design and prepare it form simulation. Design requirements determine type of the design capture tool as well as the options needed.
- Some of the options are as follows:
  - Manual netlist entry
  - Schematic capture
  - Hardware Description Language (HDL) capture (VHDL, Verilog, …)
  - State diagram entry

6.4.2 Simulation and Verification Tools

- Simulation means the computer model used for purpose of study. There are various tools used for simulation in EDA like Functional (Logic) simulation tools and Timing simulation tools.
- **Functional simulators** verify the logical behaviour of a design based on design entry. The design primitives used in this should be always characterized completely.
- **Timing simulators** on the other hand perform timing verifications at multiple stages of the design. In this type of simulation, the real behaviour of the system is verified when encountering the circuit delays and circuit elements in actual device.
- These tools are sometimes called as "back annotated" tools as it checks functioning of each part of the tool or model.

6.4.3 Layout Tools

- ASIC (application-specific integrated circuit) designers usually use these tools. Designers transform a logic representation of an ASIC into a physical representation that allows the ASIC to be manufactured.

6.4.4 Synthesis and Optimization Tools

- Synthesis tools translate abstract descriptions of functionality such as HDL into optimal physical realizations, creating netlists that can be passed to a place and route tool. Then, the designer maps the gate level description or netlist to the target design library and optimizes for speed, area or power consumption.

Summary Statistics of EDA:

- It is used to quantify properties that user has observed using visual summaries and representation of EDA data. One of the aims of EDA is to find out problems in data and to understand different variable representations such as:

1. **Range:**
   - It gives details about how variables are distributed in given data set. It means practical distribution about data. So, range is typically used to characterize data spread.
   - In statistics, the range is the spread of your data from the lowest to the highest value in the distribution. It is a commonly used measure of variability.
   - The range is calculated by subtracting the lowest value from the highest value. While a large range means high variability, a small range means low variability in a distribution. There is simple formula to calculate the range R = H-L (H: highest range L - Lowest range).
   - For example, consider the data set in Table 6.13 with age of participant and their age.

Table 6.13: Data Set for Range

Participant	Age
A	37
B	19
C	31
D	29
E	21
F	26
G	33
H	36

Highest Value: 37 **Lowest Value:** 19

Range = 37 - 19 = 18 So, the range of our data set is 18 years.

Then range is paired with measures of central tendency it can tell use the span of distribution.

2. **Central Trends:**
   - It is descriptive summary of statistics through a single value that reflects the centre of the data distribution. Central tendency is a branch of descriptive statistics. It is measured in three terms as Mean, Median and Mode:

(i) Mean: It is arithmetic average. Add up all the values and divide by the number of observations in your dataset.

Formula for mean is, $\frac{x_1 + x_2 + \ldots + x_n}{n}$ i.e. $\bar{x} = \frac{\Sigma x}{n}$

The calculation of the mean incorporates all values in the data. If you change any value, the mean changes.

(ii) Median: The median is the middle value. It is the value that splits the dataset in half. To find the median, order your data from smallest to largest, and then find the data point that has an equal number of values above it and below it.

For example, for even data set,

Consider the data set 23, 21, 18, 16, 15, 13, 12, 10, 5, 7, 2, 1

Median = $\frac{(13 + 12)}{2} = \frac{25}{2} = 12.5$

For odd data set 23, 21, 18, 16, 15, 13, 12, 10, 5, 7, 2

Median = 13

(iii) Mode: The mode is the value that occurs the most frequently in your data set. On a bar chart, the mode is the highest bar. The mode is the most frequent score in our data set. If no value repeats, the data do not have a mode. Typically, you use the mode with categorical, ordinal, and discrete data. In fact, the mode is the only measure of central tendency that you can use with categorical data—such as the most preferred flavour of ice cream. For example, consider data set 5, 5, 5, 4, 4, 3, 2, 2, 1 Mode = 5

3. Spread (Variance):

- Variance is a measure of how spread out a data set is. It is calculated as the average squared deviation of each number from the mean of a data set. For example, for the numbers 1, 2 and 3 the mean is 2 and the **variance** is 0.667.

Steps for calculating variance are as follows:

(i) Find the mean of the data set. Add all data values and divide by the sample size n.

(ii) Find the squared difference from the mean for each data value. Subtract the mean from each data value and square the result.

(iii) Find the sum of all the squared differences.

(iv) Calculate the variance.

4. Skew:

- It is the degree of asymmetry observed in a probability distribution. Distributions can exhibit right (positive) skewness or left (negative) skewness to varying degrees Data are skewed right when most of the data are on the left side of the graph and the long skinny tail extends to the right. Data are skewed left when most of the data are on the right side of the graph and the long skinny tail extends to the left.

