**EXPERMINT: 01**

**⚫ Aim: Case Study on Pharmacy Management System.**

**⚫ Theory:**

**Introduction:**

The customer goes to the shop and purchases the medicine required. So a lot of time is wasted and the person gets tired. If he wants to exchange the product, once again he goes to the shop and replaces them. The complete process depends on the physical interactions. This whole process involves lots of effort and loss of time so to make everything easier and comforting our website provides facility to get medicine at your very doorstep.

Our Application provides the following things :

* To provide login facility to the customers.
* To provide a list of all available medicines to the customers.
* To update all the list of expired medicines.
* Medicines are delivered at your door-step.

**Problem Statement:**

Pharmacy management has kept paper record in filing cabinets. Managing a very large pharmacy with records on papers will be tedious and difficult to keep track of inventories with regards to the drugs in the store, expiry date, quantity of drugs available based on the categories and their functions. The pharmacist has to order drugs to replenish the already diminishing stock. In addition, ordering of drugs is being carried out manually. Drugs are not supposed to be used after they have expired. This project work will prompt the pharmacist about drugs that are close to expiry, preventing those drugs from being sold and also providing solution to the earlier stated problems.

**Objectives:**

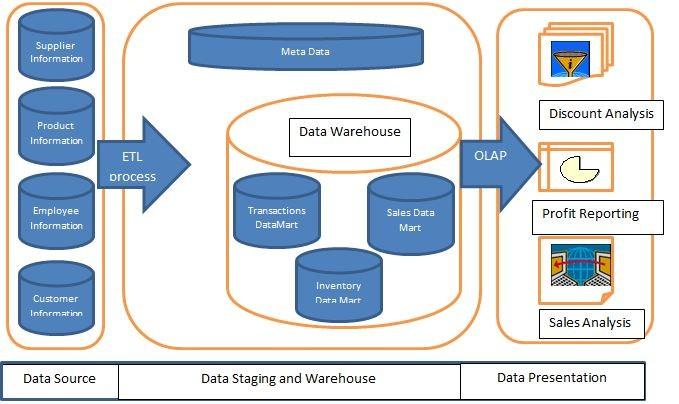
* To improve the efficiency of the system by ensuring effective monitoring of services and activities.
* Making the pharmacy organizations computerized by creating network through minimizing or eliminating wasting of time as well as removing the resources such as papers for data saving since know days is paper based, decrease malfunctioned works on the medical usage by giving correct information on each medicine.
* Minimize time and resource by which, searching the medicine data you can get the data in quickest time.

**Benefits:**

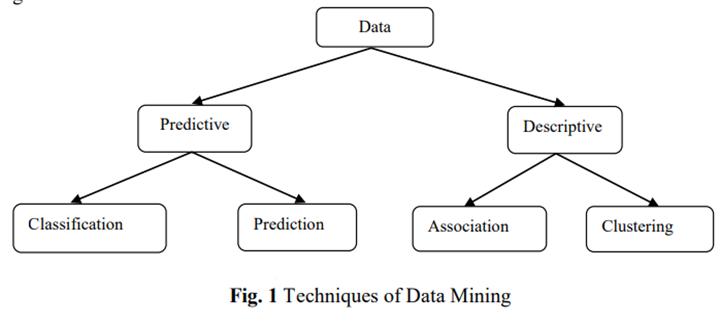
* Customer Satisfaction – PMS system ensures every transaction is accurate and processed quickly.
* Accessibility – Improved control over pharmacy operations.
* Reduced paperwork – Eliminates paper-based workflow and simplifies data storage and retrieval.
* Better productivity – Enables the users to manage patient profiles, manage processing and billing, smoothen their workflow, manage accounts receivable, manage claims, and control inventory management.
* Drug Documentation – Maintains drug profile repository with details such as the composition of medicine and substitutes
* Barcode Labels – Automates barcode labelling of medication and print expiry dates

**Architecture:-**

Construction of data warehouse starts with the relationship between multiple layers of data warehouse architecture. Data warehouse architecture for retail industry helps us to understand the basic overview which can be optimized for supermarkets. Data is obtained from multiple sources in the form of flat files, databases or sheets etc. This mainly contains data about finance, marketing, demand planning, forecasting, logistics, RnD and sales data. This data is extracted from the sources and various data cleaning operations are implemented on this data. These consist of filtering, cleaning, joining, splitting and sorting. This transformed data is then fed to data storage and aggregation layer which consists of data warehouse and several data marts. It contains operational data metadata and data marts for various domains such as product portfolio, logistic, finance and overall goals and strategies in the business.



**Models**:

There are various types of databases and information repositories on which data mining can be performed. So, to perform data mining several techniques are available to mine data from data warehouse. The main function of data mining is either to create a descriptive model or a predictive. A brief overview of these two techniques is given as follows:

**Predictive Model: -** Predictive models allow data miner to predict an unknown value of target variable. It covers following two basic data mining techniques:

• **Classification**: Classification is the process of finding a model (or function) 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.

• **Prediction** : As many real world applications in the field of data mining are required to predict the future data states based on the past and current data stored in database. Prediction is all about predicting the future state rather than a current state.

**Descriptive Model**: - Descriptive models present main characteristics of data sets in concise form. It covers following two basic data mining techniques:

• **Association**: Association is the most popular data mining techniques and fined Data Mining Predictive Descriptive Classification Prediction Association Clustering Data Mining

* **Clustering**: It is a data mining technique and a cluster is defined as a correlation of objects which are similar between them and dissimilar to the objects belonging to other cluster. This technique is used to make cluster of objects which have similar characteristics using automatic or semi-automatic means.

**⚫ Conclusion:-**

As we are well aware about the existing way of purchasing medicines, so to overcome these difficulties we are here with the solution of ordering medicines online. The main aim of developing this application is to supply the medicines all over the country by just a single click and to reduce the time consumption. It is always easier to receive medicine at your doorstep than to wait in line for some hours and waste your precious time in for doing some work that could be done just within matter of minutes.

**EXPERMINT: 02**

**⚫ Aim: Design dimensional modelling using star schema and snowflakes schema.**

**⚫ Theory:**

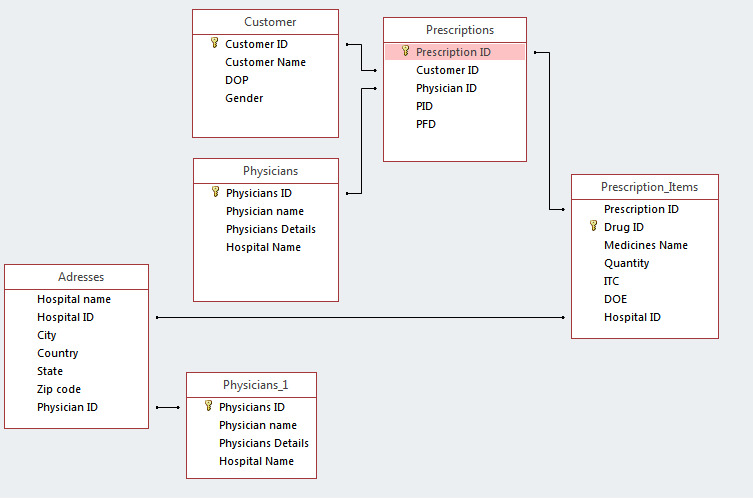
* **Star Schema:**

A star schema is a relational schema where a relational schema whose design represents a multidimensional data model. The star schema is the explicit data warehouse schema. It is known as **star schema** because the entity-relationship diagram of this schemas simulates a star, with points, diverge from a central table.

* **Fact Tables:**

A table in a star schema which contains facts and connected to dimensions. A fact table has two types of columns: those that include fact and those that are foreign keys to the dimension table. The primary key of the fact tables is generally a composite key that is made up of all of its foreign keys.

* **Dimension Tables:**

A dimension is an architecture usually composed of one or more hierarchies that categorize data. If a dimension has not got hierarchies and levels, it is called a flat dimension or list. The primary keys of each of the dimensions table are part of the composite primary keys of the fact table.

**Fig .Star Schema**

* **Snowflake schema :**

A snowflake schema is equivalent to the star schema. "A schema is known as a snowflake if one or more dimension tables do not connect directly to the fact table but must join through other dimension tables."

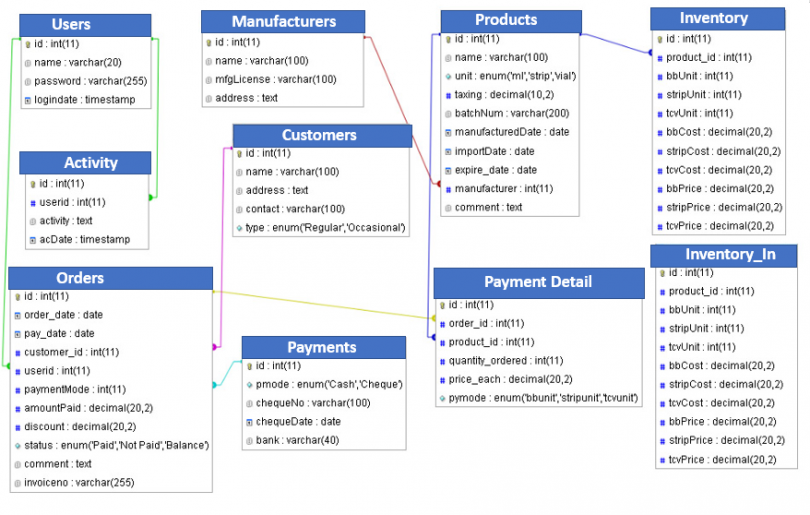
The snowflake schema is an expansion of the star schema where each point of the star explodes into more points. It is called snowflake schema because the diagram of snowflake schema resembles a snowflake. Snowflake is a method of normalizing the dimension tables in a STAR schemas. When we normalize all the dimension tables entirely, the resultant structure resembles a snowflake with the fact table in the middle.

Fig. Snowflake schema

**⚫ Conclusion: -**  we understanding the dimensional modelling using **star schema and snowflakes schema** also we’ve learned the differences between the Star Schema vs. Snowflake Schemas based on multiple criteria.

**EXPERMINT: 03**

**⚫ Aim: Implementation OLAP operations Slice, Dice, Roll up. Drill down based on experiment no.1 case study.**

**⚫ Theory:**

**Introduction:**

* **OLAP operations:**

OLAP stands for On-Line Analytical Processing. OLAP is a classification of software technology which authorizes analysts, managers, and executives to gain insight into information through fast, consistent, interactive access in a wide variety of possible views of data that has been transformed from raw information to reflect the real dimensionality of the enterprise as understood by the clients.

* **Characteristics of OLAP:**
* Fast
* Analysis
* Share
* Multidimensional
* Information
* **Benefits of OLAP**
* OLAP holds several benefits for businesses: -
* OLAP helps managers in decision-making through the multidimensional record views that it is efficient in providing, thus increasing their productivity.
* OLAP functions are self-sufficient owing to the inherent flexibility support to the organized databases.
* It facilitates simulation of business models and problems, through extensive management of analysis-capabilities.
* In conjunction with data warehouse, OLAP can be used to support a reduction in the application backlog, faster data retrieval, and reduction in query drag.
* **OLAP Operations:**

1. Roll-Up
2. Drill-Down
3. Slice
4. Dice
5. Pivot

* **Roll-Up:**

The roll-up operation (also known as drill-up or aggregation operation) performs aggregation on a data cube, by climbing down concept hierarchies, i.e., dimension reduction. Roll-up is like zooming-out on the data cubes.

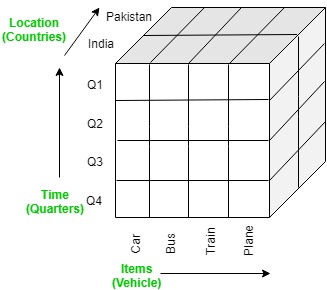


Fig. Roll-Up:

* **Drill-Down:**

The drill-down operation (also called roll-down) is the reverse operation of roll-up. Drill-down is like zooming-in on the data cube. It navigates from less detailed record to more detailed data. Drill-down can be performed by either stepping down a concept hierarchy for a dimension or adding additional dimensions.

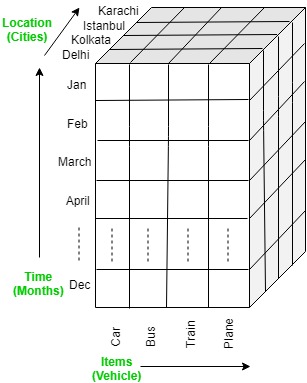


Fig. Drill Down

* **Slice:**

A slice is a subset of the cubes corresponding to a single value for one or more members of the dimension. For example, a slice operation is executed when the customer wants a selection on one dimension of a three-dimensional cube resulting in a two-dimensional site. So, the Slice operations perform a selection on one dimension of the given cube, thus resulting in a sub cube.

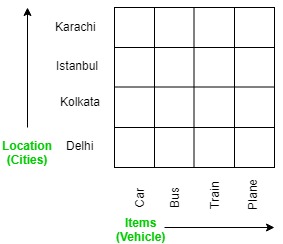


Fig. Slice

* **Dice:**

The dice operation describes a sub cube by operating a selection on two or more dimension.

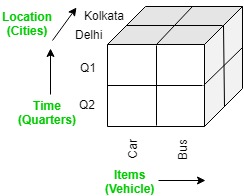
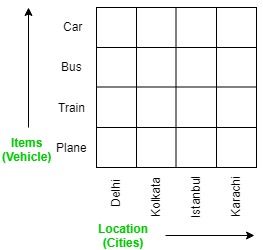


Fig. Dice

* **Pivot**:

The pivot operation is also called a rotation. Pivot is a visualization operations which rotates the data axes in view to provide an alternative presentation of the data. It may contain swapping the rows and columns or moving one of the row-dimensions into the column dimensions.



