Unit III

3 Reporting Authoring

Syllabus

Building reports with relational vs Multidimensional data models; Types of Reports - List, crosstabs, Statistics, Chart, map, financial etc; Data Grouping & Sorting, Filtering Reports, Adding Calculations to Reports, Conditional formatting, Adding Summary Lines to Reports. Drill up, drill - down, drill-through capabilities. Run or schedule report, different output forms - PDF, excel, csv, xml etc.

Contents

3.1 Introduction

3.2 Concept of Reporting

3.3 Benefits of Reporting

3.4 Building Reports with Relational and Multidimensional Data Models

3.5 Types of Reports

3.6 Operations on Data Reports

3.7 Run, Schedule and Generate Reports

3.1 Introduction

- In today's world for every business either it is small, medium or large its base is data. Data is scattered everywhere. As per the business perspective organizations know how to collect or extract data from various sources. Due to involvement of internet in every sector, data is growing tremendously. In this way normal data is converted into big data which plays very important role in online data analysis.
- Now the major challenge to all organizations is how to extract the required information from gathering data, how to understand that data and how to represent that data. Answer to all these questions is Business intelligence reporting (BI Reporting).
- If collected data is huge and from different sources then spreadsheet is not going to provide feasible solution. Because every source has created this data as per their business needs. Each of them have different naming mechanism. Mapping all these things in our single spreadsheet is very difficult, it may result in to redundancy, missing some labels etc.
- Due to these issues, business intelligence reporting comes into picture. It is very strong for providing platform for organizations to collect data effectively, transform it into required format and generate insights from data. BI reporting helps to create the smart report with the help of BI tools.

3.2 Concept of Reporting

- In BI, reporting is the process of collecting and analyzing data with the help of different modern BI tools. BI tools are capable to provide interactive data visualizations. It helps businesses organizations to extract required data and optimize organizational policies for continuous growth.

3.3 Benefits of Reporting

- Help to improve the workflow of organization
- Uses real time and historical data
- Operation optimization
- Cost optimization
- Improve procurement process
- Performance monitoring of employee
- Customer analysis and prediction

- Resource management
- Business forecasting

3.4 Building Reports with Relational and Multidimensional Data Models

3.4.1 Relational Data Model

- The basic data model is the relational data model. It is widely used in all organizations as well as in all software as a backend. It works efficiently for data storage and data processing. It is the primary and simplest model which is having all capabilities to process and store data efficiently. It basically work on ACID(atomicity, consistency, isolation and durability) properties.
- Following terminologies are frequently used to represent the relational model.
  1. **Tables -** Table formats are used to represent relation in relational data model. Table consists of rows and columns, where records are represented using rows and attributes are represented using columns. This format stores the relation among entities.
  2. **Tuple -** Tuple is a single row of a table, which consist of a single record of a relation.
  3. **Relation instance -** It is a finite set of tuples present in relation at particular instance. Relation instances do not have duplicate tuples.
  4. **Relation schema -** It describes the relation name i.e. table name along with its attributes with name.
  5. **Relation key -** It is used to identify the row in a table(relation). It consist of one or more attributes which are used to identify rows uniquely.
  6. **Attribute domain -** Every attribute has some pre-defined value scope, known as attribute domain.
- Different database software's are used to represent relational data models like ORACLE, MYSQL, MYSQL works bench, SQL server etc.

3.4.2 Multidimensional Data Model

- Relational data model is the two dimensional dataset which is combination of rows and columns. But if we have more information about same row with respect to another dimension then we need to add one more dimension in relation. Then it became three-dimensional data, similarly we can go adding number of dimensions and it is the multidimensional data model.

- We can represent student's detail like roll no, name, address, phone number, age and branch in relational model very easily in two dimension or table format. But if we want to add student's health history then we will add one more dimension with this data model. In this way the numbers of dimensions get added and data become very complex. But as we know data can be a powerful as well as valuable asset for any organization. But as it becomes complex it becomes difficult to extract some actionable insights out of it. Multidimensional data describes such a data set in a way that allows you to look at it beyond traditional two-dimensional structures.
- A multidimensional model represents the data in the cube form. It is called as data-cube. A data cube enables data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.
- The dimensions are the views or entities concerning which an organization to keep it as records. For example, a shop has created a sales data warehouse. It keep records of the store's sales with respect to item, time and location. So item, time and location become dimensions in this case. These dimensions are used to keep track of sales of items area wise and time wise. Example if store owner want to have look at weekly sales of particular item in particular location, he can do it easily if data is in multidimensional mode. Similarly we can go on adding other dimensions like item_name, brand, type etc. A multidimensional data model is organized around a central theme, for example, sales. This theme is represented by a fact table. Facts are numerical measures. The fact table contains the names of the facts or measures of the related dimensional tables.
- As shown in Fig. 3.4.1, Pid, Timeid, locid and sales are the dimensions of data set. Fact table consist of reference to all these dimensional table. Dimentional tables are nothing but details of each dimension like pid(11,12,13),Timeid(1,2,3) and locid(1,2) and sales is the actual sales count of some item.

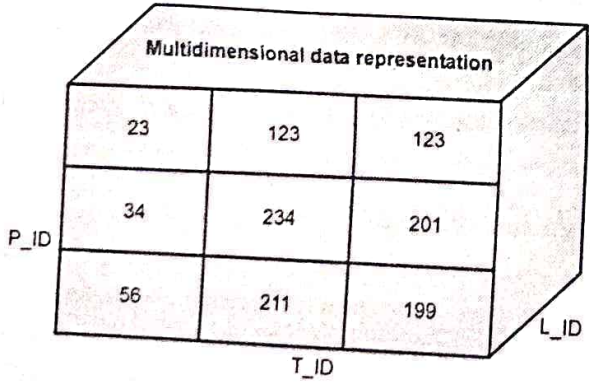


Fig. 3.4.1 Tabular and multidimensional representation of data

Tabular Representation

P_ID	T_ID	L_ID	Sales_Count
P1	T1	L1	56
P1	T2	L1	211
P1	T3	L1	199
P2	T1	L1	34
P2	T2	L1	234
P2	T3	L1	201
P3	T1	L1	23
P3	T2	L1	123
P3	T3	L1	123
P1	T1	L2	1001

- Table 3.4.1 represents the item sold in city Mumbai per quarter. This data is initially represented in two dimensions like location and time. In this 2D representation, the sales for Mumbai location organized quarter wise, further the item with its type are added and actual sales is shown in Table 3.4.1.

Location = "Mumbai"				
Time	Item sales count			
	Milk	Bread	Butter	Jam
T1	390	345	278	205
T2	299	245	120	105
T3	178	122	120	115
T4	450	378	199	170

Table 3.4.1 Location dimension

- Table 3.4.2 shows the three dimensional representation of example discussed in Table 3.4.1. Here location (Mumbai, Pune, Chennai, Delhi), Time (T1, T2, T3, T4) and item (Milk, Bread, Butter, Jam) are the three dimensions and its sales count is represented in Table 3.4.2. These 3D data are shown in the table. The 3D data of the table are represented as a series of 2D tables.

	Location = "Mumbai"				Location = "Pune"			
	Item				Item			
Time	Milk	Bread	Butter	Jam	Milk	Bread	Butter	Jam
T_1	390	345	278	205	290	245	166	200
T_2	299	245	120	105	199	245	120	105
T_3	178	122	120	115	178	122	102	115
T_4	450	378	199	170	350	322	109	97

Location = "Chennai"				Location = "Delhi"			
Item				Item			
Milk	Bread	Butter	Jam	Milk	Bread	Butter	Jam
490	344	299	250	590	544	499	450
499	445	420	305	656	645	620	605
478	422	402	315	678	622	602	315
450	422	409	397	650	622	509	397

Table 3.4.2 Three dimensional data

- Finally all these three dimensions are represented in the form of cube which is a shown in Fig. 3.4.1.

3.4.3 Advantages of Multidimensional Data Model

- Grouping of similar information effectively.
- Easy visualization of the output of complex data.
- Easily identify patterns, trends, outliers and anomalies from data as it is grouped.
- Examine relationships among data from multiple sources.
- Extract valuable information from unstructured data.
- Suitable for both structured as well as unstructured data.
- Include elaboration of relation which is helpful for non - data scientist stakeholders.
- Compare the impact of changes to variables on the target data easily.
- Process data quickly.
- Easy maintenance of data.

3.5 Types of Reports

3.5.1 List

- It is primary / basic form of report. Even though it seems to be very simple, but it is the base of reporting about data and its insights. Further by extending basic list next level analysis of data is done and it is represented in the form of graph, map, table etc.
- Example if we have student dataset which consist of students' academic performance. To report about this data, initially we will list all the students as per their final exam score like distinction, first class, second class, pass class and then fail. This is basic listing. Further by making use of it we can go for advanced analysis like qualitative and quantitative result of class or percentage of passing and failing rate.

3.5.2 Cross Tabs

- Basically cross tab is extended version of simple table. Generally to represent single categorical variable, tables or frequency tables are used to represent its count. To describe the relationship between two categorical variables, cross tabulation or cross-tab reporting style is used. In a cross tab, the categories of one variable determine the rows of the table and the categories of the other variable determine the columns. The cells of the table contain the number of times that particular combination of categories occurred. The edges or boundaries of the table represents summarized or grouped observations of that category. This type of table is also known as a Crosstab / Two-way table / Contingency table. It is a statistical tool used for categorical data. Categorical data involves values that are mutually exclusive to each other. Data is always collected in numbers, but numbers have no value unless they mean something.
- The Table is a powerful reporting tool that enables us to display detail, grouped as well as aggregated data view of our dataset. Basically table reporting type supports the Table structure, Crosstab and List structure of data. These all are optimized source of data representation for a specific data layout. Table represents detail data in grid structure, a Crosstab displays grouped data in a grid structure and a List displays detail data which doesn't follow any particular structure. It just displays data in list format.
- By design, each Table or Crosstab cell contains a text box. You can add multiple report items to a Table cell by first adding a Panel item as a container and then, the report items to the Panel. Each List cell contains a Panel. You can replace a default report item with another report item, for example, an image. As you define groups for a Table, Crosstab, or List, the Report Designer adds rows and columns to the Table in which to display grouped data.

- Examples 1, 2 and 3 show the different forms of cross tabs like (2 × 2), (4 × 2) a (3 × 2). Here Row variable are Gender (2 categories : male, female), Column varia is got placed ? (2 categories : yes, no) and table dimensions is 2 by 2.

Example 1 : Cross tab (2 × 2)

		Student got placed ?		Total
		Yes	No	
Gender	Male	120	30	150
	Female	80	20	100
Total		200	50	250

Example 2 : Cross tab (4 × 2)

		Gender		Total
		Male	Female	
Post	Manager	10	05	15
	Team lead	20	05	25
	Assistant	30	20	50
	Supervisor	80	20	100
Total		140	50	190

- Here, Row variables are Post (4 categories : manager, team lead, assistant supervisor), Column variables are Gender (2 categories : male, female) and tabl dimensions is 4 by 2. Similarly the cross tab of 2 by 3 dimension is as shown i example 3.

Example 3 : Cross tab (2 x 3)

		Post			Total
		Manager	Team lead	Assistant	
Gender	Male	05	07	10	22
	Female	03	05	15	23
Total		08	12	25	45

- In this way cross table retrieves the detail report data from the data source and also allows us to organize aggregated detail into groups. We can create cross tab using different platforms like Pandas library, SPSS tool, Alteryx, Tableau, powerBI, Excel etc.

3.5.3 Statistics

- By using statistics, the collected data can be abbreviated and represented in such a way that it can be easily understood and actionable insights can be extracted. Statistical reporting strategy uses three basic statistical tests.

1. **Descriptive statistics :** The main objective of descriptive statistics is to demonstrate a huge portion of the collected data through summary, charts and tables. It intends to give brief summary of gathered data using descriptive charts and tables. Descriptive statistics is typically used to illustrate univariate analysis of data. Descriptive statistics is associated with measures of tendency such as the mean, mode, median and measures of dispersion like variance and standard deviation. It is not directly helpful to derive certain conclusion about data or sampled population but it is a very important tool while generating statistical reports.
2. **Inferential statistics :** The main intention of Inferential statistics is to provide a more detailed and effective statistical data analysis. Inferential statistics is involved with making broader and deeper deductions and interpretations usually on the interaction between variables, cause and effect relationship and identifying the scope of the sample's representation in the population. Based on the sample, the survey will verify the hypothesis and then come up with a conclusion. Commonly used inferential statistics in data analysis are Analysis of Variance (ANOVA), T-test, Z-test, Chi-square test, linear regression and multiple regression
3. **Psychometric tests :** Psychometric tests analyze the attributes and performance of the employed survey to ensure that the survey data is reliable and valid. Example of a psychometric test is Cronbach's Alpha.

Statistical reporting tools

- Statistical reporting tools are also used in further understanding the survey data, which is a key factor in making business decisions. Among these are factor analysis, cluster analysis, gap analysis, Z-test, and U-test. In a factor analysis, the obtained data are classified into recognizable clusters. On the other hand, a cluster analysis specifies data clusters that have unique and traceable attributes. Moreover, a gap analysis correlates data and determines if the data differences are statistically important. A Z-test matches two percentile scores and determines if they are statistically important while a U-test equates median scores of axis-define groups and identifies if their differences are statistically relevant.

- These statistical reporting tools can also be supplemented with tables such as frequency tables and cross tabs. Frequency tables depict all the response choices, how many times it has been answered and the percentage of participants who chose those responses. These are beneficial especially when there are various response choices present or if there is a little disparity between the responses. When two different subgroups or subsets of data will be compared, cross tabulation or cross tab is the best way to go.

Types of statistical reporting data

- There are four types of data normally come across during statistical analysis and are presented in statistical reporting data : Categorical data, Ordinal data, Interval data, and Ratio data.
  - **Categorical data -** This type of data is a result of relative frequency statistics. Example is dividing the sum of a certain response with the total number of responses. Let us say that for a brand survey, the brand quality choice accumulated 25 responses out of 100. Therefore, it can be inferred that 25 percent of the participants prefer brand quality in choosing a brand.
  - **Ordinal data -** Ordinal data are best represented using frequency tables. These are data that have scales and ordered according to preference. For example, out of 100 respondents, 45 of them agree that the brand needs to improve its packaging. In percent, that is equivalent to 45 %.
  - **Interval data -** Encapsulating interval data can be best done when treated as an ordinal data. Averaging and standard deviation are the ideal techniques in evaluating this type of data.
  - **Ratio data -** Ratio data can be converted to a normal data using logarithms and square roots. A distinguishing characteristic of this set of data is that it has a defined zero point. Decimals and fractions are also available in a ratio data.

3.5.4 Chart

- Generally the concept of chart and graph are not similar but in excel it looks like similar. These two are different concepts and are dependent on each other. Graphs are generally a representation of numerical data. It depict the relationship between two variable, If one variable change how much it effect on other number. However, charts are the visual representations of all type of data which may or may not be related. Charts display it in the form of graphs as well as some additional data.

- The main objective behind graphs and charts is to display data in a meaningful and crisp manner with a visual representation of values that allows the intended user to easily understand and analyze the data without getting into the granular details of such data. Thus we can say that, all graphs are a type of chart, but not all charts are graphs.
- Graphs mainly focus on unprocessed or raw data and draw the conclusion related to trend overtime - related to such data. A two-dimensional graph shows the relationship between the data through a line, curve, etc., using the horizontal line along the bottom (called X-axis) and vertical line up the side (called Y-axis). A Graph is a mathematical diagram that shows the relationship between two or more sets of numbers or measurements. It allows the user to easily represent the values in the data through a visual representation. There are two types of graphs : Bar graphs and Line graphs.
- A chart is a type of representation of large sets of data, which makes it easy for the user to understand the data in better manner. Using the same helps predict existing data and forecast future data based on the present data pattern. A chart can take the form of a diagram, a picture, or a graph. We can transform datasets into a meaningful display of information using charts. An example of a simple chart is shown below.

1] Bar chart : Bar charts are used to compare data between different groups and help to track the changes in data overtime. Bar charts are most useful when there are big changes or to show how one group compares against other groups. There are different type of bar chart such as vertical bar chart, Horizontal bar chart, stacked bar charts and grouped graph chart. Following Fig. 3.5.1 represents the bar charts.

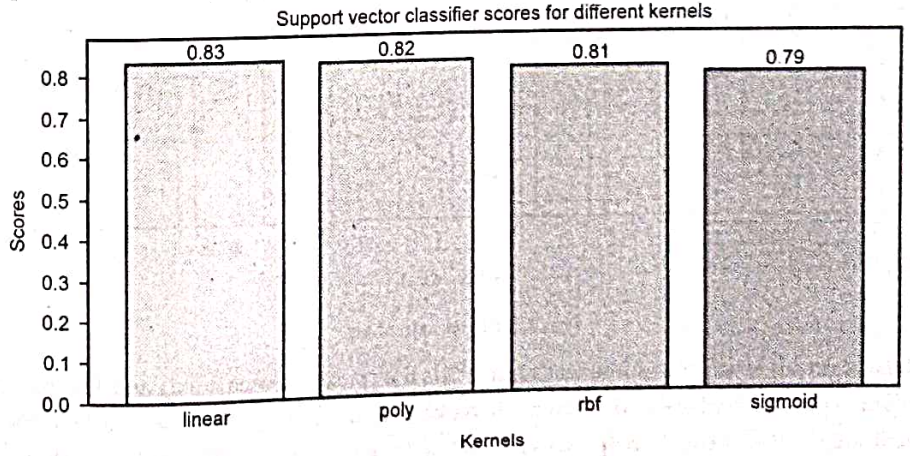


Fig. 3.5.1 Bar chart

2] Line charts : Basically a line charts are used to represent trends or progress of respective variable over time. It is suitable when input data is continuous. Fig. 3.5.2 represents line chart.

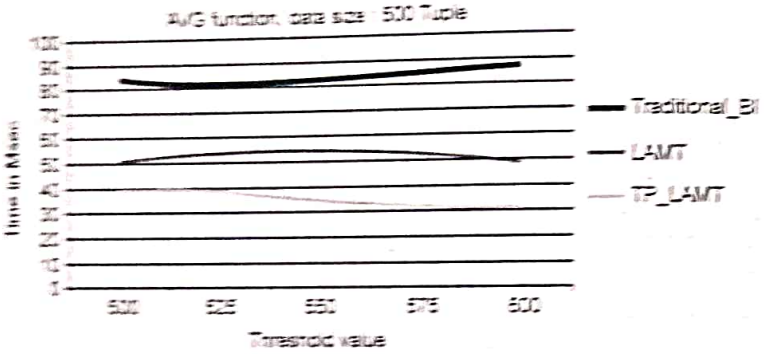


Fig. 3.5.2 Line chart

3] Dual axis chart : In dual axis chart data is plotted using one x axis and two y axis. It represents three data variables. One is a continuous set of data and the other is better suited to grouping by category. Fig. 3.5.3 represents the dual axis chart for placement statistics of any institute.

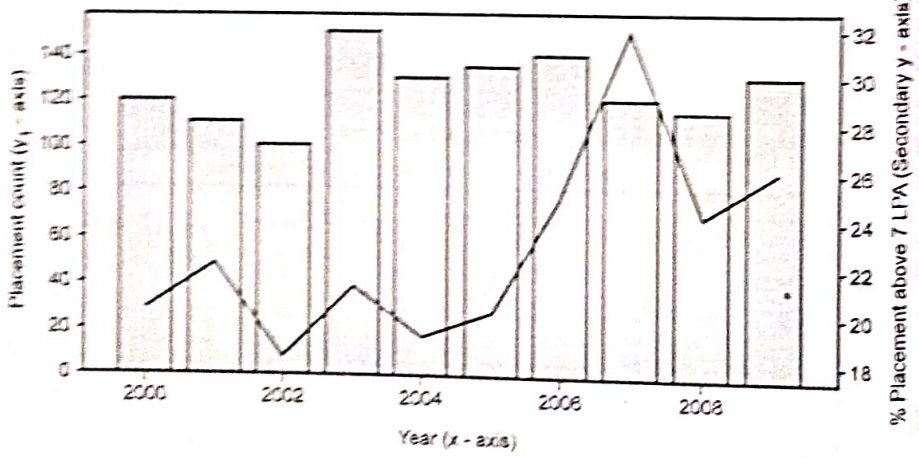


Fig. 3.5.3 Dual axis chart

4] Area chart : It is actually a line chart but it fills the space between x axis and line plot. It is filled with different color or pattern. It represents the individual as well as grouped contribution effectively. It helps to analyze both overall and individual contribution against total contribution. Fig. 3.5.4 represents the area chart.

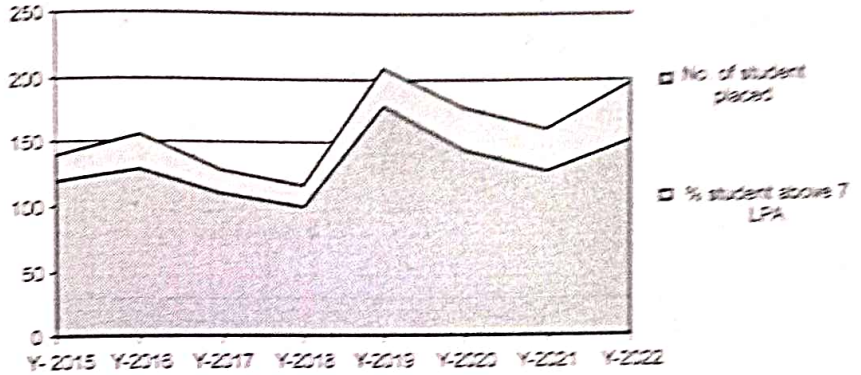


Fig. 3.5.4 Area chart

5] Mekko Chart : It is also known as Marimekko chart, this is typically used to compare measures, values, quantities and show data distribution across each one. It is similar to a stacked bar, but it also capture another dimension of our data values instead of only time like column charts. Mekko chart is used to show growth, market share, or competitor analysis. Below figure shows a representation of mekko chart for market basket analysis.

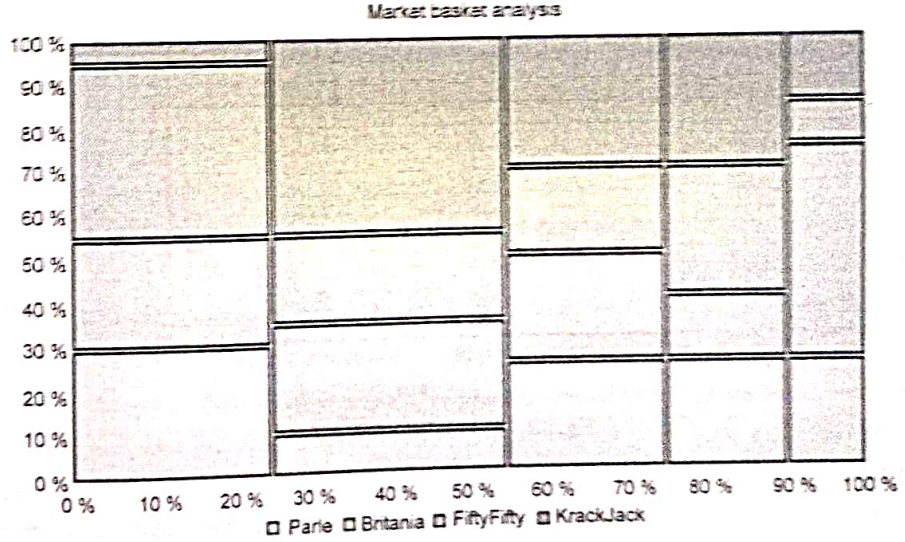


Fig. 3.5.5 Mekko chart

6] Pie chart : A pie chart represents the percentage of data distribution of any variable among the categories. It shows a static number and how categories represent part of a whole it is as shown in Fig. 3.5.6.

2] Line charts : Basically a line charts are used to represent trends or progress of respective variable over time. It is suitable when input data is continuous. Fig. 3.5.2, represents line chart.

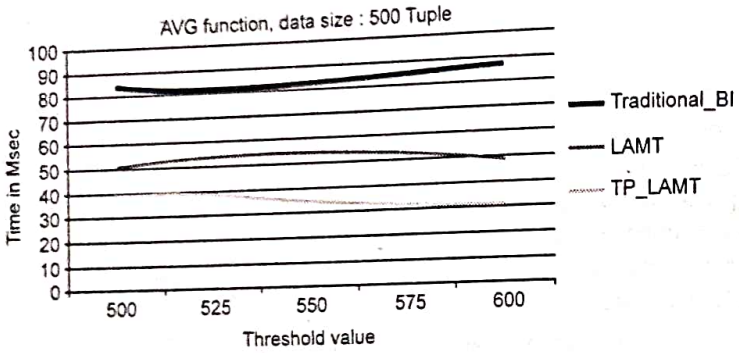


Fig. 3.5.2 Line chart

3] Dual axis chart : In dual axis chart data is plotted using one x axis and two y axis. It represents three data variables. One is a continuous set of data and the other is better suited to grouping by category. Fig. 3.5.3 represents the dual axis chart for placement statistics of any institute.

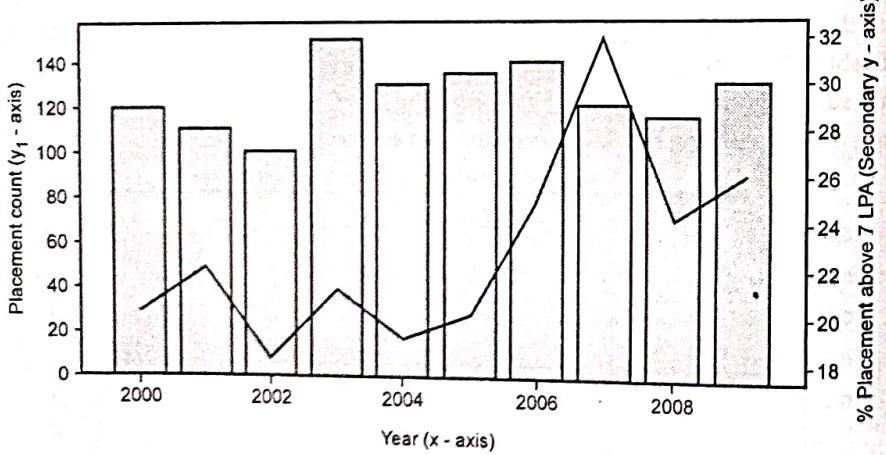


Fig. 3.5.3 Dual axis chart

4] Area chart : It is actually a line chart but it fills the space between x axis and line plot. It is filled with different color or pattern. It represents the individual as well as grouped contribution effectively. It helps to analyze both overall and individual contribution against total contribution. Fig. 3.5.4 represents the area chart.

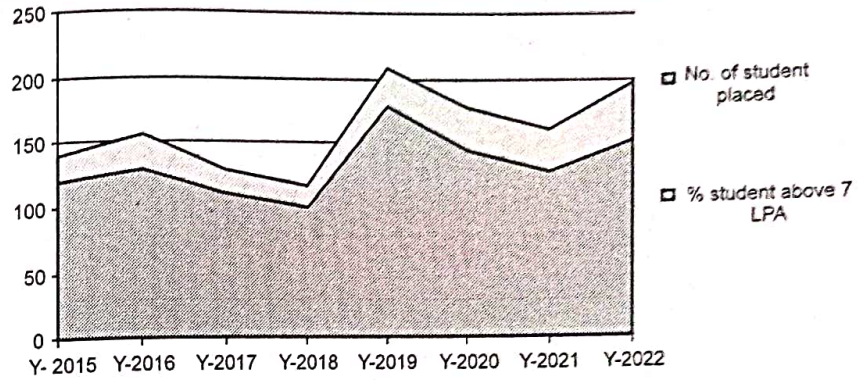


Fig. 3.5.4 Area chart

5] Mekko Chart : It is also known as Marimekko chart, this is typically used to compare measures, values, quantities and show data distribution across each one. It is similar to a stacked bar, but it also capture another dimension of our data values instead of only time like column charts. Mekko chart is used to show growth, market share, or competitor analysis: Below figure shows a representation of mekko chart for market basket analysis.

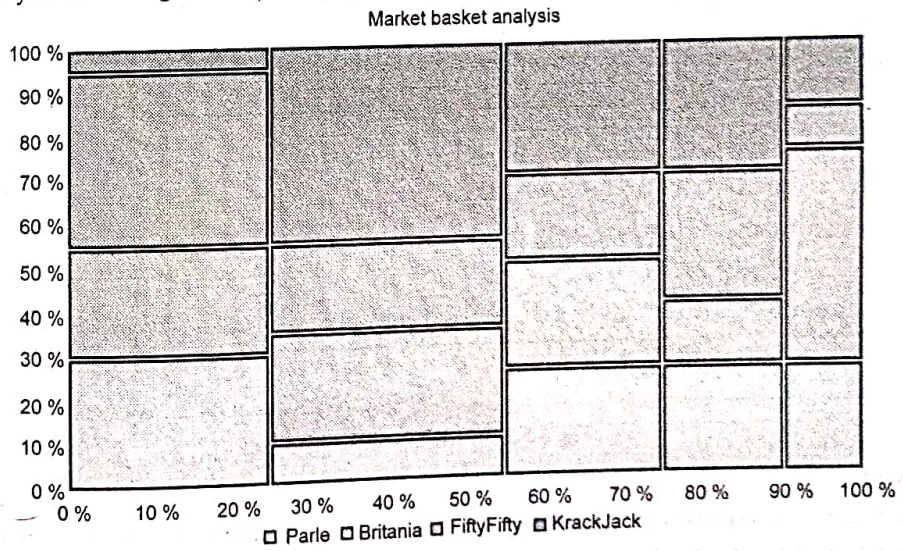


Fig. 3.5.5 Mekko chart

6] Pie chart : A pie chart represents the percentage of data distribution of any variable among the categories. It shows a static number and how categories represent part of a whole it is as shown in Fig. 3.5.6.

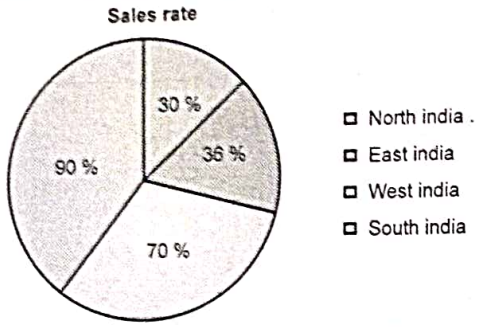


Fig. 3.5.6 Pie chart

7] Scatter plot chart : A scatter plot is also known as scatter gram chart. It is used to show the relationship between two different variables. It helps to reveal data distribution pattern. This chart is used when data contains many different data points distributed in wide range and we want to represent similarities in data distribution. Thus it is useful to find the insight from data like outlier, pattern and similarities etc. Following Fig. 3.5.7 represents the scatter plot.

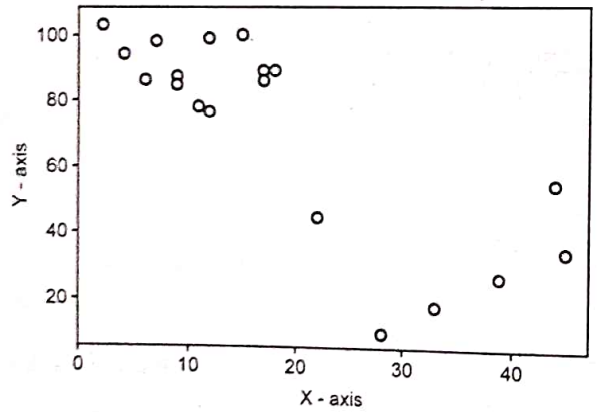


Fig. 3.5.7 Scatter plot / chart

8] Bubbled chart : It is similar like scatter plot which represents data distribution among two variable or relationship between two variables. Additionally in bubble chart third data variable is used to show the size of the bubble as per frequency of third variable. Bubble chart is represented as below in Fig. 3.5.8.

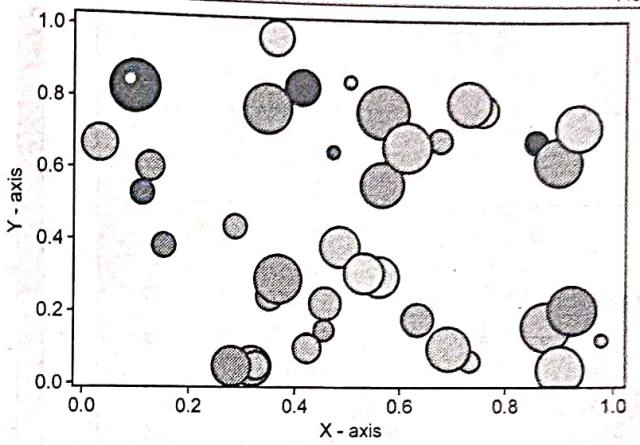


Fig. 3.5.8 Bubbled chart

3.5.5 Map

- It is another approach of data visualization which is used to analyze and represent the geographically correlated data and present it in the form of maps. This kind of data representation is clearer and more informative. Visual representation of data point's distribution is represented using map. It helps to identify the insights from data and to make proper decision.
- There are different types of maps used for data visualization. Different types of maps are heat map, point map, flow map, statistical maps, bubble map, regional map and administrative maps etc. Maps can be further divided into 2D, 3D, static maps, dynamic maps and interactive map. All these maps are often used in combination with line, point, bubble and dimensions.
- **Heat Map :** A heat map is used to represent the relationship between two data variables and provides quantity wise information, such as high, medium, low. It can be like poor to excellent. This chart displays these form of rating using different colours. Following Fig. 3.5.9 represents the head map.

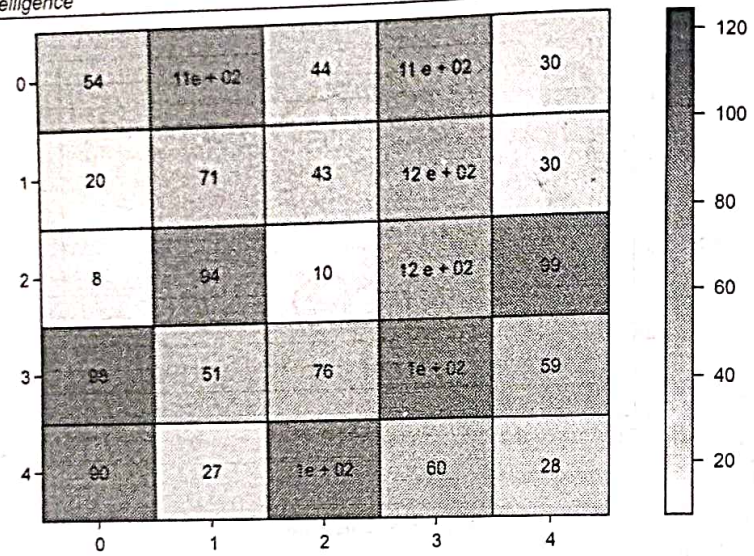


Fig. 3.5.9 Heat map

3.6 Operations on Data Reports

3.6.1 Data Grouping and Sorting

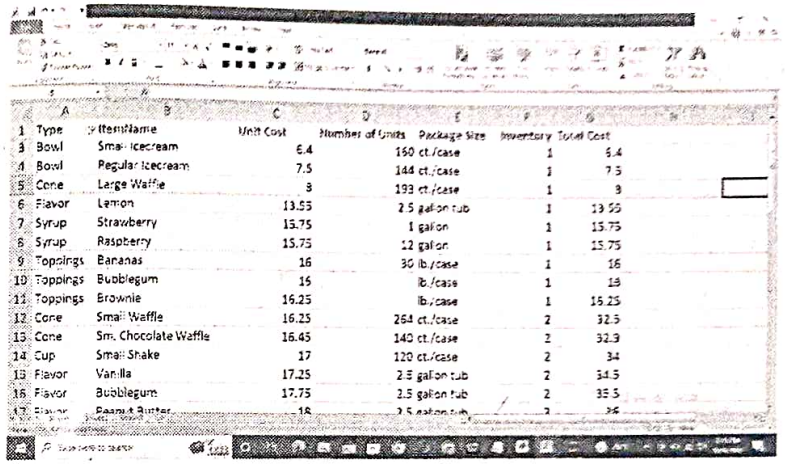
- One of the first step to become report builder or analyst is to understanding sorting and grouping concepts. Grouping and sorting helps to understand the data on which you are currently working and it allow to organize data better way.
- The purpose of grouping data is to arrange the data content of a report into various categories, while the main purpose of the sorting data is to arrange the data content into numerical or alphabetical order.
- Grouping is the process of breaking up rows of data into partitions that share common properties. Due to grouping of the document it become easy to read relevant data from report, it also removes the repeated data and summarizes the category and values as per calculation.
- Grouping is based on selected category and conditional parameters. As per the condition the data is summarized category wise.
- Following steps are followed while applying grouping and sorting.