5. **Possible Modelling Strategies:**
   - Some popular statistical model examples include Logistic Regression, Time-Series, Clustering, and Decision Trees. Supervised learning techniques include Regression models and Classification models: Common regression models include Logistic, Polynomial and Linear Regression models.
   - EDA is also used to understand the relationship between pairs of variables like correlation and covariance.

6.5 TYPES OF EXPLORATORY DATA ANALYSIS

- Exploratory data analysis is generally cross classified in two ways. First, each method is either non-graphical or graphical. And second, each method is either univariate or multivariate (usually just bivariate).
- EDA falls into four main areas:
  1. **Univariate non-graphical EDA:** Looking at one variable of interest, like age, height, income level etc. This is the simplest form of data analysis among the four options. In this type of analysis, the data that is being analysed consists of just a single variable. The main purpose of this analysis is to describe the data and to find patterns. The usual goal of univariate non-graphical EDA is to better appreciate the "sample distribution" and also to make some tentative conclusions about what population distribution(s) is/are compatible with the sample distribution. Outlier detection is also a part of this analysis.
  2. **Univariate graphical EDA:** Unlike the non-graphical method, the graphical method provides the full picture of the data. The three main methods of analysis under this type are Histogram, Stem and Leaf plot, and Box plots.

     The histogram represents the total count of cases for a range of values. Along with the data values, the stem and leaf plot show the shape of the distribution. The box plots graphically depict a summary of minimum, first quartile median, third quartile, and maximum. Graphical methods are more qualitative and include a degree of subjective analysis.
  3. **Multivariate non-graphical EDA:** Analysis of multiple variables at same time

     The multivariate non-graphical type of EDA generally depicts the relationship between multiple variables of data through cross-tabulation or statistics. Cross tabulation method is used for multivariate non-graphical data. For two variables, cross-tabulation is performed by making a two-way table with column headings that match the levels of one variable and row headings that match the levels of the other variable, then filling in the counts of all subjects that share a pair of levels.
  4. **Multivariate graphical EDA:** This type of EDA displays the relationship between two or more set of data. A bar chart, where each group represents a level of one of the variables and each bar within the group represents levels of other variables. Side by side box plot technique will be used here.

6.6 BASIC PRINCIPLES OF DATA VISUALIZATION

Data Visualization:

- It allows business users to get insight/awareness into their huge amounts of data.
- It benefits business users to identify new patterns and errors in the data. These patterns give idea and help the users to pay attention to areas that indicate progress. This process, in turn, drives the business ahead to achieve business goals to get benefits.

- Data visualization tools and techniques represent data or information in graphical form or in any visual format like chart, graph etc. It displays different trends and patterns in data to be more easily seen.
- Data visualization is another form of visual art that grabs user interests. For example, by looking at chart user can quickly understand the meaning of chart.
- Data visualization is useful for data cleaning, exploring data structure, detecting outliers and unusual groups, identifying different trends and clusters, finding local patterns, evaluating modeling output, and presenting results to stakeholders.

Principles of Data Visualization:

1. **Tell the truth about data:** It might possible that sometimes data does not show the expected variations. If you choose the wrong graph, your readers will be confused or interpret the results incorrectly. Avoid misleading methods to ensure that graphs are clear and honest. User must take some precautions while telling truth about data. Some of the following precautions can be taken for this:
   (a) Omitting the base line: Generally, the baseline for a graph should start from zero unless specified otherwise. By starting the baseline from a different number, it can bias the perception of data. This technique is used to make the difference between data points seem to be greater than actual.
   (b) Going against conventions: There are certain conventions when plotting data. A larger bar indicates a greater amount, and a bigger area represents a higher number of values. By going against the convention, the viewer is accidentally led to a wrong inference.
   (c) Cherry-picking data: Cherry-picking is when only a few data points are plotted to show a misleading trend. This is one of the most common tactics used to mislead or deceive the audience.
2. **Know your audience:** Data visualizations are used to communicate result of analysis to audience. This goal is not achieved if the message is not conveyed. A good data visualization technique should follow the following guidelines:
   (a) Display the data according to the problem under study.
   (b) Take into the considerations the background knowledge of the audience in terms of literacy, technical knowledge, language etc. in this situation use graphs charts to display data instead of numbers.
   (c) Be sensitive to ethnicity and cultural values.
3. **Select the right graph/chart:** Some charts are having good intension but more time complexity. So, choose the graph based on the kind of data and the message to be conveyed. Do not use variety of graphs as just sake of purpose. Sometimes use of numbers is essential rather than graphs or charts. For example, display pie chart for percentage instead of bar graph.
4. **Highlight the important facts:** A data visualization can encode many data points, so highlight the most important facts to convey the message faster and with more impact. User can do this by removing noise, such as unnecessary gridlines, axes and labels. Use colour, size and pattern to highlight specific data points or a focus area. For example, by using right chart it is important to use colours and text strategically.