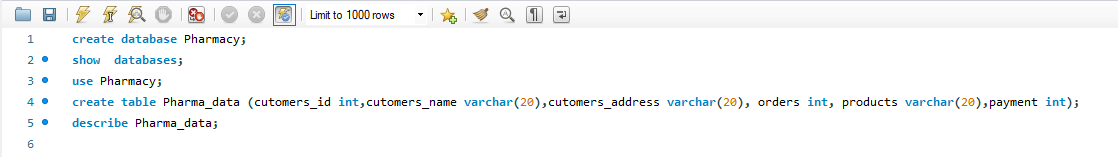
Problem Statement/Case Study

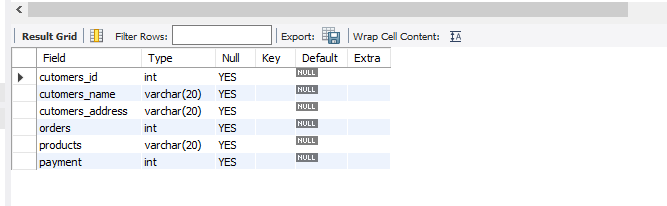
Problem Statement/Case Study

* **Problem Statement/Case Study:**

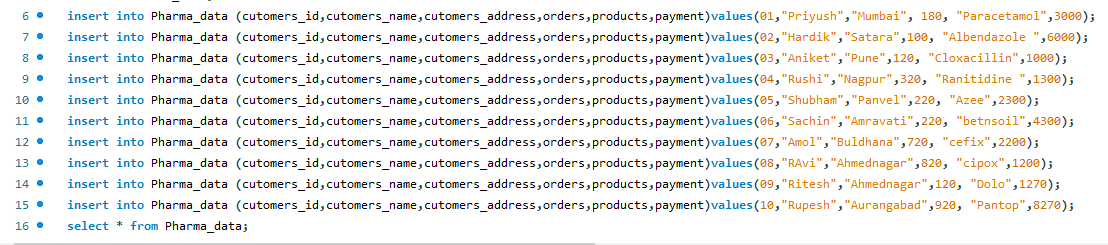
Pharmacy management has kept paper record in filing cabinets. Managing a very large pharmacy with records on papers will be tedious and difficult to keep track of inventories with regards to the drugs in the store, expiry date, quantity of drugs available based on the categories and their functions. Suppose database attribute cutomers\_id, cutomers\_name, cutomers\_address, orders, products then it olap operations is following:

* *Create database and table:*

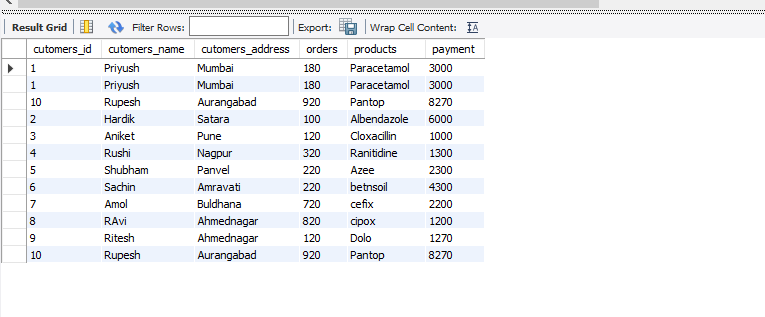




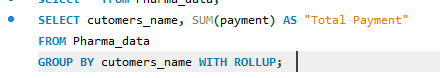
* *Insert Values :*

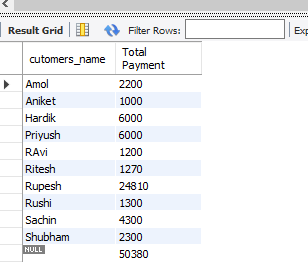


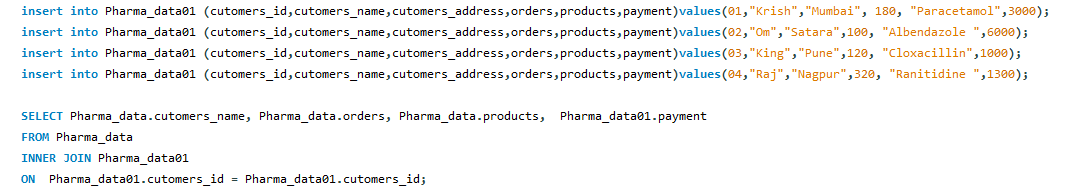


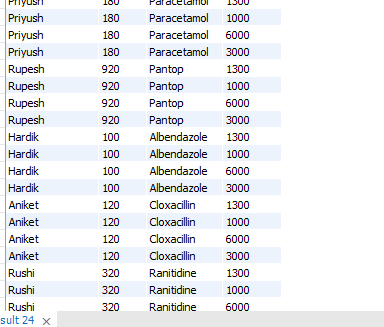


* **Roll-Up:**

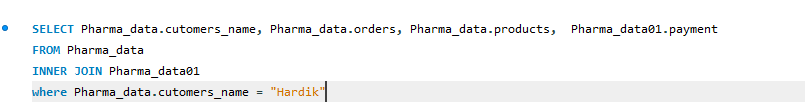


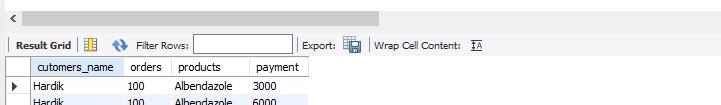


* **Slice:**

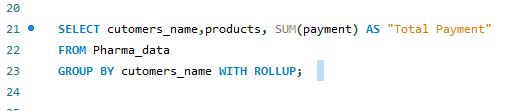


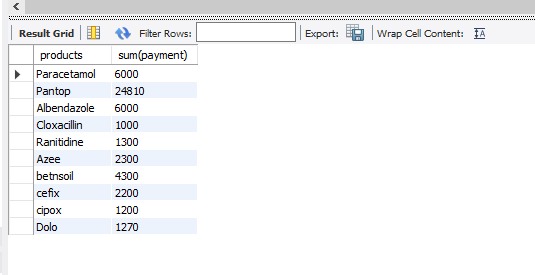
* **Dice:**





* **Drill-Down:**





**⚫ Conclusion: -**

Thus, OLAP (Online Analytical Processing) is powerful technology behind many Business

Intelligence (BI) applications that discovers data, report viewing capabilities, complex

analytical calculations, and predictive scenario, budget planning, forecast planning.It works

as it first collected the data from multiple data sources (like a spreadsheet, video, XML,

etc.) and is stored in data warehouses, which are then cleansed and organized into data

cubes on which can run the user’s queries.The five types of analytical operations against

the multidimensional databases that can perform are Roll-up, Drill-down, Slice, Dice, and

Pivot

Thus, OLAP (Online Analytical Processing) is powerful technology behind many Business

Intelligence (BI) applications that discovers data, report viewing capabilities, complex

analytical calculations, and predictive scenario, budget planning, forecast planning.It works

as it first collected the data from multiple data sources (like a spreadsheet, video, XML,

etc.) and is stored in data warehouses, which are then cleansed and organized into data

cubes on which can run the user’s queries.The five types of analytical operations against

the multidimensional databases that can perform are Roll-up, Drill-down, Slice, Dice, and

Pivot

Thus, OLAP (Online Analytical Processing) is powerful technology behind many Business Intelligence (BI) applications that discovers data, report viewing capabilities, complex analytical calculations, and predictive scenario, budget planning, forecast planning. It works as it first collected the data from multiple data sources (like a spreadsheet, video, XML, etc.) And is stored in data warehouses, which are then cleansed and organized into data cubes on which can run the user’s queries. The five types of analytical operations against the multidimensional databases that can perform are Roll-up, Drill-down, Slice, Dice, and Pivot.

**EXPERMINT: 06**

**⚫ Aim: Perform data Pre-processing task and demonstrate Classification, Clustering,** **Association algorithm on data sets using data mining tool (WEKA/R tool)**

**⚫ Theory:**

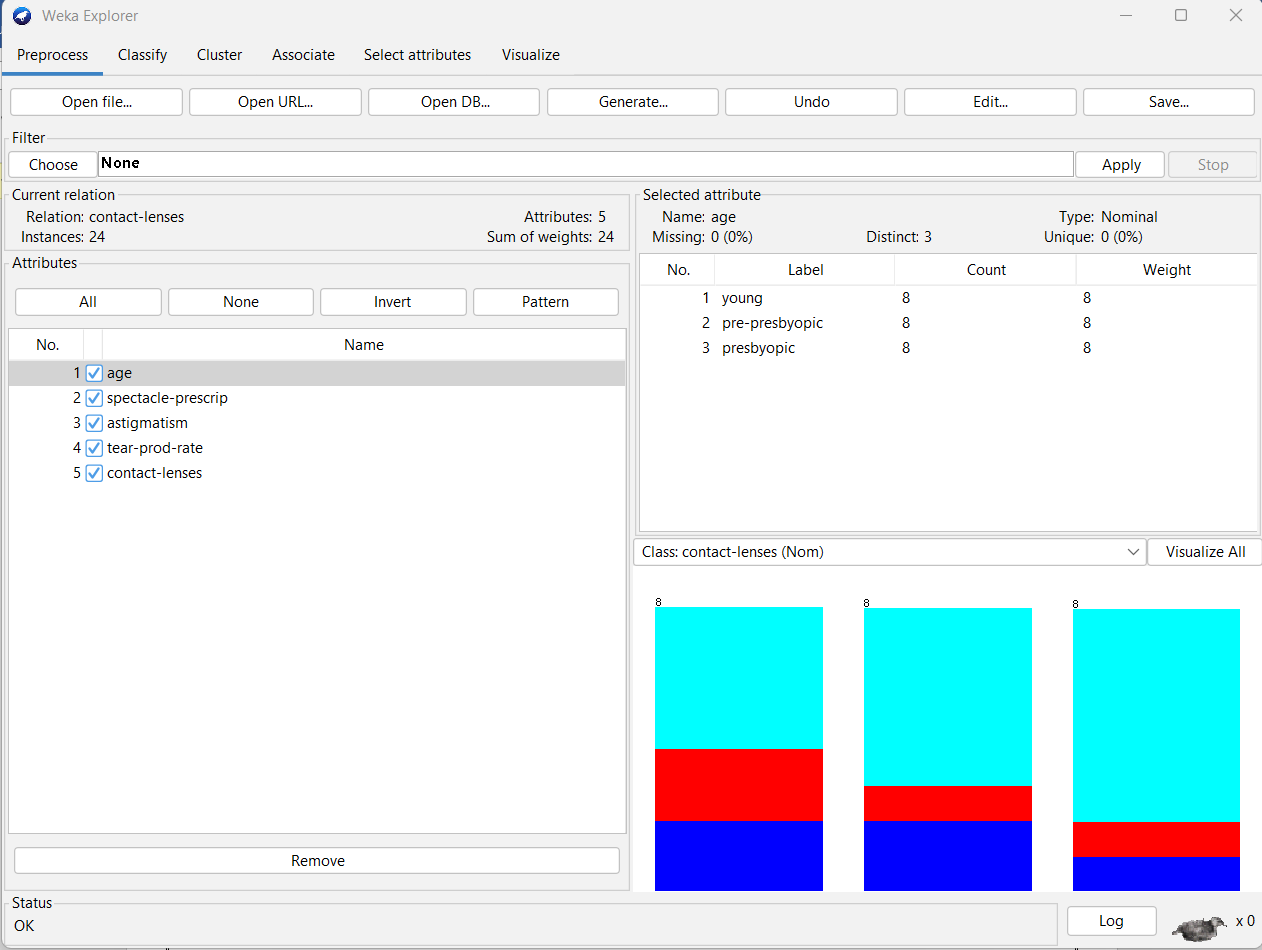
**Introduction:**

* **Data Pre-processing:**

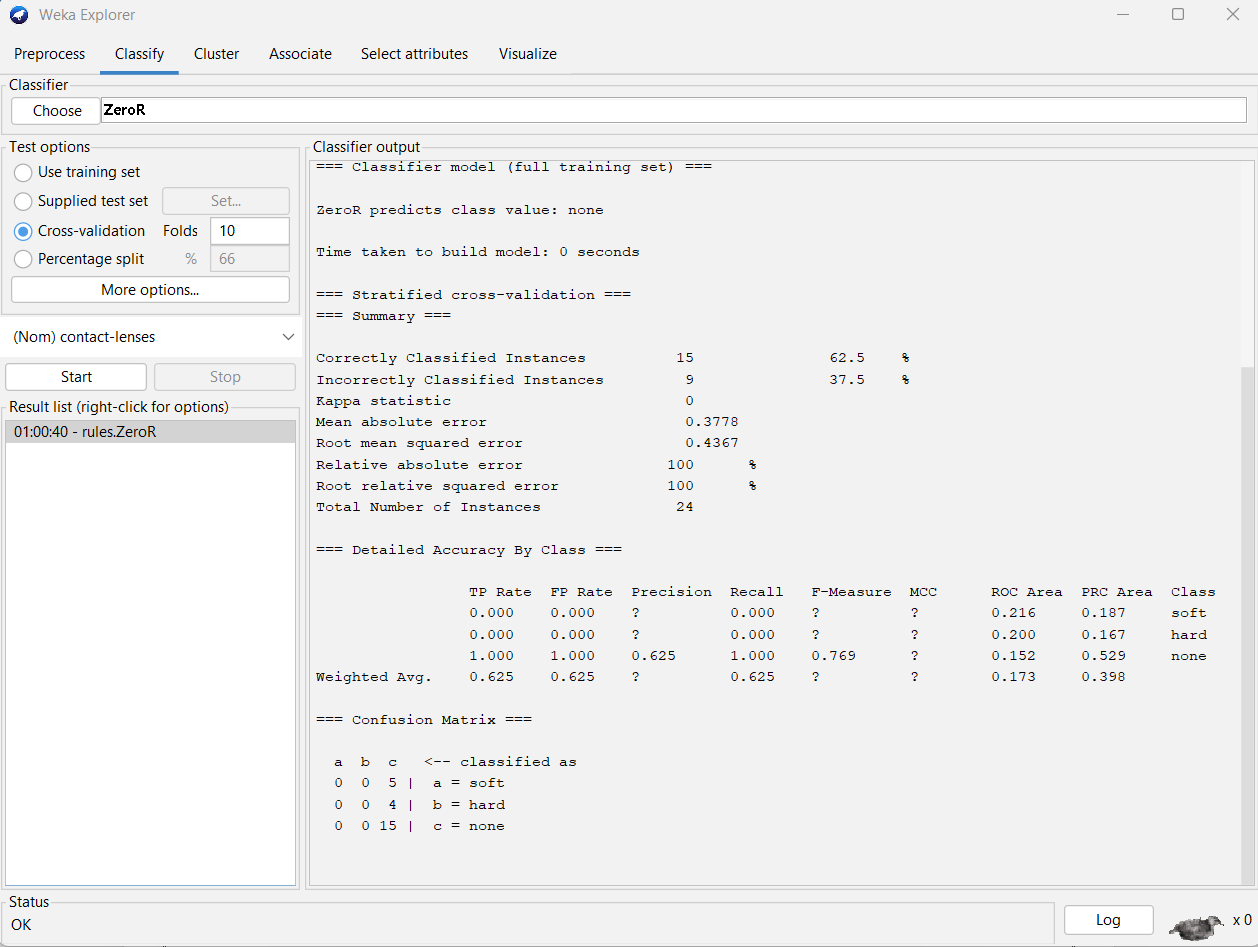
Data processing is collecting raw data and translating it into usable information. The raw data is collected, filtered, sorted, processed, analysed, stored, and then presented in a readable format. It is usually performed in a step-by-step process by a team of data scientists and data engineers in an organization.