Step 1 : Add sort operation on selected columns

Step 2 : Put all sorted data into groups

Step 3 : adding Sub-Groups

Example : Here we are using Excel tool to perform above operations. Following screenshot 3.6.1 shows the input data used in this demonstration.

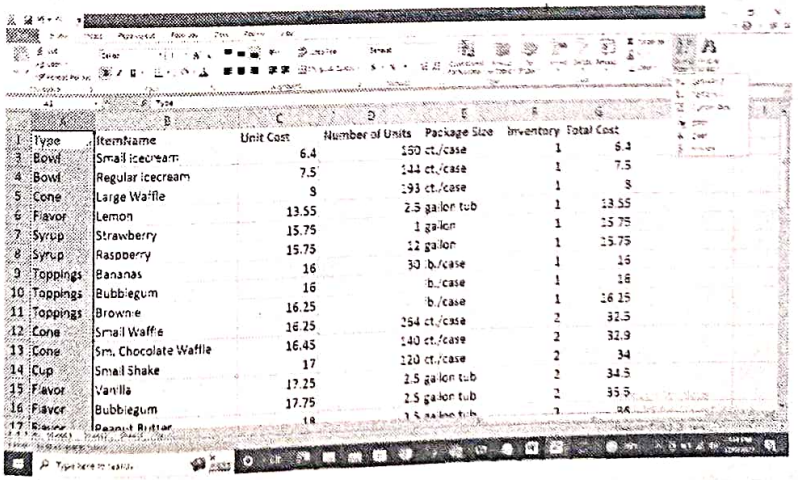


	A	B	C	D	E	F	G
1	Type	ItemName	Unit Cost	Number of Units	Package Size	Inventory	Total Cost
3	Bowl	Small Icecream	6.4		150 ct./case	1	6.4
4	Bowl	Regular Icecream	7.5		144 ct./case	1	7.5
5	Cone	Large Waffle	8		193 ct./case	1	8
6	Flavor	Lemon	13.55		2.5 gallon tub	1	13.55
7	Syrup	Strawberry	15.75		1 gallon	1	15.75
8	Syrup	Raspberry	15.75		12 gallon	1	15.75
9	Toppings	Bananas	16		30 lb./case	1	16
10	Toppings	Bubblegum	16		lb./case	1	16
11	Toppings	Brownie	16.25		lb./case	1	16.25
12	Cone	Small Waffle	16.25		264 ct./case	2	32.5
13	Cone	Sm. Chocolate Waffle	16.45		140 ct./case	2	32.9
14	Cup	Small Shake	17		120 ct./case	2	34
15	Flavor	Vanilla	17.25		2.5 gallon tub	2	34.5
16	Flavor	Bubblegum	17.75		2.5 gallon tub	2	35.5

Screenshot 3.6.1 Input data

Step 1 : Now we will perform sorting operation on it which will work in below sequence.

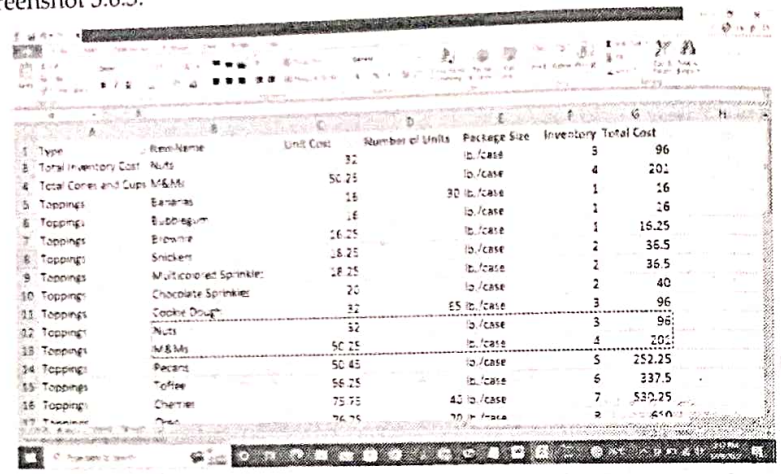
- Select a cell in the column which we want to sort, here column A is selected.
- Click the **Sort and Filter** command in the **Editing** group on the Home tab.
- Select **Sort A to Z.** Now the information in the Category column is organized in alphabetical order this is as shown in screenshot 3.6.2.
- Similarly we can sort data in descending order.
- It is applicable for both numerical and categorical data.



	A	B	C	D	E	F	G
1	Type	ItemName	Unit Cost	Number of Units	Package Size	Inventory	Total Cost
3	Bowl	Small Icecream	6.4		150 ct./case	1	6.4
4	Bowl	Regular Icecream	7.5		144 ct./case	1	7.5
5	Cone	Large Waffle	8		193 ct./case	1	8
6	Flavor	Lemon	13.55		2.5 gallon tub	1	13.55
7	Syrup	Strawberry	15.75		1 gallon	1	15.75
8	Syrup	Raspberry	15.75		12 gallon	1	15.75
9	Toppings	Bananas	16		30 lb./case	1	16
10	Toppings	Bubblegum	16		lb./case	1	16
11	Toppings	Brownie	16.25		lb./case	1	16.25
12	Cone	Small Waffle	16.25		264 ct./case	2	32.5
13	Cone	Sm. Chocolate Waffle	16.45		140 ct./case	2	32.9
14	Cup	Small Shake	17		120 ct./case	2	34
15	Flavor	Vanilla	17.25		2.5 gallon tub	2	34.5
16	Flavor	Bubblegum	17.75		2.5 gallon tub	2	35.5

Screenshot 3.6.2 Apply sort operation

- After applying sort function data set get sorted which is as shown in below screenshot 3.6.3.

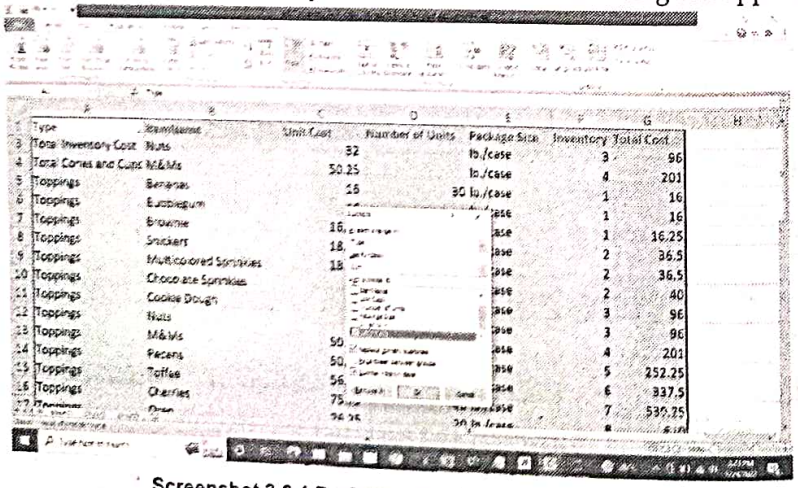


	Type	ItemName	Unit Cost	Number of Units	Package Size	Inventory	Total Cost
3	Total Inventory Cost	Nuts	32		lb./case	3	96
4	Total Cones and Cups	M&Ms	50.25		lb./case	4	201
5	Toppings	Bananas	16		30 lb./case	1	16
6	Toppings	Bubblegum	16		lb./case	1	16
7	Toppings	Brownie	16.25		lb./case	1	16.25
8	Toppings	Snickers	18.25		lb./case	2	36.5
9	Toppings	Multicolored Sprinkles	18.25		lb./case	2	36.5
10	Toppings	Chocolate Sprinkles	20		lb./case	2	40
11	Toppings	Cookie Dough	32		65 lb./case	3	96
12	Toppings	Nuts	32		lb./case	3	96
13	Toppings	M&Ms	50.25		lb./case	4	201
14	Toppings	Pecans	50.45		lb./case	5	252.25
15	Toppings	Toffee	56.25		lb./case	6	337.5
16	Toppings	Cherries	75.75		42 lb./case	7	530.25

Screenshot 3.6.3 Sorted data

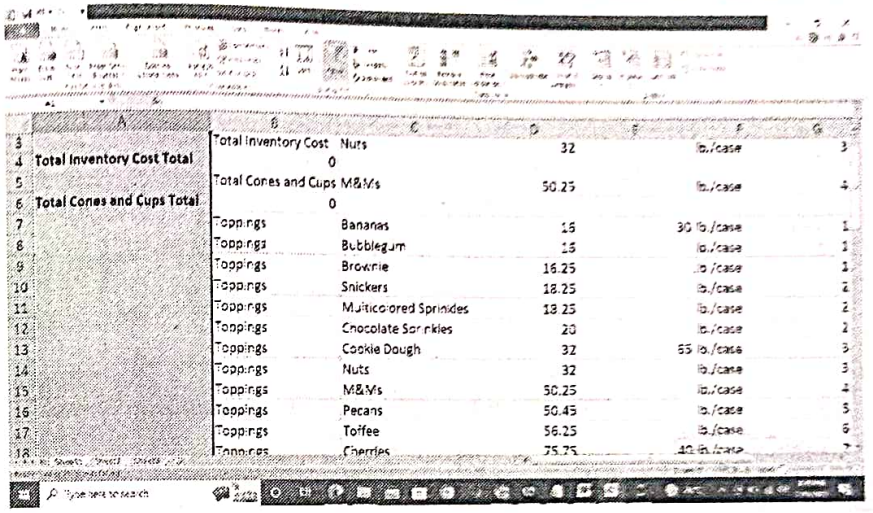
Step 2 : Grouping cells using the Subtotal command

- Grouping is a useful feature of excel which gives complete control on data to user regarding its representation. But before grouping we have to do sorting which we already did in step 1.
  - Select a column/cell on which we want to perform grouping. This is as shown in below screenshot 3.6.4.
  - Click the **Subtotal** command on the **Data** tab. The information in your spreadsheet is automatically selected and the Subtotal dialog box appears.



Screenshot 3.6.4 Perform grouping on sorted data

- After applying subtotal on Total inventory cost cell of column A, it will subtotal all rows contain Total inventory cost which is as shown in below screenshot 3.6.5.

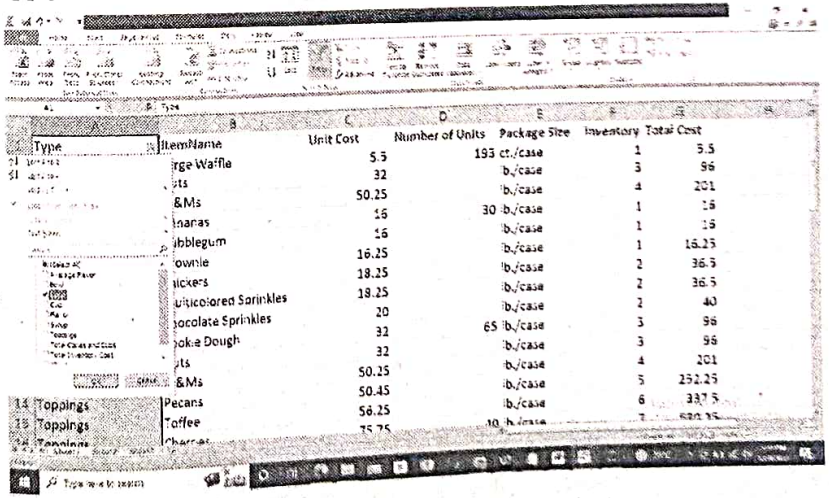


	A	B	C	D	E	F	G
3		Total Inventory Cost	Nuts	32		lb./case	3
4	Total Inventory Cost Total	0					
5		Total Cones and Cups	M&Ms	50.25		lb./case	4
6	Total Cones and Cups Total	0					
7		Toppings	Bananas	16	30	lb./case	1
8		Toppings	Bubblegum	16		lb./case	1
9		Toppings	Brownie	16.25		lb./case	1
10		Toppings	Snickers	18.25		lb./case	2
11		Toppings	Multicolored Sprinkles	18.25		lb./case	2
12		Toppings	Chocolate Sprinkles	20		lb./case	2
13		Toppings	Cookie Dough	32	65	lb./case	3
14		Toppings	Nuts	32		lb./case	3
15		Toppings	M&Ms	50.25		lb./case	4
16		Toppings	Pecans	50.45		lb./case	5
17		Toppings	Toffee	56.25		lb./case	6
18		Toppings	Cherries	75.75		42 lb./case	7

Screenshot 3.6.5 Data after grouping operation

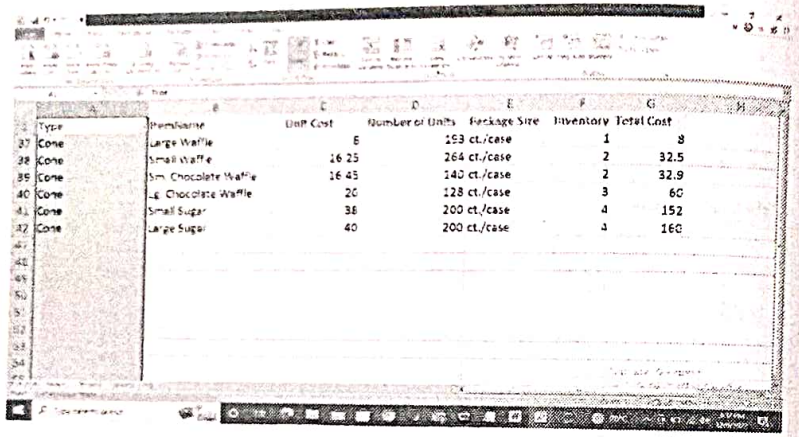
3.6.2 Filtering Reports

- Filtering is nothing but temporarily hiding data. This allows us to focus on specific spreadsheet data rows as per our requirement. In the following sequence we can apply data filtering.



Screenshot 3.6.6 Apply filtering on attributes

- Here filtering is applied on cone category from type as shown in screenshot 3.6.6. After this it will filter it out from others which is as shown below screenshot 3.6.7.



	A	B	C	D	E	F	G
1	Type	Item Name	Unit Cost	Number of Units	Package Size	Inventory	Total Cost
37	Cone	Large Waffle	8		193 ct./case	1	8
38	Cone	Small Waffle	16.25		264 ct./case	2	32.5
39	Cone	Sm. Chocolate Waffle	16.45		140 ct./case	2	32.9
40	Cone	Lg. Chocolate Waffle	20		128 ct./case	3	60
41	Cone	Small Sugar	38		200 ct./case	4	152
42	Cone	Large Sugar	40		200 ct./case	4	160

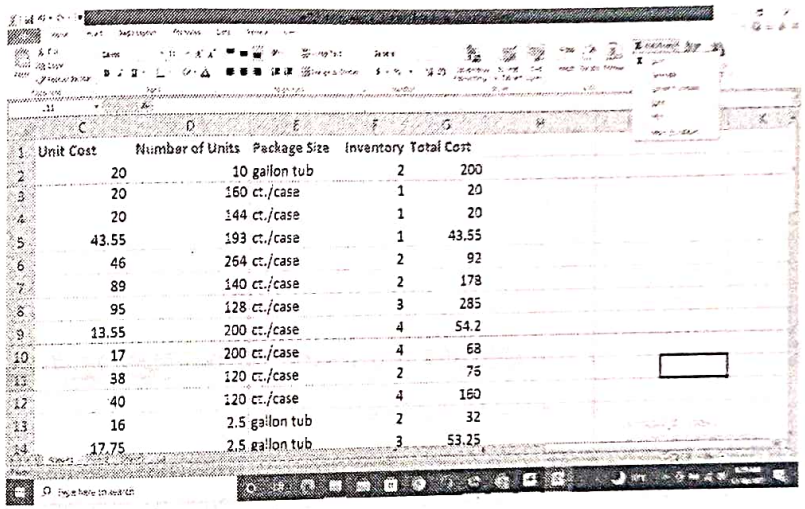
Screenshot 3.6.7 Result of filtering operation

3.6.3 Adding Calculations to Reports

- As we know reports are the combination of categorical and numerical data. To get the abstract, summarized or even detail view of available data we can add calculations or formulas. There are number of types of calculation that we can apply on any report as per our requirement. It help us to understand our data in better way. The calculations that we can perform are as below :

Calculation	Function	Description
Basic arithmetic operations	+, –, *, /, %	Addition, subtraction, multiplication, division etc.
SUM	Sum(Select cells/range)	The sum of all the numbers in the column.
AVG	AVG(Select cells /columns/range)	The average value of all the numbers in the column.
COUNT	Count(Select cells /columns/range)	The count of items in the column.
MAX	Max(Select cells /columns/range)	The highest numeric or alphabetic value in the column.
MIN	Min(Select cells /columns/range)	The lowest numeric or alphabetic value in the column.
Standard deviation	std(Select cells /columns/range)	An estimate of the standard deviation across the set of values in the column.
Variance	Var(Select cells /columns/range)	An estimate of the variance across the set of values in the column.

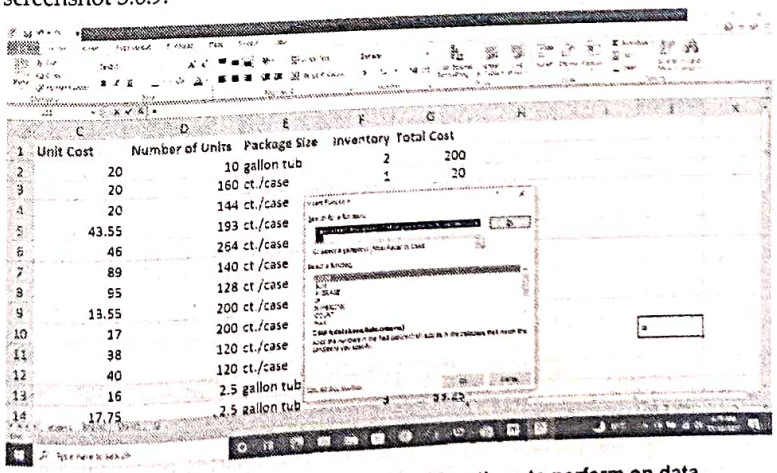
- Following are some simple steps followed in excel to apply calculations in available sheet.
- We can use AutoSum function in excel to quickly sum a column or row or numbers as per screenshot 3.6.8. Select a cell next to the numbers you want to sum, click AutoSum on the Home tab. When we select AutoSum in Excel it automatically enters a formula to sum the numbers.



	C	D	E	F	G
1	Unit Cost	Number of Units	Package Size	Inventory	Total Cost
2	20		10 gallon tub	2	200
3	20		160 ct./case	1	20
4	20		144 ct./case	1	20
5	43.55		193 ct./case	1	43.55
6	46		264 ct./case	2	92
7	89		140 ct./case	2	178
8	95		128 ct./case	3	285
9	13.55		200 ct./case	4	54.2
10	17		200 ct./case	4	68
11	38		120 ct./case	2	76
12	40		120 ct./case	4	160
13	16		2.5 gallon tub	2	32
14	17.75		2.5 gallon tub	3	53.25

Screenshot 3.6.8 Apply AutoSum on column

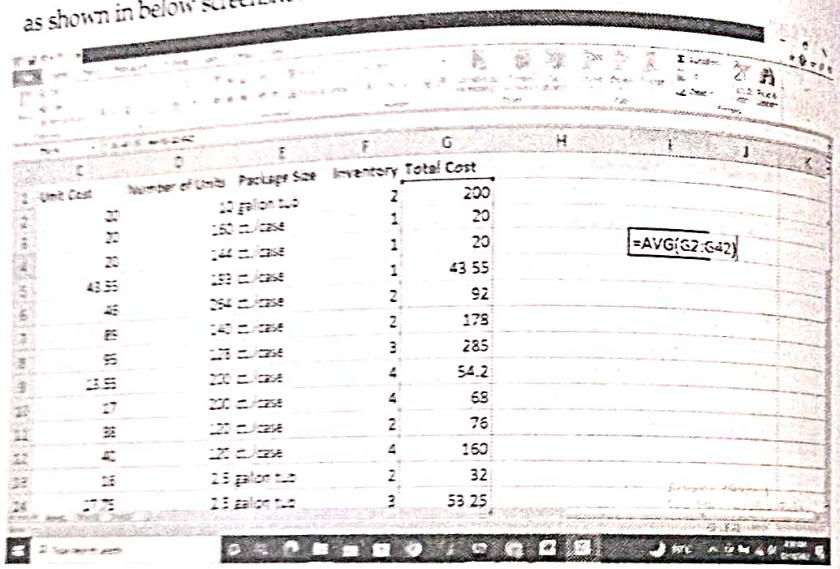
- Similarly there are more function tab is listed in AutoSum option. It gives more detail operations like AVG, COUNT, IF etc. This is as shown in below screenshot 3.6.9.



	C	D	E	F	G
1	Unit Cost	Number of Units	Package Size	Inventory	Total Cost
2	20		10 gallon tub	2	200
3	20		160 ct./case	1	20
4	20		144 ct./case		
5	43.55		193 ct./case		
6	46		264 ct./case		
7	89		140 ct /case		
8	95		128 ct /case		
9	13.55		200 ct /case		
10	17		200 ct./case		
11	38		120 ct./case		
12	40		120 ct./case		
13	16		2.5 gallon tub		
14	17.75		2.5 gallon tub		

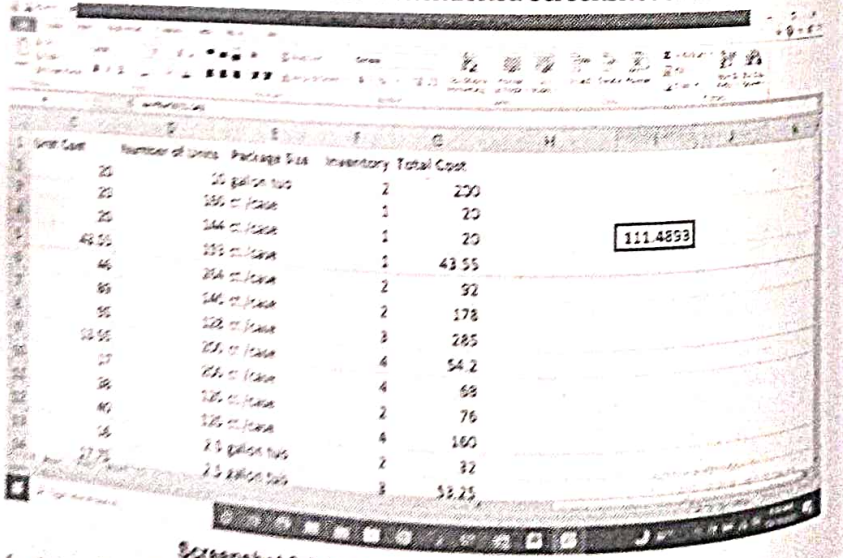
Screenshot 3.6.9 Different operations / functions to perform on data

- Another option is we can write formula instead of using AutoSum function. Thi as shown in below screenshot 3.6.10.



Screenshot 3.6.10 Writing formula in Excel

- Here we have to select any cell where we want the calculation results and p formula as shown in above screenshot 3.6.10. After entering we will get resul calculation which is as shown in below mentioned screenshot 3.6.11.

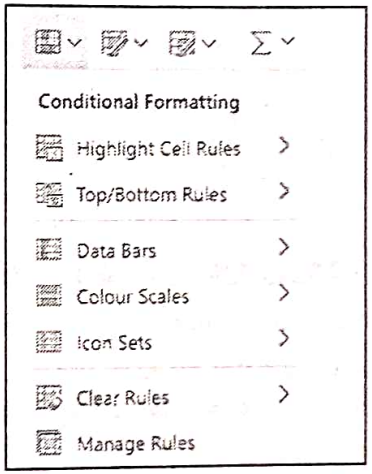


Screenshot 3.6.11 Result of formula in excel

- In this way we can perform any calculation in our report to understand it in be way.

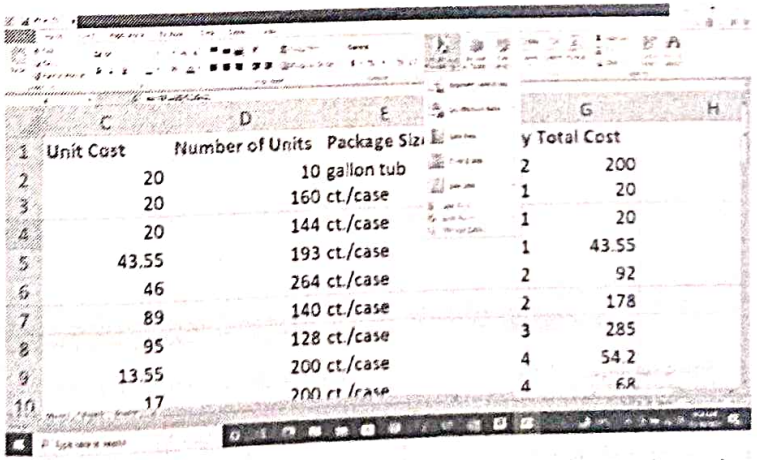
3.6.4 Conditional Formatting

- This is another interesting feature which helps analyst or user to extract the interesting data from reports. It basically works on changing the appearance of cells by highlighting them in different color or format. Conditional formatting is used to change the appearance of cell .These conditions are nothing but user specified rules like comparing with some numerical values, result of some formula and text matching. These many conditional formatting options are available in conditional formatting tab of excel which is as shown in screenshot 3.6.12.



Screenshot 3.6.12 Conditional formatting options in excel

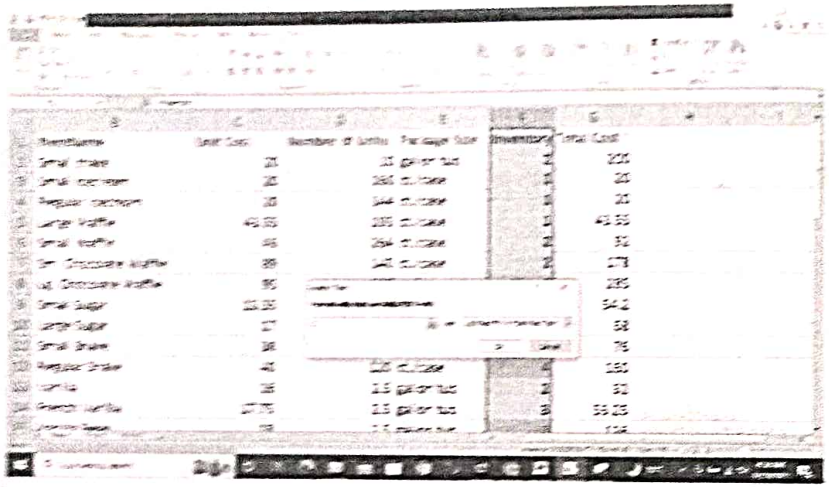
- If we select highlight cell rules then it will give us so many options which are as shown below screenshots 3.6.13.



	C	D	E	G	H
1	Unit Cost	Number of Units	Package Siz	y Total Cost	
2	20	10 gallon tub		2	200
3	20	160 ct./case		1	20
4	20	144 ct./case		1	20
5	43.55	193 ct./case		1	43.55
6	46	264 ct./case		2	92
7	89	140 ct./case		2	178
8	95	128 ct./case		3	285
9	13.55	200 ct./case		4	54.2
10	17	200 ct./case		4	68

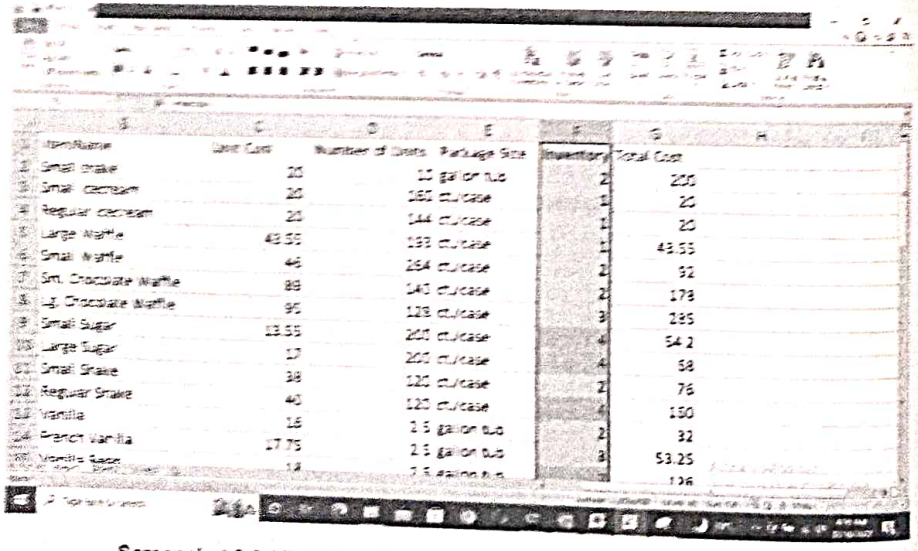
Screenshot 3.6.13 Selection of conditional formatting operation in excel

- Here we have selected F column and adding the condition as greater than 3 value. This is as shown in below screenshot 3.6.14.



Screenshot 3.6.14 Selection of Conditional formatting operation in excel

- Its result is as shown in below screenshot 3.6.15.



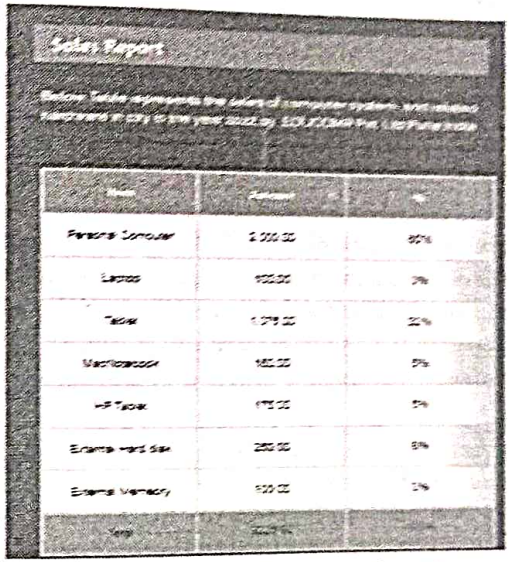
Item/Name	Unit Cost	Number of Units	Package Size	Inventory	Total Cost
Small Shake	20		10 gallon tub	2	200
Small Icecream	20		165 cs/case	1	20
Regular Icecream	20		144 cs/case	1	20
Large Waffle	43.55		193 cs/case	1	43.55
Small Waffle	46		264 cs/case	2	92
Sm. Chocolate Waffle	89		140 cs/case	2	178
Lg. Chocolate Waffle	95		128 cs/case	3	285
Small Sugar	13.55		200 cs/case	4	54.2
Large Sugar	17		200 cs/case	4	68
Small Shake	38		120 cs/case	2	76
Regular Shake	40		120 cs/case	4	160
Vanilla	16		2.5 gallon tub	2	32
French Vanilla	17.75		2.5 gallon tub	3	53.25
Vanilla Bean	18		2.5 gallon tub		126

Screenshot 3.6.15 Result of conditional formatting operation in excel

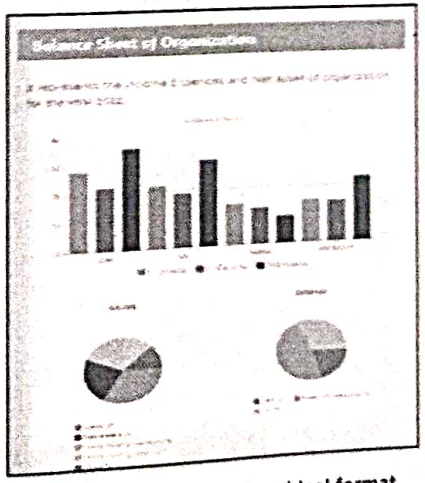
- In this way we can do conditional formatting by applying number of conditions and understand as well as represent our data more precisely.

3.6.5 Adding Summary Lines to Reports

- As discussed in above sections by making use of all these features we can prepare the summarized reports. It helps to extract the quick insights from dataset. It helps further analysis of business. There are number of tools available such as excel, PowerBI, Tableau, Power Query builder etc. Below are some screenshots of report generated in table formats as well as in graphical format as shown report 1 and report 2.



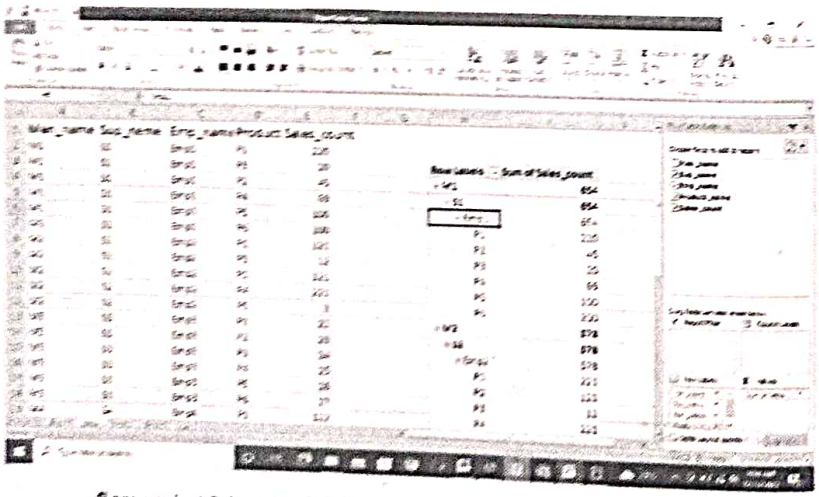
Report 3.6.1 Report in summary format



Report 3.6.2 Report in graphical format

3.6.6 Drill Up, Drill Down and Drill through Capabilities

- Drill up and drill down are used to explore the different aspects of business. It represents the level of variation of respective category either in up or down side. Drill up is like abstract view of data like yearly salary of employee and drill down is detailing of that dimension like monthly salary. Similarly we can go on expanding the details of dimensions like year, month, week, day, hour, minutes, seconds etc. For example, we can observe revenue for an entire product line and then drill down to see revenue for each individual product in the line.
- Steps used for using drill up, drill down and drill - through capabilities in excel are :
  1. Create pivot table or chart in excel.
  2. Specify the range of categories on which we want to perform drill up / down or other operations.
  3. It will show the pivot table as well as chart.
  4. As per the level of hierarchy we can drill down or drill up data. We can drill up multiple levels of a hierarchy at a time. Right-click the item we want to drill up on, click Drill Down/Drill Up and then pick the level we want to drill up to. If we have grouped items in our PivotTable, we can drill up to a group name.
  5. Finally the insights from data are extracted in summarized as well as in detail format.
- Below shown screenshot 3.6.16 represents the pivot table of main data spreadsheet. It shows the data as per hierarchy this is drill down. Similarly we can perform drill up from upper level and other capabilities as shown in sum of sales_count summary of screenshot 3.6.16.