5. **Form must follow function:** An intuitive design is more important than appealing charts, and graphs should convey the meaning of data in an easy-to-understand manner.
6. **Determine the best-suited visual:** User should understand first the volume of data in hand. User must identify the aspects that he wishes to visualize along with the information that he wishes to convey. After this, user may select the best-suited and simplest visual format for your target audience.
7. **Balance the design:** The visual elements should be equally distributed across plots, charts, colour, text, shape, and space. Symmetrical visual software should be used for best visualization of data.
8. **Focus on the key areas:** Ensure that key areas should be highlighted so that it will quickly be highlighted in front of user.
9. **Keep visualization simple:** The visualizations should be displayed must be easy to understand. Remove unwanted information. Avoid confusions. The goal of data visualization is simplicity.
10. **Incorporate interactivity:** Data visualization tools implemented in to charts and graphs.
11. **Use patterns:** Different visualizations tools will be used to display patterns in data with some similar colour.
12. **Compare aspects:** Many comparison aspects are used to display same data using different charts either horizontally or vertically or in both manners.

6.7 BENEFITS OF DATA VISUALIZATION

The benefits of data visualization are as follows:

- Data visualization helps business stakeholders analyze reports regarding sales, marketing strategies, and product interest. Based on the analysis, they can focus on the areas that require attention to increase profits. This makes the business more productive.
- Different Visualization techniques are used to take quick action on problem under solution and take necessary actions for business growth.
- It benefits business users to recognize new patterns and to find errors in the data. These patterns give idea and help the users to pay attention to areas that indicate progress. This process, in turn, drives the business ahead to achieve business goals to get benefits.
- Some visualization techniques are used understand the story behind data
- As decision making becomes easy it will be used for exploring different business insights.
- It is used for grasping the latest trends in knowledge through data.

6.8 DATA VISUALIZATION TECHNIQUES

- Data visualization gives clear idea about data and its analysis. This makes the data more natural for the human mind to understand. This makes it easier to identify trends, patterns, and outliers within large data sets.

Importance of Data Visualization:

- Data visualization simplifies the information to be presented.
- The remarkable increase in data could not be understood by everyone at single point of time. Data visualization comes handy then.

- Uses of charts /graphs and different visualization techniques show do not only show the data but also established co-relations between different data types and information.
- Sharing of data visualization with all stakeholders is easy as only importance fact about data as shared not the whole.
- With the help of different visualization techniques, interactive visualization with deep and detailed analysis of the information being presented is also possible.
- Data visualization is interactive. User can click on it and get another big picture of a particular information segment. These segments are also tailored according to the target audience and could be easily updated if the information modifies. Following Fig. 6.6 will clear idea about data visualization.

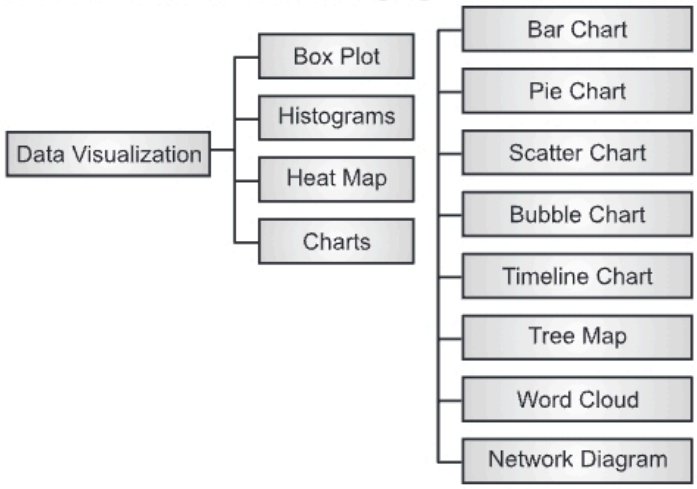


Fig. 6.6: Data Visualization Techniques

- Following are the techniques used for data visualization.

(A) Box Plot:

- Following Fig. 6.7 shows example of box plot. A box plot is a standardized way of displaying the distribution of data based on a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can tell user about outliers and what their values are. It can also tell user if data is symmetrical, how tightly data is grouped, and if and how data is skewed.