* **WEKA:**

WEKA - an open source software provides tools for data pre-processing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems.



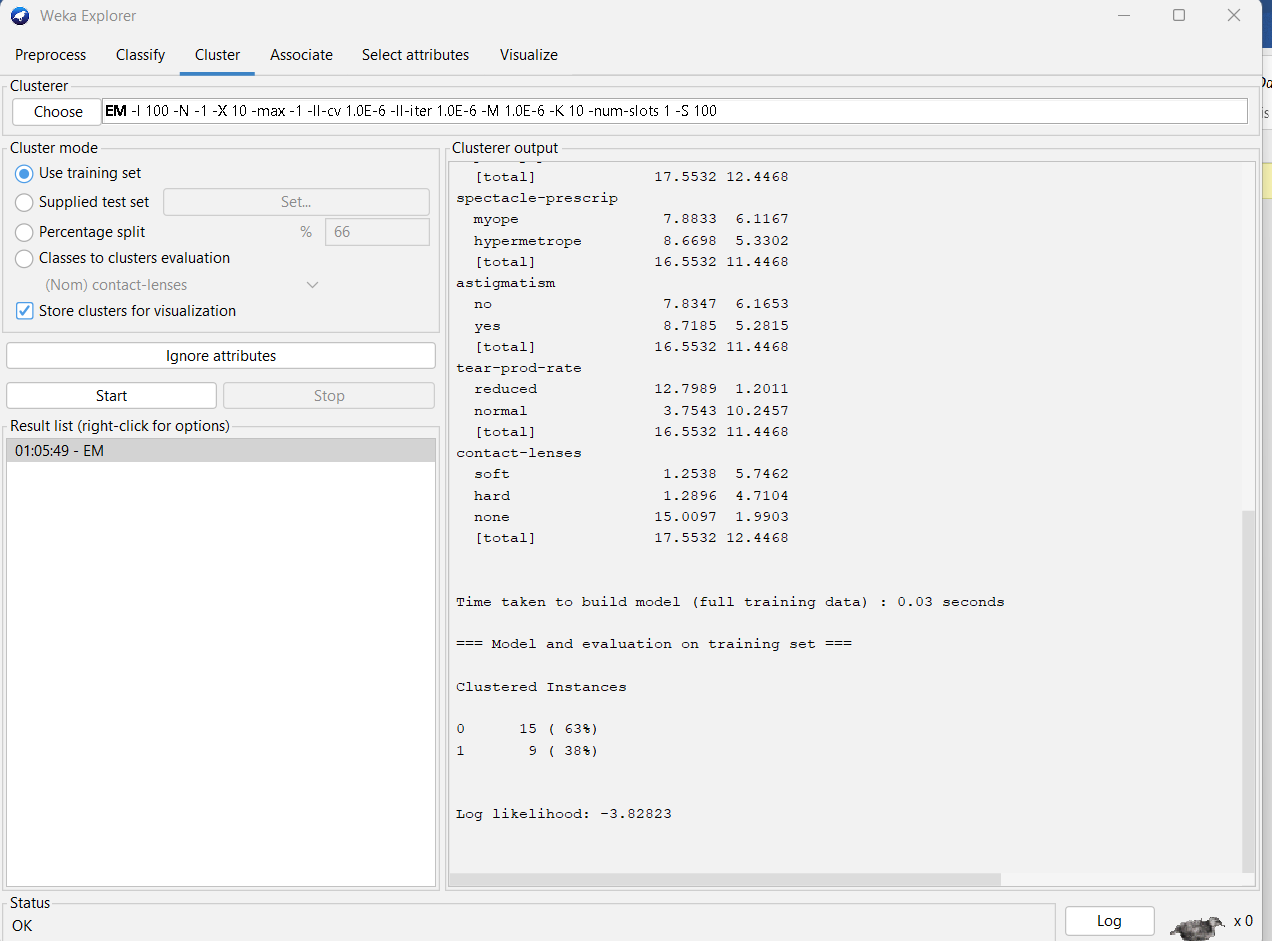
* **Classification:**

Data mining refers to the process of extracting important data from raw data. It analyses the data patterns in huge sets of data with the help of several software. Ever since the development of data mining, it is being incorporated by researchers in the research and development field.

* **Clustering:**

Clustering is an unsupervised Machine Learning-based Algorithm that comprises a group of data points into clusters so that the objects belong to the same group.

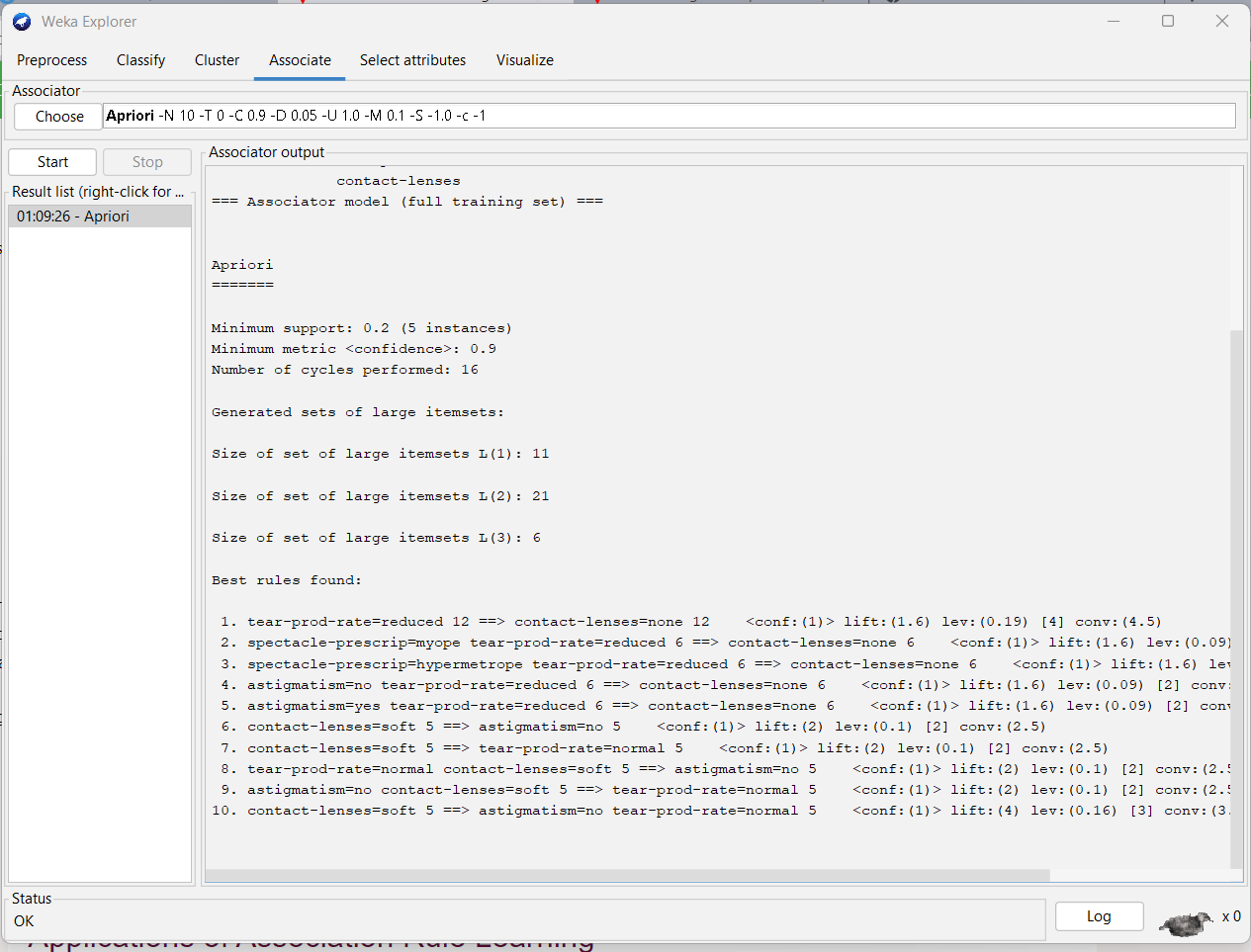
Clustering helps to splits data into several subsets. Each of these subsets contains data similar to each other, and these subsets are called clusters. Now that the data from our customer base is divided into clusters, we can make an informed decision about who we think is best suited for this product.

**Association** :

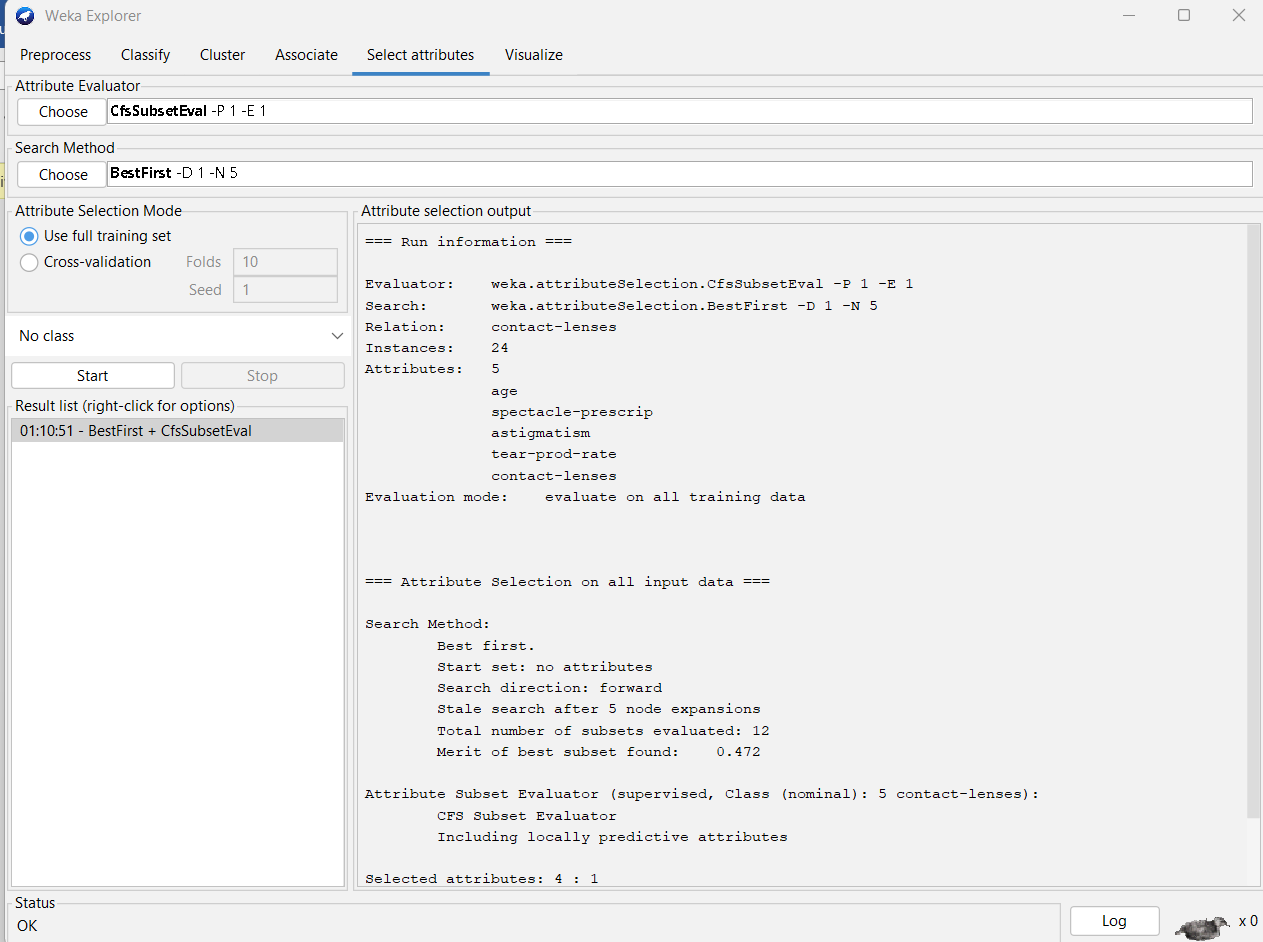
Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database.

Types of Association Rule Learning:

* Apriori Algorithm
* Eclat Algorithm
* F-P Growth Algorithm



**Attribute** :

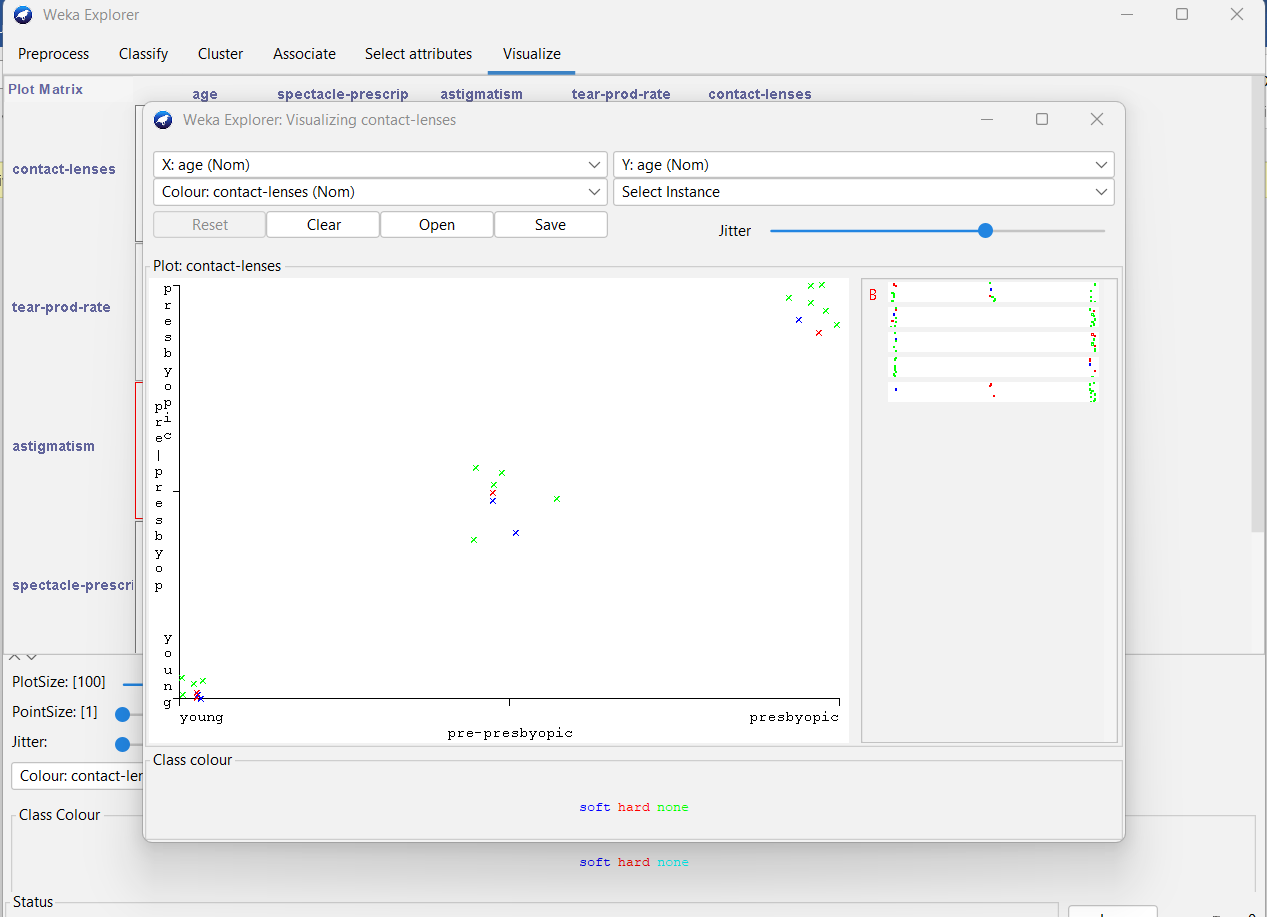


* **Data visualization**

Data visualization is a graphical representation of quantitative information and data by using visual elements like graphs, charts, and maps.