Screenshot 3.6.16 Pivot table to represent drill up/down operation

3.7 Run, Schedule and Generate Reports

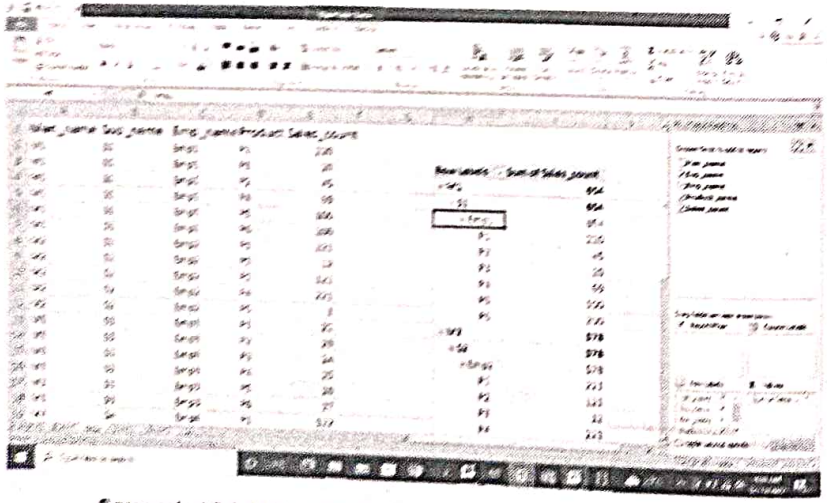
- A report represents the extracted information from data as per specified criteria. It organizes data into presentable and in an easy-to-read format. We can save it in any format like PDF, Excel, CSV and we can print it directly. As per the size and features of applications we can generate many number of reports as per requirement. For example if we are on the page of students details of any institute then for that link it can generate a report of students count ,merit and performance. After this if we switch on to the next page like placement then it will generate report for year wise placement count, package wise count, company wise count, highest package etc. In this way we can generate many reports throughout the application. The reports typically provide information related to the context of that page.
- We can schedule a report when we created it. We can choose the Schedule action and then we have to enter information such as save, print, data and time. The report is then added to the job queue and will be run at the specified time. When the report is processed, the item will be removed from the job queue.
- From the Job Queue Entries page, we are able to change some report parameters such as type of output file, run time, recurrence etc. Before editing an existing scheduled report, however, it's necessary to put the report job queue on hold :
  1. Choose the search icon, enter Job Queue Entries, then choose the related link.
  2. On the Job Queue Entries page, select the required report.
  3. Choose the Set on Hold action tab.
  4. Open and edit the scheduled report by selecting its status (On Hold).Edit the report.
  5. After editing the report options, repeat the first two steps and then select the Set Status to Ready action to resume generating the report.
- In this way run / generate report option work. Reports are generated in different formats such as PDF, Excel, CSV, XML etc. Selection of type of format to be select is generally depends upon what type of data to be represented. It also depends upon the customer requirement.

Review Questions

1. *What is relational model and how reports are generated in relational model ?*
2. *Explain with example queries how data is extracted from relational data model.*
3. *What is multidimensional data model and how reports are generated in multidimensional model ?*

3.6.6 Drill Up, Drill Down and Drill through Capabilities

- Drill up and drill down are used to explore the different aspects of business. It represents the level of variation of respective category either in up or down side. Drill up is like abstract view of data like yearly salary of employee and drill down is detailing of that dimension like monthly salary. Similarly we can go on expanding the details of dimensions like year, month, week, day, hour, minutes, seconds etc. For example, we can observe revenue for an entire product line and then drill down to see revenue for each individual product in the line.
- Steps used for using drill up, drill down and drill - through capabilities in excel are :
  1. Create pivot table or chart in excel.
  2. Specify the range of categories on which we want to perform drill up / down or other operations.
  3. It will show the pivot table as well as chart.
  4. As per the level of hierarchy we can drill down or drill up data. We can drill up multiple levels of a hierarchy at a time. Right-click the item we want to drill up on, click Drill Down/Drill Up and then pick the level we want to drill up to. If we have grouped items in our PivotTable, we can drill up to a group name.
  5. Finally the insights from data are extracted in summarized as well as in detail format.
- Below shown screenshot 3.6.16 represents the pivot table of main data spreadsheet. It shows the data as per hierarchy this is drill down. Similarly we can perform drill up from upper level and other capabilities as shown in sum of sales_count summary of screenshot 3.6.16.



Screenshot 3.6.16 Pivot table to represent drill up/down operation

3.7 Run, Schedule and Generate Reports

- A report represents the extracted information from data as per specified criteria. It organizes data into presentable and in an easy-to-read format. We can save it in any format like PDF, Excel, CSV and we can print it directly. As per the size and features of applications we can generate many number of reports as per requirement. For example if we are on the page of students details of any institute then for that link it can generate a report of students count ,merit and performance. After this if we switch on to the next page like placement then it will generate report for year wise placement count, package wise count, company wise count, highest package etc. In this way we can generate many reports throughout the application. The reports typically provide information related to the context of that page.
- We can schedule a report when we created it. We can choose the Schedule action and then we have to enter information such as save, print, data and time. The report is then added to the job queue and will be run at the specified time. When the report is processed, the item will be removed from the job queue.
- From the Job Queue Entries page, we are able to change some report parameters such as type of output file, run time, recurrence etc. Before editing an existing scheduled report, however, it's necessary to put the report job queue on hold :
  1. Choose the search icon, enter Job Queue Entries, then choose the related link.
  2. On the Job Queue Entries page, select the required report.
  3. Choose the Set on Hold action tab.
  4. Open and edit the scheduled report by selecting its status (On Hold).Edit the report.
  5. After editing the report options, repeat the first two steps and then select the Set Status to Ready action to resume generating the report.
- In this way run / generate report option work. Reports are generated in different formats such as PDF, Excel, CSV, XML etc. Selection of type of format to be select is generally depends upon what type of data to be represented. It also depends upon the customer requirement.

Review Questions

1. *What is relational model and how reports are generated in relational model ?*
2. *Explain with example queries how data is extracted from relational data model.*
3. *What is multidimensional data model and how reports are generated in multidimensional model ?*

4. *Demonstrate three dimensional data model with example and how it is represented in relational data model.*
5. *Differentiate between relational data model and multidimensional data model.*
6. *What is the concept of reporting and why it is important ?*
7. *What are the benefits of reporting ?*
8. *List out the different reporting methods.*
9. *What is the significance of cross tab explain with example ?*
10. *Justify with suitable example that list is not suitable for it but cross tab is best suited.*
11. *Describe statistics as a tool for report generation.*
12. *Explain different types of statics tool useful for report generation. Also explain which tool is suitable for which situations.*
13. *What is chart ? Explain different type of charts with example.*
14. *Explain with example the suitability of Mekko chart for particular problem.*
15. *Explain with example the suitability of bar/line/dual axis/area/ Mekko Chart/Pie/Scattered plot /Bubbled chart.*
16. *Differentiate between chart and graph.*
17. *What is the concept of map and how it helps for reporting explain with example.*
18. *What is the significance of heat map. Explain with suitable example and for which case heat map is suitable.*
19. *List out different operations on data report.*
20. *Why it is necessary to perform different operations on data report.*
21. *What is the significance of data grouping and sorting ?*
22. *When filtering reports are required ?*
23. *What is the significance of adding calculations to reports ?*
24. *What is the concept of conditional formatting explain with example.*
25. *How to perform drill up, drill down and drill through capabilities using excel or any other tool ?*
26. *What is the necessity of setting run and schedule report function ?*
27. *What are the different formats in which reports are generated ?*
28. *Explain for which case PDF/CSV/xls formats are suitable.*
29. *What are the different tools available to generate report ?*

❑❑❑

Unit IV

4 Data Preparation

Syllabus

Data validation : *Incomplete data, Data affected by noise.* ***Data transformation :*** *Standardization, Feature extraction.* ***Data reduction :*** *Sampling, Feature selection, Principal component analysis, Data discretization.* ***Data exploration : 1. Univarate analysis :*** *Graphical analysis of categorical attributes, Graphical analysis of numerical attributes, Measures of central tendency for numerical attributes, Measures of dispersion for numerical attributes, Identification of outliers for numerical attributes,* ***2. Bivariate analysis :*** *Graphical analysis, Measures of correlation for numerical attributes, Contingency tables for categorical attributes,* ***3. Multivariate analysis :*** *Graphical analysis, Measures of correlation for numerical attributes.*

Contents

4.1 Introduction to Data Preparation
4.2 Data Validation
4.3 Data Transformation
4.4 Data Reduction
4.5 Data Discretization
4.6 Data Exploration
4.7 Univariate Analysis
4.8 Bivariate Analysis
4.9 Multivariate Analysis

4.1 Introduction to Data Preparation

- Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and combining datasets to enrich data.

Introduction of Data Pre-processing

- Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Aim to reduce the data size, find the relation between data and normalized them.
- Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing.

Why Data Pre-processing ?

- Data which capture from various source is not pure. It contains some noise. It is called dirty data or incomplete data. In this data, there is lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. For example : occupation= " "
- Noisy data which contains errors or outliers. For example : Salary="-10"
- Inconsistent data which contains discrepancies in codes or names. For example : Age="51" Birthday="03/08/1998"
- Incomplete, noisy, and inconsistent data are commonplace properties of large real-world databases and data warehouses. Incomplete data can occur for a number of reasons.
- **Steps during pre-processing :**
  1. **Data cleaning :** Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
  2. **Data integration :** Data with different representations are put together and conflicts within the data are resolved.
  3. **Data transformation :** Data is normalized, aggregated and generalized.
  4. **Data reduction :** This step aims to present are duced representation of the data in a datawarehouse.
  5. **Data discretization :** Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals

4.2 Data Validation

- If collected input data is unsatisfactory due to incompleteness, noise and inconsistency, then it is not validated.

Data Cleaning

- Sometimes real-world data is incomplete, noisy, and inconsistent. Data cleaning methods are used for making useable data.
- Data cleaning tasks are as follows :
  1. Data acquisition and metadata
  2. Fill in missing values
  3. Unified date format
  4. Converting nominal to numeric
  5. Identify outliers and smooth out noisy data
  6. Correct inconsistent data
- Data cleaning is a first step in data pre-processing techniques which is used to find the missing value, smooth noise data, recognize outliers and correct inconsistent.

4.2.1 Incomplete Data

- To partially correct incomplete data one may adopt several techniques.
  a) Elimination : It is possible to discard all records for which the values of one or more attributes are missing.
  b) Inspection : One may opt for an inspection of each missing value, carried out by experts in the application domain.
  c) Identification : Conventional value might be used to encode and identify missing values, making it unnecessary to remove entire records from the given dataset.
  d) Substitution : Several criteria exist for the automatic replacement of missing data, although most of them appear somehow arbitrary.

4.2.1.1 Missing Value

These dirty data will affects on miming procedure and led to unreliable and poor output. Therefore it is important for some data cleaning routines.

How to handle noisy data in data mining ?

- Following methods are used for handling noisy data :
  1. **Ignore the tuple :** Usually done when the class label is missing. This method is not good unless the tuple contains several attributes with missing values.
  2. **Fill in the missing value manually :** It is time-consuming and not suitable for a large data set with many missing values.

3. **Use a global constant to fill in the missing value :** Replace all missing attribute values by the same constant.
4. **Use the attribute mean to fill in the missing value :** For example, suppose that the average salary of staff is Rs 65000/- . Use this value to replace the missing value for salary.
5. Use the attribute mean for all samples belonging to the same class as the given tuple
6. Use the most probable value to fill in the missing value

4.2.2 Data Affected by Noise

- **Noise :** Random error or variance in a measured variable
- For numeric values, box plots and scatter plots can be used to identify outliers. To deal with these anomalous values, data smoothing techniques are applied, which are described below.
  1. **Binning** : Using binning methods smooths sorted value by using the values around it. The sorted values are then divided into 'bins'. There are various approaches to binning. Two of them are smoothing by bin means where each bin is replaced by the mean of bin's values, and smoothing by bin medians where each bin is replaced by the median of bin's values.

Binning methods for data smoothing :

a) **In smoothing by bin means :** Each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 5, 9, and 13 in Bin is 9. Therefore, each original value in this bin is replaced by the value 9.

b) **Smoothing by bin medians** can be employed, in which each bin value is replaced by the bin median.

c) **Smoothing by bin boundaries :** The minimum and maximum bin values are stored at the boundary while intermediate bin values are replaced by the boundary value to which it is more closer.

2. **Regression :** Linear regression and multiple linear regression can be used to smooth the data, where the values are conformed to a function.
3. **Outlier analysis :** Approaches such as clustering can be used to detect outliers and deal with them.

4.3 Data Transformation

- In data transformation, the data are transformed or consolidated into forms appropriate for mining.
- Data transformation can involve the following :
  1. **Smoothing :** It removes noise from the data. Such techniques include binning, regression, and clustering.
  2. **Aggregation :** An aggregation or summary operation is applied to the data.
  3. **Generalization** of the data, where low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies.
  4. **Normalization :** The attribute data are scaled so as to fall within a small specified range.
  5. **Attribute construction :** New attributes are constructed and added from the given set of attributes to help the mining process.
- An attribute is normalized by scaling its values so that they fall within a small specified range. There are many methods for data normalization. They are min-max normalization, z-score normalization, and normalization by decimal scaling.
  a) Min-max normalization performs a linear transformation on the original data. It will scale the data between the 0 and 1.

Example :

Marks
8
10
15
20

Min : Minimum value of the given attribute. Here min is 8.

Max : Maxing value of the given attribute. Here max is 20.

V : V is the respective value of attribute.

For example :

$V_1 : 8$, $V_2 = 10$, $V_3 = 15$ and $V_4 = 20$.

New max : 1

Now min : 0

$$V' = \frac{V - Min_A}{Max_A - Min_A} (New - Max_A - New - Min_A) + New - Min_A$$

for mark 8 :

$$minmax = \frac{V - Min\ marks}{Max\ marks - Min\ marks} (New\ marks - New\ min) + New\ min$$

$$minmax = \frac{8-8}{20-8} \times (1-0) + 0$$

$$minmax = \frac{(0)}{12} \times 1$$

for mark 10 :

$$minmax = \frac{(10-8)}{20-8} \times (1-0) + 0$$

$$minmax = \frac{2}{12} \times 1$$

$$minmax = 0.25$$

for mark 15 :

$$minmax = \frac{(15-8)}{20-8} \times (1-0) + 0$$

$$minmax = \frac{(3)}{12} \times 1$$

$$minmax = 0.25$$

for mark 20 :

$$minmax = \frac{(20-8)}{20-8} \times (1-0) + 0$$

$$minmax = \frac{12}{12} \times 1$$

$$minmax = 1$$

Marks	Marks after min-max normalization
8	0
10	0.16
15	0.25
20	1

b) Decimal scaling : Decimal scaling is a data normalization technique. In this technique we move the decimal point of values of the attribute. His movement of decimal points totally depends on the maximum value among all values in the attribute.

Formula : A value V if attribute A can be obtained by normalization by the following formula.

Normalized value of attribute : $= (V^i / 10^j)$

Example :

CGPA	Formula	CGPA normalized after decimal scaling
2	2/10	0.2
3	3/10	0.3

We will check maximum value among our attribute CGPA. Here maximum value is 3 so we can convert it into decimal by dividing with 10.

Example 4.3.1 *1) Minimum salary is ₹ 20, 000 and maximum salary is ₹ 1,70,000 Map the salary ₹ 1, 00,000 in new range of ₹ (60, 000, 2,60,000) using min-max normalization method.*

2) If mean salary is ₹ 54,000 and standard deviation is ₹ 16,000 then find z score value of ₹ 73,600 salary.

Solution :

Solution 1 :

Old range = (20000, 1,70,000)

max = 1,70,000

min = 20000

New range = (60000, 260000)

new_max = 260000

new_min = 60000

V_i = 100000

$V_i' = [\{V_i - min)(max - min)\} \times \{new\_max - new\_min)\}] + new\_min$

$= [\{(80000/150000) \times 200000\}] + 60000$

$= [106666] + 60000 = 166666$

Salary ₹ 100000 in old range is equal to salary ₹ 166666 in the new range.

Solution 2 :

$$\text{mean} = ₹\ 54{,}000$$

$$\text{Standard deviation} = ₹\ 16{,}000$$

$$\text{Z-score value of } 76{,}300 = \frac{(76{,}300 - \text{mean})}{\text{Standard deviation}} = \frac{(54000)}{16000}$$

$$= \frac{54000}{16000}$$

$$= 3.375$$

Z-score value of ₹ 73,600 salary is 3.375

Example 4.3.2 *Use min-max normalization method to normalize the following group of data by setting min = 0 and max = 1, 200, 300, 400, 600, 1000.*

Solution : i) Min-max normalization by setting min = 0 and max = 1.

Original data	200	300	400	600	1000
0, 1 normalized	0	0.125	0.25	0.5	1

ii) Z-score normalization

Original data	200	300	400	600	1000
0, 1 normalized	– 1.06	– 0.7	– 0.35	0.35	1.78

Example 4.3.3 *Suppose that the minimum and maximum values for the attribute income are $ 73, 600 and $ 98,000, respectively. Normalize income value $ 73,600 to the range [0 : 0 ; 1 : 0] using min-max normalization method.*

Solution : i) The min-max normalization to transform value 73,600 onto the range [0.0, 1.0].

Given data : $\min_A = 12000$, $\max_A = 98000$, $\text{new\_min}_A = 0.0$,

$\text{new\_max}_A = 1.0$, $v = 73600$, $v' = ?$

$$v' = \frac{v - \min_A}{\max_A - \min_A} \times (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

$$v' = \frac{73600 - 12000}{98000 - 12000} \times (1.0 - 0.0) + 0.0$$

$$v' = 0.716$$

Example 4.3.4 *Consider the following group of data 200, 300, 400, 600, 1000.*

i) Use the min-max normalization to transform value 600 onto the range [0.0,1.0]

ii) Use the decimal scaling to transform value 600.

Solution : i) The min-max normalization to transform value 600 onto the range [0.0, 1.0].

Given data : $\min_A = 200$, $\max_A = 1000$, $\text{new\_min}_A = 0.0$, $\text{new\_max}_A = 1.0$, $v = 600$, $v' = ?$

$$v' = \frac{v - \min_A}{\max_A - \min_A} \times (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$

$$v' = \frac{600 - 200}{1000 - 200} \times (1.0 - 0.0) + 0.0$$

$$v' = 0.5$$

ii) Decimal scaling to transform value 600

$$v = 600, j = 3$$

$$v' = \frac{v}{10^j} = \frac{600}{10^3} = 0.6$$

4.4 Data Reduction

- Data reduction is nothing but obtaining a reduced representation of the data set that is much smaller in volume but yet produces the same analytical results.
- Data reduction is a process that reduced the volume of original data and represents it in a much smaller volume. Data reduction techniques ensure the integrity of data while reducing the data.
- Data reduction does not affect the result obtained from data mining that means the result obtained from data mining before data reduction and after data reduction is the same.
- Data reduction strategies are as follows :
  1. **Data cube aggregation :** Aggregation operations are applied to the data in the construction of a data cube.
  2. **Dimensionality reduction :** In dimensionality reduction redundant attributes are detected and removed which reduce the data set size.
  3. **Data compression :** Encoding mechanisms are used to reduce the data set size.
  4. **Numerosity reduction :** In numerosity reduction where the data are replaced or estimated by alternative.
  5. **Discretisation and concept hierarchy generation :** Where raw data values for attributes are replaced by ranges or higher conceptual levels.

- Data reduction techniques can be applied to obtain a reduced representation of data set that is much smaller in volume, yet closely maintains the integrity of original data. Mining on the reduced data set should be more efficient yet prod the same analytical results.
- Strategies for data reduction include the following :

1. **Data cube aggregation** : It is lowest level of a data cube. Summary aggregation operations are applied to the data. For example, the daily sa data may be aggregated so as to computer monthly and annual total amou This step is typically used in constructing a data cube for analysis of the data multiple granularities.
2. **Attribute subset selection** : Attribute subset selection reduces the data set si by removing irrelevant or redundant attributes. The goal of attribute subs selection is to find a minimum set of attributes such that the result probability distribution of the data classes is as close as possible to the origi distribution obtained using all attributes.

- **Irrelevant attributes :** It contains no information that is useful for the data minin task at hand. For example; Student's roll number is often irrelevant to the task predicting student marks or CGPA.
- **Redundant attributes :** Duplicate much or all of the information contained in one more other attributes. For example: purchase price of a product and the amount GST paid.

3. **Dimensionality reduction** : Dimensionality reduction is the process reducing the number of random variables under consideration, by obtaining set of principal variables.

- If the original data can be reconstructed from the compressed data without any los of information, the data reduction is called **lossless**. If, instead, we can reconstru only an approximation of the original data, then the data reduction is called lossy.
- Lossy dimensionality reduction methods are principal components analysis (PCA and wavelet transforms.
- Principal Component Analysis (PCA) is to reduce the dimensionality of a data s by finding a new set of variables, smaller than the original set of variables, retain most of the sample's information and useful for the compression and classificatio of data.

- In PCA, it is assumed that the information is carried in the variance of the features, that is, the higher the variation in a feature, the more information that feature carries.
- Hence, PCA employs a linear transformation that is based on preserving the most variance in the data using the least number of dimensions.
- A Discrete Wavelet Transform (DWT) is a transform that decomposes a given signal into a number of sets, where each set is a time series of coefficients describing the time evolution of the signal in the corresponding frequency band.

4. **Numerosity reduction :** The numerosity reduction reduces the volume of the original data and represents it in a much smaller form. This technique includes two types parametric and non-parametric numerosity reduction.

- Log-linear models is an example of parametric method and Nonparametric methods are histograms, clustering, and sampling.
- Regression and log-linear linear regression models a relationship between the two attributes by modelling a linear equation to the data set.
- Log-linear regression analysis involves using a dependent variable measured by frequency counts with categorical or continuous independent predictor variables

Clustering :

- Cluster analysis is a popular data discretization method. A clustering algorithm can be applied to discretize a numerical attribute (A) by partitioning the values of A into clusters or groups.
- Clustering takes the distribution of A into consideration, as well as the closeness of data points, and therefore is able to produce high-quality discretization results.
- Clustering can be used to generate a concept hierarchy for A by following either a top down splitting strategy or a bottom-up merging strategy, where each cluster forms a node of the concept hierarchy.

Sampling

- The basic idea of the sampling approach is to pick a random sample S of the given data D, and then search for frequent item sets in S instead of D.
- Sampling can be used as a data reduction technique because it allows a large data set to be represented by a much smaller random sample of the data. Different methods of sampling are Simple Random Sampling (SRS), Stratified Sampling, Cluster Sampling, Systematic Sampling and Multistage Sampling

4.4.1 Features Extraction

- A good feature representation is central to achieving high performance in any machine learning task.
- Consider an example of text categorization. Assume that we need to train a model for classifying a given document as spam and not spam. If we represent a document as a bag of words, the feature space consists of a vocabulary of all unique words present in all the documents in the training set.
- For a collection of 100,000 to 1,000,000 documents, we can easily expect hundreds of thousands of features. If we further extend this document model to include all possible bigrams and trigrams, we could easily get over a million features.
- A feature tree is a tree such that each internal node is labelled with a feature, and each edge emanating from an internal node is labelled with a literal. The set of literals at a node is called a split.
- Each leaf of the tree represents a logical expression, which is the conjunction of literals encountered on the path from the root of the tree to the leaf. The extension of that conjunction is called the instance space segment associated with the leaf.
- Two features are redundant if they are highly correlated, regardless of whether they are correlated with the task or not.

4.4.2 Feature Construction and Transformation

- Feature construction involves transforming a given set of input features to generate a new set of more powerful features which can then use for prediction.
- Feature construction methods may be applied to pursue two distinct goals: reducing data dimensionality and improving prediction performance.
- **Steps :**
  1. Start with an initial feature space F_0
  2. Transform F_0 to construct a new feature space F_N.
  3. Select a subset of features F_i from F_N.
  4. If some terminating criteria is achieved : Go back to step 3 otherwise set $F_T = F_i$
  5. F_T is the newly constructed feature space
- The initial feature space F_0 consists of manually constructed features that often encode some basic domain knowledge.

- The task of constructing appropriate features is often highly application specific and labour intensive. Thus, building auto-mated feature construction methods that require minimal user effort is challenging. In particular we want methods that :
  1. Generate a set of features that help improve prediction accuracy.
  2. Are computationally efficient.
  3. Are generalizable to different classifiers.
  4. Allow for easy addition of domain knowledge.
- Genetic programming is an evolutionary algorithm-based technique that starts with a population of individuals, evaluates them based on some fitness function and constructs a new population by applying a set of mutation and crossover operators on high scoring individuals and eliminating the low scoring ones.
- In the feature construction paradigm, genetic programming is used to derive a new feature set from the original one.
- Individuals are often tree like representations of features, the fitness function is usually based on the prediction performance of the classifier trained on these features while the operators can be applications specific.
- The method essentially performs a search in the new feature space and helps generate a high performing subset of features. The newly generated features may often be more comprehensible and intuitive than the original feature set, which makes GP-related methods well-suited for such tasks.
- In decision trees, the model explicitly selects features that are highly correlated with the label. In particular, by limiting the depth of the decision tree, one can at least hope that the model will be able to throw away irrelevant features.
- In the case of K-nearest neighbours, the situation is perhaps more terrible. Since KNN weighs each feature just as much as another feature, the introduction of irrelevant features can completely mess up KNN prediction.
- Feature extraction is a process that extracts a set of new features from the original features through some functional mapping.
- Transformation studies ways of mapping original attributes to new features. Different mappings can be employed to extract features.
- In general the mappings can be categorized into linear or nonlinear transformations. One could categorize transformations along two dimensions linear and labeled, linear and non labeled, nonlinear and labeled, nonlinear and non labeled.

4.4.3 Feature Selection

- Feature selection is a process that chooses a subset of features from the original features so that the feature space is optimally reduced according to a certain criterion.
- Feature selection is a critical step in the feature construction process. In text categorization problems, some words simply do not appear very often.
- Perhaps the word "groovy" appears in exactly one training document, which is positive. Is it really worth keeping this word around as a feature? It's a dangerous endeavour because it's hard to tell with just one training example if it is really correlated with the positive class, or is it just noise.
- You could hope that your learning algorithm is smart enough to figure it out. Or you could just remove it.
- There are three general classes of feature selection algorithms: filter methods, wrapper methods and embedded methods.
- The role of feature selection is as follows :
  1. To reduce the dimensionality of feature space
  2. To speed up a learning algorithm
  3. To improve the predictive accuracy of a classification algorithm
  4. To improve the comprehensibility of the learning results
- Features selection algorithms are as follows :
  1. **Instance based approaches :** There is no explicit procedure for feature subsetgeneration. Many small data samples are sampled from the data. Features are weighted according to their roles in differentiating instances of different classes for a data sample. Features with higher weights can be selected.
  2. **Nondeterministic approaches :** Genetic algorithms and simulated annealing are also used in feature selection.
  3. **Exhaustive complete approaches :** Branch and Bound evaluates estimated accuracy and ABB checks an inconsistency measure that is monotonic. Both start with a full feature set until the pre-set bound cannot be maintained.
- Feature selection methods can be classified into three main categories : Filter methods, Wrapper methods and Embedded methods.
  1. Filter method : In this method, features are filtered based on general characteristics of the dataset such as correlation with the dependent variable.

Filter method is performed without any predictive model. It is faster and usually the better approach when the number of features are huge.

2. Wrapper method : In wrapper method, the feature selection algorithm exits as a wrapper around the predictive model algorithm and uses the same model to select best features.
3. Embedded method : In embedded method, feature selection process is embedded in the learning or the model building phase. It is less computationally expensive than wrapper method and less prone to overfitting.

4.4.4 Difference between Filter, Wrapper and Embedded Methods

Filter methods	Wrapper methods	Embedded methods
Generic set of methods which do not incorporate a **specific machine learning algorithm.**	Evaluates on a specific **machine learning algorithm** to find optimal features.	Embeds (fix) features during **model building process.** Feature selection is done by observing each iteration of model training phase.
Much **faster** compared to Wrapper methods in terms of time complexity.	**High computation time** for a dataset with many features.	Sits **between Filter methods and Wrapper methods** in terms of time complexity.
Less prone to **over-fitting.**	High chances of **over-fitting** because it involves training of machine learning models with different combination of features.	Generally used to reduce **over-fitting** by **penalizing** the coefficients of a model being too large.
Examples - Correlation, Chi-square test, ANOVA, Information gain, etc.	Examples - **Forward selection, Backward elimination, Stepwise selection, etc.**	Examples - **LASSO, Elastic net, Ridge regression etc.**

4.4.5 Principal Component Analysis

- This method was introduced by Karl Pearson. It works on a condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be maximum.
- Principal Component Analysis (PCA) is to reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables, retains

most of the sample's information and useful for the compression and classification of data.

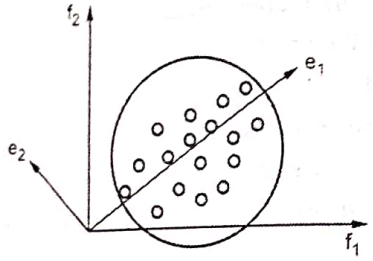


Fig. 4.4.1 PCA

- In PCA, it is assumed that the information is carried in the variance of the features, that is, the higher the variation in a feature, the more information that feature carries.
- Hence, PCA employs a linear transformation that is based on preserving the most variance in the data using the least number of dimensions.
- It involves the following steps :
  1. Construct the covariance matrix of the data.
  2. Compute the eigenvectors of this matrix.
  3. Eigenvectors corresponding to the largest Eigen values are used to reconstruct a large fraction of variance of the original data.
- The data instances are projected onto a lower dimensional space where the new features best represent the entire data in the least squares sense.
- It can be shown that the optimal approximation, in the least square error sense, of a d-dimensional random vector $x_2 < d$ by a linear combination of independent vectors is obtained by projecting the vector x onto the eigenvectors e_i corresponding to the largest eigen values l_i of the covariance matrix (or the scatter matrix) of the data from which x is drawn.
- The eigenvectors of the covariance matrix of the data are referred to as principal axes of the data and the projection of the data instances on to these principal axes are called the principal components. Dimensionality reduction is then obtained by only retaining those axes (dimensions) that account for most of the variance and discarding all others.

- In the Fig. 4.4.1, Principal axes are along the eigenvectors of the covariance matrix of the data. There are two principal axes shown in the figure, first one is closed to origin, the other is far from origin.
- If $X = X_1, X_2 \ldots X_N$ is the set of n patterns of dimension d, the sample mean of the data set is

$$m = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- The sample covariance matrix is

$$C = (X - m)(X - m)^T$$

- C is a symmetric matrix. The orthogonal basis can be calculated by finding the eigenvalues and eigenvectors.
- The eigenvectors g_i and the corresponding eigenvalues λ_i are solutions of the equation

$$C * g_i = \lambda_i * g_i \quad i = 1, \ldots, d$$

- The eigenvector corresponding to the largest eigenvalue gives the direction of the largest variance of the data. By ordering the eigenvectors according to the eigenvalues, the directions along which there is maximum variance can be found.
- If E is the matrix consisting of eigenvectors as row vectors, we can transform the data X to get Y.

$$Y = E(X - m)$$

The original data X can be got from Y as follows :

$$X = E^t Y + m$$

- Instead of using all d eigenvectors, the data can be represented by using the first k eigenvectors where $k < d$.
- If only the first k eigenvectors are used represented by E_K, then

$$Y = E_K (X - m) \quad \text{and} \quad X' = E_k^t Y + m$$

4.5 Data Discretization

Data Discretization :

- Data discretization means diving the range of continuous attribute into intervals. Actual data values are replaced by interval labels.

most of the sample's information and useful for the compression and classification of data.

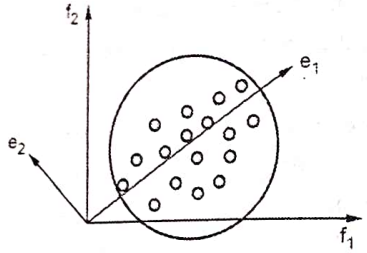


Fig. 4.4.1 PCA

- In PCA, it is assumed that the information is carried in the variance of the features, that is, the higher the variation in a feature, the more information that feature carries.
- Hence, PCA employs a linear transformation that is based on preserving the most variance in the data using the least number of dimensions.
- It involves the following steps :
  1. Construct the covariance matrix of the data.
  2. Compute the eigenvectors of this matrix.
  3. Eigenvectors corresponding to the largest Eigen values are used to reconstruct a large fraction of variance of the original data.
- The data instances are projected onto a lower dimensional space where the new features best represent the entire data in the least squares sense.
- It can be shown that the optimal approximation, in the least square error sense, of a d-dimensional random vector $x_2 < d$ by a linear combination of independent vectors is obtained by projecting the vector x onto the eigenvectors e_i corresponding to the largest eigen values l_i of the covariance matrix (or the scatter matrix) of the data from which x is drawn.
- The eigenvectors of the covariance matrix of the data are referred to as principal axes of the data and the projection of the data instances on to these principal axes are called the principal components. Dimensionality reduction is then obtained by only retaining those axes (dimensions) that account for most of the variance and discarding all others.

- In the Fig. 4.4.1, Principal axes are along the eigenvectors of the covariance matrix of the data. There are two principal axes shown in the figure, first one is closed to origin, the other is far from origin.
- If $X = X_1, X_2 \ldots X_N$ is the set of n patterns of dimension d, the sample mean of the data set is

$$m = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- The sample covariance matrix is

$$C = (X - m)(X - m)^T$$

- C is a symmetric matrix. The orthogonal basis can be calculated by finding the eigenvalues and eigenvectors.
- The eigenvectors g_i and the corresponding eigenvalues λ_i are solutions of the equation

$$C * g_i = \lambda_i * g_i \quad i = 1, \ldots, d$$

- The eigenvector corresponding to the largest eigenvalue gives the direction of the largest variance of the data. By ordering the eigenvectors according to the eigenvalues, the directions along which there is maximum variance can be found.
- If E is the matrix consisting of eigenvectors as row vectors, we can transform the data X to get Y.

$$Y = E(X - m)$$

The original data X can be got from Y as follows :

$$X = E^t Y + m$$

- Instead of using all d eigenvectors, the data can be represented by using the first k eigenvectors where $k < d$.
- If only the first k eigenvectors are used represented by E_K, then

$$Y = E_K (X - m) \quad \text{and} \quad X' = E_k^t Y + m$$

4.5 Data Discretization

Data Discretization :

- Data discretization means diving the range of continuous attribute into intervals. Actual data values are replaced by interval labels.