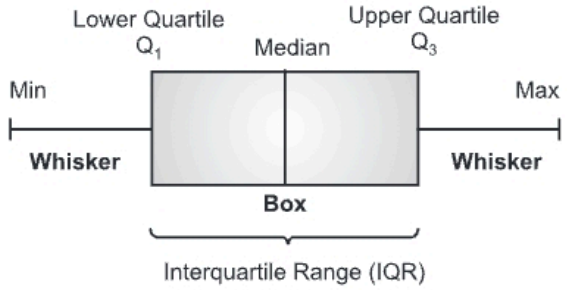


Fig. 6.7: Box Plot

- A box plot is a graph that gives you a good indication of how the values in the data are spread out.

- Box plots are very simple in comparison to a histogram or density plot.
- They have the advantage of taking up less space, which is useful when comparing distributions between many groups or datasets. For some distributions/datasets, user will find that more information is needed than the measures of central tendency (median, mean, and mode). User need to have information on the variability or dispersion of the data.

Table 6.14: Five-number Summary of Box Plot

Minimum	Q1 – 1.5*IQR
First quartile (Q1/25th Percentile)	The middle number between the smallest number (not the "minimum") and the median of the dataset.
Median (Q2/50th Percentile)	The middle value of the dataset.
Third quartile (Q3/75th Percentile)	The middle value between the median and the highest value (not the "maximum") of the dataset.
Maximum	Q3 + 1.5*IQR
Interquartile Range (IQR)	25th to the 75th percentile

(B) Histograms:

- A histogram is a graphical display of data using bars of different heights. In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range. A histogram displays the shape and spread of continuous sample data.

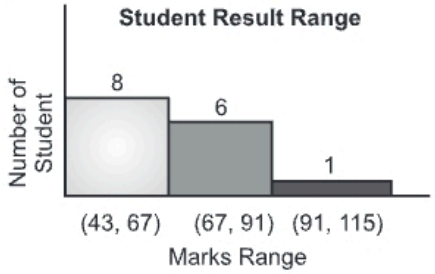


Fig. 6.8: Histogram for Student Result Range

- Histogram discovers and shows the underlying frequency distribution (shape) of a set of continuous data.

Histogram Vs Bar Chart:

- The major difference is that a histogram is only used to plot the frequency of score occurrences in a continuous data set that has been divided into classes, called bins. Bar charts, on the other hand, can be used for a lot of other types of variables including ordinal and nominal data sets.

(C) Heat Map:

- A heat map uses colour the way a bar graph uses height and width as a data visualization tool. For example, if user looking at a web page and he want to know which areas get the most attention, a heat map shows you in a visual way that's easy to assimilate and make decisions from. It is a graphical representation of data where the individual values contained in a matrix are represented as colours.

- It is useful for two purposes: For visualizing correlation tables and For visualizing missing values in the data.
- Excel assigns red colour to the lowest value and the green colour to the highest value, and all the remaining values get a colour based on the value. So, there is a gradient with different shades of the three colours based on the value. Following fig will clear idea about heat map.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2019	522	456	123	586	195	114	412	365	135	125	186	153
2020	589	452	145	452	652	485	875	965	352	145	465	485
2021	458	365	154	856	862	624	359	369	145	175	115	412

Fig. 6.9: Heat Map

(D) Charts:

(a) Line Chart: It is simplest technique. A line plot is used to plot the relationship or dependence of one variable on another. For Example, Number of books sold per month.

Fig. 6.10: Line Chart

(b) Bar Chart: It is used for comparing the quantities of different categories or groups. Values of a category are represented with the help of bars and they can be configured with vertical or horizontal bars, with the length or height of each bar representing the value. For example, following bar chart shows favourite colour of student.

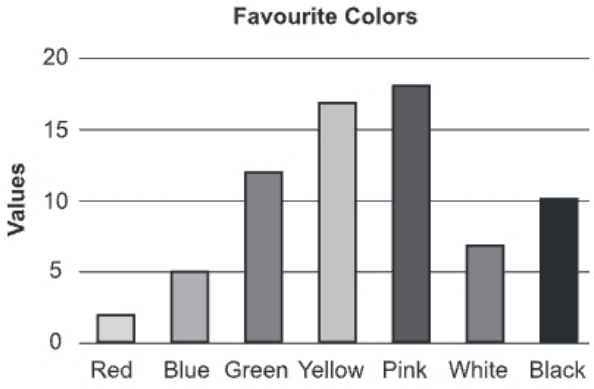


Fig. 6.11: Bar Chart

(c) **Pie Chart:** It is a circular statistical graph which decides slices to illustrate numerical proportion. Here the arc length of each slide is proportional to the quantity it represents. As a rule, they are used to compare the parts of a whole and are most effective when there are limited components and when text and percentages are included to describe the content.