Data visualization convert large and small data sets into visuals, which is easy to understand and process for humans.

Data visualization tools provide accessible ways to understand outliers, patterns, and trends in the data.



**⚫ Conclusion: -** We understanding data Pre-processing task and demonstrate Classification, Clustering, Association algorithm on data sets using data mining tool (WEKA/R tool).

**EXPERMINT: 09**

**⚫ Aim: Implementation of Association Rule Mining algorithm (Apriori)**

Aim: Implementation of Association Rule Mining algorithm (Apriori)

**⚫ Theory:**

**Introduction:**

* **Apriori algorithm:**

Apriori algorithm refers to an algorithm that is used in mining frequent products sets and relevant association rules. Generally, the apriori algorithm operates on a database containing a huge number of transactions. For example, the items customers but at a Big Bazar.

Algorithm:

The following are the main steps of the algorithm:

1. Calculate the support of item sets (of size k = 1) in the transactional database (note that

support is the frequency of occurrence of an itemset). 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.

**Components of Apriori algorithm**

The given three components comprise the apriori algorithm.

* Support
* Confidence
* Lift

**Steps for Apriori Algorithm**

Below are the steps for the apriori algorithm:

Step-1: Determine the support of item sets in the transactional database, and select the minimum support and confidence.

Step-2: Take all supports in the transaction with higher support value than the minimum or selected support value.

Step-3: Find all the rules of these subsets that have higher confidence value than the threshold or minimum confidence.

Step-4: Sort the rules as the decreasing order of lift.

Let's take an example to understand this concept.

We have already discussed above; you need a huge database containing a large no of transactions. Suppose you have 4000 customers transactions in a Big Bazar. You have to calculate the Support, Confidence, and Lift for two products, and you may say Biscuits and Chocolate. This is because customers frequently buy these two items together.

Out of 4000 transactions, 400 contain Biscuits, whereas 600 contain Chocolate, and these 600 transactions include a 200 that includes Biscuits and chocolates. Using this data, we will find out the support, confidence, and lift.

**Support**

Support refers to the default popularity of any product. You find the support as a quotient of the division of the number of transactions comprising that product by the total number of transactions. Hence, we get

Support (Biscuits) = (Transactions relating biscuits) / (Total transactions)

= 400/4000 = 10 percent.

**Confidence**

Confidence refers to the possibility that the customers bought both biscuits and chocolates together. So, you need to divide the number of transactions that comprise both biscuits and chocolates by the total number of transactions to get the confidence.

Hence,

Confidence = (Transactions relating both biscuits and Chocolate) / (Total transactions involving Biscuits)

= 200/400

= 50 percent.

It means that 50 percent of customers who bought biscuits bought chocolates also.

**Lift**

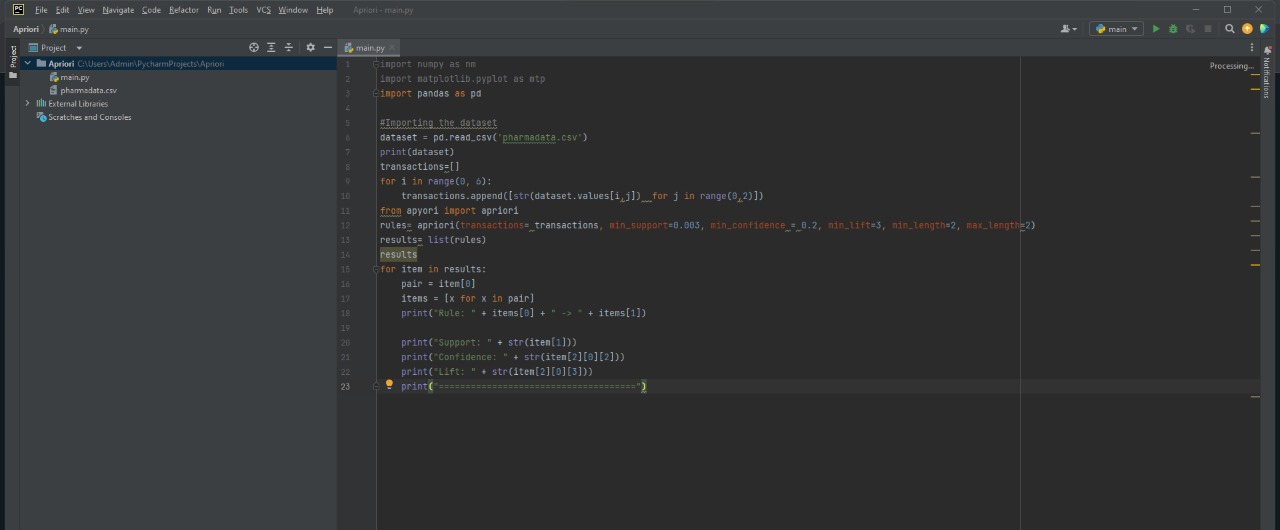
Consider the above example; lift refers to the increase in the ratio of the sale of chocolates when you sell biscuits. The mathematical equations of lift are given below.

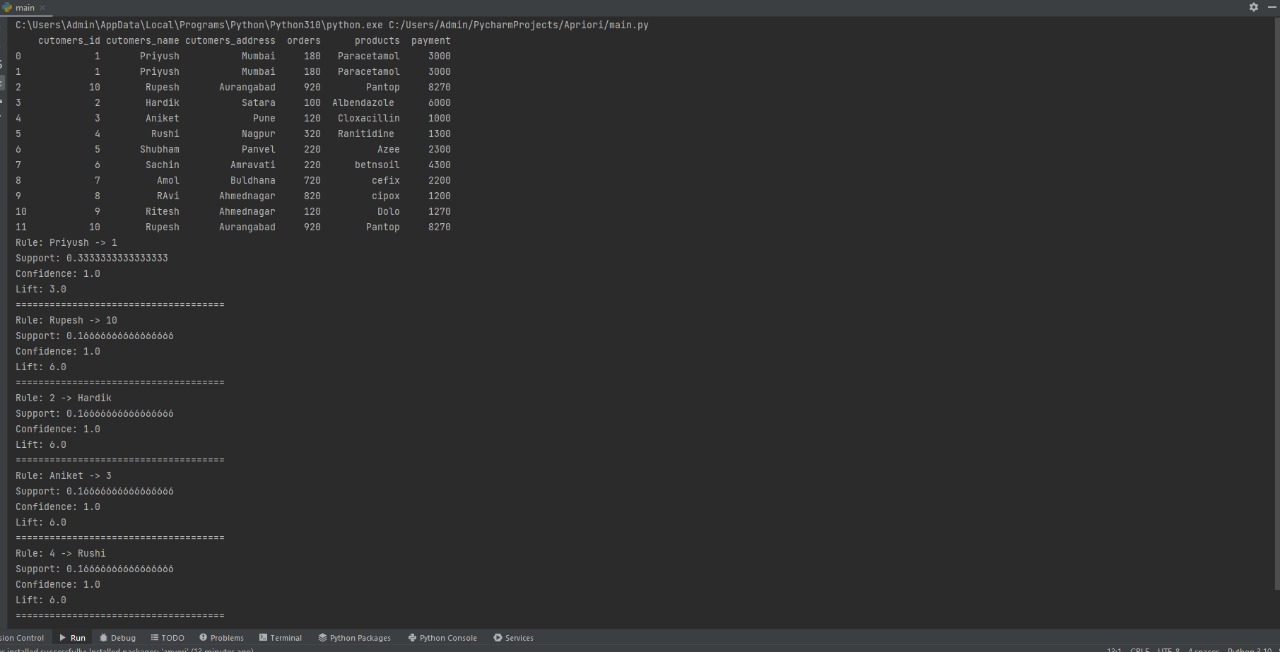
Lift = (Confidence (Biscuits - chocolates)/ (Support (Biscuits)

AD

= 50/10 = 5

**Implementation:**

****

****

**Advantages of Apriori Algorithm**

* It is used to calculate large item sets.
* Simple to understand and apply.

**Disadvantages of Apriori Algorithms**

* Apriori algorithm is an expensive method to find support since the calculation has to pass through the whole database.
* Sometimes, you need a huge number of candidate rules, so it becomes computationally more expensive.

**⚫ Conclusion: -**

Thus, we successfully implemented Association Rule Mining algorithm (Apriori)

**EXPERMINT: 08**

**⚫ Aim: Implementation of a hierarchical clustering method.**

**⚫ Theory:**

* **Hierarchical clustering**

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabelled datasets into a cluster and also known as hierarchical cluster analysis or HCA.

In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram.

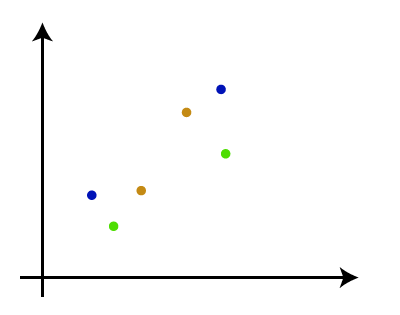
Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.

The hierarchical clustering technique has two approaches:

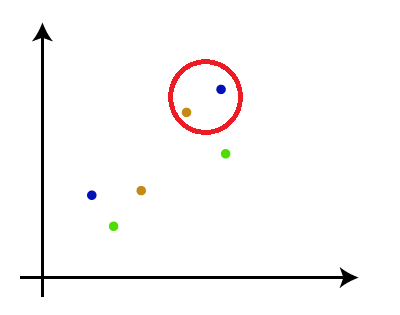
* **Agglomerative**: Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
* **Divisive**: Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

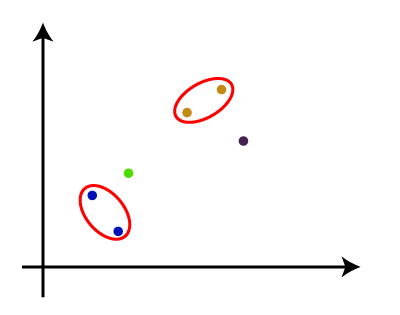
**How the Agglomerative Hierarchical clustering Work**?

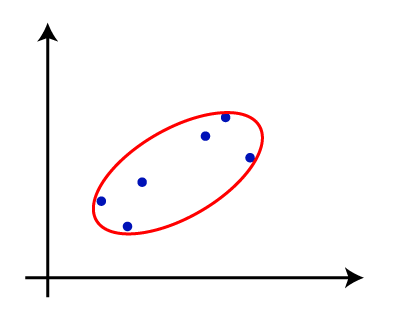
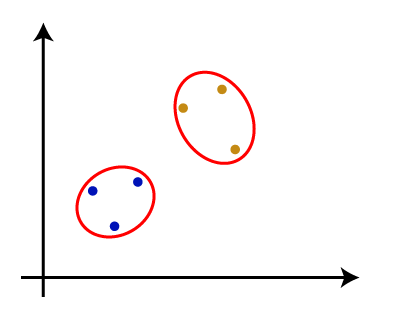
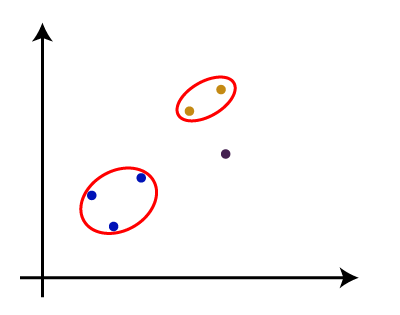
**Step-1**: Create each data point as a single cluster. Let's say there are N data points, so the number of clusters will also be N.



**Step-2:** Take two closest data points or clusters and merge them to form one cluster. So, there will now be N-1 clusters.



**Step-3**: Again, take the two closest clusters and merge them together to form one cluster. There will be N-2 clusters.