- It reduces the number of values for a given continuous attribute. Some classification algorithms only accept categorical attributes. It helps to a concise, easy-to-use, knowledge-level representation of mining results.
- Data discretization techniques can be categorized based on class information and which direction it proceeds. Class information is divided into two types: supervised and unsupervised discretization. Categorized based on which direction it proceeds are of two types : top-down and bottom-up.
- Discretization techniques can be classified as supervised and unsupervised discretization. Supervised discretization uses class information and unsupervised discretization does not use class information.
- **Top-down** : If the process starts by first finding one or a few points to split the entire attribute range, and then repeats this recursively on the resulting intervals. It is also called splitting.
- **Bottom-up** : It starts by considering all of the continuous values as potential split-points, removes some by merging neighbourhood values to form intervals, and then recursively applies this process to the resulting intervals. It is also called merging.
- Discretization can be performed recursively on an attribute to provide a hierarchical or multiresolution partitioning of the attribute values, known as a concept hierarchy. Concept hierarchies are useful for mining at multiple levels of abstraction.
- Discretization and concept hierarchy generation for numerical data uses following methods :

a) Binning b) Histogram analysis
c) Clustering analysis d) Entropy-based discretization
e) Segmentation by natural partitioning

4.5.1 Concept Hierarchy Generation for Categorical Data

- Categorical data are discrete data. Categorical attributes have a finite number of distinct values, with no ordering among the values.
- Example : geographic location, job category, and item type.
- Various methods are used for the generation of concept hierarchies for categorical data :

a) Specification of a partial ordering of attributes explicitly at the schema level by users or experts

- Example : A relational database or a dimension location of a data warehouse may contain the following group of attributes: street, city, province or state, and country.
- A user or expert can easily define a concept hierarchy by specifying ordering of the attributes at the schema level.
- A hierarchy can be defined by specifying the total ordering among these attributes at the schema level, such as: *street < city < province or state < country*

b) Specification of a portion of a hierarchy by explicit data grouping

- We can easily specify explicit groupings for a small portion of intermediate-level data.
- For example, after specifying that area and country form a hierarchy at the schema level, a user could define some intermediate levels manually, such as: {India, Maharashtra, Pune} < SPPU.

c) Specification of a set of attributes, but not of their partial ordering

- A user may specify a set of attributes forming a concept hierarchy, but omit to explicitly state their partial ordering.
- The system can then try to automatically generate the attribute ordering so as to construct a meaningful concept.
- Example : Suppose a user selects a set of location-oriented attributes, street, country, state, and city, from the database, but does not specify the hierarchical ordering among the attributes.
- Fig. 4.5.1 shows automatic generation of a schema concept hierarchy based on the number of distinct attribute values.

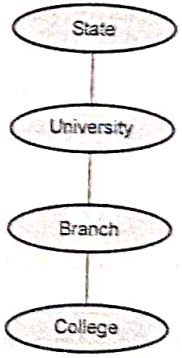


Fig. 4.5.1 : Automatic generation of a schema concept hierarchy

4.6 Data Exploration

- Exploratory Data Analysis (EDA) is a general approach to exploring datasets by means of simple summary statistics and graphic visualizations in order to gain a deeper understanding of data.

- EDA is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers user need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis or check assumptions.
- With EDA, following functions are performed :
  1. Describe of user data
  2. Closely explore data distributions
  3. Understand the relations between variables
  4. Notice unusual or unexpected situations
  5. Place the data into groups
  6. Notice unexpected patterns within groups
  7. Take note of group differences
- Exploratory data analysis is majorly performed using the following methods :
  1. Univariate analysis : Provides summary statistics for each field in the raw data set (or) summary only on one variable. Ex : CDF, PDF, Box plot
  2. Bivariate analysis is performed to find the relationship between each variable in the dataset and the target variable of interest (or) using two variables and finding relationship between them. Ex : Boxplot, Violin plot.
  3. Multivariate analysis is performed to understand interactions between different fields in the dataset (or) finding interactions between variables more than 2.

4.7 Univariate Analysis

- Univariate analysis is used to study the behavior of each attribute, considered as an entity independent of the other variables of the dataset

4.7.1 Graphical Analysis of Categorical Attributes

- Categorical attribute is one that has a specific value from a limited selection of values. The number of values is usually fixed.
- Categorical features can only take on a limited and usually fixed, number of possible values. For example, if a dataset is about information related to users, then we will typically find features like country, gender, age group, etc. Alternatively, if the data we are working with is related to products, we will find features like product type, manufacturer, seller and so on.

- The most natural representation for the graphical analysis of a categorical attribute is a vertical bar chart, which indicates along the vertical axis or ordinate the empirical frequencies, that is the number of observations of the dataset corresponding to each of the values assumed by the attribute.
- A bar chart is a way of summarizing a set of categorical data. The bar chart displays data using a number of bars, each representing a particular category. The height of each bar is proportional to a specific aggregation.
- Fig. 4.7.1 shows vertical and horizontal bar chart for a categorical attribute

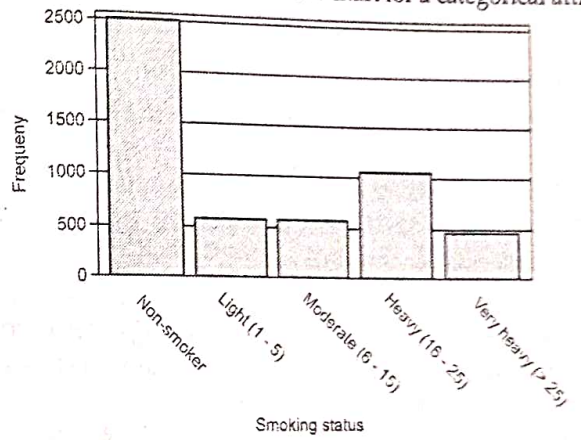


Fig. 4.7.1 Vertical bar chart

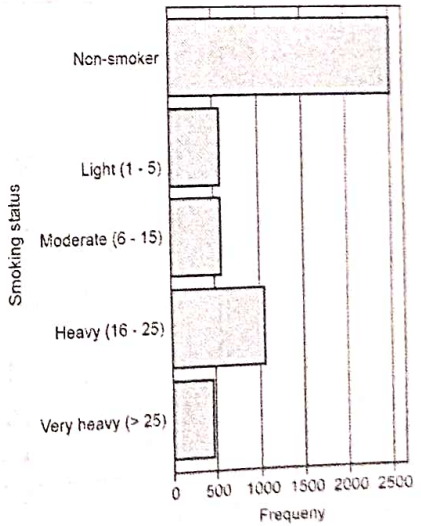


Fig. 4.7.2 Horizontal bar chart

- The categories could be something like an age group or a geographical location. It is also possible to color or split each bar into another categorical column in the data which enables you to see the contribution from different categories to each bar or group of bars in the bar chart.
- Advantages to horizontal bar charts
  1. Long labels for the categories are easier to display and read.
  2. Many categories are easier to display.
  3. Labels for many bars are easier to display without collision.

4.7.2 Graphical Analysis of Numerical Attributes

- For discrete numerical attributes assuming a finite and limited number of values, it is possible to resort to a bar chart representation, just as in the case of categorical attributes. In the presence of continuous or discrete attributes that might assume infinite distinct values, this type of representation cannot be used, as it would require an infinite number of vertical bars.
- In a histogram, the data are grouped into ranges (e.g. 10 – 19, 20 – 29) and then plotted as connected bars. Each bar represents a range of data. The width of each bar is proportional to the width of each category and the height is proportional to the frequency or percentage of that category.
- Fig. 4.7.3 shows histogram.

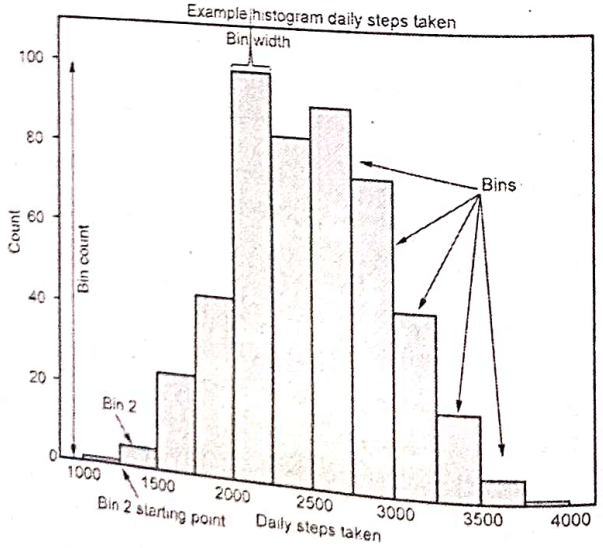


Fig. 4.7.3 Histogram

- It provides a visual interpretation of numerical data by showing the number of data points that fall within a specified range of values called "bins".
- Histograms can display a large amount of data and the frequency of the data values. The median and distribution of the data can be determined by a histogram. In addition, it can show any outliers or gaps in the data

4.7.3 Measuring the Central Tendency

- We look at various ways to measure the central tendency of data, include : Mean, Weighted mean, Trimmed mean, Median, Mode and Midrange.

1. Mean :

- The mean of a data set is the average of all the data values. The sample mean $\bar{x}$ is the point estimator of the population mean μ.

$$\text{Sample mean } \bar{x} = \frac{\text{Sum of the values of the n observations}}{\text{Number of observations in the sample}} = \frac{\Sigma x_i}{n}$$

$$\text{Population mean } \mu = \frac{\text{Sum of the values of the N observations}}{\text{Number of observations in the population}} = \frac{\Sigma x_i}{n}$$

2. Median :

- The median of a data set is the value in the middle when the data items are arranged in ascending order. Whenever a data set has extreme values, the median is the preferred measure of central location.
- The median is the measure of location most often reported for annual income and property value data. A few extremely large incomes of property values can inflate the mean.
- For an off number of observations :

  7 observations = 26, 18, 27, 12, 14, 29, 19

  Numbers in ascending order = 12, 14, 18, 19, 26, 27, 29
- The median is the middle value.

  Median = 19
- For an even number of observations :

  8 observations = 26, 18, 29, 12, 14, 27, 30, 19

  Numbers in ascending order = 12, 14, 18, 19, 26, 27, 29,30

The median is the average of the middle two values.

$$\text{Median} = \frac{(19 + 26)}{2} = 22.5$$

2. **Mode :**

- The mode of a data set is the value that occurs with greatest frequency. The greatest frequency can occur at two or more different values. If the data have exactly two modes, the data have exactly two modes, the data are bimodal. If the data have more than two modes, the data are multimodal.
- **Weighted mean :** Sometimes, each value in a set may be associated with a weight, the weights reflect the significance, importance, or occurrence frequency attached to their respective values.
- **Trimmed mean :** A major problem with the mean is its sensitivity to extreme (e.g., outlier) values. Even a small number of extreme values can corrupt the mean. The trimmed mean is the mean obtained after cutting off values at the high and low extremes.
- For example, we can sort the values and remove the top and bottom 2 % before computing the mean. We should avoid trimming too large a portion (such as 20 %) at both ends as this can result in the loss of valuable information.
- **Holistic measure** is a measure that must be computed on the entire data set as a whole. It cannot be computed by partitioning the given data into subsets and merging the values obtained for the measure in each subset.

4.7.4 Measuring the Dispersion of Data

- An **outlier** is an observation that lies an abnormal distance from other values in a random sample from a population.
- **First quartile (Q_1) :** The first quartile is the value, where 25 % of the values are smaller than Q_1 and 75 % are larger.
- **Third quartile (Q_3) :** The third quartile is the value, where75 % of the values are smaller than Q_3 and 25 % are larger.
- The **box plot** is a useful graphical display for describing the behavior of the data in the middle as well as at the ends of the distributions. The box plot uses the median and the lower and upper quartiles. If the lower quartile is Q_1 and the upper quartile is Q_3, then the difference $(Q_3 - Q_1)$ is called the interquartile range or IQ.
- **Range :** Difference between highest and lowest observed values

Variance :

- The variance is a measure of variability that utilizes all the data. It is based on the difference between the value of each observation (x_i) and the mean ($\bar{x}$) for a sample, μ for a population).
- The variance is the average of the squared between each data value and the mean.

$$\text{Sample variance : } S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$\text{Population variance : } \sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Standard deviation :

- The standard deviation of a data set is the positive square root of the variance. It is measured in the same in the same units as the data, making it more easily interpreted than the variance.
- The standard deviation is computed as follows :

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

$$\text{Sample standard deviation} = S = \sqrt{S^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Difference between Standard deviation and Variance

Sr. No.	Standard deviation	Variance
1.	Standard deviation is a measure of dispersion of the values of a data set from their mean.	It is the statistical measure of how far the numbers are spread in a data set from their average.
2.	It is a common term in statistical theory to calculate central tendency	Variance is primarily used for statistical probability distribution to measure volatility from the mean
3.	It measures the absolute variability of the dispersion	It helps determine the size of the data spread.
4.	It is calculated by taking the square root of the variance.	It is calculated by taking the average of the squared deviation of each value in the data set from the mean
5.	The standard deviation is symbolized by the Greek letter sigma "σ" as in lower case sigma	The notation for the variance of a variable is "σ^2" sigma squared

6.	$\sigma = \sqrt{\sum (x - M)^2 / n}$ where M = mean, x = a values in a data set, and n = number of values	$\sigma^2 = \sum (x - M)^2 / n$ where M = mean, x = each value in the data set, n = number of values in the data set
7.	Used in finance sector as a measure of market and security volatility.	Used in asset allocation

4.7.5 Identification of Outliers for Numerical Attributes

- A second way to identify outliers is based on the use of box plots, sometimes called box-and-whisker plots, in which the median and the lower and upper quartiles are represented on the axis where the observations are placed.
- A box plot is a type of chart often used in exploratory data analysis to visually show the distribution of numerical data and skewness through displaying the data quartiles or percentile and averages

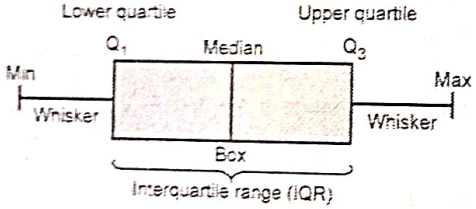


Fig. 4.7.4

1. Minimum score : The lowest score, excluding outliers.
2. Lower quartile : 25 % scores fall below the lower quartile value.
3. Median : The median marks the mid-point of the data and is shown by the line that divides the box into two parts.
4. Upper quartile : 75 % of the scores fall below the upper quartile value.
5. Maximum score : The highest score, excluding outliers.
6. Whiskers : The upper and lower whiskers represent scores outside the middle 50 %.
7. The interquartile range : This is the box plot showing the middle 50 % of scores.

- Boxplots are also extremely useful for visually checking group differences. Suppose we have four groups of scores and we want to compare them by teaching method. Teaching method is our categorical grouping variable and score is the continuous outcome variable that the researchers measured.

- Fig. 4.7.5 shows boxplot.

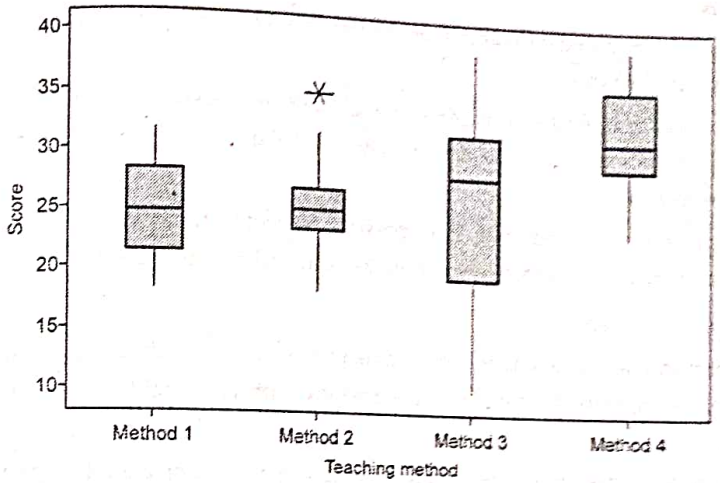


Fig. 4.7.5 Boxplot of score

4.8 Bivariate Analysis

4.8.1 Graphical Analysis

1. Scatter diagram

- Also called scatter plot, X-Y graph.
- While working with statistical data it is often observed that there are connections between sets of data. For example the mass and height of persons are related, the taller the person the greater his/her mass.
- To find out whether or not two sets of data are connected, scatter diagrams can be used.
- Scatter diagram shows the relationship between children's age and height.
- A scatter diagram is a tool for analyzing relationship between two variables. One variable is plotted on the horizontal axis and the other is plotted on the vertical axis.
- The pattern of their intersecting points can graphically show relationship patterns. Commonly a scatter diagram is used to prove or disprove cause-and-effect relationships.
- While scatter diagram shows relationships, it does not by itself prove that one variable causes other. In addition to showing possible cause and effect relationships, a scatter diagram can show that two variables are from a common cause that is unknown or that one variable can be used as a surrogate for the other.

2. Loess plots

- Loess plots are based on scatter plots and can therefore be applied in turn to pairs of numerical attributes.
- It is a popular tool used in regression analysis that creates a smooth line through a time plot or scatter plot to help you to see relationship between variables and foresee trends.
- Starting from a scatter plot, it is possible to add a trend curve to express the functional relationship between the attribute a_k and the attribute a_j.
- It is typically used for :
  a) Fitting a line to a scatter plot or time plot where noisy data values, sparse data points or weak interrelationships interfere with your ability to see a line of best fit.
  b) Linear regression where least squares fitting doesn't create a line of good fit or is too labor - intensive to use.
  c) Data exploration and analysis in the social sciences, particularly in elections and voting behavior.

3. Level curves

- Level curves can only be used for numerical attributes. They highlight the value of a third numerical attribute a_z as the attributes a_j and a_k placed on the axes of the plot vary.
- Connecting to each other the points in the plot that share the value of the third attribute, possibly by using some form of numerical interpolation, curved lines are obtained representing the geometric locus of the points for which the attribute a_z assumes a given value.

4.8.2 Measures of Correlation for Numerical Attributes

- When one measurement is made on each observation, uni - variate analysis is applied. If more than one measurement is made on each observation, multivariate analysis is applied. Here we focus on bivariate analysis, where exactly two measurements are made on each observation.
- The two measurements will be called X and Y. Since X and Y are obtained for each observation, the data for observation is the pair (X, Y).
- Some examples
  1. Height (X) and weight (Y) are measured for each individual in a sample.

  2. Stock market valuation (X) and quarterly corporate earnings (Y) are recorded for each company in a sample.
  3. A cell culture is treated with varying concentrations of a drug and the growth rate (X) and drug concentrations (Y) are recorded for each trial.
  4. Temperature (X) and precipitation (Y) are measured on a given day at a set of weather stations.
- There is difference in bivariate data and two sample data. In two sample data, the X and Y values are paired and there are not necessarily the same number of X and Y values.
- **Correlation** refers to a relationship between two or more objects. In statistics, the word correlation refers to the relationship between two variables. Correlations exists between two variables when one of them is related to the other in some way.
- **Examples :** One variable might be the number of hunters in a region and the other variable could be the deer population. Perhaps as the number of hunters increases, the deer population decreases. This is an example of a **negative correlation :** As one variable increases, the other decreases.
- A **positive correlation** is where the two variables react in the same way, increasing or decreasing together. Temperature in Celsius and Fahrenheit has a positive correlation.
- The term "correlation" refers to a measure of the strength of association between two variables.
- **Covariance** is the extent to which a change in one variable corresponds systematically to a change in another. Correlation can be thought of as a standardized covariance.
- The correlation coefficient r is a function of the data, so it really should be called the **sample correlation coefficient.** The (sample) correlation coefficient r estimates the population correlation coefficient ρ.
- If either the X_i or the Y_i values are constant (i.e. all have the same value), then one the sample standard deviations is zero and therefore the correlation is not defined.

4.8.3 Contingency Tables for Categorical Attributes

- A contingency table is a tabular representation of categorical data. A contingency table usually shows frequencies for particular combinations of values of two discrete random variable.

- When dealing with a pair of categorical attributes a_j and a_k, let

$$V = \{v_1, v_2, \ldots, v_j\}, U = \{u_1, u_2, \ldots, u_k\}$$

denote the sets of distinct values respectively assumed by each of them.

- A contingency table is defined as a matrix T whose generic element t_{rs} indicates the frequency with which the pair of values $\{x_{ij} = v_r\}$ and $\{x_{ik} = u_s\}$ appears in the records of the dataset D.
- Example : Suppose we have two categorical variables : Gender (male or female) and Handedness (right or left handed). Assume that we conduct a simple random sampling and obtain a size 100 data. We can then summarize our data using the following 2 × 2 table :

	Right - handed	Left-handed
Male	43	9
Female	44	4

- Such a table is called a 2 × 2 contingency table.
- Sometimes we may see people augmented the table with the total sums :

	Right - handed	Left - handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

- The contingency table elegantly summarizes the information about our data and may be one of the most common data analysis tools.

4.9 Multivariate Analysis

- When measuring several response variables, multivariate statistical techniques, such as multivariate analysis of variance, are often more powerful in detecting differences among populations than traditional Univariate techniques. The increased power of multivariate techniques is achieved by utilizing the correlation among the various response variables measured on a single experimentation unit.
- Multivariate data analysis techniques are appropriate when more than one response is measured on an experimentation unit.
- Traditionally, multivariate data analysis techniques are considered when a variety of response variables are measured on individual experimentation units which together quantify level and blood pressure are measured on each subject in the trial.

- To quantify the difference between the two treatments, Univariate statistical techniques could be employed, where each response variable is statistically analyzed. However, the univariate analysis approach ignores the potential correlation among response variables.

4.9.1 Graphical Analysis

- Using a matrix of scatter plots we can :
  a) Look at all of the relationships between pairs of variables in one group of plots
  b) Describe relationships among three or more variables
- Here, we have a matrix of scatterplots for quarter-root transformed data on all variables. Note that each variable appears to be positively related to the remaining variables. However, the strength of that relationship depends on which pair of variables is considered. For example, quarter-root iron is strongly related to quarter-root protein, but the relationship between calcium and vitamin C is not very strong.
- Fig. 4.9.1 shows scatter plot matrix

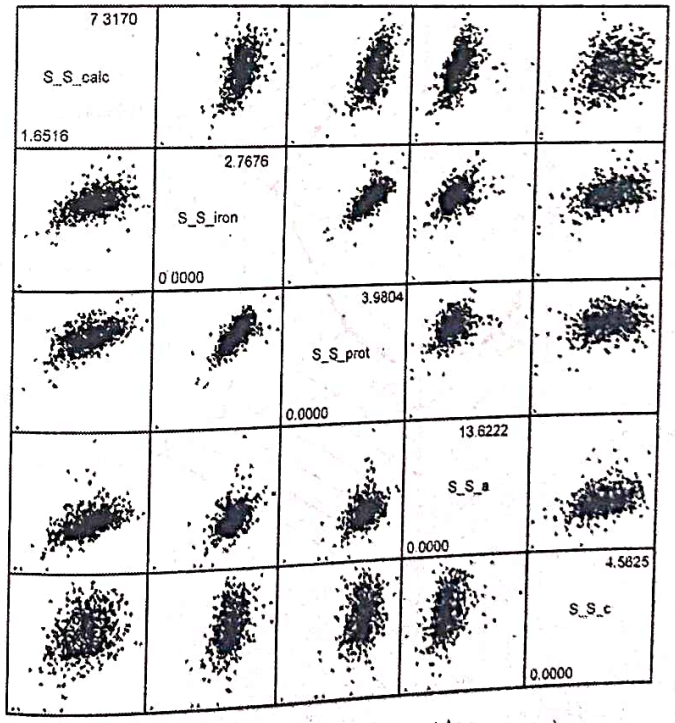


Fig. 4.9.1 Scatter plot matrix

Spider web chart :

- Spider web chart, also called radar chart, a graphical method to represent multivariate data in the form of a two - dimensional chart of three or more quantitative variables. It is useful for rating an item or items along three or more axes, e.g. the cost, quality of faculty, campus facilities and student life for three different colleges.
- They are used to plot one or more groups of values over multiple common variables. They do this by giving an axis for each variable and these axes are arranged radially around a central point and spaced equally. The data from a single observation are plotted along each axis and connected to form a polygon.
- Multiple observations can be placed in a single chart by displaying multiple polygons, overlaying them and reducing the opacity of each polygon.
- Fig. 4.9.2 shows spider web chart.

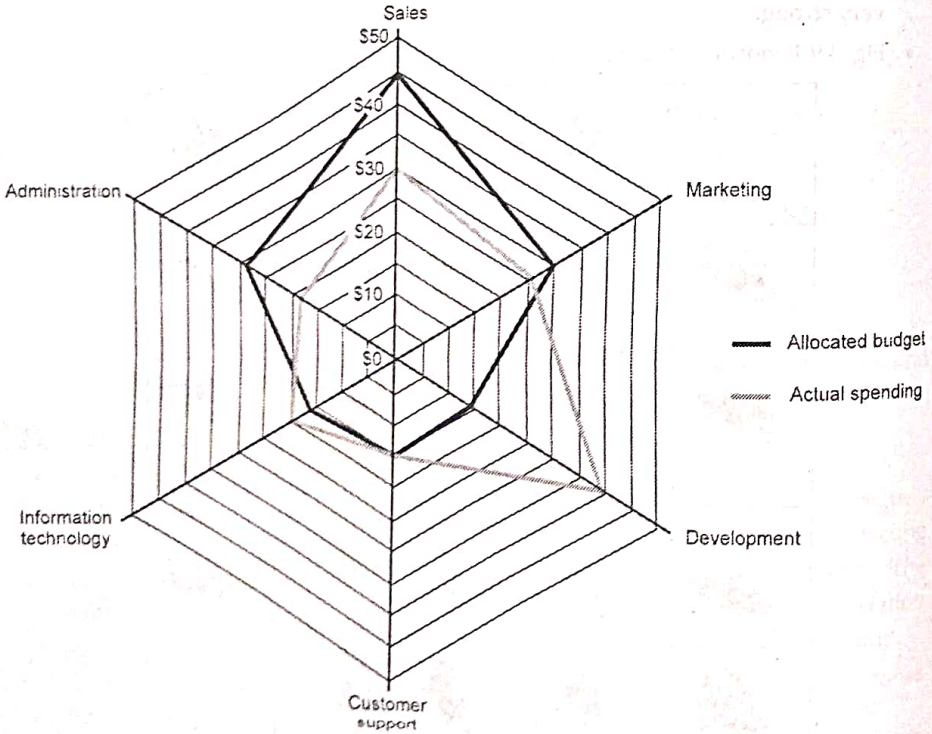


Fig. 4.9.2 Spider web chart

- Benefits of using spider web charts.
  a) Make concentrations of strengths and deficiencies visible.
  b) Clearly display the important categories.
  c) Define full performance in each category.
  d) Offer vivid and visual description.

4.9.2 Measures of Correlation for Numerical Attributes

- For multivariate analysis of numerical attributes, covariance and correlation matrices are calculated among all pairs of attributes.
- Let V and R be the two n × n square matrices whose elements are represented by the covariance values and correlations. Both matrices V and R are symmetric and positive definite.
- Notice that the covariance matrix V contains on its main diagonal the sample covariance values of each single attribute and for this reason it is also called the variance – covariance matrix.

❑❑❑

Notes

Unit V

5 Impact of Machine Learning in Business Intelligence Process

Syllabus

Classification : Classification problems, Evaluation of classification models, Bayesian methods. Logistic regression. Clustering : Clustering methods, Partition methods, Hierarchical methods. Evaluation of clustering models. Association Rule : Structure of Association Rule, Apriori Algorithm.

Contents

5.1 Classification Problem

5.2 Evaluation of Classification Models

5.3 Bayesian Methods

5.4 Logistic Regression

5.5 Clustering

5.6 Partition Methods

5.7 Hierarchical Methods

5.8 Evaluation of Clustering Models

5.9 Frequent Item-set Mining Methods

5.10 Improving Apriori Efficiency

5.11 Mining Frequent Itemset without Candidate Generation

5.12 Mining Various Kind of Association Rules

5.1 Classification Problem

- Classification predicts categorical labels (classes), prediction models continuous-valued functions. Classification is considered to be supervised learning.
- Classifies data based on the training set and the values in a classifying attribute and uses it in classifying new data. **Prediction** means models continuous-valued functions, i.e., predicts unknown or missing values.
- Preprocessing of the data in preparation for classification and prediction can involve data cleaning to reduce noise or handle missing values, relevance analysis to remove irrelevant or redundant attributes and data transformation, such as generalizing the data to higher level concepts or normalizing data.
- Fig. 5.1.1 shows the classification.

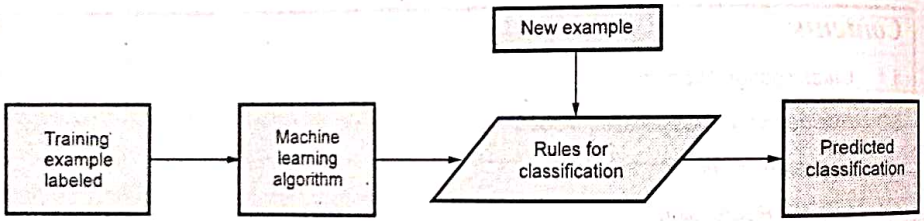


Fig. 5.1.1 Classification

Aim : To predict categorical class labels for new samples.

Input : Training set of samples, each with a class label.

Output : Classifier is based on the training set and the class labels.

- Prediction is similar to classification. It constructs a model and uses the model to predict unknown or missing value.
- Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data.
- Classification and prediction may need to be preceded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process.

- Numeric prediction is the task of predicting continuous values for given input. For example, we may wish to predict the salary of college employee with 15 years of work experience, or the potential sales of a new product given its price.
- Some of the classification methods like back - propagation, support vector machines, and k - nearest - neighbor classifiers can be used for prediction.
- Fig. 5.1.2 shows flow diagram for probabilistic structure of the learning process for classification. It may clarify the probability assumptions concerning the three components of a classification problem : A generator of observations, a supervisor of the target class and a classification algorithm.

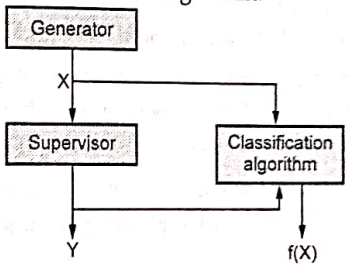


Fig. 5.1.2 Probabilistic structure of the learning process for classification

1. **Generator :** The task of the generator is to extract random vectors x of examples according to an unknown probability distribution $P_x(x)$.
2. **Supervisor :** The supervisor returns for each vector x of examples the value of the target class according to a conditional distribution $P_{y|x}(y|x)$ which is also unknown.
3. **Algorithm :** A classification algorithm A_F, also called a classifier, chooses a function $f^* \in F$ in the hypothesis space so as to minimize a suitably defined loss function.

- The development of a classification model consists of three main phases : Training, test and prediction phase.

1. **Training phase :** During the training phase, the classification algorithm is applied to the examples belonging to a subset T of the dataset D, called the **training set**, in order to derive classification rules that allow the corresponding target class y to be attached to each observation x.
2. **Test phase :** In the test phase, the rules generated during the training phase are used to classify the observations of D not included in the training set, for which the target class value is already known. To assess the accuracy of the

5.1 Classification Problem

- Classification predicts categorical labels (classes), prediction models continuous-valued functions. Classification is considered to be supervised learning.
- Classifies data based on the training set and the values in a classifying attribute and uses it in classifying new data. **Prediction** means models continuous-valued functions, i.e., predicts unknown or missing values.
- Preprocessing of the data in preparation for classification and prediction can involve data cleaning to reduce noise or handle missing values, relevance analysis to remove irrelevant or redundant attributes and data transformation, such as generalizing the data to higher level concepts or normalizing data.
- Fig. 5.1.1 shows the classification.

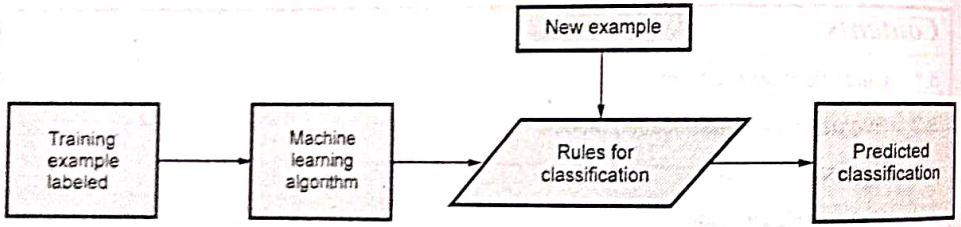


Fig. 5.1.1 Classification

Aim : To predict categorical class labels for new samples.

Input : Training set of samples, each with a class label.

Output : Classifier is based on the training set and the class labels.

- Prediction is similar to classification. It constructs a model and uses the model to predict unknown or missing value.
- Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data.
- Classification and prediction may need to be preceded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process.

- Numeric prediction is the task of predicting continuous values for given input. For example, we may wish to predict the salary of college employee with 15 years of work experience, or the potential sales of a new product given its price.
- Some of the classification methods like back - propagation, support vector machines, and k - nearest - neighbor classifiers can be used for prediction.
- Fig. 5.1.2 shows flow diagram for probabilistic structure of the learning process for classification. It may clarify the probability assumptions concerning the three components of a classification problem : A generator of observations, a supervisor of the target class and a classification algorithm.

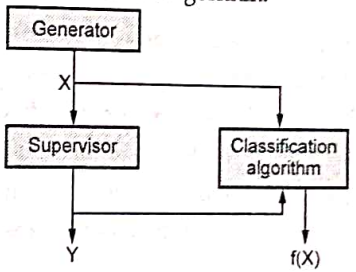


Fig. 5.1.2 Probabilistic structure of the learning process for classification

1. **Generator :** The task of the generator is to extract random vectors x of examples according to an unknown probability distribution $P_x(x)$.
2. **Supervisor :** The supervisor returns for each vector x of examples the value of the target class according to a conditional distribution $P_{y|x}(y|x)$ which is also unknown.
3. **Algorithm :** A classification algorithm A_F, also called a classifier, chooses a function $f^* \in F$ in the hypothesis space so as to minimize a suitably defined loss function.

- The development of a classification model consists of three main phases : Training, test and prediction phase.

1. **Training phase :** During the training phase, the classification algorithm is applied to the examples belonging to a subset T of the dataset D, called the **training set**, in order to derive classification rules that allow the corresponding target class y to be attached to each observation x.
2. **Test phase :** In the test phase, the rules generated during the training phase are used to classify the observations of D not included in the training set, for which the the target class value is already known. To assess the accuracy of the

classification model, the actual target class of each instance in the test set $V = D - T$ is then compared with the class predicted by the classifier.

3. **Prediction phase :** The prediction phase represents the actual use of the classification model to assign the target class to new observations that will be recorded in the future. A prediction is obtained by applying the rules generated during the training phase to the explanatory variables that describe the new instance.

5.1.1 Taxonomy of Classification Models

- Categories of classification models are heuristic models, separation models, regression models and probabilistic models.

1. **Heuristic models :** It makes use of classification procedures based on simple and intuitive algorithms. Nearest neighbor method is based on this concept. It uses divide-and-conquer scheme.
2. **Separation models :** Separation models divide the attribute space R^n into H disjoint regions $\{S_1, S_2, \ldots, S_H\}$, separating the observations based on the target class.

   Example : Discriminant analysis, perceptron methods, neural networks and support vector machines.
3. **Probabilistic models :** A hypothesis is formulated regarding the functional form of the conditional probabilities $P_{x|y}(x|y)$ of the observations given the target class, known as class-conditional probabilities.
4. **Regression model :** Regression is a data mining function that predicts a number. Profit, sales, mortgage rates, house values, square footage, temperature or distance could all be predicted using regression techniques. For example, a regression model could be used to predict the values of a data warehouse based on web-marketing, number of data entries, size and other factors. Regression analysis is a good choice when all of the predictor variables are continuous valued as well.