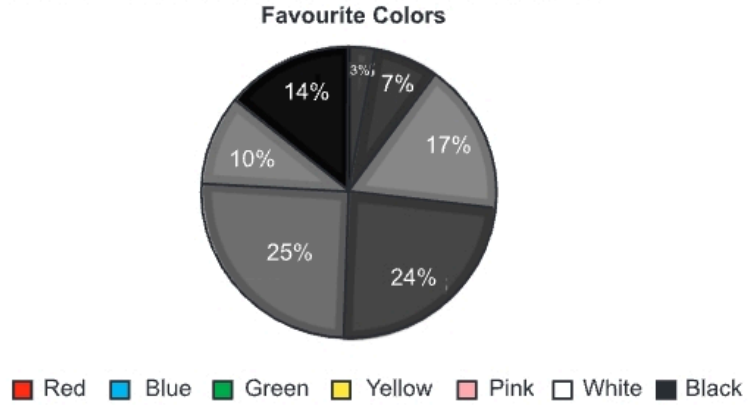


Fig. 6.12: Pie Chart

(d) **Scatter Chart:** A Scatter plot/chart is a two-dimensional plot representing the joint variation of two data items. Each marker (symbols such as dots, squares and plus signs) represents an observation. The marker position indicates the value for each observation. When you assign more than two measures, a scatter plot matrix is produced that is a series scatter plot displaying every possible pairing of the measures that are assigned to the visualization. Scatter plots are used for examining the relationship, or correlations, between X and Y variables.

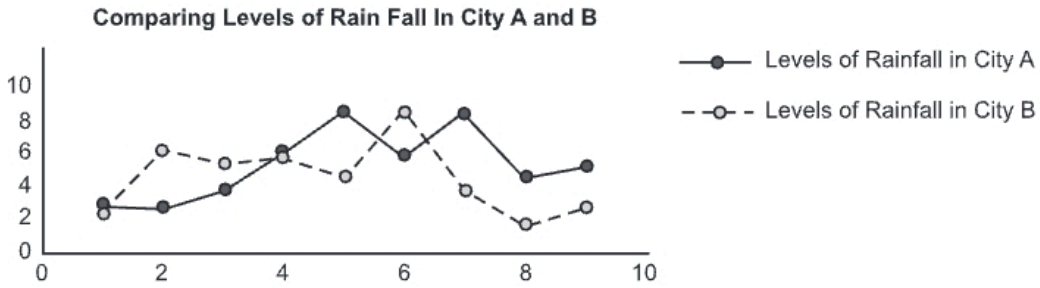


Fig. 6.13: Scatter Plot

(e) **Bubble Chart:** It is a variation of scatter chart in which the data points are replaced with bubbles, and an additional dimension of data is represented in the size of the bubbles.

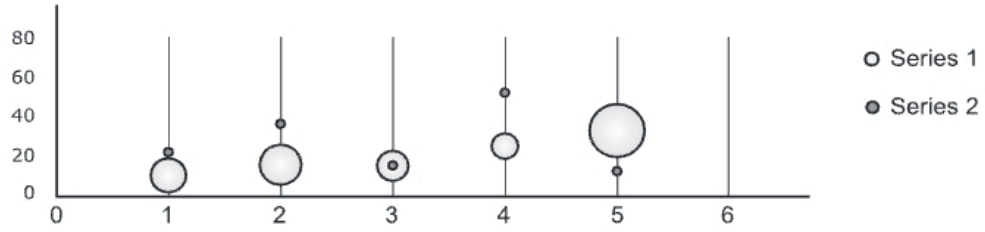


Fig. 6.14: Bubble Chart

(f) Timeline Chart: Timeline charts illustrate events, in chronological order. For example, the progress of a project, advertising campaign, acquisition process — in whatever unit of time the data was recorded. For example, week, month, year, quarter. It shows the chronological sequence of past or future events on a timescale.