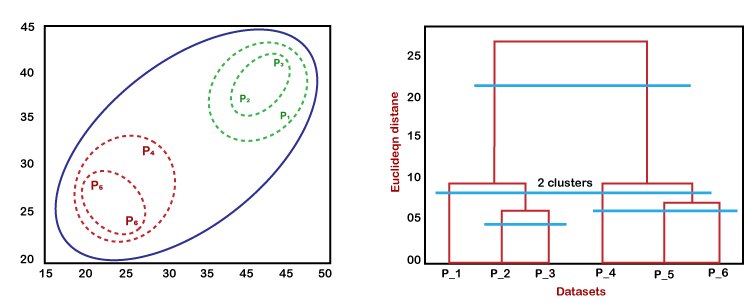
**Step-4:** Repeat Step 3 until only one cluster left. So, we will get the following clusters. Consider the below images:

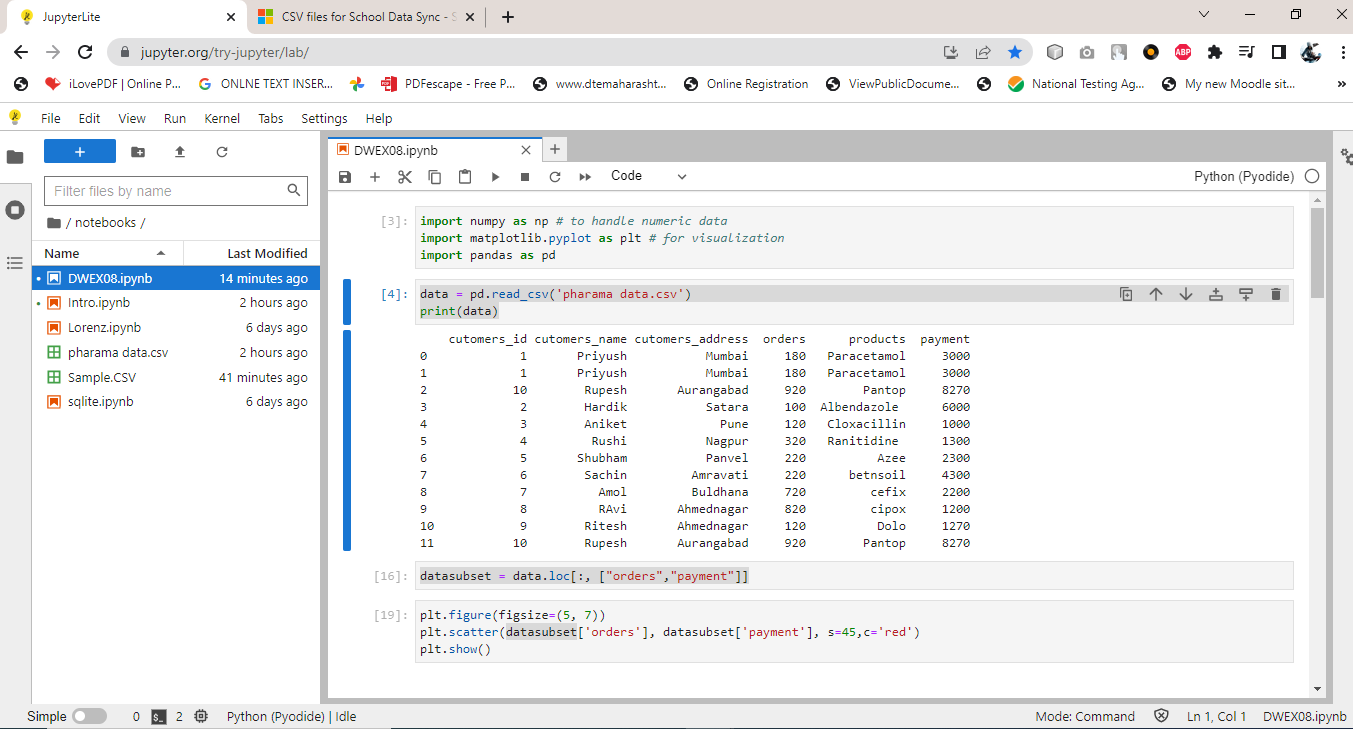
**Step-5:** Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.

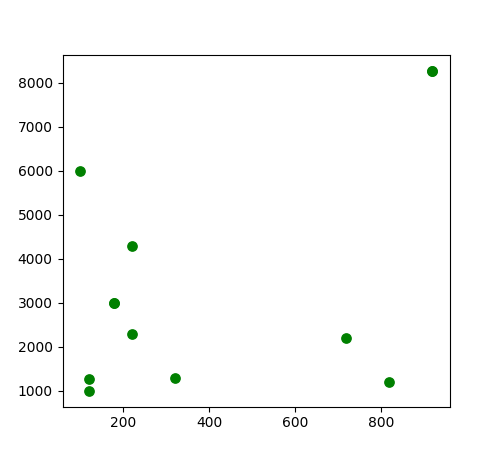
**Woking of Dendrogram in Hierarchical clustering**

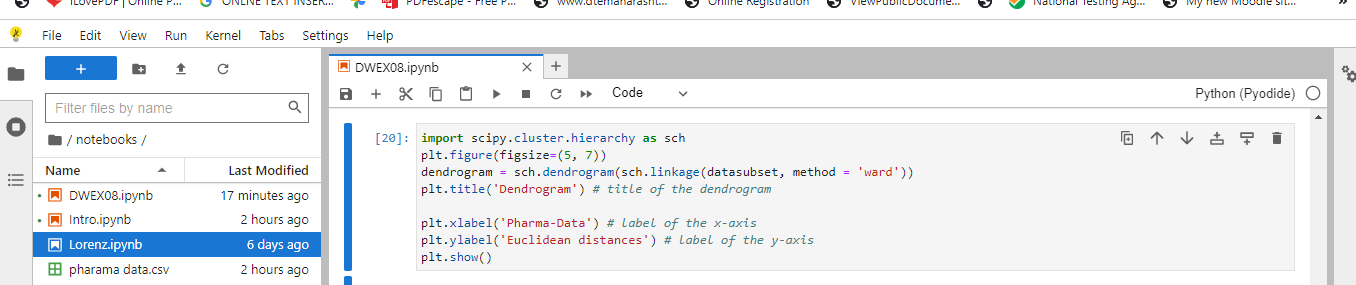
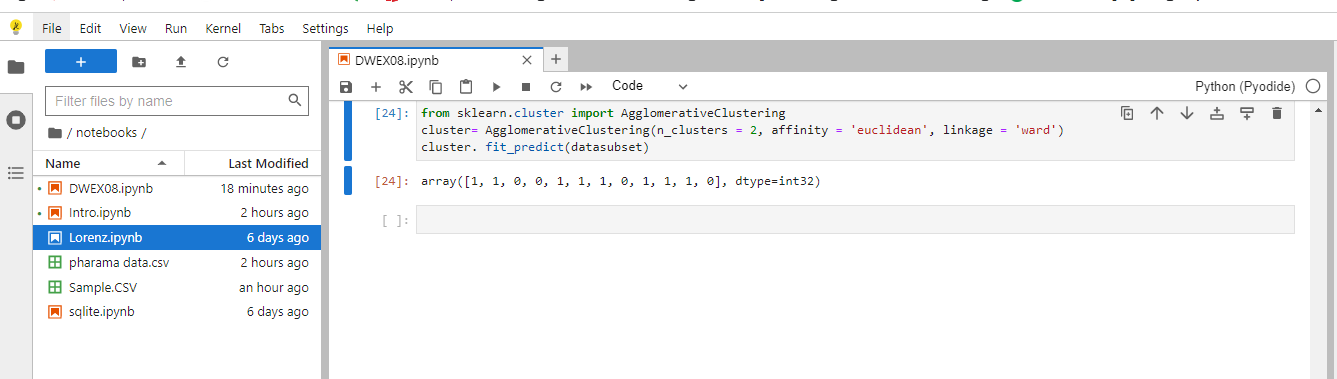
The dendrogram is a tree-like structure that is mainly used to store each step as a memory that the HC algorithm performs. In the dendrogram plot, the Y-axis shows the Euclidean distances between the data points, and the x-axis shows all the data points of the given dataset.

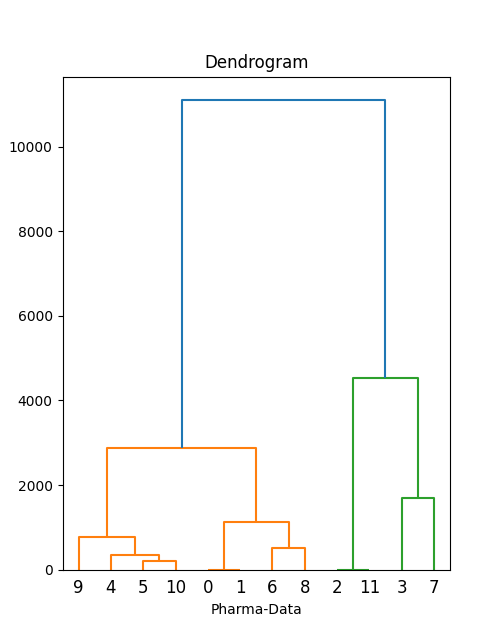
The working of the dendrogram can be explained using the below diagram:



**Program**:







**⚫ Conclusion: -** Thus we learnt and implemented hierarchical clustering algorithm

**EXPERMINT: 05**

**⚫ Aim: Implementation of Data Discretization (any one) & Visualization (any one).**

**⚫ Theory:**

**Data discretization:**

Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy. In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss. There are two forms of data discretization first is supervised discretization, and the second is unsupervised discretization.

**Some Famous techniques of data discretization:**

* Histogram analysis
* Binning
* Cluster Analysis
* Data discretization using decision tree analysis

**Why is Discretization important?**

As we know, an infinite of degrees of freedom mathematical problem poses with the continuous data. For many purposes, data scientists need the implementation of discretization. It is also used to improve signal noise ratio.

**Data Visualization: -**

Data visualization is a graphical representation of quantitative information and data by using visual elements like graphs, charts, and maps. Data visualization convert large and small data sets into visuals, which is easy to understand and process for humans.Data visualization tools provide accessible ways to understand outliers, patterns, and trends in the data.

In the world of Big Data, the data visualization tools and technologies are required to analyze vast amounts of information. Data visualizations are common in your everyday life, but they always appear in the form of graphs and charts. The combination of multiple visualizations and bits of information are still referred to as Info graphics.

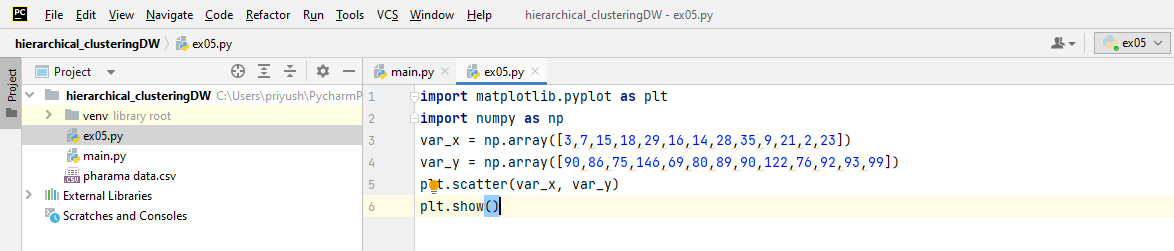
**Scatter Plot:-** A scatter plot (also called a scatterplot, scatter graph, scatter chart, scattergram, or scatter diagram) is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. If the points are coded (color/shape/size), one additional variable can be displayed.

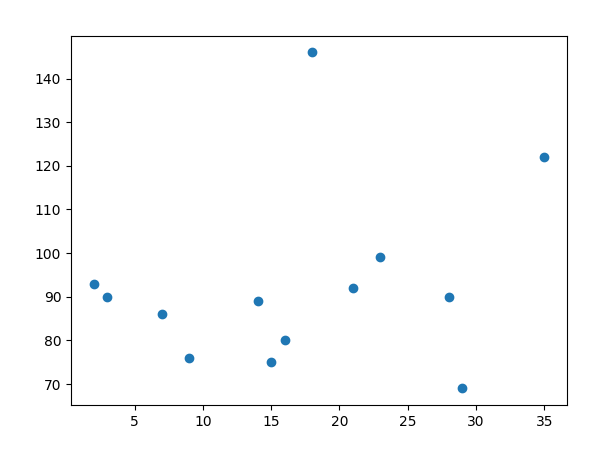
**Bar chart: -** A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart

**Pie chart**: - A pie chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents.

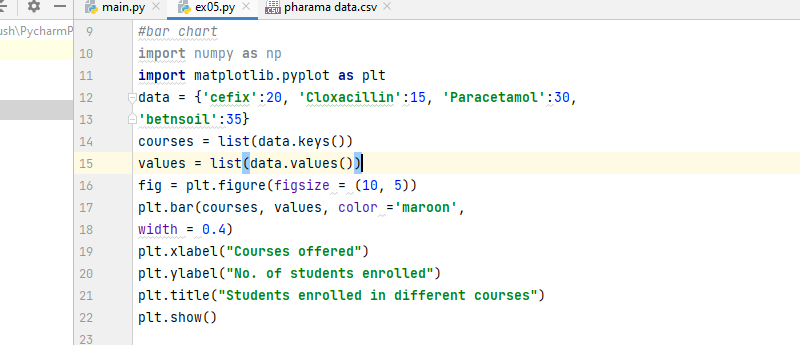
**Histogram**: - A histogram is a graphical representation that organizes a group of data points into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins.

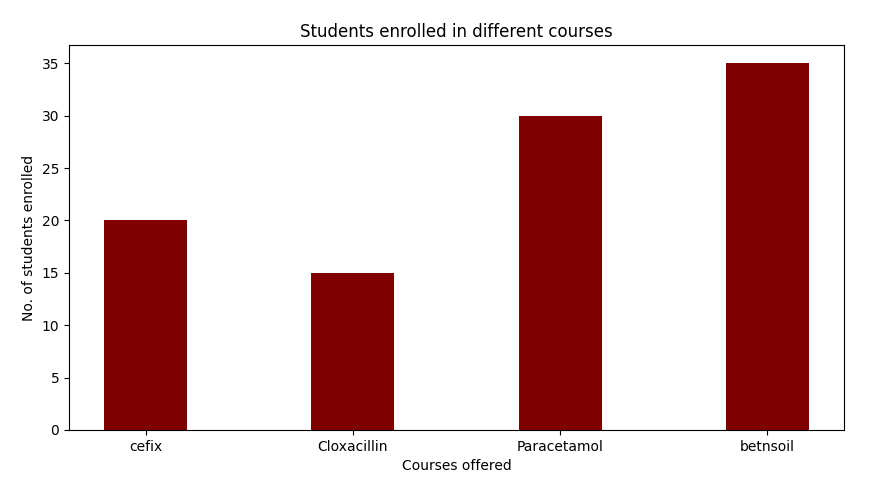
***Scatter Plot:***

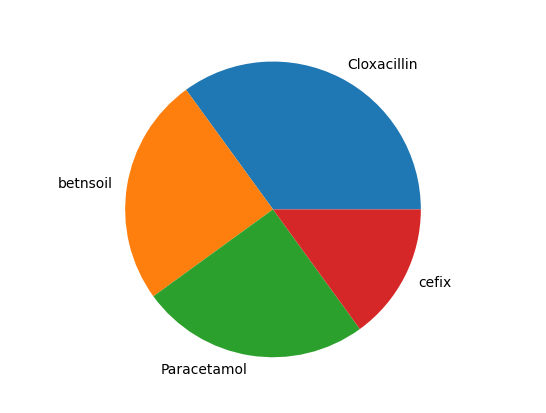
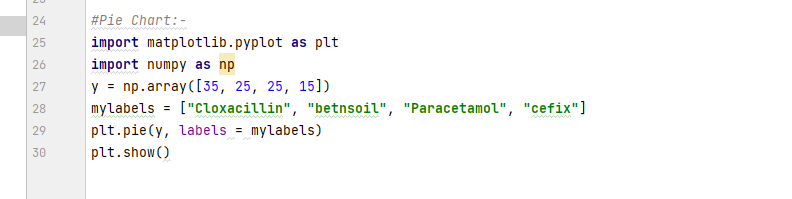




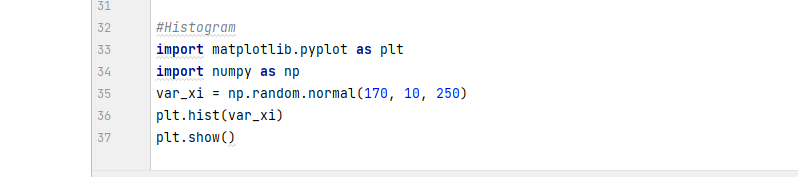
***Bar chart***

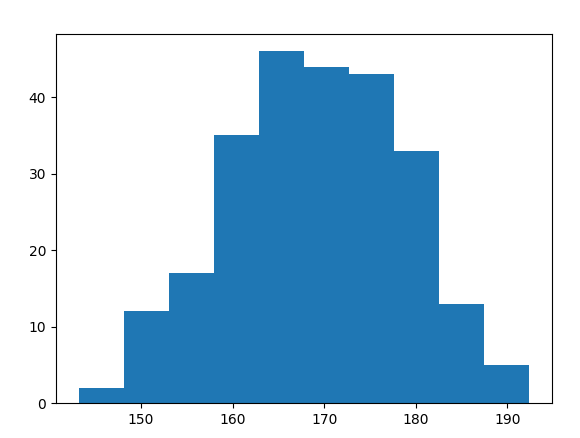




***Pie chart***

***Histogram:***





**⚫ Conclusion:**

We implemented binning as a method of **data** **discretization** and various **visualization** method such as scatter plot, bar chart, pie chart and histogram as a method of data visualization.