Review Questions

1. *Define classification. With a neat figure, explain the general approach for solving classification model.*
2. *With neat block diagram, explain general approach to solve classification problem.*

5.2 Evaluation of Classification Models

- Classification methods can be evaluated based on following criteria :

1. **Accuracy :** The accuracy of a model is an indicator of its ability to predict the target class for future observations. Based on their accuracy values, it is also possible to compare different models in order to select the classifier associated with the best performance.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

- For binary classification, accuracy can also be calculated in terms of positives and negatives.
- The evaluation measure most used in practice is the accuracy rate. It evaluates the effectiveness of the classifier by its percentage of correct predictions.

$$\text{Accuracy rate} = \frac{|\text{True negatives}| + |\text{True positives}|}{|\text{False negatives}| + |\text{False positives}| + |\text{True negatives}| + |\text{True positives}|}$$

2. **Robustness :** A classification method is robust if the classification rules generated, as well as the corresponding accuracy, do not vary significantly as the choice of the training set and the test set varies, and if it is able to handle missing data and outliers.
3. **Scalability :** The scalability of a classifier refers to its ability to learn from large datasets and it is inevitably related to its computation speed.

5.2.1 Holdout Method

- The data is split into two different datasets labelled as a training and a testing dataset. This can be a 60/40 or 70/30 or 80/20 split. This technique is called **the hold-out** validation technique.
- Suppose we have a database with house prices as the dependent variable and two independent variables showing the square footage of the house and the number of rooms.
- Now, imagine this dataset has 30 rows. The whole idea is that you build a model that can predict house prices accurately.
- To 'train' your model, or see how well it performs, we randomly subset 20 of those rows and fit the model.
- The second step is to predict the values of those 10 rows that we excluded and measure how well our predictions were.

- As a rule of thumb, experts suggest to randomly sample 80 % of the data into the training set and 20 % into the test set.
- Training set : Used to train the classifier.

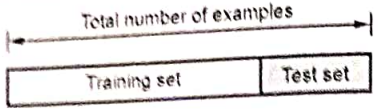


Fig. 5.2.1

- The holdout method has two basic drawbacks :
  1. It requires extra dataset
  2. It is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an "unfortunate" split.

5.2.2 Cross-validation

- Cross-validation is a technique for evaluating estimating performance by training several machine learning models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting, i.e., failing to generalize a pattern.
- In general, machine learning involves deriving models from data, with the aim of achieving some kind of desired behaviour, e.g., prediction or classification.
- Fig. 5.2.2 shows cross-validation.

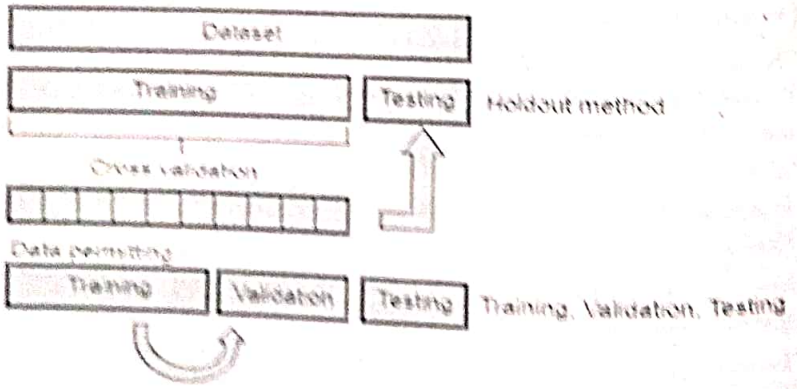


Fig. 5.2.2 Cross validation

- But this generic task is broken down into a number of special cases. When training is done, the data that was removed can be used to test the performance of the learned model on "new" data. This is the basic idea for a whole class of model evaluation methods called **cross validation.**
- Types of cross validation methods are holdout, K-fold and Leave-one-out.
- The holdout method is the simplest kind of cross validation. The data set is separated into two sets, called the training set and the testing set. The function approximate fits a function using the training set only.
- The K-fold cross validation is one way to improve over the holdout method. The data set is divided into k subsets, and the holdout method is repeated k times.
- Each time, one of the k subsets is used as the test set and the other k – 1 subsets are put together to form a training set. Then the average error across all k trials is computed.
- Leave-one-out cross validation is K-fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set.
- That means that N separate times, the function approximate is trained on all the data except for one point and a prediction is made for that point.
- Cross-validation ensures non-overlapping test sets.

K-fold cross-validation :

- In this technique, k – 1 folds are used for training and the remaining one is used for testing as shown in Fig. 5.2.3.

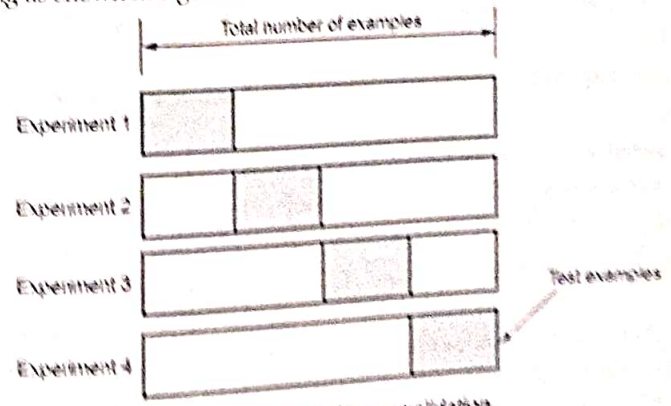


Fig. 5.2.3 K-fold cross validation

- The advantage is that entire data is used for training and testing. The error rate of the model is average of the error rate of each iteration.
- This technique can also be called a form the repeated hold-out method. The error rate could be improved by using stratification technique.

5.2.3 Confusion Matrix

- A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. The numbers displayed give the frequency of each data point.
- The confusion matrix for binary classification shown below :

True class	Predicted class	
	Positive	Negative
Positive	True negative	False negative
Negative	False positive	True negative

- A confusion matrix contains information about actual and predicted classification done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. Confusion matrix is also called a contingency table.
  1. **False positives :** Examples predicted as positive, which are from the negative class.
  2. **False negatives :** Examples predicted as negative, whose true class is positive.
  3. **True positives :** Examples correctly predicted as pertraining to the positive class.
  4. **True negatives :** Examples correctly predicted as belonging to the negative class.
- Binary classification accuracy metrics quantify the two types of correct predictions and two types of errors. Typical metrics are accuracy (ACC), precision, recall, false positive rate, F1-measure. Each metric measures a different aspect of the predictive model.
- Accuracy (ACC) measures the fraction of correct predictions. Precision measures the fraction of actual positives among those examples that are predicted as positive. Recall measures how many actual positives were predicted as positive. F1-measure is the harmonic mean of precision and recall many actual positives were predicted as positive. F1-measure is the harmonic mean of precision and recall.

5.2.4 ROC Curve

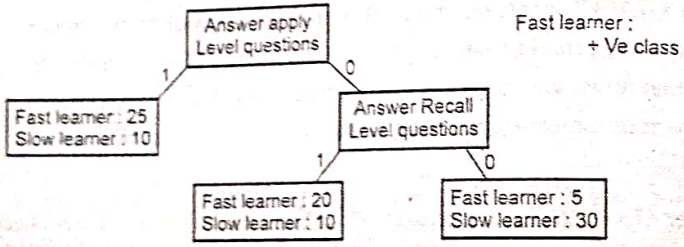
- Binary classification accuracy metrics quantify the two types of correct predictions and two types of errors. Typical metrics are accuracy (ACC), precision, recall, false positive rate, F1-measure. Each metric measures a different aspect of the predicative model.
- Accuracy (ACC) measures the fraction of correct predictions. Precision measures the fraction of actual positives among those examples that are predicted as positive. Recall measures how many actual positives were predicted as positive. F1-measure is the harmonic mean of precision and recall.

ROC Curve

- Receiver Operating Characteristics (ROC) graphs have long been used in signal detection theory to depict the tradeoff between hit rates and false alarm fates over noisy channel. Recent years have seen an increase in the use of ROC graphs in the machine learning community.
- An ROC plot plots true positive rate on the Y-axis false positive rate on the X-axis; a single contingency table corresponds to a single point in an ROC plot.
- The performance of a ranker can be assessed by drawing a piecewise linear curve in an ROC plot, known as an ROC curve. The curve starts in (0, 0), finishes in (1, 1) and is monitorically non-decreasing in both axes.
- A useful technique for organizing classifiers and visualizing their performance. Especially useful for domains with skewed class distribution and unequal classification error costs.
- It allows to create ROC curve and a complete sensitivity/specificity report. The ROC curve is a fundamental tool for diagnostic test evaluation.
- In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100 Specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve is a measure of how well a parameter can distinguish between two diagnostic groups.
- Each point on an ROC curve connecting two segments corresponds to the true and false positive rates achieved on the same test set by the classifier obtained from the ranker by splitting the ranking between those two segments.
- An ROC curve is convex if the slopes are montonically non-increasing when moving along the curve from (0, 0) to (1, 1). A concavity in an ROC curve, i.e., two

or more adjacent segments with increasing slopes, indicates a locally worse than random ranking. In this we would get better ranking performance by joining the segments involved in the concavity, thus creating a coarser classifier.

Example 5.2.1 *i) Find contingency table ii) Find recall iii) Precision iv) Negative recall v) False positive rate*



Solution : Contingency table

	Predicted			Total
Faster Learner	25	20	5	50
Slow Learner	10	10	30	50
Total	35	30	35	100

Actual

$$\text{Precision} = \frac{\text{True Positive}}{\text{Acutal Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \text{ or } \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

Calculate precision and recall

$$\text{Precision} = 25/35 = 0.714$$

$$\text{Recall} = 25/30 = 0.833$$

$$\text{False positive rate} = (\text{False positive}) / (\text{false positive + true negative})$$

$$= 10/(10 + 30) = 0.25$$

Example 5.2.2 *Consider following confusion matrix and calculate following i) Sensitivity of classifier ii) Specificity of classifier.)*

Confusion Matrix		Predicted		Total
Actual	+	8	10	18
	-	4	8	12
Total		12	18	30

Solution : Given data : TP = 8, FN = 10, FP = 4, TN = 8

- Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR)

$$\text{Sensitivity (SN)} = \frac{TP}{TP + FN} = \frac{8}{8+4} = 0.444$$

- Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called True Negative Rate (TNR).

$$\text{Specificity (SP)} = \frac{TN}{TN + FN} = \frac{8}{8+4} = 0.666$$

Example 5.2.3 *Consider the following 3-class confusion matrix. Calculate precision and recall per class. Also calculate weighted average precision and recall for classifier.*

	Predicted		
Actual	*15*	*2*	*3*
	7	*15*	*8*
	2	*3*	*45*

Solution :

	Predicted			
Actual	15	2	3	20
	7	15	8	30
	2	3	45	50
	24	20	56	100

$$\text{Classifier Accuracy} = \frac{15+15+45}{100} = \frac{75}{100} = 0.75$$

Calculate per-class precision and recall :

$$\text{First class} = \frac{15}{24} = 0.63 \quad \text{and} \quad \frac{15}{20} = 0.75$$

$$\text{Second class} = \frac{15}{20} = 0.75 \quad \text{and} \quad \frac{15}{30} = 0.50$$

$$\text{Third class} = \frac{45}{56} = 0.8 \quad \text{and} \quad \frac{45}{50} = 0.9$$

5.3 Bayesian Methods

- Bayesian methods belong to the family of probabilistic classification models. They explicitly calculate the posterior probability that a given observation belongs to a specific target class by means of Bayes theorem once the prior probability and the class conditional probabilities are known.
- Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class.
- Bayesian classification is based on Bayes' theorem. Studies comparing classification algorithms have found a simple Bayesian classifier known as the **naive Bayesian classifier** to be comparable in performance with decision tree and selected neural network classifier to be comparable in performance with decision tree and selected neural network classifiers.
- Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.
- Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Baye's theorem with strong independence assumptions between the features.
- It is highly scalable, requiring a number of parameters linear in the number of variables in a learning problem.

A Naive Bayes classifier is a program which predicts a class value given a set of attributes.

- For each known class value,
  1. Calculate probabilities for each attribute, conditional on the class value.
  2. Use the product rule to obtain a joint conditional probability for the attributes.
  3. Use Bayes rule to derive conditional probabilities for the class variable.
- Once this has been done for all class values, output the class with the highest probability.
- Native Bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.
- The probability of a class value given a value of an attribute is called the **conditional probability**. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class.

- Baye's theorem is a result in probability theory that relates conditional probabilities. If A and B denote two events, P(A|B) denotes the conditional probability of A occurring, given that B occurs. The two conditional probabilities P(A|B) and P(B|A) are in general different.
- Bays theorem gives a relation between P(A|B) and P(B|A). An important application of Baye's theorem is that it gives a rule how to update or revise the strengths of evidence - based beliefs in light of new evidence posteriori.
- A **prior probability** is an initial probability value originally obtained before any additional information that is later obtained.
- A **posterior probability** is value that has been revised by using additional information that is later obtained.
- Suppose that $B_1, B_2, \ldots, B_3 \ldots B_n$ partition the outcomes of an experiment and that A is another event. For any number, k, with $1 \leq k \leq n$, we have the formula :

$$P(B_k/A) = \frac{(A/B_k) \cdot P(B_k)}{\sum_{i=1}^{n} P(A/B_i) \cdot P(A/B_i)}$$

5.3.1 Bayesian Networks

- Bayesian belief networks represent the full joint distribution over the variables more compactly with a smaller number of parameters.
- It take advantage of conditional and marginal independences among random variables
- A and B are independent then P(A, B) = P(A)P(B)
- A and B are conditionally independent given C

  $P(A, B \mid C) = P(A \mid C)P(B \mid C)$

  $P(A \mid C, B) = P(A \mid C)$
- **Example : Alarm system example.**
- Assume your house has an alarm system against burglary. You live in the seismically active area and the alarm system can get occasionally set off by an earthquake.
- You have two neighbors, Mary and John, who do not know each other. If they hear the alarm they call you, but this is not guaranteed.
- We want to represent the probability distribution of events : Burglary, Earthquake, Alarm, Mary calls and John calls.

Causal relations :

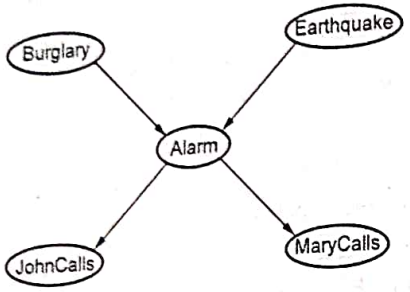


Fig. 5.3.1

Directed acyclic graph :

- Nodes = Random variables

  Burglary, Earthquake, Alarm, Mary calls and John calls

- Links = Direct (causal) dependencies between variables.

The chance of Alarm is influenced by Earthquake, The chance of John calling is affected by the Alarm.

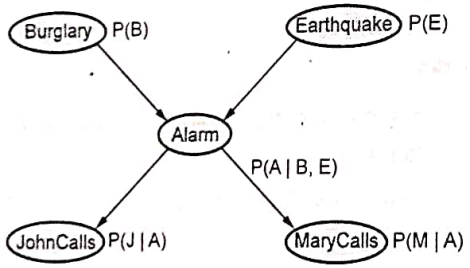


Fig. 5.3.2

Local conditional distributions : Relate variables and their parents

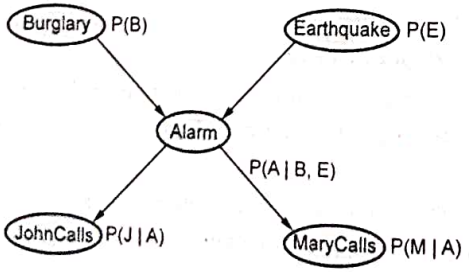


Fig. 5.3.3

Bayesian belief network :

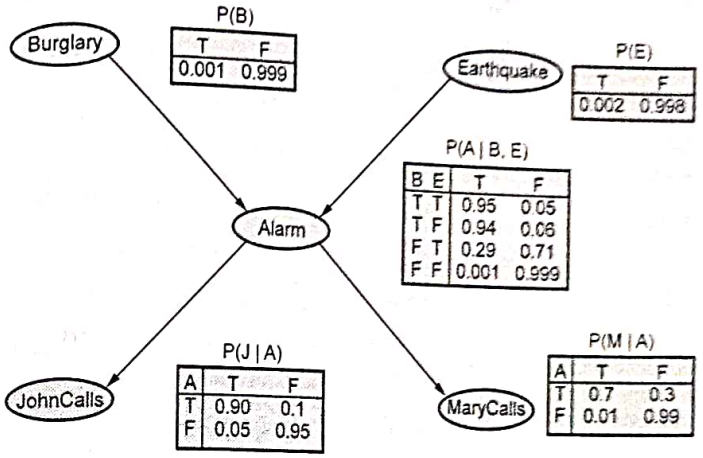


Fig. 5.3.4

Example 5.3.1 *For the given dataset apply Naive Bayes algorithm and predict the outcome for the car = {Red, Domestic, SUV}*

Color	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Domestic	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

Solution : • We want to classify a Red Domestic SUV.

- We need to calculate the probabilities.

P(Red | Yes), P(SUV | Yes), P(Domestic | Yes), P(Red | No), P(SUV | No) and P(Domestic | No) and multiply them by P(Yes) and P(No) respectively.

Yes :

Red :

n = 5

n_c = 3

p = 0.5

m = 3

SUV :

n = 5

n_c = 1

p = 0.5

m = 3

Domestic :

n = 5

n_c = 2

p = 0.5

m = 3

No :

Red :

n = 5

n_c = 2

p = 0.5

m = 3

SUV :

n = 5

n_c = 3

p = 0.5

m = 3

Domestic :

n = 5

n_c = 3

p = 0.5

m = 3

Looking at P(Red/Yes), we have 5 cases where v_j = Yes and in 3 of those cases a_i = Red. So for P(Red/Yes), n = 5 and = 3. Note that all attribute are binary (two possible values). We are assuming no other information so, p = 1/(number-of-attribute-values) = 0.5 for all of our attributes. Our m value is arbitrary, (We will use m = 3) but consistent for all attributes.

$$P(\text{Red} \mid \text{Yes}) = \frac{3 + 3 * 0.5}{5 + 3} = 0.56$$

$$P(\text{Red} \mid \text{No}) = \frac{2 + 3 * 0.5}{5 + 3} = 0.43$$

$$P(\text{SUV} \mid \text{Yes}) = \frac{1 + 3 * 0.5}{5 + 3} = 0.31$$

$$P(\text{SUV} \mid \text{No}) = \frac{3 + 3 * 0.5}{5 + 3} = 0.56$$

$$P(\text{Domestic} \mid \text{Yes}) = \frac{2 + 3 * 0.5}{5 + 3} = 0.43$$

$$P(\text{Domestic} \mid \text{No}) = \frac{3 + 3 * 0.5}{5 + 3} = 0.56$$

We have P(Yes) = 0.5 and P(No) = 0.5,

For v = Yes, we have

P(Yes) * P(Red | Yes) * P(SUV | Yes) * P(Domestic | Yes)

= 0.5 * 0.56 * 0.31 * 0.43 = 0.37

and for v = No, we have

P(No) * P(Red | No) * P(SUV | No) * P(Domestic | No)

= 0.5 * 0.43 * 0.56 * 0.56 * 0.69

Since 0.069 > 0.037, our example gets classified as 'NO'.

Example 5.3.2 *Consider following dataset and predict the class of new instance X using Navie Bayes Classification algorithm.*

Tid	Refund	Material Status	Taxable Amount	Evade
1	Yes	Single	125 K	No
2	No	Married	100 K	No
3	No	Single	70 K	No
4	Yes	Married	120 K	No
5	No	Divorced	95 K	Yes
6	No	Married	60 K	No
7	Yes	Married	220 K	No
8	No	Single	85 K	Yes
9	No	Married	75 K	No
10	No	Single	90 K	Yes

X = (Refund = No, Martial status = Married, Income = 120 K)

Ans. :

P(Refund = Yes|No) = 3/7

P(Refund = No|No) = 4/7

P(Refund = Yes|Yes) = 0

P(Refund = No|Yes) = 1

P(Marital Status = Single|No) = 2/7

P(Marital Status = Divorced|No) = 1/7

P(Marital Status = Married|No) = 4/7

P(Marital Status = Single|Yes) = 2/7

P(Marital Status = Divorced|Yes) = 1/7

P(Marital Status = Married|Yes) = 0

For taxable income :

If class = No : sample mean = 110

sample variance = 2975

If class = Yes : sample mean = 90

sample variance = 25

$$P(X|Class = No) = P(Refund = No|Class = No) \times P(Married|Class = No) \times P(Income = 120\ K|Class = No)$$

$$= 4/7 \times 4/7 \times 0.0072 = 0.0024$$

$$P(X|Class = Yes) = P(Refund = No|\ Class = Yes) \times P(Married\ |\ Class = Yes) \times P(Income=120\ K|\ Class=Yes)$$

$$= 1 \times 0 \times 1.2 \times 10^{-9} = 0$$

Since $P(X|No)P(No) > P(X|Yes)P(Yes)$

Therefore $P(No|X) > P(Yes|X) \Rightarrow Class = No$

5.4 Logistic Regression

- Logistic regression is a form of regression analysis in which the outcome variable is binary or dichotomous. A statistical method used to model dichotomous or binary outcomes using predictor variables.
- Logistic regression is one of the supervised learning algorithms.
- The binary logistic regression model is given by,

$$P(Y) = \frac{e^z}{1 + e^z}$$

Where $Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k$ (X_1, X_2, X_3 are independent variable)

- The logistic regression function is rewritten as,

$$\frac{P(Y = 1)}{1 - P(Y = 1)} = e^z$$

$$ln\left(\frac{P(Y)}{1 - P(Y)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

- Multiple linear regression is an extension of linear regression, which allows a response variable, y to be modeled as a linear function of two or more predictor variables.
- Logistic function is similar to a multiple linear regression model. Such models are called Generalized Linear Models (GLM), in GLM the errors do not follow normal distribution and there exists a transformation function of the outcome variable that takes a linear function.

5.5 Clustering

What is cluster analysis ?

- Given a set of objects, place them in groups such that the objects in a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups.
- Cluster analysis can be a powerful data-mining tool for any organisation that needs to identify discrete groups of customers, sales transactions, or other types of behaviors and things. For example, insurance providers use cluster analysis to detect fraudulent claims, and banks use it for credit scoring.
- Cluster analysis uses mathematical models to discover groups of similar customers based on the smallest variations among customers within each group.
- Cluster is a group of objects that belong to the same class. In other words the similar object are grouped in one cluster and dissimilar are grouped in other cluster.
- Clustering is a process of partitioning a set of data in a set of meaningful subclasses. Every data in the subclass shares a common trait. It helps a user understand the natural grouping or structure in a data set.
- Various types of clustering methods are partitioning methods, Hierarchical clustering, Fuzzy clustering, Density-based clustering and Model-based clustering.
- Cluster analysis is process of grouping a set of data-objects into clusters.
- **Desirable properties of a clustering algorithm are as follows :**
  1. Scalability (in terms of both time and space)
  2. Ability to deal with different data types
  3. Minimal requirements for domain knowledge to determine input parameters
  4. Interpretability and usability.

- Clustering of data is a method by which large sets of data are grouped into clusters of smaller sets of similar data. Clustering can be considered the most important unsupervised learning problem.

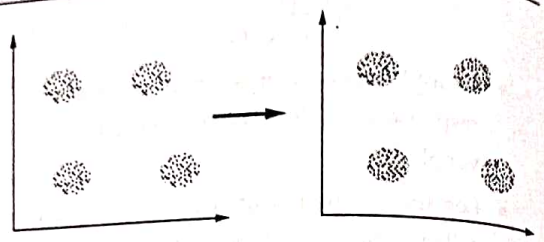
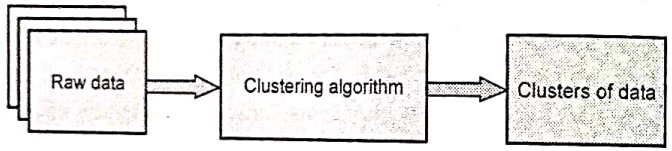
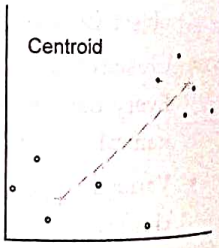


Fig. 5.5.1 Cluster

- A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Fig. 5.5.1 shows cluster.
- In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance : two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called **distance-based clustering**.



- Clustering means grouping of data or dividing a large data set into smaller data sets of some similarity.
- A clustering algorithm attempts to find natural groups of components or data based on some similarity. Also, the clustering algorithm finds the centroid of a group of data sets.
- To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster.



- **Cluster centroid :** The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the clusters. Each cluster has a well defined centroid.
- **Distance :** The distance between two points is taken as a common metric to as see the similarity among the components of a population. The commonly used distance

measure is the euclidean metric which defines the distance between two points $p = (p_1, p_2, ...)$ and $q = (q_1, q_2, ...)$ is given by :

$$d = \sum_{i=1}^{k} (p_i - q_i)^2$$

- The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering ? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.
- Clustering analysis helps construct meaningful partitioning of a large set of objects. Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing, etc.
- Clustering algorithms may be classified as listed below :
  1. Exclusive clustering
  2. Overlapping clustering
  3. Hierarchical clustering
  4. Probabilistic clustering.
- A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

Examples of Clustering Applications

1. **Marketing :** Help marketers discover distinct groups in their customer bases and then use this knowledge to develop targeted marketing programs.
2. **Land use :** Identification of areas of similar land use in an earth observation database.
3. **Insurance :** Identifying groups of motor insurance policy holders with a high average claim cost.
4. **Urban planning :** Identifying groups of houses according to their house type, value, and geographical location.
5. **Seismology :** Observed earth quake epicenters should be clustered along continent faults.

- Clustering of data is a method by which large sets of data are grouped into clusters of smaller sets of similar data. Clustering can be considered the most important unsupervised learning problem.

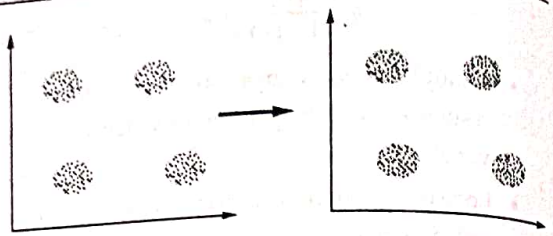
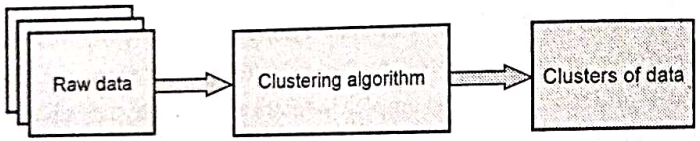
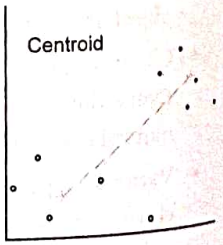


Fig. 5.5.1 Cluster

- A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Fig. 5.5.1 shows cluster.
- In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance : two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called **distance-based clustering**.



- Clustering means grouping of data or dividing a large data set into smaller data sets of some similarity.
- A clustering algorithm attempts to find natural groups of components or data based on some similarity. Also, the clustering algorithm finds the centroid of a group of data sets.
- To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster.



- **Cluster centroid :** The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the clusters. Each cluster has a well defined centroid.
- **Distance :** The distance between two points is taken as a common metric to as see the similarity among the components of a population. The commonly used distance

measure is the euclidean metric which defines the distance between two points $p = (p_1, p_2, \ldots)$ and $q = (q_1, q_2, \ldots)$ is given by :

$$d = \sum_{i=1}^{k} (p_i - q_i)^2$$

- The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering ? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.
- Clustering analysis helps construct meaningful partitioning of a large set of objects. Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing, etc.
- Clustering algorithms may be classified as listed below :
  1. Exclusive clustering
  2. Overlapping clustering
  3. Hierarchical clustering
  4. Probabilistic clustering.
- A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

Examples of Clustering Applications

1. **Marketing :** Help marketers discover distinct groups in their customer bases and then use this knowledge to develop targeted marketing programs.
2. **Land use :** Identification of areas of similar land use in an earth observation database.
3. **Insurance :** Identifying groups of motor insurance policy holders with a high average claim cost.
4. **Urban planning :** Identifying groups of houses according to their house type, value, and geographical location.
5. **Seismology :** Observed earth quake epicenters should be clustered along continent faults.

5.5.1 Typical Requirements of Clustering in Data Mining

1. **Scalability :** Many clustering algorithms work well on small data sets.
2. **Ability to deal with different types of attributes :** Many algorithms are designed to cluster interval-based data.
3. **Discovery of clusters with arbitrary shape :** Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures.
4. **Minimal requirements for domain knowledge to determine input parameters :** Many clustering algorithms require users to input certain parameters in cluster analysis.
5. **Ability to deal with noisy data :** Most real-world databases contain outliers or missing, unknown or erroneous data. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.
6. **Incremental clustering and insensitivity to the order of input records :** Some clustering algorithms cannot incorporate newly inserted data into existing clustering structures.
7. **High dimensionality :** A database or a data warehouse can contain several dimensions or attributes.
8. **Constraint-based clustering :** Real-world applications may need to perform clustering under various kinds of constraints.
9. **Interpretability and usability :** Users expect clustering results to be interpretable, comprehensible and usable.

5.5.2 Problems with Clustering

1. Current clustering techniques do not address all the requirements adequately;
2. Dealing with large number of dimensions and large number of data items can be problematic because of time complexity;
3. The effectiveness of the method depends on the definition of "distance";
4. If an obvious distance measure doesn't exist we must "define" it, which is not always easy, especially in multi-dimensional spaces;
5. The result of the clustering algorithm can be interpreted in different ways.

5.5.3 Types of Clusters

- Type of clusters are as follows :
  a) Well - separated clusters

  b) Prototype - based clusters
  c) Contiguity - based clusters
  d) Density - based clusters.

a) Well - separated clusters :

- A cluster is a set of points such that any point in a cluster is closer to every other point in the cluster than to any point not in the cluster.
- Fig. 5.5.2 shows well-separated cluster.

Fig. 5.5.2 Well-separated cluster

- Sometimes a threshold is used to specify that all the objects in a cluster must sufficiently close to one another. Definition of a cluster is satisfied only when the data contains natural clusters.

b) Prototype - based cluster

- A cluster is a set of objects such that an object in a cluster is closer (more similar) to the prototype or "center" of a cluster, than to the center of any other cluster. Prototype based clusters can also be referred to as "Center-Based" Clusters.
- The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster. Fig. 5.5.3 shows 4 center-based clusters.

Fig. 5.5.3 4 Center-based clusters

- If the data is numerical, the prototype of the cluster is often a centroid i.e., the average of all the points in the cluster.
- If the data has categorical attributes, the prototype of the cluster is often a medoid i.e., the most representative point of the cluster.

5.5.1 Typical Requirements of Clustering in Data Mining

1. **Scalability :** Many clustering algorithms work well on small data sets.
2. **Ability to deal with different types of attributes :** Many algorithms are designed to cluster interval-based data.
3. **Discovery of clusters with arbitrary shape :** Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures.
4. **Minimal requirements for domain knowledge to determine input parameters :** Many clustering algorithms require users to input certain parameters in cluster analysis.
5. **Ability to deal with noisy data :** Most real-world databases contain outliers or missing, unknown or erroneous data. Some clustering algorithms are sensitive to such data and may lead to clusters of poor quality.
6. **Incremental clustering and insensitivity to the order of input records :** Some clustering algorithms cannot incorporate newly inserted data into existing clustering structures.
7. **High dimensionality :** A database or a data warehouse can contain several dimensions or attributes.
8. **Constraint-based clustering :** Real-world applications may need to perform clustering under various kinds of constraints.
9. **Interpretability and usability :** Users expect clustering results to be interpretable, comprehensible and usable.

5.5.2 Problems with Clustering

1. Current clustering techniques do not address all the requirements adequately;
2. Dealing with large number of dimensions and large number of data items can be problematic because of time complexity;
3. The effectiveness of the method depends on the definition of "distance";
4. If an obvious distance measure doesn't exist we must "define" it, which is not always easy, especially in multi-dimensional spaces;
5. The result of the clustering algorithm can be interpreted in different ways.

5.5.3 Types of Clusters

- Type of clusters are as follows :
  a) Well - separated clusters

  b) Prototype - based clusters
  c) Contiguity - based clusters
  d) Density - based clusters.

a) Well - separated clusters :

- A cluster is a set of points such that any point in a cluster is closer to every other point in the cluster than to any point not in the cluster.
- Fig. 5.5.2 shows well-separated cluster.

Fig. 5.5.2 Well-separated cluster

- Sometimes a threshold is used to specify that all the objects in a cluster must sufficiently close to one another. Definition of a cluster is satisfied only when the data contains natural clusters.

b) Prototype - based cluster

- A cluster is a set of objects such that an object in a cluster is closer (more similar) to the prototype or "center" of a cluster, than to the center of any other cluster. Prototype based clusters can also be referred to as "Center-Based" Clusters.
- The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster. Fig. 5.5.3 shows 4 center-based clusters.

Fig. 5.5.3 4 Center-based clusters

- If the data is numerical, the prototype of the cluster is often a centroid i.e., the average of all the points in the cluster.
- If the data has categorical attributes, the prototype of the cluster is often a medoid i.e., the most representative point of the cluster.

- Objects in the cluster are closer to the prototype of the cluster than to the prototype of any other cluster.
- K-Means and K-Medoids are the examples of Prototype Based Clustering algorithm.

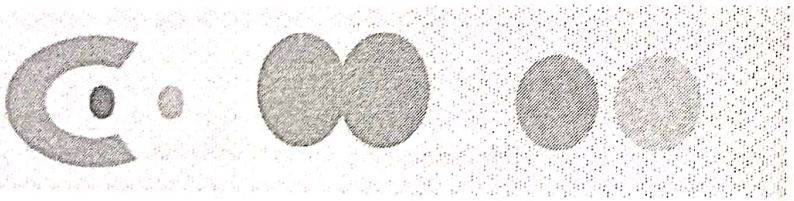
c) Contiguity - based clusters

- A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

Fig. 5.5.4 8 contiguous clusters

d) Density - based

- A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
- Used when the clusters are irregular or intertwined, and when noise and outliers are present.



5.5.4 Desired Features of Cluster Analysis

- Features are as follows :

1. **Scalability :** Data-mining problems can be large and therefore a cluster-analysis method should be able to deal with large problems gracefully. Ideally, performance should be linear with data-size.
2. **Only one scan of the dataset :** For large problems, data must be stored on disk, so cost of I/O disk can become significant in solving the problem.
3. **Ability to stop and resume :** For large dataset, cluster-analysis may require huge processor-time to complete the task. In such cases, the task should be able to be stopped and then resumed when convenient.

4. **Minimal input parameters :** The method should not expect too much guidance from the data-mining analyst.
5. **Robustness :** Most data obtained from a variety of sources has errors. Therefore, the method should be able to deal with noise, outlier and missing values gracefully.
6. **Ability to discover different cluster-shapes :** Clusters appear in different shapes and not all clusters are spherical. So the method should be able to discover cluster-shapes other than spherical.
7. **Different data types :** Many problems have a mixture of data types, for e.g. numerical, categorical and even textual. Therefore, the method should be able to deal with numerical, boolean and categorical data.
8. **Result independent of data input order :** The method should not be sensitive to data input-order.

Review Questions

1. *What is cluster analysis ? Explain different types of clusterings.*
2. *What is cluster analysis ? Discuss the different types of clusters with examples.*

5.6 Partition Methods

- Partitioning clustering are clustering methods used to classify observations, within a data set, into multiple groups based on their similarity. The algorithms require the analyst to specify the number of clusters to be generated.
- Commonly used partitioning methods are k-means, k-medoids.