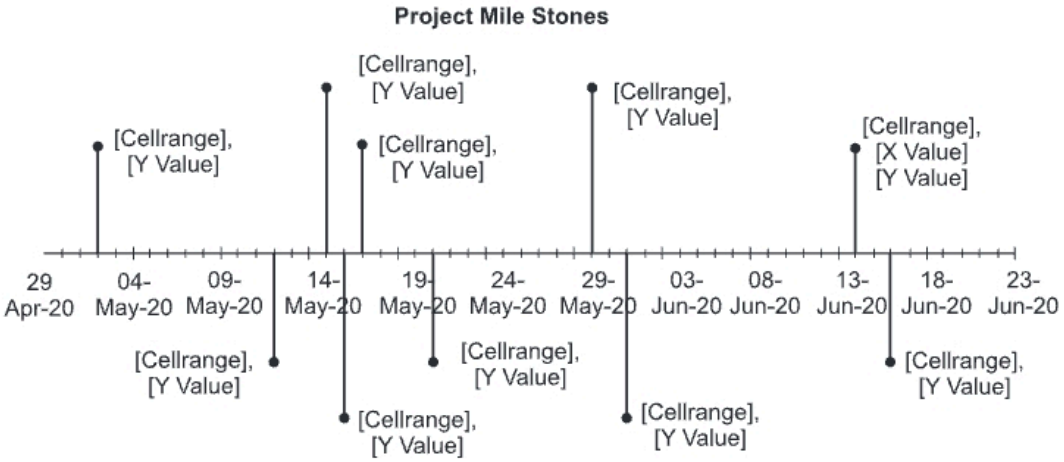


Fig. 6.15: Timeline Chart

(g) Tree Map: A tree map is a visualization that displays hierarchically organized data as a set of nested rectangles.

In this Map, parent elements being tiled with their child elements. The sizes and colours of rectangles are proportional to the values of the data points they represent. A leaf node rectangle has an area proportional to the specified dimension of the data. Depending on the choice, the leaf node is coloured, sized or both according to chosen attributes. They make efficient use of space, so display thousands of items on the screen simultaneously..

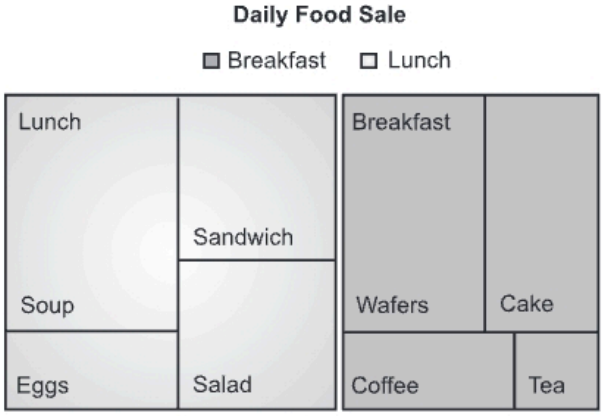


Fig. 6.16: Tree Map

(h) Word Cloud:

- The variety of big data brings challenges because semi-structured, and unstructured data require new visualization techniques. A word cloud visual represents the frequency of a word within a body of text with its relative size in the cloud. This technique is used on unstructured data as a way to display high or low-frequency words.

Fig. 6.17: Word Cloud

(i) Network Diagram:

- This technique can be used for semi-structured or unstructured data. Network diagrams represent relationships as nodes (individual actors within the network) and ties (relationships between the individuals). They are used in many applications. For example, for analysis of social networks or mapping product sales across geographic areas.

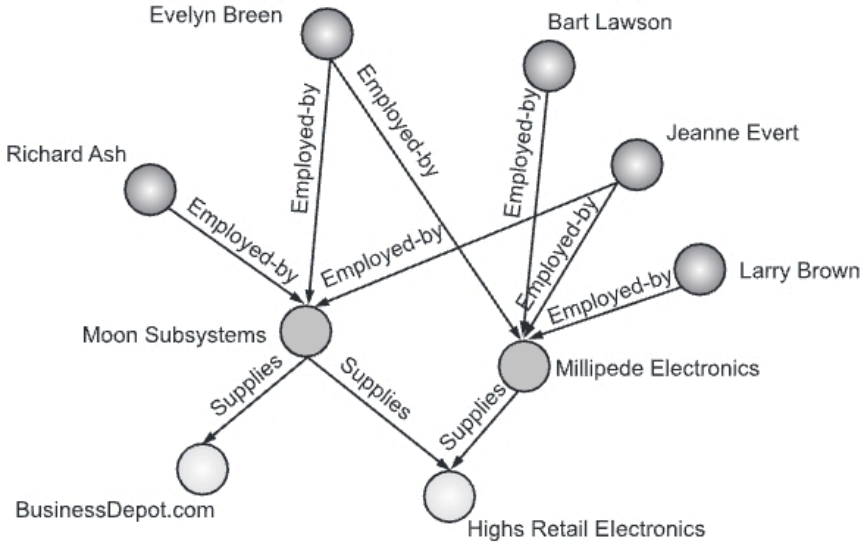


Fig. 6.18: Network Diagram

6.9 TOOLS FOR DATA VISUALIZATION

- There are various vast number of tools available in market for Data Visualization. Using these tools, data and information can be generated and read easily and quickly. Many data visualization tools range from simple to complex and from intuitive to obtuse.
  - **Tableau Desktop:** A business intelligence tool which helps you in visualizing and understanding user data.
  - **Zoho Reports:** Zoho Reports is self-service business intelligence (BI) and analytics tool that enables user to design intuitive data visualizations.
  - **MATLAB:** A detailed data analysis tool that has an easy-to-use tool interface and graphical design options for visuals.
  - **Microsoft Power BI:** Developed by Microsoft, this is a suite of business analytics tools that allows you to transform information into visuals.
  - **Sisense:** A BI platform that allows you to visualize the information to make better and more informed business decisions.