**EXPERMINT: 04**

**⚫ Aim: Implementation of Bayesian algorithm.**

**⚫ Theory:**

**Naïve Bayes Classifier Algorithm**

* Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
* It is mainly used in text classification that includes a high-dimensional training dataset.
* Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
* It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

**Bayes' Theorem:**

* Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
* The formula for Bayes' theorem is given as:

Naïve Bayes Classifier Algorithm

**Where,**

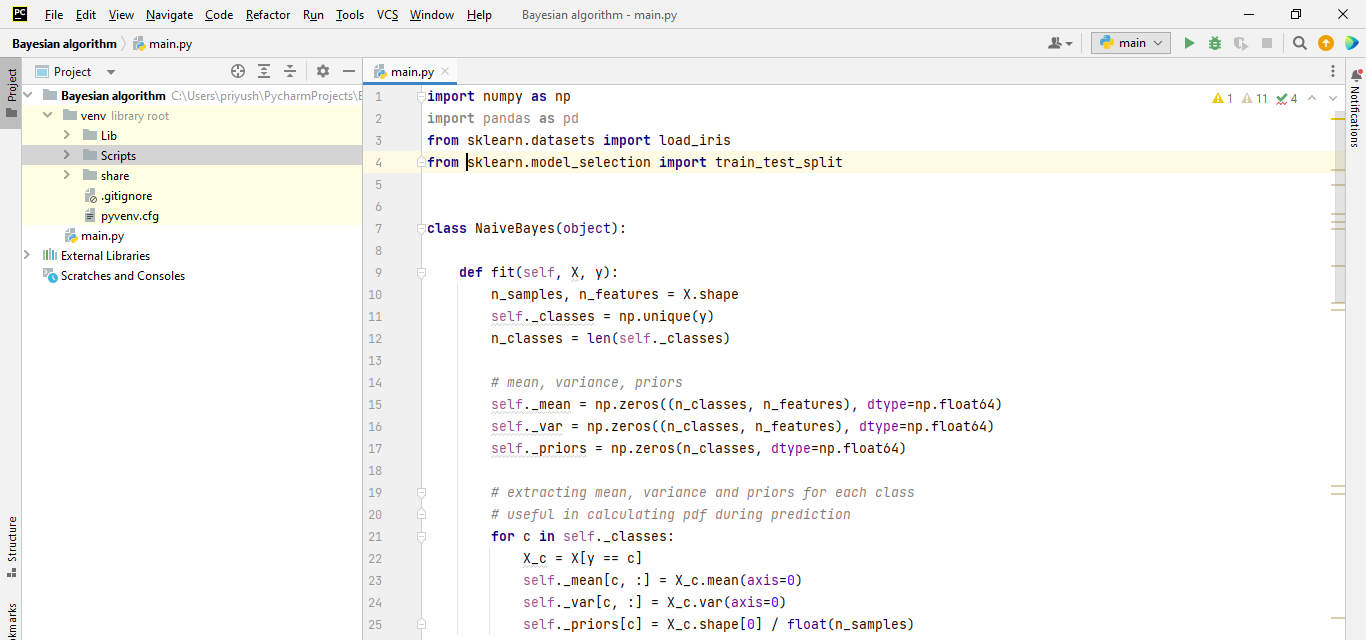
**P(A|B) is Posterior probability**: Probability of hypothesis A on the observed event B.

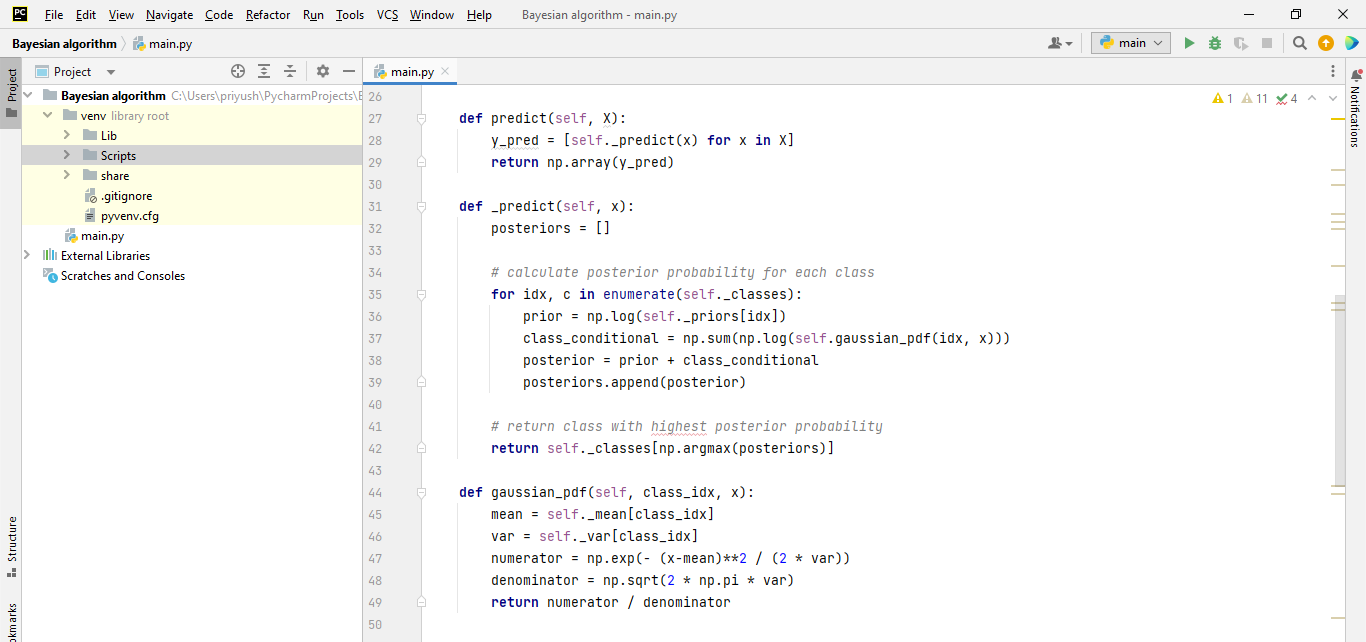
**P(B|A) is Likelihood probability**: Probability of the evidence given that the probability of a hypothesis is true.

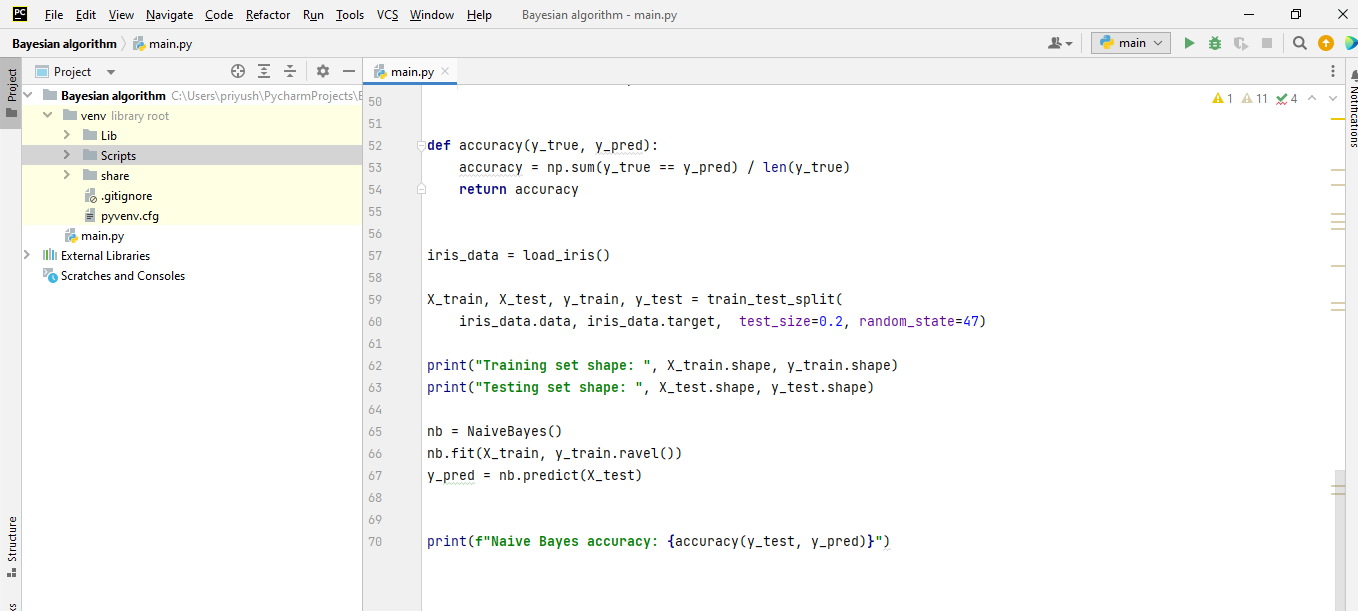
**P(A) is Prior Probability**: Probability of hypothesis before observing the evidence.

**P(B) is Marginal Probability**: Probability of Evidence.

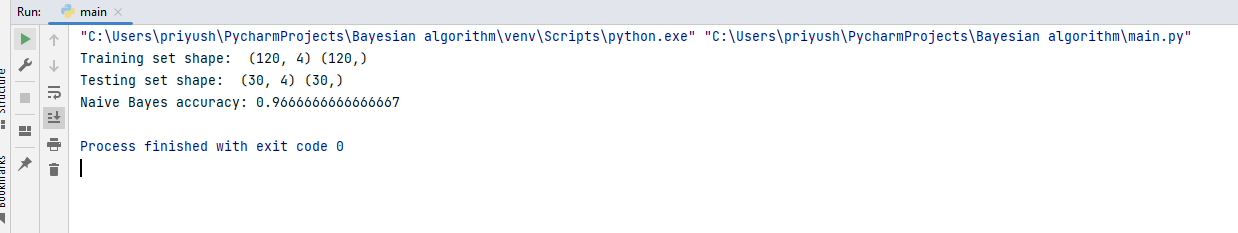
**Code**:







**Output :**



**Advantages of Naïve Bayes Classifier:**

* Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
* It can be used for Binary as well as Multi-class Classifications.
* It performs well in multi-class predictions as compared to the other Algorithms.
* It is the most popular choice for text classification problems.

**Disadvantages of Naïve Bayes Classifier:**

* Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

**Applications of Naïve Bayes Classifier:**

* It is used for Credit Scoring.
* It is used in medical data classification.
* It can be used in real-time predictions because Naïve Bayes Classifier is an eager learner.
* It is used in Text classification such as Spam filtering and Sentiment analysis.

**⚫ Conclusion:**

We implemented Naive **Bayes** **algorithms** are mostly used in sentiment analysis, spam filtering, recommendation systems etc. They are fast and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent.

**EXPERMINT: 07**

**⚫ Aim: Implementation of Clustering algorithm (K-means/K-medoids).**

**⚫ Theory:**

**K-Means Clustering:**

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

* Determines the best value for K center points or centroids by an iterative process.
* Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.



**How does the K-Means Algorithm Work?**

* The working of the K-Means algorithm is explained in the below steps:
* Step-1: Select the number K to decide the number of clusters.
* Step-2: Select random K points or centroids. (It can be other from the input dataset).
* Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.
* Step-4: Calculate the variance and place a new centroid of each cluster.
* Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.
* Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.
* Step-7: The model is ready.

**Advantages of k-means:**

**Simple**: It is easy to implement k-means and identify unknown groups of data from complex data sets.

**Flexible**: K-means algorithm can easily adjust to the changes. If there are any problems, adjusting the cluster segment will allow changes to easily occur on the algorithm.

**Suitable** **in a large dataset**: K-means is suitable for a large number of datasets and it’s computed much faster than the smaller dataset. It can also produce higher clusters.

**Efficient**: The algorithm used is good at segmenting the large data set. Its efficiency depends on the shape of the clusters. K-means works well in hyper-spherical clusters.

**Time complexity**: K-means segmentation is linear in the number of data objects thus increasing execution time. It doesn’t take more time in classifying similar characteristics in data like hierarchical algorithms.