5.6.1 K-Mean Clustering

- K-Means clustering is heuristic method. Here each cluster is represented by the center of the cluster. "K" stands for number of clusters, it is typically a user input to the algorithm; some criteria can be used to automatically estimate K.
- This method initially takes the number of components of the population equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart.
- Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The centroid's position is recalculated everytime a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters.

- Given K, the K-means algorithm consists of four steps :
  1. Select initial centroids at random.
  2. Assign each object to the cluster with the nearest centroid.
  3. Compute each centroid as the mean of the objects assigned to it.
  4. Repeat previous 2 steps until no change.
- The $x_1, ..., x_N$ are data points or vectors of observations. Each observation (vector x_i) will be assigned to one and only one cluster. The C(i) denotes cluster number for the observation. K-means minimizes within-cluster point scatter :

$$W(C) = \frac{1}{2} \sum_{K-1}^{K} \sum_{C(i)-K} \sum_{C(i)-K} \| x_i - x_j \|^2$$

$$= \sum_{K-1}^{K} Nk \sum_{C(i)-K} \| x_i - m_K \|^2$$

where

m_K is the mean vector of the K^{th} cluster.

N_K is the number of observations in K^{th} cluster.

K-Means Algorithm Properties

1. There are always K clusters.
2. There is always at least one item in each cluster.
3. The clusters are non-hierarchical and they do not overlap.
4. Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

The K-Means Algorithm Process

1. The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.
2. For each data point.
   a. Calculate the distance from the data point to each cluster.
   b. If the data point is closest to its own cluster, leave it where it is.
   c. If the data point is not closest to its own cluster, move it into the closest cluster.

3. Repeat the above step until a complete pass through all the data points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
4. The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intracluster distances and cohesion.

- K-means algorithm is iterative in nature. It converges, however only a local minimum is obtained. It works only for numerical data. This method easy to implement.
- **Advantages of K-Means Algorithm :**
  1. Efficient in computation
  2. Easy to implement
- Weaknesses
  1. Applicable only when mean is defined.
  2. Need to specify K, the number of clusters, in advance.
  3. Trouble with noisy data and outliers.
  4. Not suitable to discover clusters with non-convex shapes.

5.6.2 K-Medoids

- The K-medoids algorithm is a clustering algorithm related to the K-means algorithm and the medoidshift algorithm. K-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into K clusters known a priori. A useful tool for determining K is the silhouette.
- The most common realisation of K-medoid clustering is the Partitioning Around Medoids (PAM) algorithm. PAM uses a greedy search which may not find the optimum solution, but it is faster than exhaustive search.

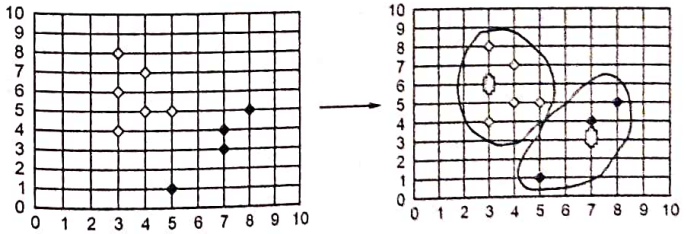


Fig. 5.6.1

- Given K, the K-means algorithm consists of four steps :
  1. Select initial centroids at random.
  2. Assign each object to the cluster with the nearest centroid.
  3. Compute each centroid as the mean of the objects assigned to it.
  4. Repeat previous 2 steps until no change.
- The $x_1, ..., x_N$ are data points or vectors of observations. Each observation (vector x_i) will be assigned to one and only one cluster. The C(i) denotes cluster number for the observation. K-means minimizes within-cluster point scatter :

$$W(C) = \frac{1}{2} \sum_{K-1}^{K} \sum_{C(i)-K} \sum_{C(i)-K} \| x_i - x_j \|^2$$

$$= \sum_{K-1}^{K} Nk \sum_{C(i)-K} \| x_i - m_K \|^2$$

where

m_K is the mean vector of the K^{th} cluster.

N_K is the number of observations in K^{th} cluster.

K-Means Algorithm Properties

1. There are always K clusters.
2. There is always at least one item in each cluster.
3. The clusters are non-hierarchical and they do not overlap.
4. Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

The K-Means Algorithm Process

1. The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.
2. For each data point.

   a. Calculate the distance from the data point to each cluster.

   b. If the data point is closest to its own cluster, leave it where it is.

   c. If the data point is not closest to its own cluster, move it into the closest cluster.

3. Repeat the above step until a complete pass through all the data points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
4. The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intracluster distances and cohesion.

- K-means algorithm is iterative in nature. It converges, however only a local minimum is obtained. It works only for numerical data. This method easy to implement.
- **Advantages of K-Means Algorithm :**
  1. Efficient in computation
  2. Easy to implement
- Weaknesses
  1. Applicable only when mean is defined.
  2. Need to specify K, the number of clusters, in advance.
  3. Trouble with noisy data and outliers.
  4. Not suitable to discover clusters with non-convex shapes.

5.6.2 K-Medoids

- The K-medoids algorithm is a clustering algorithm related to the K-means algorithm and the medoidshift algorithm. K-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into K clusters known a priori. A useful tool for determining K is the silhouette.
- The most common realisation of K-medoid clustering is the Partitioning Around Medoids (PAM) algorithm. PAM uses a greedy search which may not find the optimum solution, but it is faster than exhaustive search.

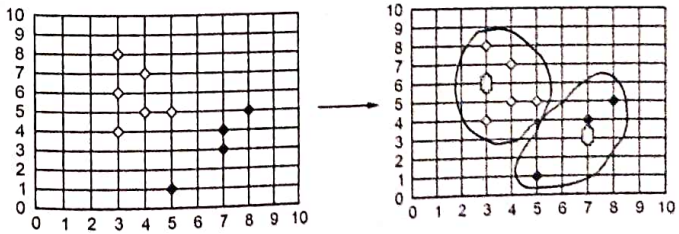


Fig. 5.6.1

- Instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster.
- A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the given data set.

1. The algorithm begins with arbitrary selection of the K objects as medoid points out of n data points (n > K)
2. After selection of the k medoid points, associate each data object in the given data set to most similar medoid. The similarity here is defined using distance measure that can be euclidean distance, manhattan distance or minkowski distance
3. Randomly select nonmedoid object O'
4. Compute total cost, S of swapping initial medoid object to O'
5. If S < 0, then swap initial medoid with the new one (if S < 0 then there will be new set of medoids)
6. Repeat steps 2 to 5 until there is no change in the medoid.

5.6.3 Difference between K-means and k-medoids

Sr. No.	K-means	k-medoids
1.	K-means attempts to minimize the total squared error.	k-medoids minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster.
2.	It is more efficient than k-medoids.	Comparatively less efficient.
3.	Sensitive to outliers.	Not sensitive to outliers.
4.	Convex shape is required.	Convex shape is not required.
5.	Efficient for separated clusters.	Efficient for separated clusters and small data sets.

Review Questions

1. *Find all 3 item itemsets from this set with minimum support = 2.*

Trans_id	Item list
T1	{K, A, D, B}
T2	{D, A, C, E, B}
T3	{C, A, B, E}
T4	{B, A, D}

2. *Explain silhouettes.*
3. *With an example, explain feature as a split and feature as a predictor.*

5.7 Hierarchical Methods

- This method use distance matrix as clustering criteria. This method does not require the number of clusters K as an input, but needs a termination condition. Hierarchical clustering is a widely used data analysis tool.
- The idea is to build a binary tree of the data that successively merges similar groups of points. Visualizing this tree provides a useful summary of the data.
- Hierarchical clustering arranges items in a hierarchy with a tree like structure based on the distance or similarity between them. The graphical representation of the resulting hierarchy is a tree-structured graph called a dendrogram.
- Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion.

Agglomerative hierarchical clustering

- This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied. Most hierarchical clustering methods belong to this category.
- Initially, AGNES places each objects into a cluster of its own. The clusters are then merged step-by-step according to some criterion. For example, cluster C1 and C2 may be merged if an object in C_1 and object in C_2 form the minimum Euclidean distance between any two objects from different clusters.

- Instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster.
- A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the given data set.

1. The algorithm begins with arbitrary selection of the K objects as medoid points out of n data points (n > K)
2. After selection of the k medoid points, associate each data object in the given data set to most similar medoid. The similarity here is defined using distance measure that can be euclidean distance, manhattan distance or minkowski distance
3. Randomly select nonmedoid object O'
4. Compute total cost, S of swapping initial medoid object to O'
5. If S < 0, then swap initial medoid with the new one (if S < 0 then there will be new set of medoids)
6. Repeat steps 2 to 5 until there is no change in the medoid.

5.6.3 Difference between K-means and k-medoids

Sr. No.	K-means	k-medoids
1.	K-means attempts to minimize the total squared error.	k-medoids minimizes the sum of dissimilarities between points labeled to be in a cluster and a point designated as the center of that cluster.
2.	It is more efficient than k-medoids.	Comparatively less efficient.
3.	Sensitive to outliers.	Not sensitive to outliers.
4.	Convex shape is required.	Convex shape is not required.
5.	Efficient for separated clusters.	Efficient for separated clusters and small data sets.

Review Questions

1. *Find all 3 item itemsets from this set with minimum support = 2.*

Trans_id	Item list
T1	{K, A, D, B}
T2	{D, A, C, E, B}
T3	{C, A, B, E}
T4	{B, A, D}

2. *Explain silhouettes.*
3. *With an example, explain feature as a split and feature as a predictor.*

5.7 Hierarchical Methods

- This method use distance matrix as clustering criteria. This method does not require the number of clusters K as an input, but needs a termination condition. Hierarchical clustering is a widely used data analysis tool.
- The idea is to build a binary tree of the data that successively merges similar groups of points. Visualizing this tree provides a useful summary of the data.
- Hierarchical clustering arranges items in a hierarchy with a tree like structure based on the distance or similarity between them. The graphical representation of the resulting hierarchy is a tree-structured graph called a dendrogram.
- Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion.

Agglomerative hierarchical clustering

- This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied. Most hierarchical clustering methods belong to this category.
- Initially, AGNES places each objects into a cluster of its own. The clusters are then merged step-by-step according to some criterion. For example, cluster C1 and C2 may be merged if an object in C_1 and object in C_2 form the minimum Euclidean distance between any two objects from different clusters.

- In the agglomerative hierarchical approach, we start by defining each data point to be a cluster and combine existing clusters at each step. Here are four different methods for doing this :
  1. **Single linkage :** Smallest pairwise distance between elements from each cluster
  2. **Complete linkage :** Largest distance between elements from each cluster
  3. **Average linkage :** The average distance between elements from each cluster
  4. **Centroid linkage :** Distance between cluster means.

Divisive Hierarchical Clustering

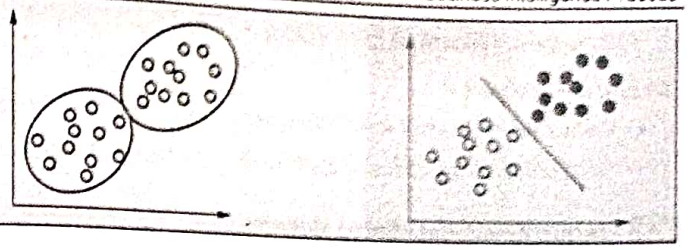
- This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster. It subdivides the clusters into smaller and smaller pieces, until each object form a cluster on its own or until it satisfies certain termination conditions, such as a desired number of cluster or the diameter of each cluster is within a certain threshold.

Sr. No.	Agglomerative	Divisive Hierarchical Clustering
1.	Initially each item in its own cluster.	Initially all items in one cluster.
2.	Iteratively clusters are merged together.	Large clusters are successively divided.
3.	Bottom up.	Top down.

5.7.1 Difference between Clustering vs Classification

Sr. No.	Clustering	Classification
1.	This function maps the data into one of several clusters which is the grouping of data items based on the similarities between them.	This model function classifies the data into one of several predefined categorical classes.
2.	Involved in unsupervised learning.	Involved in supervised learning.
3.	Training sample is not provided.	Training sample is provided.
4.	The number of cluster is not known before clustering, These are identified after the completion of clustering.	The number of classes is known before classification as there is predefined output based on input data.
5.	Data is not labeled.	Labeled data points.
6.	Asks how can I group this set of items ?	Asks what class does this item belong to ?
7.	Unknown number of classes.	Known number of classes.
8.	Used to understand data.	Used to classify future observations.

9.



Review Question

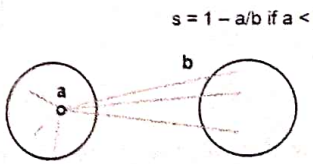
1. *Explain single linkage, complete linkage and average linkage.*

5.8 Evaluation of Clustering Models

- Silhouette refers to a method of interpretation and validation of clusters of data.
- Silhouettes are a general graphical aid for interpretation and validation of cluster analysis. This technique is available through the silhouette function. In order to calculate silhouettes, two types of data are needed :
  1. The collection of all distances between objects. These distances are obtained from application of dist function on the coordinates of the elements in mat with argument method.
  2. The partition obtained by the application of a clustering technique.
- For each element, a silhouette value is calculated and evaluates the degree of confidence in the assignment of the element :
  1. Well-clustered elements have a score near 1.
  2. Poorly-clustered elements have a score near – 1.
- Thus, silhouettes indicate the objects that are well or poorly clustered. Silhouette coefficient combines ideas of both cohesion and separation, but for individual points, as well as clusters and clustering's.
- For an individual point, I

  a = Average distance of i to the points in the same cluster

  b = min (average distance of i to points in another cluster)



Silhouette coefficient of i :

Silhouette coefficient

- **Cohesion :** Measures how closely related are objects in a cluster.
- **Separation :** Measure how distinct or well-separated a cluster is from other clusters.

5.9 Frequent Item-set Mining Methods

- Frequent item set mining is a method for market basket analysis. It aims at finding regularities in the shopping behaviour of customers of on-line shop, super-marks etc.
- **Market basket analysis** is an example of frequent item set mining. The purpose of market basket analysis is to determine what products customers purchase together. It takes its name from the idea of customers throwing all their purchases into a shopping cart (a "market basket") during grocery shopping.
- Market basket analysis is a technique which identifies the strength of association between pairs of products purchased together and identify patterns of co-occurrence. A co-occurrence is when two or more things take place together.
- One basket tells you about what one customer purchased at one time. Fig. 5.9.1 shows shopping cart.

Fig. 5.9.1 Basket

- Market basket analysis creates If-Then scenario rules, for example, if item A is purchased then item B is likely to be purchased.

- The rules are probabilistic in nature or, they are derived from the frequencies of co-occurrence in the observations.
- Frequency is the proportion of baskets that contain the items of interest. The rules can be used in pricing strategies, product placement, and various types of cross-selling strategies.
- Market basket analysis takes data at transaction level, which lists all items bought by a customer in a single purchase.
- The technique determines relationships of what products were purchased with which other product(s). These relationships are then used to build profiles containing If-Then rules of the items purchased.
- The rules could be written as : **If {A} Then {B}**
- The If part of the rule (the {A} above) is known as the antecedent and the THEN part of the rule is known as the consequent (the {B} above).
- The antecedent is the condition and the consequent is the result. The association rule has three measures that express the degree of confidence in the rule, Support, Confidence, and Lift.
- For example, you are in a supermarket to buy milk. Based on the analysis, are you more likely to buy apples or cheese in the same transaction than somebody who did not buy milk ?
- In the following table, there are nine baskets containing varying combinations of milk, cheese, apples, and bananas.

Basket	Product 1	Product 2	Product 3
1	Milk	Cheese	
2	Milk	Apples	Cheese
3	Apples	Banana	
4	Milk	Cheese	
5	Apples	Banana	
6	Milk	Cheese	Banana
7	Milk	Cheese	
8	Cheese	Banana	
9	Cheese	Milk	

- **Support** : Support is the number of transactions that include items in the {A} and {B} parts of the rule as a percentage of the total number of transactions. It is a measure of how frequently the collection of items occur together as a percentage of all transaction.

$$\text{Support} = \frac{A + B}{\text{Total}}$$

- **Confidence :** Confidence of the rule is the ratio of the number of transactions that include all items in {B} as well as the number of transactions that include all items in {A} to the number of transactions that include all items in {A}.

$$\text{Confidence} = \frac{A + B}{A}$$

- **Lift or lift ratio** : It is the ratio of confidence to expected confidence. Expected confidence is the confidence divided by the frequency of B. The Lift tells us how much better a rule is at predicting the result than just assuming the result in the first place. Greater lift values indicate stronger associations.

$$\text{Lift ratio} = \frac{(A + B) / A}{(B / \text{Total})} = \frac{\text{Confidence}}{(B / \text{Total})}$$

- **Leverage :** Leverage measures the difference in the probability of X and Y appearing together compared to statistical independence.

$$\text{Leverage } (X \rightarrow Y) = \text{Support } (X \wedge Y) - \text{support}(X) * \text{Support}(Y)$$

- *Leverage* = 0 if X and Y are statistically independent
- *Leverage* > 0 indicates degree of usefulness of rule
- **Conviction :** The conviction of a rule is defined as :

$$\text{conv } (X \rightarrow Y) = \frac{1 - \text{supp}(Y)}{1 - \text{conf } (X \rightarrow Y)}$$

- The conviction of the rule **X=>Y** can be interpreted as the ratio of the expected frequency that **X** occurs without **Y** if **X** and **Y** were independent divided by the observed frequency of incorrect predictions

5.9.1 Application of Market Basket Analysis

1. **Retail :** In retail, market basket analysis can help determine what items are purchased together, purchased sequentially, and purchased by season. This can assist retailers to determine product placement and promotion optimization. Does it make sense to sell soda and chips or soda and crackers ?

2. **Telecommunications :** Market basket analysis can be used to determine what services are being utilized and what packages customers are purchasing. They can use that knowledge to direct marketing efforts at customers who are more likely to follow the same path. Telecommunications these days is also offering TV and Internet.
3. **Banks :** In financial, market basket analysis can be used to analyze credit card purchases of customers to build profiles for fraud detection purposes and cross-selling opportunities.
4. **Insurance :** Market basket analysis can be used to build profiles to detect medical insurance claim fraud. By building profiles of claims, you are able to then use the profiles to determine if more than one claim belongs to a particular claimee within a specified period of time.
5. **Medical :** Market basket analysis can be used for com or bid conditions and symptom analysis, with which a profile of illness can be better identified. It can also be used to reveal biologically relevant associations between different genes or between environmental effects and gene expression.

5.9.2 Frequent Itemsets, Closed Itemsets and Association Rules

1. A set of items is referred to as an itemset. An itemset is a unordered set of distinct items. An itemset that contains k items is a k-itemset.
2. The set {computer, antivirus software} is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known as the frequency, support count, or count of the itemset.
3. Frequent itemsets that cannot be extended with any item without making them infrequent are called maximal frequent itemsets. Exact support counts of the subsets cannot be directly derived from support of the maximal frequent itemset.

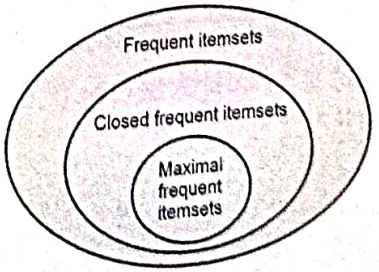


Fig. 5.9.2

Closed Itemsets :

- An alternative approach is to try to retain some of the support information in the compacted representation.
- A closed itemset is an itemset whose all immediate supersets have different support count.
- A closed frequent itemset is a closed itemset that satisfies the minimum support threshold.
- Maximal frequent itemsets are closed by definition.
- An itemset X is closed in a data set S if there exists no proper super-itemset Y such that Y has the same support count as X in S. An itemset X is a closed frequent itemset in set S if X is bothclosed and frequent in S.
- An itemset is closed if none of its immediate supersets has the same support as the itemset.
- Closed **itemset example 1 :**

TID	Items
1	{A, B}
2	{B, C, D}
3	{A, B, C, D}
4	{A, B, D}
5	{A, B, C, D}

Itemset	Support
{A}	4
{B}	5
{C}	3
(D)	4
{A, B}	4
{A, C}	2
{A, D}	3
{B, C}	3
{B, D}	4
{C, D}	3

Itemset	Support
{A, B, C}	2
{A, B, D}	3
{A, C, D}	2
{B, C, D}	3
{A, B, C, D}	2

- Closed itemset are :

{B}, {A, B}, {B, D}, {A, B, D}, {B, C, D}, {A, B, C, D}

- **Closed itemset example 2 :**

TID	Items
100	a, c, d, e, f
200	a, b, e
300	c, e, f
400	a, c, d, f
500	c, e, f

Total frequent itemsets : 20

{a}, {c}, {d}, {e}, {f}, {a, c}, {a, d}, {a, e}, {a, f}, {c, d}, {c, e}, {c, f}, {d, f}, {e, f}, {a, c, d}, {a, c, f}, {a·d·f}, {c, d, f}, {c·e·f}, {a, c, d, f}

Closed frequent Itemssets :

{a, c, d, f}, {c, e, f}, {a, e}, {c, f}, {a}, {e}

5.9.3 Association Rule

- Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories.
- Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a transactional database, relational database or other information repository.
- An example of an association rule would be "If a customer buys a dozen eggs, he is 80 % likely to also purchase milk."
- Association rule mining can be viewed as a two-step process :
  1. **Find all frequent item sets :** By definition, each of these item sets will occur atleast as frequently as a predetermined minimum support count, min sup.
  2. **Generate strong association rules from the frequent item sets :** By definition, these rules must satisfy minimum support and minimum confidence
- An association rule is commonly understood to be an expression of the form : $X \Rightarrow Y$ where X and Y are sets of items such that $X \cap Y = \theta$
- The association rule $X \Rightarrow Y$ means that transactions containing items from set X tend to contain items from set Y.
- Association rules show attribute value conditions that occur frequently together in a given data set. A typical example of association rule mining is market basket analysis.
- Data is collected using bar-code scanners in supermarkets. Such market basket databases consist of a large number of transaction records.
- Each record lists all items bought by a customer on a single purchase transaction. Managers would be interested to know if certain groups of items are consistently purchased together.

- They could use this data for adjusting store layouts, for cross-selling, for promotions, for catalog design, and to identify customer segments based on buying patterns.
- Association rules provide information of this type in the form of if-then statements. These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature.
- In addition to the antecedent (if) and the consequent (then), an association rule has two numbers that express the degree of uncertainty about the rule.
- In association analysis, the antecedent and consequent are sets of items (called itemsets) that are disjoint (do not have any items in common).
- The first number is called the support for the rule. The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule.
- The other number is known as the confidence of the rule. Confidence is the ratio of the number of transactions that include all items in the consequent, as well asthe antecedent (the support) to the number of transactions that include all items in the antecedent.

5.9.4 The Apriori Algorithm

- The Apriori algorithm is an influential algorithm for mining frequent itemsets for boolean association rules.
- Innovative way to find association rules on large scale, allowing implication outcomes that consist of more than one item. It based on minimum support threshold.
- Let $I = \{i_1, i_2, ..., i_n\}$ be a set of n binary attributes called items. Let $D = \{t_1, t_2, ..., t_n\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I.
- A rule is defined as an implication of the form $X \rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. The sets of items X and Y are called antecedent (left-hand-side or LHS) and consequent(right-hand-side or RHS) of the rule respectively.
- The Apriori algorithm is used for mining frequent itemsets and devising association rules from a transactional database. The parameters "support" and "confidence" are used.
- Support refers to items' frequency of occurrence; confidence is a conditional probability.

- To improve the efficiency of level-wise generation of frequent itemsets, an important property is used called Apriori property which helps by reducing the search space.
- Major components of Apriori algorithm are **Support, Confidence and Lift**
- The following are the main steps of the Apriori algorithm :
  1. Calculate the support of item sets (of size k = 1) in the transactional database. This is called generating the candidate set.
  2. Prune the candidate set by eliminating items with a support less than the given threshold.
  3. Join the frequent itemsets to form sets of size k + 1, and repeat the above sets until no more itemsets can be formed. This will happen when the set(s) formed have a support less than the given support.

Pseudocode of Apriori algorithm

L_1= {frequent items};
for (k= 2; L_{k-1} !=∅; k++) **do begin**
  C_k= candidates generated from L_{k-1}(that is: cartesian product L_{k-1} x L_{k-1} and eliminating any
  k-1 size itemset that is not frequent);
  foreach transaction t in database **do**
    increment the count of all candidates in
    C_k that are contained in t
  L_k = candidates in C_k with min_sup
  end
return $\cup_k L_k$;

5.9.5 Limitations of Apriori Algorithm

1. Needs several iterations of the data.
2. Uses a uniform minimum support threshold.
3. Difficulties to find rarely occuring events.
4. Some competing alternative approaches focus on partition and sampling.

Example 5.9.1 *Generate frequent itemsets and generate association rules based on it using aprori algorithm. Minimum support is 50 % and minimum confidence is 70 %.*

TID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

Solution : Apriori algorithm :

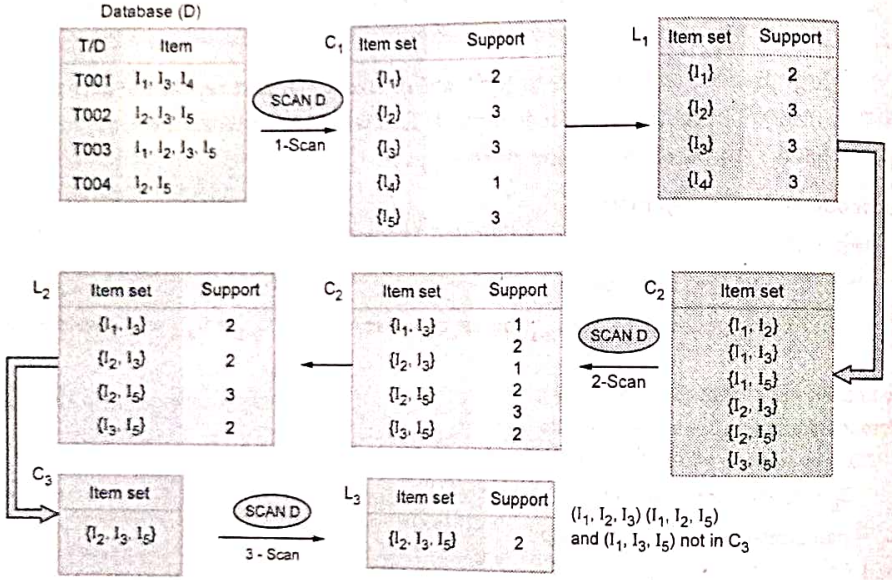


Fig. 5.9.3

Example 5.9.2 *Using Apriori algorithm, generate frequent item sets (min_sup >= 33.3 %) for the following transaction database.*

Trans_id	Itemlist
T_1	{A, B, D, K}
T_2	{A, B, C, D, E}
T_3	{A, B, C, E}
T_4	{B, D}
T_5	{A, C}
T_6	{B, D}

Solution :

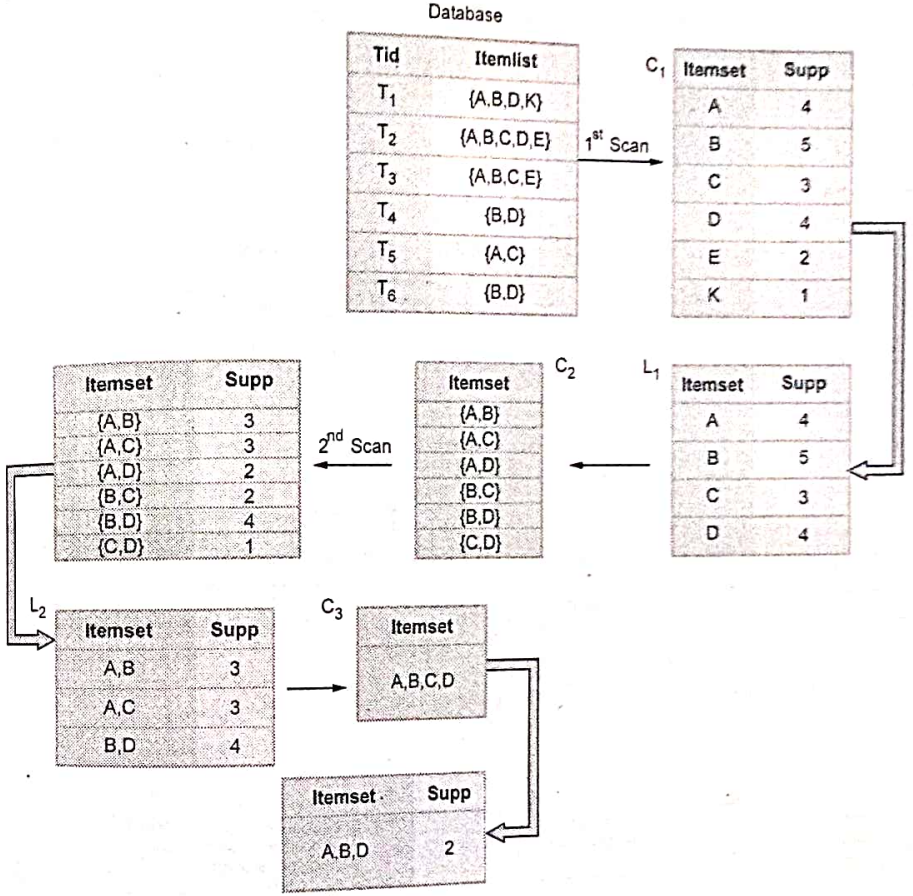


Fig. 5.9.4

Example 5.2.3 *State the Apriori property. Generate candidate itemsets, frequent itemsets and association rules using Apriori algorithm on the following data set with minimum support count is 2.*

	TID	List of items_IDs
1.	T100	11, 12, 15
2.	T200	12, 14
3.	T300	12, 13
4.	T400	11, 12, 14

5.	T500	11, 13
6.	T600	12, 13
7.	T700	11, 13
8.	T800	11, 12, 13, 15
9.	T900	11, 12, 13

Solution :

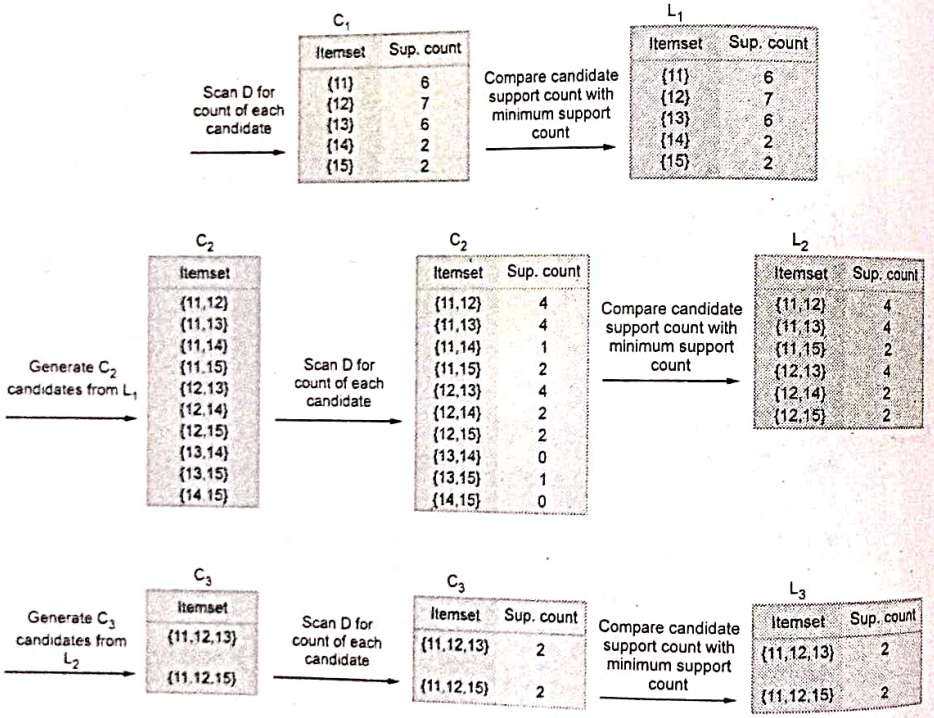


Fig. 5.9.5

Example 5.9.4 *Consider the following set of frequent 3-itemsets: {1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5}. Assume that there are only five items in the data set.*

i) List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori algorithm.

ii) List all candidate 4-itemsets that survive the candidate pruning setup of the Apriori algorithm.

Solution :

Supports for 1-Itemssets :

Item	Support
1	5
2	5
3	6
4	4
5	4

i) List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori : .

{1, 2, 3, 4}, {1, 2, 3, 5}, {1, 2, 4, 5,}, {2, 3, 4, 5} {2, 3, 4, 6}

ii) List all candidate 4-itemsets that survive the candidate pruning setp of the Apriori algorithm.

{1, 2, 3, 4}

Example 5.9.5 *A database has five transactions. Let minimum support is 60 %.*

TID	*Items*
1	*Butter, Milk*
2	*Butter, Dates, Balloon, Eggs*
3	*Milk, Dates, Balloon, Cake*
4	*Butter, Milk, Dates, Balloon*
5	*Butter, Milk, Dates, Cake*

Find all the frequent item sets using Apriori algorithm. Show each step.

Solution :

- Database is scanned once to generate frequent 1 – itemsets. To do this, use absolute support, where duplicate values are counted only once per TID.

Itemset	Support	Support %
{Butter}	4	80 %
{Milk}	4	80 %
{Dates}	4	80 %
{Balloon}	3	60 %
{Eggs}	1	20 %
{Cake}	2	40 %

5.	T500	11, 13
6.	T600	12, 13
7.	T700	11, 13
8.	T800	11, 12, 13, 15
9.	T900	11, 12, 13

Solution :

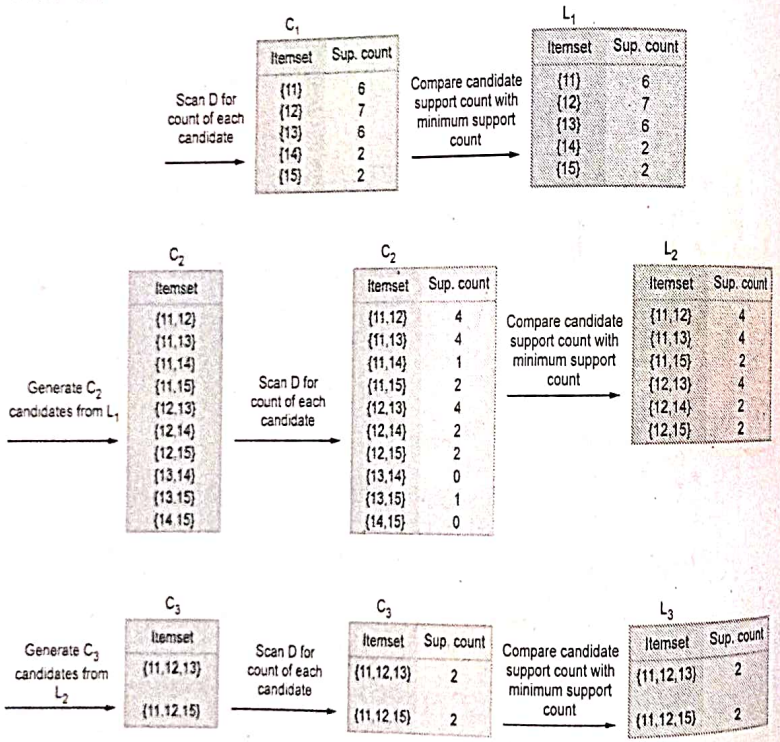


Fig. 5.9.5

Example 5.9.4 *Consider the following set of frequent 3-itemsets: {1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5}. Assume that there are only five items in the data set.*

i) List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori algorithm.

ii) List all candidate 4-itemsets that survive the candidate pruning setup of the Apriori algorithm.