6.10 DIFFERENTIAL DATA VISUALIZATION EXAMPLES

- **Government Budget:** Government budgets are always tough to understand as they number and more numbers. A recent example is a colour-coded tree map that was designed by The White House during Barack Obama's presidency, which visually broke down the US's 2016 the budget for better understanding and put government programs in context.
- **World Population:** A world map showing the population density is presented by data visualization.
- **Profit and Loss:** Business companies often resort to pie charts or bar graphs showing their annual profit or loss margin.
- **Films and Dialogues:** The makers of popular sitcom 'FRIENDS' used a pie chart during shooting to ensure that every six characters have an equal number of jokes and dialogues.
- **Anscombe's Quartet:** It is one of the most well-known and popular, which has four data sets of identical descriptive statistics, but they appear different when graphed.
- **Sports Analytics:** Sports broadcasting channels uses visualization techniques to convey the right information to the audience by just glancing through them for fraction of seconds.

6.11 LIST OF METHODS TO VISUALIZE DATA

- **Column Chart:** It is called vertical bar chart where each category is represented by a rectangle. The height of the rectangle is proportional to the values that are plotted.
- **Bar Graph:** It has rectangular bars in which the lengths are proportional to the values which are represented.
- **Stacked Bar Graph:** It is a bar style graph that has various components stacked together so that apart from the bar, the components can also be compared to each other.
- **Stacked Column Chart:** It is like a stacked bar; however, the data is stacked horizontally.
- **Area Chart:** It combines the line chart and bar chart to show how the numeric values of one or more groups change over the progress of a feasible area.
- **Dual Axis Chart:** It combines a column chart and a line chart and then compares the two variables.

- **Line Graph:** The data points are connected through a straight line; therefore, creating a representation of the changing trend.
- **Mekko Chart:** It can be called a two-dimensional stacked chart with varying column widths.
- **Pie Chart:** It is a chart where various components of a data set are presented in the form of a pie which represents their proportion in the entire data set.
- **Waterfall Chart:** With the help of this chart, the increasing effect of sequentially introduced positive or negative values can be understood.
- **Bubble Chart:** It is a multi-variable graph that is a hybrid of Scatter Plot and a Proportional Area Chart.
- **Scatter Plot Chart:** It is also called a scatter chart or scatter graph. Dots are used to denote values for two different numeric variables.
- **Bullet Graph:** It is a variation of a bar graph. A bullet graph is used to swap dashboard gauges and meters.
- **Funnel Chart:** The chart determines the flow of users with the help of a business or sales process.
- **Heat Map:** It is a technique of data visualization that shows the level of instances as colour in two dimensions.

6.12 ADVANTAGES AND DISADVANTAGES OF EDA

Advantages of EDA:

- It gives up valuable insights into the data.
- Visualization is an effective tool to detect outlier.
- It helps us with feature selection.

Disadvantages of EDA:

- If not performed properly EDA can misguide a problem
- EDA is not effective when we deal with high dimensional data.

Summary

- Exploratory data analysis (EDA) is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods.
- It can also help determine if the statistical techniques you are considering for data analysis are appropriate.
- There are different types of dimensionalities of data sets namely univariate, bivariable and multivariate data set.
- EDA consists of different steps like distinguish attributes, univariate analysis, detect interactions among attributes (bivariate analysis, multivariate analysis), detect aberrant and missing values, detect outliers and feature engineering.
- There are different types of EDA as univariate non-graphical, univariate graphical, multivariate non-graphical, multivariable graphical.
- Users have to follow basic principles for data visualization for better understanding by its stake holders.
- Different techniques such as box plot, histogram, heat map, chats(bar chart, pie chart, scatter chart, bubble chart, timeline chart, tree map, word cloud, network diagram) will be used depend on type of data.