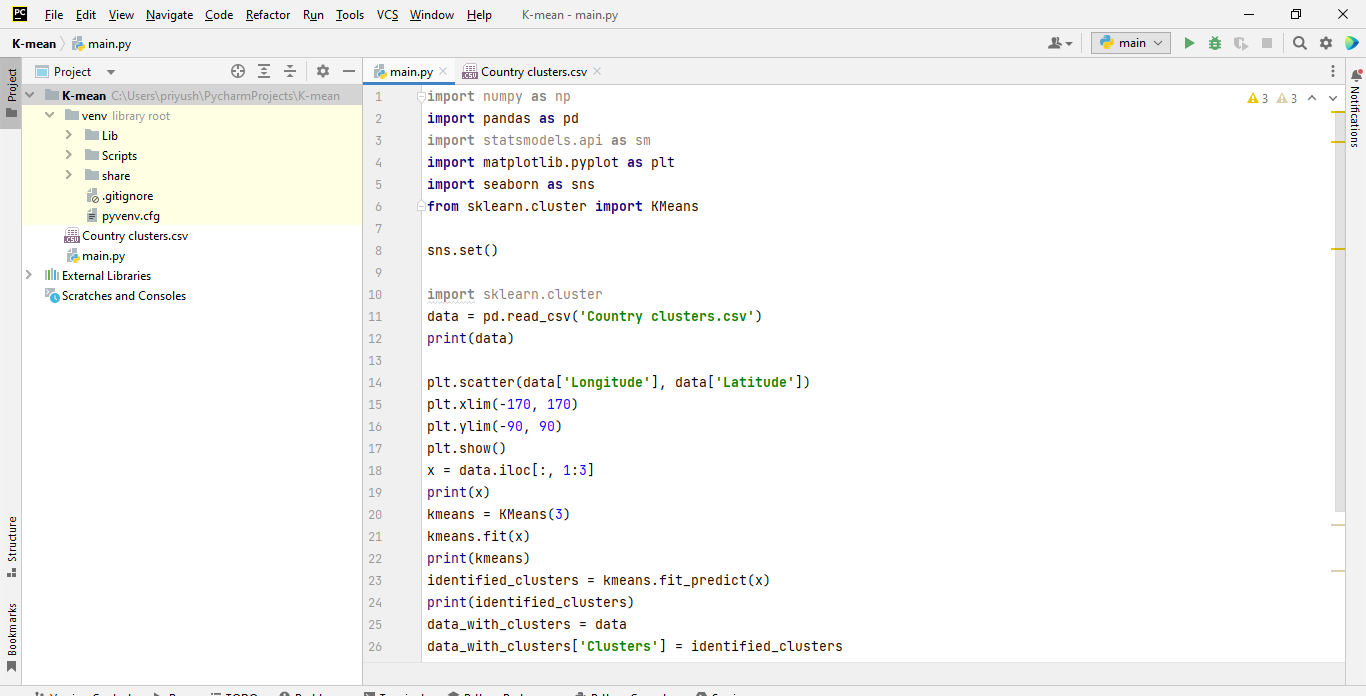
**Disadvantages of k-means:**

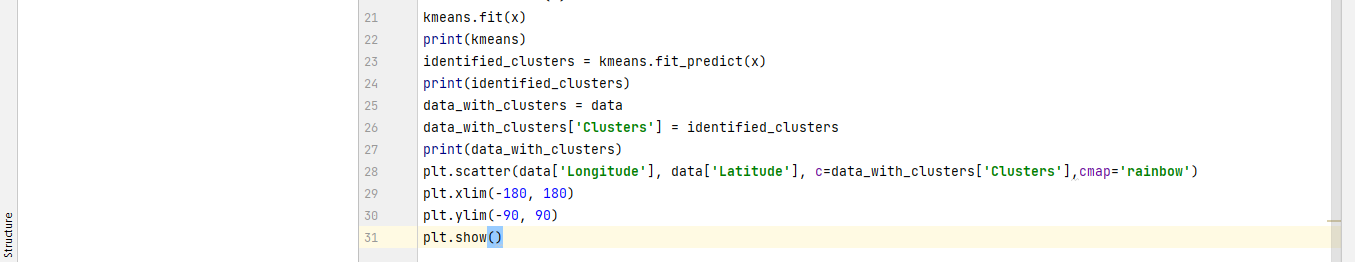
**No-optimal set of clusters:** K-means doesn’t allow the development of an optimal set of clusters and for effective results, you should decide on the clusters before.

**Lacks consistency**: K-means clustering gives varying results on different runs of an algorithm. A random choice of cluster patterns yields different clustering results resulting in inconsistency.

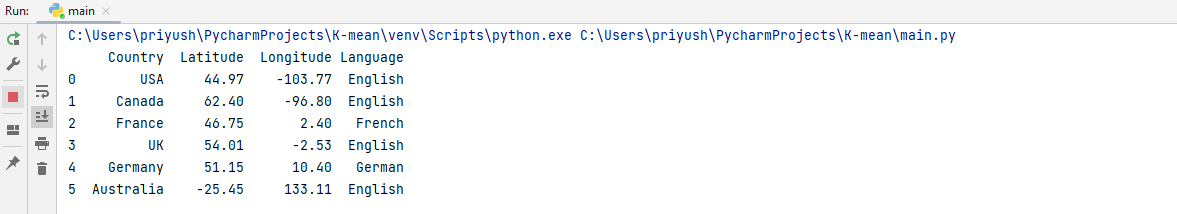
**Uniform effec**t: It produces clusters with uniform sizes even when the input data has different sizes.

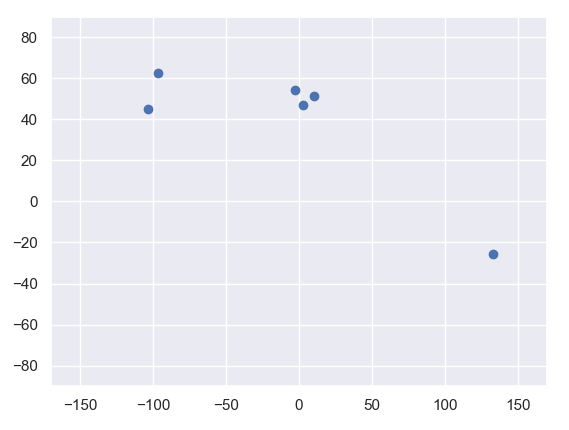
**Source code:**

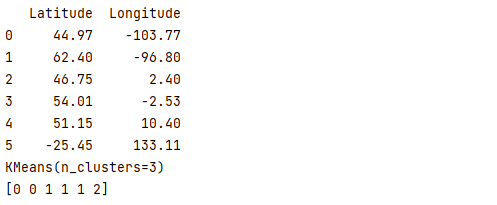


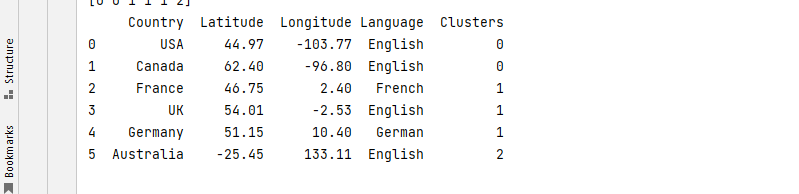


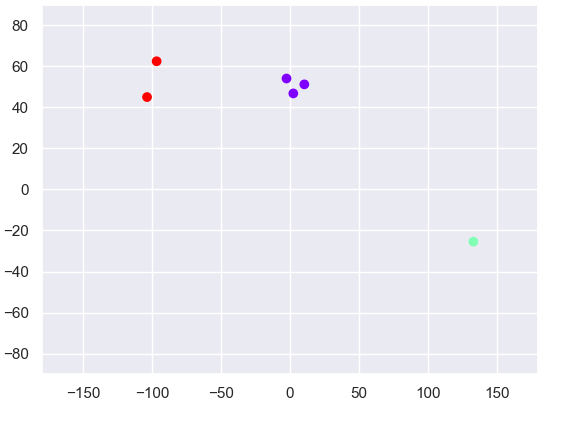
**Output**











|  |  |
| --- | --- |
| Advantages K-Means Algorithm | Disadvantages K-Means Algorithm |
| High Performance | Result repeatability |
| Unlabeled Data | Manual Work |
| Easy to Use | Spherical Clustering Only |

**⚫ Conclusion:** K-means clustering is the unsupervised machine learning algorithm that is part of a much deep pool of data techniques and operations in the realm of Data Science. It is the fastest and most efficient algorithm to categorize data points into groups even when very little information is available aboutdata**.**

**EXPERMINT: 10**

**⚫ Aim: Implementation of Page rank/HITS** **algorithm**

**⚫ Theory:**

**PageRank algorithm**

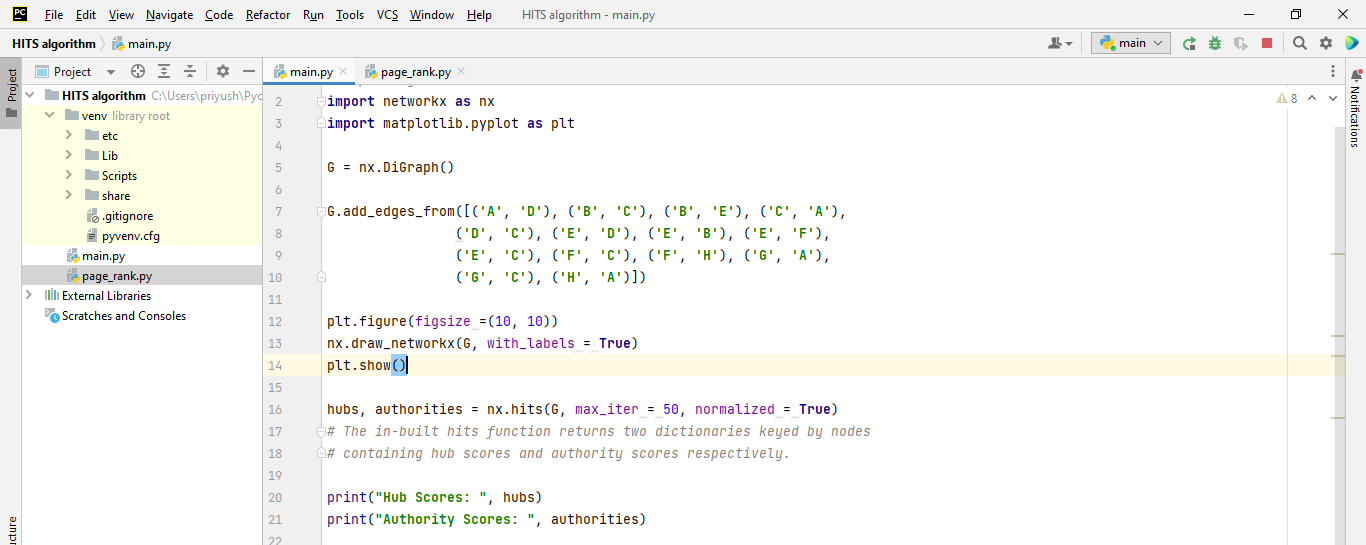
PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

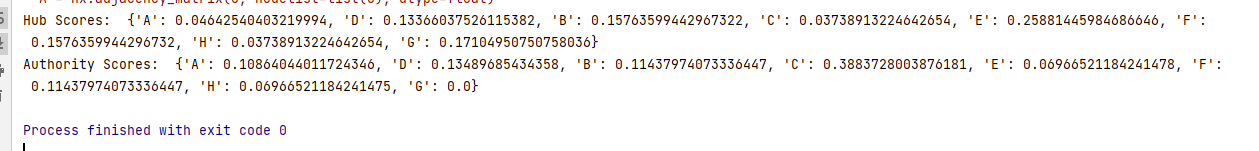
PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google.

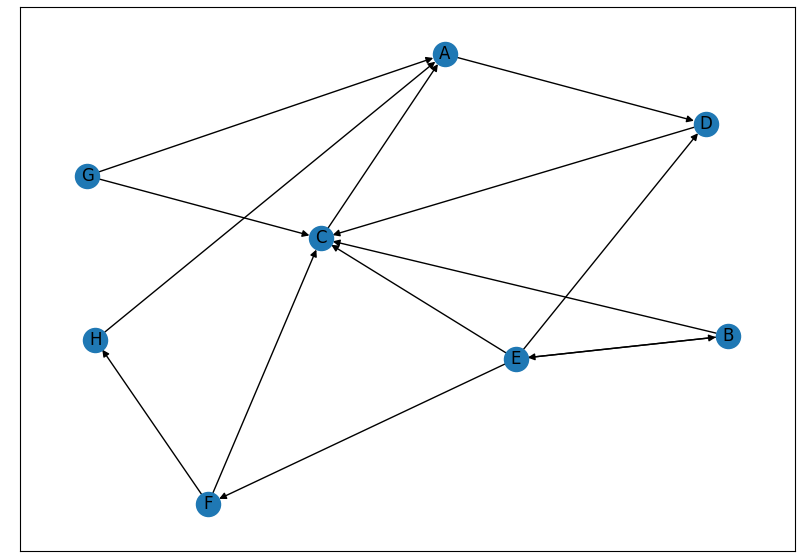
**Algorithm:**

The PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank can be calculated for collections of documents of any size. It is assumed in several research papers that the distribution is evenly divided among all documents in the collection at the beginning of the computational process. The PageRank computations require several passes, called “iterations”, through the collection to adjust approximate PageRank values to more closely reflect the theoretical true value.

**Source code:**







**⚫ Conclusion:** We implemented Page rank/HITS algorithm and So various ranking algorithm such as PageRank, HITS are available that helps the users to navigate in the results. These ranking method uses by search engine that sort and displayed the result to users. So users can easily find the best result.

**VIVA**

**A Data Warehousing (DW):**

A Data Warehousing (DW) is process for collecting and managing data from varied sources to

provide meaningful business insights. A Data warehouse is typically used to connect and analyze

business data from various sources.

A Data Warehousing (DW) is process for collecting and managing data from varied sources to provide meaningful business insights. A Data warehouse is typically used to connect and analyze business data from various sources**.**

“A data warehouse is an electronic storage of an organization’s historical data for the purpose of reporting, analysis and data mining or knowledge discovery.”

**Characteristic**

* Subject oriented
* Integrated
* Time -variant
* Non volatile

Advantages of Data Warehouse (DWH)

* To alleviate the burden on the manufacturing system, the Data Warehouse helps to combine multiple data sources.
* The data warehouse helps to decrease the overall research and reporting turnaround time.
* Restructuring and convergence make documentation and review simpler for the customer.
* The data warehouse allows users to access confidential data from a single location from several sources. It also saves time for users to access data from various sources.
* A large amount of historical information is stored in a data center. This helps users to compare different times and trends to construct possible predictions.

Disadvantages of Data Warehouse (DWH)

* Data centers are high-quality maintenance systems. The data warehouse could be impacted by any reorganization of the business processes and the source systems, resulting in high maintenance costs.
* The data warehouse may sound basic, but it's just too complex for average people.
* Despite the best intentions of project management, the scope of the data storage project will begin to grow.
* At this point, warehouse customers can have various business rules in place.

Benefit DW:

* Scalability
* Save time
* Analysis, trend, predication
* Improves the decision-making process
* Provides competitive advantage

**Metadata:**

Metadata is simply defined as data about data. The data that is used to represent other data is known as metadata. For example, the index of a book serves as a metadata for the contents in the book.