Solution :

Supports for 1-Itemssets :

Item	Support
1	5
2	5
3	6
4	4
5	4

i) List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori : .

{1, 2, 3, 4}, {1, 2, 3, 5}, {1, 2, 4, 5,}, {2, 3, 4, 5} {2, 3, 4, 6}

ii) List all candidate 4-itemsets that survive the candidate pruning setp of the Apriori algorithm.

{1, 2, 3, 4}

Example 5.9.5 *A database has five transactions. Let minimum support is 60 %.*

TID	Items
1	Butter, Milk
2	Butter, Dates, Balloon, Eggs
3	Milk, Dates, Balloon, Cake
4	Butter, Milk, Dates, Balloon
5	Butter, Milk, Dates, Cake

Find all the frequent item sets using Apriori algorithm. Show each step.

Solution :

- Database is scanned once to generate frequent 1 – itemsets. To do this, use absolute support, where duplicate values are counted only once per TID.

Itemset	Support	Support %
{Butter}	4	80 %
{Milk}	4	80 %
{Dates}	4	80 %
{Balloon}	3	60 %
{Eggs}	1	20 %
{Cake}	2	40 %

- The total number of TID is 5, so minimum support of 60% is equivalent to 3/5. Thus itemsets with 1 or 2 support counts are eliminated.

Itemset	Support	Support %
{Butter}	4	80 %
{Milk}	4	80 %
{Dates}	4	80 %
{Balloon}	3	60 %

- Now, database is scanned second time to generate frequent 2 – itemsets. Using absolute support, each combination is counted per TID, and combinations that are below support value of 3 are eliminated.

Itemset	Support	Support %
{Butter, Milk}	3	60 %
{Butter, Dates}	3	60 %
{Butter, Balloon}	2	40 %
{Butter, Cake}	1	20 %
{Butter, Eggs}	1	20 %
{Milk, Dates}	3	60 %
{Milk, Balloon}	2	40 %
{Milk, Cake}	2	40 %
{Dates, Balloon}	3	60 %
{Dates, Eggs}	1	20 %
{Dates, Cake}	2	40 %
{Balloon, Cake}	1	20 %
{Balloon, Eggs}	1	20 %
{Eggs, Cake}	0	0

2 - itemset results, consolidated :

Itemset	Support	Support %
{Butter, Milk}	3	60 %
{Milk, Dates}	3	60 %
{Dates, Dates}	3	60 %
{Balloon, Balloon}	3	60 %

To generate frequent 3 – itemsets :

Itemset	Support	Support %
{Butter, Milk, Dates}	2	40 %
{Milk, Dates, Balloon}	2	40 %

5.9.6 Challenges of Frequent Pattern Mining

Challenges :

1. Multiple scans of transaction database
2. Huge number of candidates
3. Tedious workload of support counting for candidates.

- Improving Apriori : General ideas
  1. Reduce number of transaction database scans
  2. Shrink number of candidates
  3. Facilitate support counting of candidates.

5.10 Improving Apriori Efficiency

- Apriori algorithm is a classical algorithm of association rule mining and widely used for generating frequent item sets. This classical algorithm is inefficient due to so many scans of database. And if the database is large, it will take too much time to scan the database. To overcome these limitations, researchers have made a lot of improvements to the Apriori.
- Techniques for improving efficiency of Apriori algorithm are as follows :
  1. **Hash based technique** : It is used to reduce the size of the candidate k-itemsets(C_k) fork > 1. These techniques work by creating a dictionary (hash table) that stores the candidate item sets as keys, and the number of appearances as the value. Initialization start with zero and Increment the counter for each item set that you see in the data.

2. **Transaction reduction :** In this approach, the number of transactions to be scanned is greatly reduced when comparing to the original Apriori algorithm by reducing the number of similar transactions in the database and this results in reduction of time.
3. **Partitioning :** The partitioning algorithm divides the transactional dataset D into "n" non-overlapping partitions, D1, D2...Dn. The algorithm reduces the number of datasets process to two phases. During the first phase, the algorithm finds all item sets in each partition. Those local frequent item sets are collected into the global candidate item sets. During the second phase, these global item sets are counted to determine if they are large across the entire dataset. The partitioning algorithm improves the performance of finding frequent item sets and also provide several advantages. Small partitions might be fit into main memory than large one. Because the size of each partition is small, the algorithm might reduce the size of candidate item sets.
4. **Sampling :** The basic idea of the sampling approach is to pick a random sample S of the given data D, and then search for frequent itemsets in S instead of D.

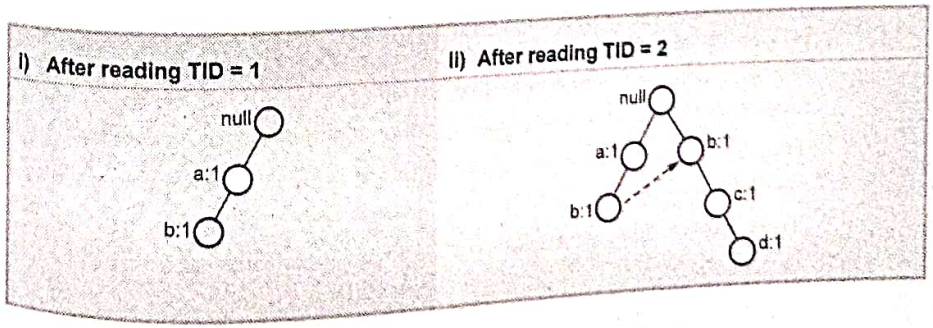
5.11 Mining Frequent Itemset without Candidate Generation

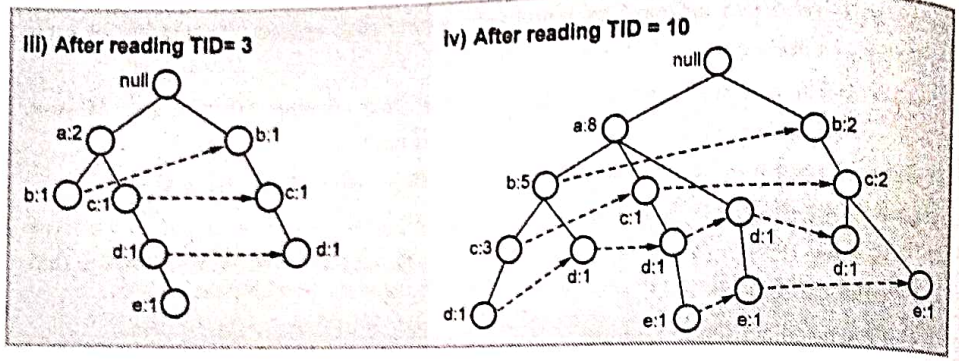
- FP-Growth Algorithm was introduced by Han, Pei and Yin in 2000 to eliminate the candidate generation of Apriori algorithm.
- FP-growth algorithm using a root-like data structure and divide and conquer strategy to find candidate, this makes the FP-Growth algorithm as an efficient algorithm to find rules.
- FP growth algorithm represents the database in the form of a tree called a frequent pattern tree or FP tree. This tree structure will maintain the association between the itemsets.
- A FP-tree is a compact data structure that represents the data set in tree form. Each transaction is read and then mapped onto a path in the FP-tree. This is done until all transactions have been read. Different transactions that have common subsets allow the tree to remain compact because their paths overlap.
- The construction of a FP-tree is subdivided into three major steps :
  1. Scan the data set to determine the support count of each item, discard the infrequent items and sort the frequent items in decreasing order.
  2. Scan the data set one transaction at a time to create the FP-tree. For each transaction :

   - If it is a unique transaction form a new path and set the counter for each node to 1.
   - If it shares a common prefix itemset then increment the common itemset node counters and create new nodes if needed.
  3. Continue this until each transaction has been mapped unto the tree.
- FP-growth algorithm can reduce memory and time used to find association rules because the FP-growth algorithm only needs to scan the database two times to find rules candidates.
- FP-tree construction example :

Transaction data set :

TID	Items
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}





5.11.1 Advantages and Disadvantages of FP-Growth

Advantages :

1. This algorithm needs to scan the database only twice when compared to Apriori which scans the transactions for each iteration.
2. The pairing of items is not done in this algorithm and this makes it faster.
3. The database is stored in a compact version in memory.
4. It is efficient and scalable for mining both long and short frequent patterns.

Disadvantages :

1. FP Tree is more cumbersome and difficult to build than Apriori.
2. It may be expensive.
3. When the database is large, the algorithm may not fit in the shared memory.

5.11.2 Difference between FP-Growth and Apriori Algorithm

Sr. No.	FP-growth	Apriori algorithm
1.	FP-Growth algorithm using a root-like data structure anddivide and conquer strategy to find candidate.	Apriori algorithm using a Brute-force strategyto find data patterns by scanning the database repeatedly.
2.	FP-growth algorithm works better with a small dataset.	Apriori algorithm works better with a big dataset.
3.	There is no candidate generation.	Apriori algorithm uses candidate generation.
4.	FP growth generates pattern by constructing a FP tree.	Apriori generates pattern by pairing the items into singletons, pairs and triplets.
5.	Scan the database only two times.	Multiple scan for generating candidate sets.
6.	It is tree based algorithm.	It is array based algorithm.
7.	FP Growth uses a depth-first search.	Apriori uses a breadth-first search.

5.12 Mining Various Kind of Association Rules

- Association rules generated from mining data at multiple levels of abstraction are called multiple-level or multilevel association rules.
- Multilevel association rules can be mined efficiently using concept hierarchies under a support-confidence framework.
- When a uniform minimum support threshold is used, the search procedure is simplified. The method is also simple in that users are required to specify only one minimum support threshold.
- For each level, any algorithm for discovering frequent itemsets may be used, such as Apriori or its variations are as follows :
  1. Using uniform minimum support for all levels (referred to as uniform support)
  2. Using reduced minimum support at lower levels (referred to as reduced support)
  3. Using item or group-based minimum support (referred to as group based support)

❑❑❑

Notes

Unit VI

6 BI Applications

Syllabus

Tools for Business Intelligence, Role of analytical tools in BI, Case study of Analytical Tools: WEKA, KNIME, Rapid Miner, R;
Data analytics, Business analytics, ERP and Business Intelligence, BI and operation management, BI in inventory management system, BI and human resource management, BI Applications in CRM, BI Applications in Marketing, BI Applications in Logistics and Production, Role of BI in Finance, BI Applications in Banking, BI Applications in Telecommunications, BI in salesforce management.

Contents

6.1 Tools for Business Intelligence
6.2 Case Study of Analytical Tools
6.3 Data Analytics
6.4 Business Analytics
6.5 How Various forms of BA are Supported in Practice
6.6 ERP and Business Intelligence
6.7 BI and Operation Management
6.8 BI in Inventory Management System
6.9 BI and Human Resource Management
6.10 BI Applications in CRM
6.11 BI Applications in Marketing
6.12 BI Applications in Logistics and Production
6.13 Role of BI in Finance
6.14 BI Applications in Banking
6.15 BI Applications in Telecommunications
6.16 BI in Salesforce Management

6.1 Tools for Business Intelligence

- Business intelligence tools are types of application software that collect and process large amounts of unstructured data from internal and external systems. By utilizing modern and professional BI tools, each challenge can be addressed promptly by any business user, without the need for massive IT involvement.
- These tools step up into collecting, analyzing, monitoring and predicting future business scenarios by creating a clear perspective of all the data a company manages. Identifying trends, enabling self-service analytics, utilizing powerful visualizations and offering professional BI dashboards are becoming the standard in business operations, strategic development and ultimately, indispensable tools in increasing profit.
- BI tools are types of software used to gather, process, analyze and visualize large volumes of past, current and future data in order to generate actionable business insights, create interactive reports and simplify the decision-making processes.
- These business intelligence platforms include key features such as data visualization, visual analytics, interactive dash boarding and KPI scorecards. Additionally, they enable users to utilize automated reporting and predictive analytics features based on self-service all of this in one single solution, which makes the analysis process efficient and accessible.

6.1.1 Role of Analytical Tools in BI

- Business intelligence tools can be used by all teams at a company, including sales, marketing and customer support. Team members and executives can both make use of BI tools' output. Data engineers and data analysts can also make use of the convenience of a BI tool when performing their own investigations.
- Examples of how business intelligence is used include :
  a) Visualize the volume of visitors and users on a website over time
  b) Track potential customers through a sales pipeline
  c) Measure performance of business metrics against benchmarks and goals
  d) Evaluate performance of marketing campaigns and experiments
  e) Segment users by demographic characteristics
  f) Generate reports for team and executive decision-making.
- While business intelligence tools also collect and display aggregate data, business analytics tools go a step further to not only report the results of the data, but explain

why the results occurred to help identify weaknesses, fix potential problem areas, alert decision makers to unforeseen events and even forecast future results based on decisions the company might make. This gives organizations the understanding and confidence to achieve business goals, keep the company competitive and increase overall customer satisfaction.

6.2 Case Study of Analytical Tools

- A case study analysis requires user to investigate a business problem, examine the alternative solutions and propose the most effective solution using supporting evidence.

6.2.1 WEKA

- WEKA stands for Waikato Environment for Knowledge Learning. It was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from different domains.
- WEKA is a data mining visualization tool which contains collection of machine learning algorithms for data mining tasks. It is an open source software issued under the GNU General Public License. It provides result information in the form of chart, tree, table etc.
- The GUI Chooser application allows user to run five different types of applications as listed below :
  a) Explorer b) Experimenter c) KnowledgeFlow
  d) Workbench e) Simple CLI
- We can download Weka from the official website http://www.cs.waikato.ac.nz/ml/weka/.
- Execute the following commands at the command prompt to set the WEKA environment variable for Java, as follows :

setenv WEKAHOME /usr/local/weka/weka-3-0-2

setenv CLASSPATH $WEKAHOME/weka.jar:$CLASSPATH

- Once the download is completed, run the exe file and choose the default set-up.

Features :

1. **Data preprocessing :** It is cleaning of data while data gathering and selection phase. It removes/adds default value to missing fields and resolve conflicts.
2. **Data classification and prediction :** It classifies data based on relationship between things and predicts data label. **Example :** Bank, based on available

data of loan, classifies and predicts customer label 'risky' or 'safe'.

3. **Clustering :** Group of related data into cluster, used to discover distinct group. For example we have data of weather and based on that we want to decide whether to play outside or not, in such case, using Weka tool we can visualize overall data and can make decision according to the charts.
4. **Associate :** Association rules highlight all the associations and correlations between items of a dataset.

- WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values.
- The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces.
- WEKA is a collection of tools for regression, clustering, association, data pre-processing, classification and visualisation. Fig. 6.2.1 shows starting with WEKA.

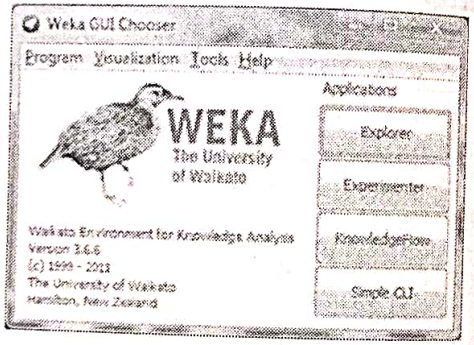


Fig. 6.2.1 Starting with WEKA

- There are four options available on this initial screen.
  1. **Explorer -** The graphical interface used to conduct experimentation on raw data
  2. **Simple CLI -** Provides users without a graphic interface option the ability to execute commands from a terminal window.
  3. **Experimenter -** This option allows users to conduct different experimental variations on data sets and perform statistical manipulation.
  4. **Knowledge flow -** Basically the same functionality as explorer with drag and drop functionality. The advantage of this option is that it supports incremental learning from previous results.
- Fig. 6.2.2 shows the opening screen with the available option.

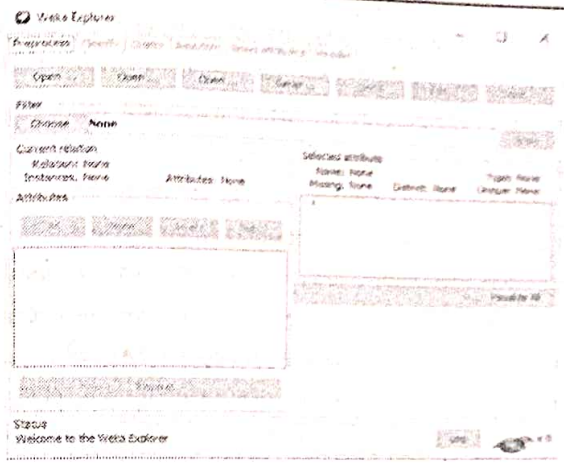


Fig. 6.2.2 Opening screen with the available option

- At first there is only the option to select the pre-process tab in the top left corner. This is due to the necessity to present the data set to the application so it can be manipulated. After the data has been pre-processed the other tabs become active for use.
- Weka uses the attribute relation file format for data analysis, by default. But some of the formats that Weka supports are CSV, ARFF and database using ODBC.
- There are six tabs :
  1. **Pre-process -** Used to choose the data file to be used by the application
  2. **Classify -** Used to test and train different learning schemes on the pre-processed data file under experimentation.
  3. **Cluster -** Used to apply different tools that identify clusters within the data file.
  4. **Association -** Used to apply different rules to the data file that identify association within the data.
  5. **Select attributes -** Used to apply different rules to reveal changes based on selected attributes inclusion or exclusion from the experiment.
  6. **Visualize -** Used to see what the various manipulation produced on the data set in a 2D format, in scatter plot and bar graph output.

6.2.2 KNIME

- KNIME analytics platform is an open source software that allows users to access, blend, analyze and visualize data, without any coding. In 2004, a team of expert

software engineers from the University of Konstanz in Baden-Württemberg, Germany, developed the innovative KNIME business analytics platform.

- It integrates various components for data mining and machine learning via its modular data pipelining concept.
- KNIME is written in Java and based on Eclipse, which comprises an integrated development environment and an extensible plug-in system. The software makes designing data science workflows and reusable components accessible to everyone.
- KNIME analytics platform lets users combine data from different sources. It can open and combine data from simple text formats such as CSV, XLS, XML, PDF, or JSON. The software can also work with unstructured data types such as images, documents and networks, as well as time series data. It can connect to a host of databases and data warehouses to integrate data.
- Access and retrieve data from cloud apps, repositories and services such as Salesforce, SharePoint, Azure, AWS, Google Sheets, or Twitter.
- KNIME's platform enables you to build machine learning models using visual nodes. Use the models for classification, regression, or clustering. Fig. 6.2.3 shows KNIME server.

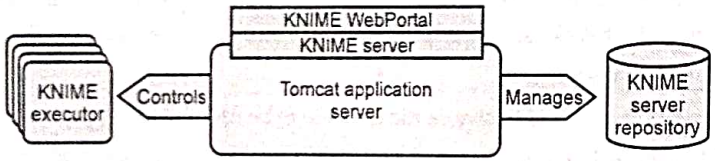


Fig. 6.2.3 KNIME server

- KNIME server is a Java enterprise application and the KNIME WebPortal a standard Java web application, both installed on a Tomcat application server, the box in the middle of Fig. 6.2.3. Users can log in to the server and the server will authenticate against any authentication source provided by Tomcat.
- One of the main tasks of KNIME server is to manage and control the server's repository. Workflows uploaded to the server go through the server application and are stored in the repository which is just a folder on the server's file system. Access to the stored workflows is controlled in KNIME server and access rights for the workflows can be manipulated from KNIME explorer once the client side server extensions are installed.

- Workflow execution on the server is carried out by a KNIME executor. The KNIME executor is a persistent headless instance of a normal KNIME analytics platform application.
- It is important to note that workflows can only be successfully loaded and executed on the server, if the executor has the required features installed and is of the same version (or newer) than the KNIME Analytics Platform version that was used to create the workflow.
- KNIME provides a collection of nodes which can assist with predictive analytics. These are easily output as charts and graphs.
- KNIME analytics platform is also integrated with different open-source projects such as H2O, ScikitLearn and Keras machine learning algorithms. These integrations support deep learning and wrappers for calling codes. It also provides nodes to help users run Python, Java, Perl and other coding scripts. It supports different web-based reporting techniques and offers a complete design of workflow.
- Pros of KNIME :
  1. No license fee
  2. Easy to understand and learn
  3. Open architecture.
- Cons of KNIME
  1. User interface is not that efficient
  2. Lack of learning resources.

6.2.3 RapidMiner

- RapidMiner is a free of charge, open source software tool for data and text mining. In addition to Windows operating systems, RapidMiner also supports Macintosh, Linux and UNIX systems. It is available as a stand-alone application for data/text analysis and as a data/text mining engine for the integration into your own products.
- Rapid Miner provides its own collection of datasets but it also provides options to set up a database in the cloud for storing large amounts of data. You can store and load the data from Hadoop, Cloud, RDBMS, NoSQL etc. Apart from this, you can load your CSV data very easily and start using it as well.
- The standard implementation of procedures like data cleaning, visualization, pre-processing can be done with drag and drop options without having to write even a

single line of code. Rapid miner provides a wide range of machine learning algorithms in classification, clustering and regression as well.

- There are many products of RapidMiner that are used to perform multiple operations. Some of the products are RapidMiner Studio, RapidMiner Auto Model, RapidMiner Turbo Prep, RapidMiner Go, RapidMiner Server and RapidMiner Radoop.
- With RapidMiner Studio, one can access, load and analyze both traditional structured data and unstructured data like text, images and media. It can also extract information from these types of data and transform unstructured data into structured.
- RapidMiner Auto Model solves prediction, clustering and outlier's problem. The auto model provides an evaluation of data, offers relevant models for problem-solving and once the calculations are completed, it compares the results of these models.
- RapidMiner Turbo Prep ensures to assemble every piece of important data together, eliminates worthless data, transforms the remaining data into a consistent and useful format, and presents the result.
- RapidMiner Go helps user to understand different model types through a series of charts and visualizations and easily get the models into production.
- In RapidMiner server, version management and shared repositories help in collaborating, creating interactive apps, and visualizing results locally or remotely using HTML5 charts and maps.
- RapidMiner Radoop is designed to eliminate the complexity of data science on Hadoop and Spark. Now, it is very easy to code machine learning for Hadoop and Spark, create predictive models with the help of RapidMiner Studio visual workflow designer.

6.2.4 R

- R is a programming language for statistical computing and statistical graphics. It is a popular data mining tool for scientists, researchers and students. R has strong object-oriented programming facilities which allow the extension of the software as soon as one has mastered the R language.
- For usage of R as a BI production tool, the package DBI offers an interface to relational database systems. For big data, a number of solutions are provided. The package data.table is a fast tabulation tool as long as the data fit in the memory.

- Basically R is command line oriented, but a number of GUIs exist. For the development, RStudio offers an IDE for data mining the Rattle GUI can be used, and Revolution Analytics provides a visual studio-based IDE.
- The R programming language gets used by many quantitative analysts as a programming tool since it is useful for data importing and cleaning.
- R packages are defined as collections of R functions, sampled data, documentation, and compiled code. These elements are stored in a directory called "library" within the R environment and are installed by default during installation.
- **Features :**
  a) R is an open-source tool
  b) R is also a cross-platform compatible language
  c) R is a great visualization tool
  d) R is used for data science and machine learning tasks.
- The reason why R should be used in data analysis is because it helps in processing large number of commands together, saves all the data and progress on work and enables analysts to easily edit small mistakes so that they do not have to go through different commands to retrace their steps and find the mistake and then fix it.
- R is designed for statistical computing, with thousands of packages that contain the implementations of almost every available statistical methods, making it meet the first characteristic perfectly.
- R can handle more data points than Excel by default. You can even write R code in a parallel fashion and use inexpensive commodity computers in the cloud to analyze large datasets, like Amazon Web Services (AWS) instances. Finally, user can easily integrate R with web technologies to make analytic web apps.
- Data analysis using R is increasing the efficiency in data analysis, because data analytics using R, enables analysts to process data sets that are traditionally considered large data-sets.

6.3 Data Analytics

- Data analytics enables organizations to analyze all their data (real-time, historical, unstructured, structured, qualitative) to identify patterns and generate insights to inform and in some cases, automate decisions, connecting intelligence and action.
- Data analytics allows organizations to digitally transform their business and culture, becoming more innovative and forward-thinking in their decision-making. Going beyond traditional KPI monitoring and reporting to finding hidden patterns

in data, algorithm-driven organizations are the new innovators and business leaders.

- With collaborative data analytics, companies empower everyone to contribute to business success-from data engineers and data scientists, to developers and business analysts and even business professionals and business leaders. Collaborative data analytics also encourages those both inside and outside an organization to connect and collaborate. For instance, data scientists can work closely with a customer to help them solve their problems in real time using the highly collaborative UI of today's modern analytics.

6.3.1 Data Analytic Lifecycle

- The data analytic lifecycle is designed for big data problems and data science projects. With six phases the project work can occur in several phases simultaneously. The cycle is iterative to portray a real project. Work can return to earlier phases as new information is uncovered.
- According to Dietrich (2013), it is a cyclical life cycle that has iterative parts in each of its six steps :

1) Discovery 2) Pre-processing data
3) Model planning 4) Model building
5) Communicate results 6) Operationalize.

- Fig. 6.3.1 shows data analytic lifecycle.

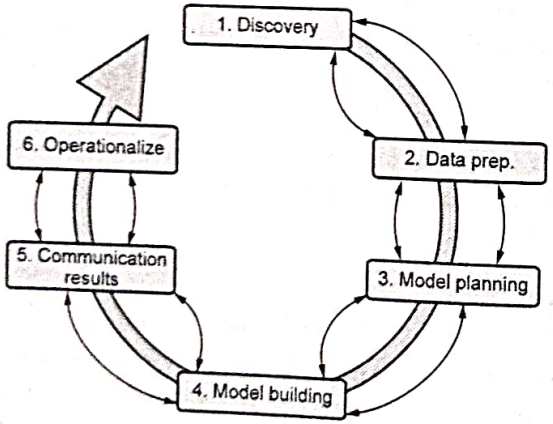


Fig. 6.3.1 Data analytic life cycle

6.3.2 Phase 1 : Discovery

- This phase is all about defining the data's purpose and how to achieve it by the end of the data analytics lifecycle. The stage consists of identifying critical objectives a business is trying to discover by mapping out the data.
- During this process, the team learns about the business domain and checks whether the business unit or organization has worked on similar projects to refer to any learnings.
- In this phase, the team also evaluates technology, people, data and time. For example, while dealing with a small dataset, the team can use excel.
- Phase 1 contains following process :
  a) Learning the business domain
  b) Resources
  c) Framing the problem
  d) Identifying key stakeholders
  e) Interviewing the analytics sponsor
  f) Developing initial hypotheses
  g) Identifying potential data sources.

6.3.3 Phase 2 : Data Preparation

- This stage involves collecting, processing and cleaning data. Here the focus shifts from business requirements to data requirements. In this early phase, data is collected but not analyzed.
- The data preparation phase is generally the most iterative and the one that teams tend to underestimate most often.

a) Preparing the analytic sandbox :

- Create the analytic sandbox. It also called a **workspace.** It allows the team to explore data without interfering with live production data.
- Sandbox collects all kinds of data. The sandbox allows organizations to undertake ambitious projects beyond traditional data analysis and BI to perform advanced predictive analytics.

b) Performing ETLT (Extract, Transform, Load, Transform) :

- The team needs to execute Extract, Load and Transform (ELT) to get data into the sandbox.

- Extract, Transform, Load (ETL) : It transforms the data based on a set of business rules before loading it into the sandbox.
- Extract, Load, Transform (ELT) : It loads the data into the sandbox and then transforms it based on a set of business rules.
- Extract, Transform, Load, Transform (ETLT) : It is the combination of ETL and ELT and has two transformation levels.

c) Learning about the data :

- Data is captured through three main ways :
  - i. **Data acquisition :** Obtaining existing data from outside sources.
  - ii. **Data entry :** Creating new data values from data inputted within the organization.
  - iii. **Signal reception :** Capturing data created by devices.

d) Data Conditioning :

- Data conditioning includes cleaning data, normalizing datasets and performing transformations. It is often viewed as a preprocessing step prior to data analysis; it might be performed by data owner, IT department, DBA, etc.
- Best to have data scientists involved and data science teams prefer more data than too little.

e) Common tools for data preparation :

- Hadoop can perform parallel ingest and analysis.
- Alpine miner provides a graphical user interface for creating analytic workflows.
- OpenRefine is a free, open source tool for working with messy data.
- Similar to OpenRefine, data wrangler is an interactive tool for data cleansing and transformation.

6.3.4 Phase 3 : Model Planning

- The team determines the methods, techniques and workflow it intends to follow for the subsequent model building phase.
- The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.
- Activities to consider :
  - a) Assess the structure of the data - This dictates the tools and analytic techniques for the next phase.

  - b) Ensure the analytic techniques enable the team to meet the business objectives and accept or reject the working hypotheses.
  - c) Determine if the situation warrants a single model or a series of techniques as part of a larger analytic workflow.
  - d) Research and understand how other analysts have approached this kind or similar kind of problem.

a) Data exploration and variable selection

- Explore the data to understand the relationships among the variables to inform selection of the variables and methods. A common way to do this is to use data visualization tools.
- Often, stakeholders and subject matter experts may have ideas. For example, some hypothesis that led to the project.
- Aim for capturing the most essential predictors and variables. This often requires iterations and testing to identify key variables.
- If the team plans to run regression analysis, identify the candidate predictors and outcome variables of the model.

b) Model selection

- The main goal is to choose an analytical technique or several candidates, based on the end goal of the project. We observe events in the real world and attempt to construct models that emulate this behavior with a set of rules and conditions.
- A model is simply an abstraction from reality. Determine whether to use techniques best suited for structured data, unstructured data or a hybrid approach.
- Teams often create initial models using statistical software packages such as R, SAS or Matlab. Which may have limitations when applied to very large datasets.
- The team moves to the model building phase once it has a good idea about the type of model to try.

c) Common tools for the model planning phase

- R programming language has a complete set of modeling capabilities. It contains about 5000 packages for data analysis and graphical presentation.
- SQL analysis services can perform in-database analytics of common data mining functions, involved aggregations and basic predictive models.
- SAS/ACCESS provides integration between SAS and the analytics sandbox via multiple data connections.

6.3.5 Phase 4 : Model Building

- The team develops datasets for testing, training and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase.
- Building a model involves two phases :
  a) **Design the model :** Identify a suitable model. This step can involve a number of different modeling techniques to identify a suitable model. These may include decision trees, regression techniques and neural networks.
  b) **Execute the model :** The model is run against the data to ensure that the model fits the data.
- Common commercial tools for the model building phase :
  a. SAS enterprise miner used for building enterprise-level computing and analytics.
  b. SPSS modeler (IBM) provides enterprise-level computing and analytics.
  c. Matlab is a high-level language for data analytics, algorithms, data exploration.
  d. Alpine miner provides GUI frontend for backend analytics tools.
  e. STATISTICA and MATHEMATICA is popular data mining and analytics tools.

6.3.6 Phase 5 : Communicate Results

- This phase aims to determine whether the project results are a success or failure and start collaborating with significant stakeholders.
- The team identifies the vital findings of their analysis, measures the associated business value and creates a summarized narrative to convey the stakeholder's results.
- Communicate and document the key findings and major insights derived from the analysis. This is the most visible portion of the process to the outside stakeholders and sponsors.

6.3.7 Phase 6 : Operationalize

- This final phase moves data from the sandbox into a live environment. Data is monitored and analyzed to see if the generated model is creating the expected results. If the results are not as expected, you can return to any of the preceding phases to tweak the data.

- The team communicates the benefits of the project more broadly and sets up a pilot project to deploy the work in a controlled way. Risk is managed effectively by undertaking small scope, pilot deployment before a wide-scale rollout.
- During the pilot project, the team may need to execute the algorithm more efficiently in the database rather than with in-memory tools like R, especially with larger datasets.
- To test the model in a live setting, consider running the model in a production environment for a discrete set of products or a single line of business. Monitor model accuracy and retrain the model if necessary.
- Key outputs from successful analytics project
  a) Business user tries to determine business benefits and implications.
  b) Project sponsor wants business impact, risks, ROI.
  c) Project manager needs to determine if project completed on time, within budget, goals met.
  d) Business intelligence analyst needs to know if reports and dashboards will be impacted and need to change.
  e) Data engineer and DBA must share code and document.
  f) Data scientist must share code and explain model to peers, managers, stakeholders.

Review Questions

1. *Explain different phases of data analytics life cycle.*
2. *Explain data analytic life cycle.*
3. *Draw data analytics lifecycle and give brief description about all phases.*
4. *Why communication is important in data analytics lifecycle projects ?*
5. *Demonstrate the overview of data analytics life cycle.*

6.4 Business Analytics

- Business Analytics (BA) is the iterative, methodical exploration of an organization's data, with an emphasis on statistical analysis. Business analytics is used by companies that are committed to making data-driven decisions.
- Business analytics combines the fields of management, business and computer science. The business aspect requires both a high-level understanding of the business as well as the practical limitations that exist. The analytical part requires an understanding of data, statistics and computer science.

- Business analytics is the process of making sense of gathered data, measuring business performance and producing valuable conclusions that can help companies make informed decisions on the future of the business, through the use of various statistical methods and techniques.
- Business analytics utilizes big data, statistical analysis and data visualization to implement organization changes. Predictive analytics is an important aspect of this work as it involves available data to create statistical models.
- These models can be used to predict outcomes and inform decision making. By learning from existing data, business analytics can make concrete recommendations to solve problems and improve businesses.
- Companies use Business Analytics (BA) to make data-driven decisions. The insight gained by BA enables these companies to automate and optimize their business processes. In fact, data-driven companies that utilize business analytics achieve a competitive advantage because they are able to use the insights to :
  1. Conduct data mining.
  2. Complete statistical analysis and quantitative analysis to explain why certain results occur.
  3. Test previous decisions using A/B testing and multivariate testing.
  4. Make use of predictive modeling and predictive analytics to forecast future results.
- Challenges with developing and implementing business analytics are as follows :
  1. **Executive ownership** - Business analytics requires buy-in from senior leadership and a clear corporate strategy for integrating predictive models.
  2. **IT involvement** - Technology infrastructure and tools must be able to handle the data and business analytics processes.
  3. Available production data vs. cleansed modeling data - Watch for technology infrastructure that restrict available data for historical modeling and know the difference between historical data for model development and real-time data in production.
  4. Project Management Office (PMO) - The correct project management structure must be in place in order to implement predictive models and adopt an agile approach.
  5. End user involvement and buy-in - End users should be involved in adopting business analytics and have a stake in the predictive model.

  6. Change management - Organizations should be prepared for the changes that business analytics bring to current business and technology operations.
- Data-driven decision-making process uses the following steps :
  1. Identify the problem or opportunity for value creation.
  2. Identify primary as well secondary data sources.
  3. Pre-process the data for issues such as missing and incorrect data. Generate derived variables and transform the data if necessary. Prepare the data for analytics model building.
  4. Divide the data sets into subsets training and validation data sets.
  5. Build analytical models and identify the best model(s) using model performance in validation data.
  6. Implement solution/Decision/Develop product.