Check Your Understanding

1. HDL stands for _____ .
   (a) Hardware Description Language (b) Hardware Definition Language
   (c) Hard Description Language (d) Heavy Description Language
2. Which of the following data Visualization tool is created be Microsoft?
   (a) Indo BI (b) MATLAB
   (c) Mi BI (d) Microsoft Power BI
3. In _____ dimensionality data set measurement is made on one variable per subject.
   (a) Univariate (b) Bivariate
   (c) Multivariate (d) Manyvariate
4. IDA means _____ .
   (a) Intimate Data Analysis (b) Init Data Analysis
   (c) Initial Data Analysis (d) It Data Analysis
5. _____ step of EDA is also called as variable identification.
   (a) Different Attributes (b) Data Attributes
   (c) Dealing Attributes (d) Distinguish Attributes
6. _____ step in EDA allows the determination of attribute combinations/patterns for successive analysis.
   (a) Univariate Data Analysis (b) Feature engineering
   (c) Data Analysis (d) Dealing Attributes
7. Which of the following is the measure of central tendency?
   (a) Mean (b) Range
   (c) IQR (d) Histogram
8. Which of the following is the measure of dispersion?
   (a) Standard Deviation (b) Bar Plot
   (c) Mode (d) Mean
9. _____ type of chart can be used for visualization of categorical variables.
   (a) Bar chart (b) Box Plot
   (c) Histogram (d) Tree Map
10. _____ method is used to find out relationship between two variables in two Categorical variables.
    (a) Three-way table (b) Table
    (c) Rows (d) Two-way table
11. _____ method is used to display two ay table in visual form.
    (a) Bar chart (b) Histogram
    (c) Pie chart (d) Stacked Column Chart
12. _____ plot will be used for bivariable analysis between two continuous variables.
    (a) Pie chart (b) Scatter
    (c) Histogram (d) Column
13. In which of the following two stages there is possibility of missing values in data.
    (a) Data Extraction, Data Mining (b) Data Extraction, Data Collection
    (c) Data Localization, Data Mining (d) Knowledge Extraction, Noise Removal

14. KNN stands for _____ .
    (a) K-Near Node (b) K-Near Nothing
    (c) K-Nearest Neighbour (d) K-Narrow Node
15. _____ is the final step of EDA.
    (a) Feature Engineering (b) Knowledge Extraction
    (c) Distinguish Attributes (d) Detect Outliers
16. _____ is the well-known method for variable transformation.
    (a) Data Visualization technique (b) Data Reduction
    (c) Logarithm (d) Data binning
17. _____ visualization technique displays hierarchically organized data as a set of nested rectangles, parent elements being tiled with their child elements.
    (a) Bar Chart (b) Pie Chart
    (c) Histogram (d) Tree Maps
18. IQR stands for _____ .
    (a) Interquartile Range (b) In quartile Range
    (c) In Quotient Range (d) In Quad Range

Answers

1. (a)	2. (d)	3. (a)	4. (c)	5. (d)	6. (a)	7. (a)	8. (a)	9. (a)	10. (d)
11. (d)	12. (b)	13. (b)	14. (c)	15. (a)	16. (c)	17. (d)	18. (a)		

Practice Questions

Q.I Answer the following questions in short.

1. What is EDA?
2. List the reasons of using EDA.
3. List of the different steps of EDA.
4. Explain Continuous variables with example.
5. Explain Categorical variables with example.
6. Why is chi square test used?
7. What is the purpose of scatter plot?
8. What is the purpose of Z-test?
9. List of the methods used to treat missing values.
10. List the different Data Visualization Techniques.
11. Which are the two basic types of data visualization?
12. Which is the best visualization tool?

Q.II Answer the following questions.

1. What are different dimensionalities of Data Sets?
2. Comment the statement 'IDA is part of EDA'.
3. Differentiate between EDA and CDA.
4. Explain steps of EDA with neat diagram.

5. What is stacked column chart? Explain with example?
6. Explain with example how missing values are corrected?
7. Explain different types of outliers.
8. Explain the two steps of Feature Engineering.
9. State the different reasons of variable transformation.
10. What are the different methods used for variable transformation?
11. Explain variable or feature creation with simple example.
12. Explain different techniques used for variable creation.
13. What are different types of Exploratory Data Analysis?
14. State the basic principles of Data Visualization.
15. Explain different benefits of Data Visualization.
16. Explain the following Data Visualization Techniques.
    (a) Box Plot
    (b) Histograms
    (c) Heat Maps
    (d) Bar Charts
    (e) Pie Chart
    (f) Scatter Chart
    (g) Bubble Chart
    (h) Timeline chart
    (i) Tree Maps
    (j) Word cloud
    (k) Network Diagram
18. What are different tools of Data Visualization?
19. Explain any two Data Visualization examples in real Life.

Q.III Define the terms.

1. Bivariate Data Set
2. Categorical Variables
3. Chi-Square test
4. Outliers
5. Central trends-Skew

✍ ✍ ✍