6.4.1 Why Business Analytics

- Business analysts are the people that have the needed knowledge, skills and sources of information to decide on the direction the business needs to take to succeed in the future.
- Commercial organizations use business analytics in order to :
  1. Analyze data from multiple sources
  2. Use advanced analytics and statistics to find hidden patterns in large datasets
  3. Monitor KPIs and react to changing trends in real-time
  4. Justify and revise decisions based on up-to-date information.

6.4.2 Differences between Business Intelligent and Data Science

Sr. No.	Business Intelligent (BI)	Data Science
1.	BI tends to provide reports, dashboards and queries on business questions for the current period or in the past.	Data science tends to use disaggregated data in a more forward-looking, exploratory way, focusing on analyzing the present and enabling informed decision about the future.
2.	BI systems make it easy to answer questions related to quarter-to-date revenue, progress toward quartetly targets and understand how much of a given product was sold in a prior quarter or year.	Data science tends to be more exploratory in nature and may use scenario optimization to deal with more open-ended questions.
3.	BI helps monitor the current state of business data to understand the historical performance of a business.	Data science, as used in business is basically data-driven, where many interdisciplinary sciences are applied together to extract meaning.

4.	BI is designed to handle static and highly structured data.	Data science can handle high-speed, high-volume and complex, multi-structured data from a wide variety of data sources.

6.4.3 Difference between Business Intelligence and Business Analytics

Business Intelligence	Business Analytics
What does it do ?	
Reports on what happened in the past or what is happening in now, in current time.	Investigate why it happened and predict what may happen in future.
How is it achieved ?	
Basic querying and reporting : OLAP cubes, slice and dice, drill-down.	Applying statistical and mathematical techniques.
Interactive display options - Dashboards, scorecards, charts, graphs, alerts.	Identifying relationships between key data variables.
	Reveal hidden patterns in data.
What does your business gain ?	
• Dashboards with "how are we doing" information.	• Response to "what do we do next" ?
• Standard reports and preset KPIs.	• Proactive and planned solutions for unknown circumstances.
• Alert mechanisms when something goes wrong .	• The ability to adapt and respond to changes and challenges.

6.5 How Various forms of BA are Supported in Practice

- The four types of analytics are usually implemented in stages and no one type of analytics is said to be better than the other. They are interrelated and each of these offers a different insight. With data being important to so many diverse sectors-from manufacturing to energy grids, most of the companies rely on one or all of these types of analytics.
- With the right choice of analytical techniques, big data can deliver richer insights for the companies. Before diving deeper into each of these, let's define the four types of analytics :

6.5.1 Descriptive Analytics

- It simple method and used in first phase of analytics, involves gathering, organizing tabulating and depicting data then the characteristics of what we are studying.
- The descriptive model shows relationships between the customer and product/service with the acquired data. This model can be used to organize a customer by their personal preferences for example.
- Descriptive statistics are useful to show things like, total stock in inventory, average dollars spent per customer and year over year change in sales.
- Common examples of descriptive analytics are reports that provide historical insights regarding the company's production, financials, operations, sales, finance, inventory and customers.
- While business intelligence tries to make sense of all the data that's collected each and every day by organizations of all types, communicating the data in a way that people can easily grasp often becomes an issue.
- Data visualization evolved because data displayed graphically allows for an easier comprehension of the information, validating the old adage, "a picture is worth a thousand words."
- In business, proper data visualization provides a different approach to show potential connections, relationships, etc. which are not as obvious in data that's non-visual.
- A business intelligence dashboard is an information management tool that is used to track KPIs, metrics and other key data points relevant to a business, department or specific process.
- Through the use of data visualizations, dashboards simplify complex data sets to provide users with at a glance awareness of current performance.
- Dashboards provide sleek, real-time visibility to your team.
- Combining business intelligence data with dashboards gives your team the at-a-glance view of their performance that they need to run smoothly.
- BI dashboards must be designed carefully though. If the data being fed into the visualizations is not reliable, no matter how easy the dashboard itself is to read and analyze, the dashboard will be useless.
- The goal of BI dashboards is to help business individuals make more informed decisions by enabling companies to gather, analyze, build dashboards and create reports on their most important and business-driving data.

6.5.2 Predictive Analytic

- Predictive analytics helps your organization predict with confidence what will happen next so that you can make smarter decisions and improve business outcomes.
- The purpose of the predictive model is finding the likelihood different samples will perform in a specific way.
- The predictive model typically calculates live transactions multiple times to help evaluate the benefit of a customer transaction.
- Predictive models typically utilize a variety of variable data to make the prediction. The variability of the component data will have a relationship with what it is likely to predict.
- Predictive analytics can be used throughout the organization, from forecasting customer behavior and purchasing patterns to identifying trends in sales activities.
- They also help forecast demand for inputs from the supply chain, operations and inventory.
- Process involved in predictive analytics.
  1. **Project definition :** Identify what shall be the outcome of the project, the deliverables, business objectives and based on that go towards gathering those data sets that are to be used.
  2. **Data collection :** This is more of the big basket where all data from various sources are binned for usage. This gives a picture about the various customer interactions as a single view item.
  3. **Analysis :** Here the data is inspected, cleansed, transformed and modelled to discover if it really provides useful information and arriving at conclusion ultimately.
  4. **Statistics :** This enables to validate if the findings, assumptions and hypothesis are fine to go ahead with and test them using statistical model.
  5. **Modelling :** Through this accurate predictive models about the future can be provided. From the options available the best option could be chosen as the required solution with multi model evaluation.

6. **Deployment :** Through the predicative model deployment an option is created to deploy the analytics results into everyday effective decision. This way the results, reports and other metrics can be taken based on modelling.
7. **Monitoring :** Models are monitored to control and check for performance conformance to ensure that the desired results are obtained as expected.

Examples of predictive analytics :

1. **Retail :** Probably the largest sector to use predictive analytics, retail is always looking to improve its sales position and forge better relations with customers. One of the most ubiquitous examples is Amazon's recommendations. When you make a purchase, it puts up a list of other similar items that other buyers purchased.
2. **Weather :** Weather forecasting has improved by leaps and bounds thanks to predictive analytics models. Today's five-day forecast is as accurate as a one-day forecast from the 1980s. Forecasts as long as nine to 10 days are now possible, and more important, 72-hour predictions of hurricane tracks are more accurate than 24-hour forecasts from 40 years ago.
3. **Social media analysis :** Online social media is a fundamental shift of how information is being produced, particularly as relates to businesses. Tracking user comments on social media outlets enables companies to gain immediate feedback and the chance to respond quickly. Nothing makes a local business jump like a bad review on yelp or makes a merchant respond like a bad review on Amazon. This means collecting and sorting through massive amounts of social media data and creating the right models to extract the useful data.
4. **Health care :** Usage of predictive analytics in the health care domain can aid to determine and prevent cases and risks of those developing certain health related complications like diabetics, asthma and other life threatening ailments. Through the administering of predictive analytics in health care better clinical decisions can be made.
5. **Fraud detection :** Predictive analytics can aid to spot inaccurate credit application, deviant transactions leading to frauds both online and offline, identity thefts and false insurance claims saving financial and insurance institutions of lots of security issues and damages to their operations.

6.5.3 Prescriptive Analytic

- This model suggests a course of action. Prescriptive analytics assists users in finding the optimal solution to a problem or in making the right choice/decision among several alternatives.

- The prescriptive model utilizes an understanding of what has happened, why it has happened and a variety of "what-might-happen" analysis to help the user determine the best course of action to take.
- A prescriptive analysis is typically not just with one individual response but is, in fact, a host of other actions.
- An example of this is a traffic application helping you choose the best route home and taking into account the distance of each route, the speed at which one can travel on each road and crucially, the current traffic constraints.
- Another example might be producing an exam time-table such that no students have clashing schedules.
- Larger companies are successfully using prescriptive analytics to optimize production; scheduling and inventory in the supply chain to make sure that are delivering the right products at the right time and optimizing the customer experience.
- Operations Research (OR) techniques form the core of prescriptive analytics.
- With known parameters, prescriptive analytics not only can anticipate what will happen and when, but it also explains why it will happen. It can automatically improve prediction accuracy and inform the best next step because it can continually take in new data to re-predict and re-prescribe.
- Fig. 6.5.1 shows relation between all analysis.

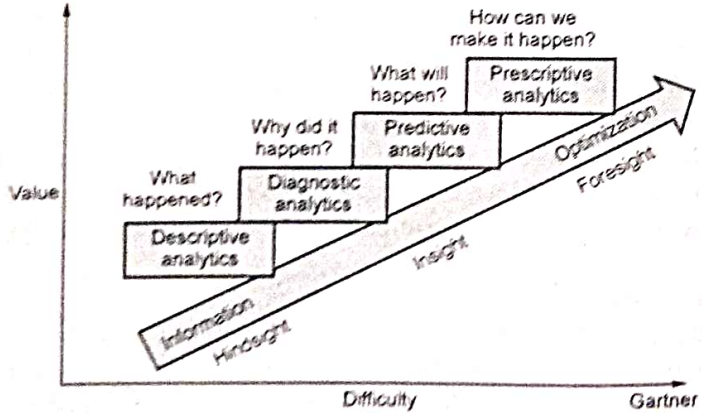


Fig. 6.5.1 Relation between descriptive, predictive and prescriptive analytics

- Organisations can take advantage of the following benefits :
  1. Helps make decisions for the future before decisions have to be made.
  2. Can assist in mitigating risk.

3. Continuously processes new data to give better options.
4. Improve operations - optimise planning, reduce inefficiencies, etc.
5. Optimise production.
6. Schedule inventory and optimise supply chain.

6.5.4 Diagnostic Analytic

- A set of techniques for determine what has happened and why. It uses methods such as correlations factors for the interpretation of factors that contributed to the outcome.
- Here, it is important to verify the possible correlation dependencies that may exist between the input variables. It can be calculated with different way and the less distant to 1 is the result, the more correlated the variables are.
- Using descriptive analytics, user can condense big data into smaller, more useful representations of information. It allows user to understand what happened in the past to inform current decisions.
- Diagnostic analytics takes a deeper look at data to understand the root causes of the events. It is helpful in determining what factors and events contributed to the outcome. It mostly uses probabilities, likelihoods and the distribution of outcomes for the analysis.
- In a time series data of sales, diagnostic analytics would help to understand why the sales have decrease or increase for a specific year or so. However, this type of analytics has a limited ability to give actionable insights. It just provides an understanding of causal relationships and sequences while looking backward.
- A few techniques that uses diagnostic analytics include attribute importance, principle components analysis, sensitivity analysis and conjoint analysis. Training algorithms for classification and regression also fall in this type of analytics.

6.5.5 Difference between Descriptive, Predictive and Prescriptive Analytic

Sr. No.	Descriptive model	Predictive model	Prescriptive model
1.	It use data aggregation and data mining to provide insight into the past and answer.	Use statistical models and forecasts techniques to understand the future and answer.	Use optimization and simulation algorithms to advice on possible outcomes and answer.
2.	What has happened ?"	What could happened ?	What should we do ?
3.	Descriptive analytics is the analysis of past or historical	Predictive analytics predicts future trends.	Prescriptive analytics showcases viable

	data to understand trends and evaluate metrics over time.		solutions to a problem and the impact of considering a solution on future trend.
4.	Examples of tools used : Data aggregation and data mining.	Examples of tools used : Machine learning, statistical models and simulation.	Examples of tools used : Optimization and heuristics.
5.	Used when user want to sumarize results for all or part of your business.	Used when user want to make an educated guess at likely results.	Used when user have importance interdependent, complex or time-sensitive decisions to make.
6.	**Limitation :** Snapshot of the past, often with limited ability to help guide decisions.	**Limitation :** Guess at the future, helps inform low complexity decisions.	**Limitation :** Most effective where user have some control over what is being modeled

6.6 ERP and Business Intelligence

- Enterprise Resource Planning (ERP) is business process management software that allows an organization to use a system of integrated applications to manage the business and automate many back office functions related to technology, services and human resources.
- ERP software typically integrates all facets of an operation including product planning, development, manufacturing, sales and marketing in a single database, application and user interface.
- ERP characteristics :
  1. **Flexibility :** An ERP system should be flexible to respond to the changing needs of an enterprise.
  2. **Modular and open :** ERP system has to have open system architecture.
  3. **Comprehensive :** It should be able to support variety of organizational functions and must be suitable for a wide range of business organizations.
  4. **Best business practices :** It must have a collection of the best business processes applicable worldwide. An ERP package imposes its own logic on a company's strategy, culture and organization.

6.6.1 ERP Features

- **Features of ERP :**
  1. ERP provides multi-platform, multi-facility, multi-mode manufacturing, multi-currency, multi-lingual facilities.
  2. It supports strategic and business planning activities, operational planning and execution activities, creation of materials and resources.
  3. ERP covering all functional areas like manufacturing, selling and distribution, payables, receivables, inventory, accounts, human resources, purchases etc.
  4. ERP performs core activities and increases customer service, thereby augmenting the corporate image.
  5. ERP bridges the information gap across organizations.
  6. ERP provides complete integration of systems not only across departments but also across companies under the same management.
  7. ERP provides intelligent business tools like decision support system.
- **Different functional areas of ERP :** ERP is designed to facilitate the sharing of information across functions to eliminate inconsistency and duplication of effort. In selecting an enterprise resource planning platform, organizations should consider the various ERP modules that align with their strategic, economic and technical goals.

6.6.2 Functional Area of ERP

- Let's take a closer look at some of those functional areas :
  1. **Marketing/Sales :** Sales and marketing departments can track the customer experience from presale activities, which begin with contacting the customer, through the actual dispatch of the customer's order.
- Tasks related to customer visits, expenses, shipping, invoicing, forecasting and competitor analysis can be automated and/or enhanced through an ERP system.
- Employees can contact customers, follow up on invoices and track orders. Additionally, sales and marketing personnel can monitor their individual goals, which also can be collated and analyzed by managers and business partners.
  2. **Customer Relationship Management (CRM) :** ERP platforms also can incorporate CRM modules to focus on how a business communicates with its customers. This may include departments such as sales and marketing and call center support, as well as functions such as customer interaction data, sales pipeline management, lead prioritization and customer retention.

3. **Supply chain management :** ERP modules supporting supply chain management may feature functions for purchasing, product configuration, supplier scheduling, goods inspections, claims processing, warehousing and more. There are also related modules to manage order processing and distribution tasks.
4. **Manufacturing :** Engineering, scheduling capacity, quality control, workflow and product life management are among the core functions that can fall within an ERP system's manufacturing module.
5. **Accounting/Finance :** By automating and streamlining tasks related to budgeting, cost and cash management, activity-based costing and other accounting/finance functions, ERP systems can provide businesses with real-time data and insights on performance while also ensuring compliance with relevant financial regulations.
6. **Human resources :** Human resources modules within an enterprise resource planning system may include tools and dashboards to gather and interpret data on training, recruiting, payroll, benefits, retirement and diversity management. HR managers also can monitor and measure key performance indicators (KPIs) for individual employees, job roles and departments.

6.6.3 ERP Benefits

- **Benefits of ERP :**

1. Improves timeliness of information by permitting posting daily instead of monthly.
2. Greater accuracy of information with detailed content, better presentation, satisfactory for the auditors.
3. Improved cost control.
4. Faster response and follow-up on customers.
5. More efficient cash collection, say, material reduction in delay in payments by customers.
6. Better monitoring and quicker resolution of queries.
7. Enables quick response to change in business operations and market conditions.
8. Helps to achieve competitive advantage by improving its business process.
9. Improves supply-demand linkage with remote locations and branches in different countries.
10. Provides a unified customer database usable by all applications.

6.7 BI and Operation Management

- Operations Management (OM) is the administration of business practices to create the highest level of efficiency possible within an organization. It is concerned with converting materials and labor into goods and services as efficiently as possible to maximize the profit of an organization. Operations management teams attempt to balance costs with revenue to achieve the highest net operating profit possible.
- Operations management involves utilizing resources from staff, materials, equipment and technology. Operations managers acquire, develop and deliver goods to clients based on client needs and the abilities of the company.
- Operations management handles various strategic issues, including determining the size of manufacturing plants and project management methods and implementing the structure of information technology networks. Other operational issues include the management of inventory levels, including work-in-process levels and raw materials acquisition, quality control, materials handling and maintenance policies.
- A critical function of operations management relates to the management of inventory through the supply chain. This process is known as **operations and supply chain management (OSCM)**. To be an effective operations management professional, one must be able to understand the processes that are essential to what a company does and get them to flow and work together seamlessly. The coordination involved in setting up business processes in an efficient way requires a solid understanding of logistics.
- Operations managers are involved in coordinating and developing new processes while revaluating current structures. Operations Management (OM) is concerned with controlling the production process and business operations in the most efficient manner possible. OM professionals attempt to balance operating costs with revenue to maximize net operating profit.
- Business Process Redesign (BPR), which is focused on analyzing and designing workflow and business processes within a company. The goal of BPR is to help companies dramatically restructure the organization by designing the business process from the ground up.
- Reconfigurable manufacturing systems, designed to incorporate accelerated change in structure, hardware and software components. This allows systems to adjust rapidly to the capacity to which they can continue production and how efficiently they function in response to market or intrinsic system changes.

- Operational Intelligence (OI) is an approach to data analysis that enables decisions and actions in business operations to be based on real-time data as it is generated or collected by companies. Typically, the data analysis process is automated and the resulting information is integrated into operational systems for immediate use by business managers and workers.

6.8 BI in Inventory Management System

- The "inventory" stands for a complete list of goods owned or stored either to resell or as a raw material for producing the final product and then, in turn, sell the final product. "Inventory is defined as those stocks used to support production, such as raw material and work in process, supporting activities, such as maintenance, repair and operating supplies and finally customer service in the form of finished goods and spare parts."
- Inventory management system is all about ordering the right quantity of products and keeping track of all the company's goods and storing them in an appropriate facility for easy retrieval while selling them.
- BI in inventory control uses software to transform your stock data into actionable insights that support the daily operations of your warehouse. Dealing with inventory paperwork can be time-consuming. The process can be tiring as well especially when done manually. But with an inventory management software, you can reduce the entire process into one central point.
- However, data-driven inventory management systems are quickly becoming the new norm. These systems give businesses the ability to track their inventory in real-time, identify patterns and trends and make more informed decisions about stocking levels.
- Business Intelligence (BI) tools are designed to help users collect, analyze and present data in a variety of different formats. These tools can be used for a wide range of tasks, including data analysis, report building and data visualization.
- For inventory management systems, this data includes :
  1. **Sales data :** This data includes information about the products that have been sold, such as the product name, description, quantity and price. This data can be used to track inventory levels and sales trends.
  2. **Purchase orders :** This data includes information about the products that have been ordered, such as the product name, description and quantity. This data can be used to track inventory levels and purchase trends.

3. **Supplier data :** This data includes information about the suppliers of the products, such as the supplier name, contact information and website. This data can be used to track supplier information and trends.
4. **Shipping data :** This data includes information about the products that have been shipped, such as the product name, description, quantity and shipping date. This data can be used to track inventory levels and shipping trends.

6.9 BI and Human Resource Management

- Human resources management is one of the key areas of the organization strategy, which affects the achievement of its business objectives. That means that contemporary HR managers have to apply the new information technology tools and methods of data analysis which enable to find the relationships between people and organization's outcomes more effectively.
- Applying business intelligence in HR helps in executing extensive manpower assessments, preparing account reports, employee performance reports, evaluating wages, staffing, available jobs and termination rules. This eventually helps the organization in making advanced choices that joins the staff with the corporate goals.
- BI is used in HR to enhance outcomes across all divisions of the firm right from applicant selection, performance assessment, value control, maintenance and profitability. The platform collects the significant data and converts it into commercial acumen that assists the extensive organizational plan.
- **Application of Business Intelligence in HRM :**
  1. BI helps in acquiring right candidate in shortest span of time using economical measures.
  2. It helps in organizing staff and segmenting those who are consistent performers.
  3. It offers prospects for expansion such as training, counseling and on site experience, etc.
  4. It assists in retaining top performer by locating the significant talent within the firm.
  5. It helps in supporting the appraisal and incentives process in accordance to corporate objective.
  6. It helps in supervising major metrics like value per staff, revenue, staffing, demographics and training efficiency.

7. Inspecting opportunities for enhancement in context of enrollment, abrasion and maintenance.
8. Reducing the managerial weight of labor-intensive methods engaged with spreadsheets.

6.10 BI Applications in CRM

- Customer relationship management is a term that refers to practices, strategies and technologies that companies use to analyze customer interactions. Also, it includes proceeding data with the goal of improving business relationships.
- In fact, by integrating CRM into the systems, we add a personal touch to the communication with the end-user. The successful company in this crowded market is the one who puts their customers first.
- BI helps with reporting, forecasting and predictive analytics. In business intelligence organizations use software applications to analyze raw data. This information is useful to any organization as it can help with critical decision making. In other words, it can help cut costs, improve decision making and identify new business opportunities.
- With CRM data, BI can help unlock various trends around marketing. We get to serve our customers based on their personas, buying behavior and industry influence. This way we can engage better with them and influence their opinions in favor of your brand.
- By doing a BI analysis, a business can recognize customer behavior when it comes to interacting with them. Building on this report, a CRM system can proactively engage with these customers through e-mails, calls, sales meetings and provide enough relevant information to upgrade their status from a casual buyer to a loyal and frequent customer.
- For example, if a BI analysis report states that a customer viewed a product on the website but decided not to buy it, the business can automate e-mails that offer a cheaper or a better alternative. This lets the customer know that the company is looking out for them and makes them more likely to purchase again.

Create Loyalty Programs :

- A business can utilize a BI analysis to figure out which of their customers are profitable to the company and provide high value to their business. When the CRM system takes over, the company can work closely with these customers and provide them with loyalty benefits, which is not only beneficial for providing more business

through probable referrals, but also increases customer satisfaction and the customer retention rate.

- For example, if a customer is seen to be regularly purchasing novels from a bookstore, the managers can provide them with a membership card that may give a flat 10 % discount on every third purchase. This makes the customer happy and more likely to recommend the bookstore to their acquaintances through the word-of-mouth.
- Customer service and customer satisfaction are the backbone of customer relationships. If an organization can accurately monitor and measure customer service factors and customer satisfaction, it is easier to make appropriate corrections and ensure customer retention, good client references and new customer acquisition.
- To identify the problems affecting customer service, one must understand delays in delivery, response time for information requests, the number and nature of complaints and claims and numerous other factors.
- To effectively manage customer relationships, the organization must also set priorities to serve key customers and to understand buying behavior and customer satisfaction for various market segments, client profiles, products and services.
- Team members must work with consistent goals and expectations to satisfy service expectations and be accountable for established objectives. Where results are declining or customer satisfaction is poor, the organization must quickly assess the root cause of the problem and adapt processes, training, policies and procedures to ensure that critical business initiatives are not put at risk.

6.11 BI Applications in Marketing

- Business intelligence can provide valuable insight into a company's operations and using a BI solution to gather information about the marketing department.
- Business intelligence can provide useful information to refine and target advertising and other marketing strategies on virtually any platform, whether in print, on TV or over the Internet.
- Here are a few ways business intelligence can helps to marketing department :

1. **Look at the right data :** Business intelligence software can help you look at nearly any facet of your business. With so much information at your disposal, it may be a challenge to determine which metrics are the most important for your marketing department.

- The Huffington post suggested starting with the basics - The number of site visitors, how long they spend browsing the site and which pages they viewed. User can also look at how people are arriving at your site, whether they found it through an organic search, a social media channel or other avenues.
- All of these insights can helps to find strengths and weaknesses that may shape new initiatives and lead to improvements in site conversions and an uptick in sales.

2. Analyze customers and their behavior

- Marketing data can provide a clear picture of customers, making it easier to tailor online services. For instance, if the majority of site traffic is coming from landing pages and blog posts, we can concentrate our efforts on those aspects, while spending less time on another part of site experience that is not seeing as much activity.
- This information can also be used to drive the content. Analyzing the popularity of blog posts can help clue to which topics are more engaging, so we can refine content strategies to improve site visits and conversions.

3. Track progress of external strategies

- Social media offers an easy and effective way for companies to reach their audience, and business intelligence can improve the return on investment.
- The Huffington post suggested tracking data such as Facebook likes and twitter retweets to get a better understanding of how effective different posts are.

6.12 BI Applications in Logistics and Production

- BI applications can be designed for the analysis of transport costs, route planning and scheduling, performance analysis of drivers and vehicles, the analysis of delivery time cycles depending on various factors, capacity planning in line with expected demand trends, evaluation of the carrier that provides the service and analysis of the causes and consequences.
- BI provides business information and the analysis of key business processes, quality decision-making at different management levels and improvement of the performance in the business system. The role of BI in the process of decision making is shown in Fig. 6.12.1.

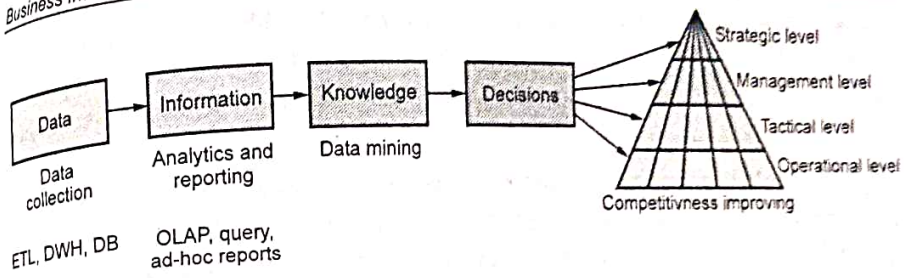


Fig. 6.12.1 The role of BI in the process of decision making

- In logistics, a large volume of data is produced, based on several operations which are performed daily and include a large number of actors, like clients, logistics operators, transport, warehousing, administrations, customs and ports.
- All whole information sometimes signifies the loss of valuable information, which can be decisive in determining the profitability of an operation, often because of difficulty gathering the data, complexity of reporting, information hosted across multiple platforms and manually updating information.
- The difficulty in generating updated, reliable and global reports often results in the lack of awareness of the costs associated with every operation. Business intelligence aims to solve these issues to provide a clear and easy overview of the company's evolution.
- **Applications of Business Intelligence in Logistics :**

1. **Accuracy and clarity of information :** We can find the status and evolution of the whole business at a glance, without the need of consulting different sources and cross data with the help of spreadsheets. It also shortens the time it takes to train employees.
2. **Information updates :** Automatically, business intelligence tools are updated whenever new data exists. It minimizes the reporting times as well as updates.
3. **More agile and responsive :** In case if something happens in the operation which will affect the business, alerts are received immediately, which helps us accelerate decision making.
4. **Fewer bottlenecks :** With appropriate permissions, the information can be easily accessible by the users, So that it is not required to request every area manager the data necessary to generate status reports.
5. **Broader context of information :** Through the creation of evolutionary reports, it is easy to understand the data we manage. Moreover, these kinds of tools offer visual information like shipment tracking maps.

6. **Collaborative work :** Business intelligence enables collaborative work, allowing all users to be involved in preparing and analyzing status reports. It helps the users easily find the problem and learn more about them until they find its source.

- **Transport Management :** BI applications can be designed for the analysis of transport costs, route planning and scheduling, performance analysis of drivers and vehicles, the analysis of delivery time cycles depending on various factors, capacity planning in line with expected demand trends, evaluation of the carrier that provides the service and analysis of the causes and consequences.
- Business intelligence and big data analysis play a crucial role in transportation companies. Supply chain data could help improve the efficiency and speed of the supply chain.
- In the travel industry, business intelligence and big data analysis are also useful in enriching inventory management. However, to take advantage of the benefits of BI, the companies need to use the BI tools and techniques that are specific to their industry.
- Travel companies can successfully and effectively address existing and future needs using the latest BI tools.
- In advance, companies are prepared for any changes possible in the market if they make careful use of BI information. Visualizing the insights of big data and BI enables transportation companies to make business decisions and strategies that are more effective in growing the business.
- Dashboards and scorecards are the most advanced version of BI applications because they include a large collection of data relating to different business processes and the visualization of results in a way that is fastest to present new information and knowledge to the user.
- Dashboards are most commonly used to monitor and manage processes and activities, ensuring monitoring changes in real time. Balanced scorecards is an approach of monitoring business performances from different aspects, which enables users to monitor the current situation and provides the possibility to warn of future changes.

6.13 Role of BI in Finance

- Technology is transforming the banking and finance industry. Thanks to the Internet and the proliferation of mobile devices and apps, today's financial

institutions face mounting competition, changing client demands and the need for strict control and risk management in a highly dynamic market.

- At the same time, technology has given rise to powerful business intelligence tools. Tools that the banking and finance industry can use to leverage customer data for insights that can lead to smarter management practices and better business decisions. To that end, here is a look at some of the ways banking and finance institutions are using Business Intelligence (BI) solutions to drive profitability, reduce risk and create competitive advantage.
- Improved operational efficiencies : In today's ultracompetitive marketplace financial institutions need to be as lean and efficient as possible. Using BI solutions to analyze operational processes, organizations can reduce ongoing costs and maximize existing resources and expertise.
- For example, by analyzing the performance of customer-facing employees, such as sales personnel, tellers and account managers, organizations can discover ways to improve and enhance the customer experience at the point-of-contact.
- Improved products and services : BI solutions allow organizations to track individual revenue streams to better determine which products and services are profitable and which are not. But the benefits do not stop there.
- Business intelligence solutions also enable financial organizations to analyze vast amounts of customer data to gain insights about customer needs and sentiments regarding banking that can be used to improve products and services.
- As an example, perhaps it is learned that customers want a quicker, easier way to track and analyze their earning and spending patterns. Institutions, may be able to send more timely alerts to customers. Or they are looking for a smoother and less complicated application and funding process. Armed with these kinds of insights, organizations can develop new and improved financial products and services to better meet customer needs and in turn create a competitive edge.
- **Improved marketing :** Using BI, marketers can analyze CRM data based on a range of criteria to uncover the most profitable customer profile. In addition, the customer base can be analyzed to identify and develop new cross-sell and up-sell opportunities and to carry out more targeted online marketing campaigns. This presents a major advantage, as research shows that it costs five times more to sell financial products and services to new customers than to existing customers.
- **Improved customer retention :** BI applications can help financial institutions identify and pursue those customers that are the most profitable. BI also plays an

important role in improving customer retention and loyalty. Using business analytics tools and techniques, organizations can discover the reasons why customers switch to a competing institution. They can then implement new processes to help reduce customer churn. The ability to track customer habits, preferences and behaviors also allows organizations to tailor their products and services in ways that meet needs, solve problems and promote customer retention and loyalty.

- **Developing new investment strategies :** Asset managers are utilizing new data sets to develop new strategies for investing. By developing models around social media, investors can gain specific insight on sentiment and develop trading signals. Other research analysts are using satellite imagery to understand global supply of commodities like oil and gas or triangulating consumer spend based on the number of cars in shopping center parking lots. Whole new categories of investing are emerging from leveraging analytics and BI applications.
- **Risk reduction :** The ability to track customer transaction histories allows institutions to quickly detect and reduce the incidents of fraudulent activities, the most notable being credit card fraud.

6.14 BI Applications in Banking

- BI in banking evolved through manual systems to management information systems with computerization. Banks had efficient transaction recording systems before computerization also. The manual systems too had effectively provided the necessary reports for management and regulatory requirements.
- These reports were manually consolidated at lower offices and final reports were presented at head office level. These manual systems worked well till the scale of operations of the banks were small.
- As the banks grew in size and expanded geographically the number of branch network grew leaps and bounds and so the, the volume of transactions became quite large and manual operations became time consuming and error prone.
- To cater the load of operations from all bank branches spread across geographies the banks have started using computers and slowly banks have become fully automated.
- Business intelligence tools can be used by banks for historical analysis, performance budgeting, business performance analytics, employee performance measurement,

executive dashboards, marketing and sales automation, product innovation, customer profitability, regulatory compliance and risk management.

- Banks can analyze their historical performance over time to be able to plan for the future. The key performance indicators include deposits, credit, profit, income, expenses; number of accounts, branches, employees etc.

Absolute figures and growth rates are required for this analysis.

A CRM helps a bank with the following :

1. Find customers
2. Get to know them
3. Communicate with them
4. Ensure they get what they want (not what the bank offers)
5. Retain them regardless of profitability
6. Make them profitable through cross-sell and up-sell
7. Covert them into influencers
8. Strive continuously to increase their lifetime value for the bank.

- Operational BI embeds analytical processes with the operational business structure to support near real-time decision making and collaboration. This characteristic fundamentally changes the way how data is used, where it exists and how it is accessed.
- Risk management is a process in which a bank methodologically manages all the risk processing phases (identification, analysis, measurement, control and reporting) posing a threat to the achievement of its goals and individual business activities, so that the achieved risk level should not endanger the bank's safe and stable operation.
- Some of the risks faced by banks include credit risks, market risks, interest rate risks, foreign change risks, liquidity risks, operational risks, reputational risks, etc.

6.15 BI Applications in Telecommunications

- Telecommunication companies today are operating in highly competitive and challenging environment. Huge volume of data is generated from various operational systems and these are used for solving many business problems that required urgent handling. These data include call detail data, customer data and network data.

- Data mining methods and business intelligence technology are widely used for handling the business problems in this industry. The main application areas of BI and data mining in telecommunication industry include fraud detection, network fault isolation and improving market effectiveness.
- Network fault isolation and prediction telecommunication networks are comprised of highly complex configurations of hardware and software. Since the industry requires optimum network efficiency and reliability, most of the network elements have the capability of self-diagnosis and generating status and alarm messages.
- Expert systems were developed to handle alarms. Network fault isolation in the telecommunication industry is a quiet tedious task because of the following reasons. Huge volume of data A single fault can generate different unrelated alarms.
- Hence alarm correlation has an important role in predicting network faults. A proactive rapid response is very much essential for maintaining the reliability of the network. Data mining techniques like classification, neural network and sequence analysis can be used for identifying network faults.
- The Telecommunication Alarm Sequence Analysis (TASA) is a data mining tool which support fault identification by searching for recurrent patterns of algorithms. This information can be used to generate a rule based alarm correlation system, which can be used for identifying faults in real time.
- Time weaver is a genetic algorithm which has the capability to operate directly on the raw network level time series data. This algorithm will identify patterns that will successfully predict the target event.
- Bayesian belief networks can also be used to identify the network faults. Standard classification tools can be used to generate rules to predict future failures but it has several draw backs. Most importance drawback of this is that some information will be lost in reformulation process.

6.16 BI in Salesforce Management

- Salesforce is a company that makes cloud-based software designed to help businesses find more prospects, close more deals and wow customers with amazing service.
- Salesforce is a cloud-based Business Intelligence (BI) system. It provides an interactive platform to access and share business trends. And helps to uncover new business opportunities. Salesforce bi tool allows businesses to analyze critical sales

and service-specific data, generate business-relevant insights and make reformed decisions.

- Salesforce business intelligence tool seamlessly integrates with other salesforce products such as salesforce CRM, to share contact information and data insights across applications. BI dashboard can be used to track and monitor critical business metrics and KPIs.
- Salesforce enables to auto-generate slides and presentations with data visuals and key points using a one-click storytelling feature. Salesforce analytics and intelligence solution can be used by companies operating in multiple industries.
- Salesforce business intelligence tool build trust by safeguarding employee and customer health. This enables to manage health-related interactions and workplace planning on a single platform. And boost healthcare and community responsiveness now and for future crises.
- Salesforce applications of business intelligence tools connect any system, whether in the cloud or on-premises, on a unified platform. It allows teams to discover and reuse integration assets to build upon prior projects and address new business priorities. It enables your entire ecosystem to create new revenue opportunities and customer experiences with packaged APIs.

❑❑❑