Role of metadata:

* Metadata acts as a directory.
* This directory helps the decision support system to locate the contents of the data warehouse.
* Metadata helps in decision support system for mapping of data when data is transformed from operational environment to data warehouse environment.
* Metadata helps in summarization between current detailed data and highly summarized data.
* Metadata also helps in summarization between lightly detailed data and highly summarized data.
* Metadata is used for query tools.
* Metadata is used in extraction and cleansing tools.
* Metadata is used in reporting tools.
* Metadata is used in transformation tools.
* Metadata plays an important role in loading functions.

A **Data Mart** is a subset of (DW)a directorial information store, generally oriented to a specific purpose or primary data subject which may be distributed to provide business needs.

|  |  |  |
| --- | --- | --- |
| S.NO | Data Warehouse | Data Mart |
| 1. | Data warehouse is a Centralised system. | While it is a decentralised system. |
| 2. | In data warehouse, lightly denormalization takes place. | While in Data mart, highly denormalization takes place. |
| 3. | Data warehouse is top-down model. | While it is a bottom-up model. |
| 4. | To built a warehouse is difficult. | While to build a mart is easy. |
| 5. | In data warehouse, Fact constellation schema is used. | While in this, Star schema and snowflake schema are used. |
| 6. | Data Warehouse is flexible. | While it is not flexible. |
| 7. | Data Warehouse is the data-oriented in nature. | While it is the project-oriented in nature. |
| 8. | Data Ware house has long life. | While data-mart has short life than warehouse. |
| 9. | In Data Warehouse, Data are contained in detail form. | While in this, data are contained in summarized form. |
| 10. | Data Warehouse is vast in size. | While data mart is smaller than warehouse. |
| 11. | It collects data from various data sources. | It generally stores data from a data warehouse. |
| 12. | Long time for processing the data because of large data. | Less time for processing the data because of handling only a small amount of data. |
| 13. | Complicated design process of creating schemas and views. | Easy design |

| **S.N     o** | **ER Modeling** | **Dimensional Modeling** |
| --- | --- | --- |
| 1 | It is transaction-oriented. | It is subject-oriented. |
| 2 | Entities and Relationships. | Fact Tables and Dimension Tables. |
| 3 | Few levels of granularity. | Multiple levels of granularity. |
| 4 | Real-time information. | Historical information. |
| 5 | It eliminates redundancy. | It plans for redundancy. |
| 6 | High transaction volumes using few records at a time. | Low transaction volumes using many records at a time. |
| 7 | Highly Volatile data. | Non-volatile data. |
| 8 | Physical and Logical Model. | Physical Model. |
| 9 | [Normalization](https://www.geeksforgeeks.org/introduction-of-database-normalization/) is suggested. | [De-Normalization](https://www.geeksforgeeks.org/denormalization-in-databases/) is suggested. |
| 10 | [OLTP](https://www.geeksforgeeks.org/on-line-transaction-processing-oltp-system-in-dbms/) Application. | [OLAP](https://www.geeksforgeeks.org/olap-operations-in-dbms/)Application. |
| **Ex** | The application is used for buying products from e-commerce websites like Amazon. | Application to analyze buying patterns of the customer of the various cities over the past 10 years. |

**Fact**

It is something measurable

A Fact Table contains

1. Measurements/facts
2. Foreign key to dimension table

**dimension table :**

* A **dimension table** contains dimensions of a fact.
* They are joined to fact table via a foreign key.

**fact less fact table**

fact less fact table is means only the key available in the fact there is no measure available .

**Star Schema:** In a star schema, the fact table will be at the center and is connected to the dimension tables.

* Adv 1. The tables are completely in a denormalized structure.
* Simplest DW schema
* Easy to understand.

**Snowflake Schema :** A snowflake schema is an extension of star schema where the dimension tables are connected to one or more dimensions.

* The performance of SQL queries is a bit less when compared to star schema as more number of joins are involved.
* Hybrid
* Data redundancy is low .

|  |  |  |
| --- | --- | --- |
| S.NO | Star Schema | Snowflake Schema |
| 1. | In star schema, The fact tables and the dimension tables are contained. | While in snowflake schema, The fact tables, dimension tables as well as sub dimension tables are contained. |
| 2. | Star schema is a top-down model. | While it is a bottom-up model. |
| 3. | Star schema uses more space. | While it uses less space. |
| 4. | It takes less time for the execution of queries. | While it takes more time than star schema for the execution of queries. |
| 5. | In star schema, Normalization is not used. | While in this, Both normalization and denormalization are used. |
| 6. | It’s design is very simple. | While it’s design is complex. |
| 7. | The query complexity of star schema is low. | While the query complexity of snowflake schema is higher than star schema. |
| 8. | It’s understanding is very simple. | While it’s understanding is difficult. |
| 9. | It has less number of foreign keys. | While it has more number of foreign keys. |
| 10. | It has high data redundancy. | While it has low data redundancy. |

**ETL is a process in Data Warehousing**

ETL is a process in Data Warehousing and it stands for **Extract, Transform and Load**. It is a process in which an ETL tool **extracts** the data from various data source systems, **transforms** it in the staging area, and then finally, **loads** it into the Data Warehouse system.

**Extraction:**

The first step of the ETL process is extraction. In this step, data from various source systems is extracted which can be in various formats like relational databases, No SQL, XML, and flat files into the staging area.

**Transformation:**

The second step of the ETL process is transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format.

**Loading**:

The third and final step of the ETL process is loading. In this step, the transformed data is finally loaded into the data warehouse.

**OLAP stands for On-Line Analytical Processing:**

OLAP stands for **On-Line Analytical Processing.** OLAP is a classification of software technology which authorizes analysts, managers, and executives to gain insight into information through **fast, consistent, interactive access in** a wide variety of possible views of data that has been transformed from raw information to reflect the real dimensionality of the enterprise as understood by the clients.

**Types of OLAP Servers**

We have four types of OLAP servers −

* Relational OLAP (ROLAP)
* Multidimensional OLAP (MOLAP)
* Hybrid OLAP (HOLAP)
* Specialized SQL Servers

OLAP Operations

Since OLAP servers are based on multidimensional view of data, we will discuss OLAP operations in multidimensional data.

Here is the list of OLAP operations −

* Roll-up
* Drill-down
* Slice and dice
* Pivot (rotate)

### **Roll-up**

Roll-up performs aggregation on a data cube in any of the following ways −

* By climbing up a concept hierarchy for a dimension
* By dimension reductio

### **Roll-up**

Roll-up performs aggregation on a data cube in any of the following ways −

* By climbing up a concept hierarchy for a dimension
* By dimension reductio

## Slice

A **slice** is a subset of the cubes corresponding to a single value for one or more members of the dimension.

## Dice

The dice operation describes a sub cube by operating a selection on two or more dimension.

## Pivot

The pivot operation is also called a rotation. Pivot is a visualization operations which rotates the data axes in view to provide an alternative presentation of the data.

Adv olap

* OLTP offers accurate forecast for revenue and expense.
* It provides a solid foundation for a stable business /organization due to timely modification of all transactions.
* OLTP makes transactions much easier on behalf of the customers.
* It broadens the client base for an organization by speeding up and simplifying individual processes.
* OLTP provides support for bigger databases.
* Partition of data for data manipulation is easy.
* We need OLTP to use the tasks which are frequently performed by the system.
* When we need only a small number of records.

Dis olap

* If the OLTP system faces hardware failures, then online transactions get severely affected.
* OLTP systems allow multiple users to access and change the same data at the same time, which many times created an unprecedented situation.
* If the server hangs for seconds, it can affect to a large number of transactions.
* OLTP required a lot of staff working in groups in order to maintain inventory.
* Online Transaction Processing Systems do not have proper methods of transferring products to buyers by themselves.

Application OLTP:

* Financial
* Sales
* Business

**2 Introduction to Data Mining, Data Exploration and Data Pre-processing**

**Data Mining:**

The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called **Data Mining.**

**KDD:** The term KDD stands for Knowledge Discovery in Databases. It refers to the broad procedure of discovering knowledge in data and emphasizes the high-level applications of specific Data Mining techniques.

1. ***Data Cleaning***: Data cleaning is defined as removal of noisy and irrelevant data from collection.
   * Cleaning in case of ***Missing values***.
   * Cleaning ***noisy*** data, where noise is a random or variance error.
   * Cleaning with ***Data discrepancy detection*** and ***Data transformation tools***.
2. ***Data Integration***: Data integration is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse).
   * Data integration using ***Data Migration tools***.
   * Data integration using ***Data Synchronization tools***.
   * Data integration using ***ETL***(Extract-Load-Transformation) process.
3. ***Data Selection***: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
   * Data selection using ***Neural network***.
   * Data selection using ***Decision Trees***.
   * Data selection using ***Naive bayes***.
   * Data selection using ***Clustering***, ***Regression***, etc.
4. ***Data Transformation***: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.

Data Transformation is a two step process:

* + ***Data Mapping***: Assigning elements from source base to destination to capture transformations.
  + ***Code generation***: Creation of the actual transformation program.

1. ***Data Mining***: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
   * Transforms task relevant data into ***patterns***.
   * Decides purpose of model using ***classification*** or ***characterization***.
2. ***Pattern Evaluation***: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures.
   * Find ***interestingness score*** of each pattern.
   * Uses ***summarization*** and ***Visualization*** to make data understandable by user.
3. ***Knowledge representation***: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
   * Generate ***reports***.
   * Generate ***tables***.
   * Generate ***discriminant rules***, ***classification rules***, ***characterization rules***, etc.

**Data Visualization** refers to the visual representation of data with the help of comprehensive charts, images, lists, charts, and other visual objects.

**Data reduction** techniques ensure the integrity of data while reducing the data. Data reduction is a process that reduces the volume of original data and represents it in a much smaller volume. Data reduction techniques are used to obtain a reduced representation of the dataset that is much smaller in volume by maintaining the integrity of the original data. By reducing the data, the efficiency of the data mining process is improved, which produces the same analytical results.

**Data Mining Applications**

Here is the list of areas where data mining is widely used −

* Financial Data Analysis
* Retail Industry
* Telecommunication Industry
* Biological Data Analysis
* Other Scientific Applications
* Intrusion Detection
* An **attribute** is an object’s property or characteristics. For example. A person’s hair colour, air humidity etc.
* An attribute set defines an **object**. The **object** is also referred to as a record of the instances or entity.

Different **types of attributes or data types:**

1. **Nominal Attribute:**   
   Nominal Attributes only provide enough attributes to differentiate between one object and another. Such as Student Roll No., Sex of the Person.
2. **Ordinal Attribute:**   
   The ordinal attribute value provides sufficient information to order the objects. Such as Rankings, Grades, Height
3. **Binary Attribute:**   
   These are 0 and 1. Where 0 is the absence of any features and 1 is the inclusion of any characteristics.
4. **Numeric attribute:** It is quantitative, such that quantity can be measured and represented in integer or real values ,are of two types  
   **5. Interval Scaled attribute:**   
   It is measured on a scale of equal size units, these attributes allow us to compare such as temperature in C or F and thus values of attributes have ordered.  
    **Ratio Scaled attribute:**   
   Both differences and ratios are significant for Ratio. For eg. age, length, and Weight.

**3 Classification + 4**

**Classifiers Of Machine Learning:**

1. Decision Trees
2. Bayesian Classifiers
3. Neural Networks
4. K-Nearest Neighbour
5. Support Vector Machines
6. Linear Regression
7. Logistic Regression

**What is a Decision Tree Algorithm?**

A decision tree is a tree in which every node is either a leaf node or a decision node. This tree takes an input an object and outputs some decision. All Paths from root node to the leaf node are reached by either using AND or OR or BOTH. The tree is constructed using the regularities of the data. The decision tree is not affected by Automatic Data Preparation.

**What Is Naive Bayes Algorithm?**

Naive Bayes Algorithm is used to generate mining models. These models help to identify relationships between input columns and the predictable columns. This algorithm can be used in the initial stage of exploration. The algorithm calculates the probability of every state of each input column given predictable columns possible states. After the model is made, the results can be used for exploration and making predictions.

**Clustering Algorithm?**

**Clustering algorithm is used to group sets of data with similar characteristics also called as cluster**s. These clusters help in making faster decisions, and exploring data. The algorithm first identifies relationships in a dataset following which it generates a series of clusters based on the relationships. The process of creating clusters is iterative. The algorithm redefines the groupings to create clusters that better represent the data**.**

**agglomerative clustering**

 In agglomerative clustering, each data point act as an individual cluster and at each step, data objects are grouped in a bottom-up method. Initially, each data object is in its cluster. At each iteration, the clusters are combined with different clusters until one cluster is formed.

**Divisive Hierarchical Clustering**

Divisive hierarchical clustering is exactly the opposite of Agglomerative Hierarchical clustering. In Divisive Hierarchical clustering, all the data points are considered an individual cluster, and in every iteration, the data points that are not similar are separated from the cluster.

### **Advantages of Hierarchical clustering**

* It is simple to implement and gives the best output in some cases.
* It is easy and results in a hierarchy, a structure that contains more information.
* It does not need us to pre-specify the number of clusters

### **Disadvantages of hierarchical clustering**

* It breaks the large clusters.
* It is Difficult to handle different sized clusters and convex shapes.
* It is sensitive to noise and outliers.
* The algorithm can never be changed or deleted once it was done previously

**5) Mining frequent patterns and associations**

**Market basket analysis:**

Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves Analyzing large data sets, such as purchase history, to reveal product groupings and products that are likely to be purchased together.

**frequent item set:**

An itemset consists of two or more items. An itemset that occurs frequently is called a frequent itemset. Thus frequent itemset mining is a data mining technique to identify the items that often occur together. For Example, Bread and butter, Laptop and Antivirus software, etc

**Association rule**

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is a Market Based Analysis.

**Support** is an indication of how frequently the items appear in the data.

**Confidence** indicates the number of times the if-then statements are found true.

**Apriori algorithm**

Apriori algorithm refers to an algorithm that is used in mining frequent products sets and relevant association rules.

* It used to solve frequent itemset problem.
* It user bottom -up approach
* Apriori algorithm helps the customers to buy their products with ease and increases the sales performance of the particular store.

The given three components comprise the Apriori algorithm.

* Support
* Confidence
* Lift

**Advantages of Apriori Algorithm**

* It is used to calculate large itemset.
* Simple to understand and apply.

**Disadvantages of Apriori Algorithms**

* Apriori algorithm is an expensive method to find support since the calculation has to pass through the whole database.
* Sometimes, you need a huge number of candidate rules, so it becomes computationally more expensive.

**Mining Frequent Itemset without candidate generation i.e. FP TREE**

**FP TREE :**

* frequent patterns
* The FP-Growth Algorithm is an alternative way to find frequent item sets without using candidate generations, thus improving performance.
* FP TREE Is tree like structure which consist of one root node called null and sub tree
* Sub consists of 1item name 2count 3node-link

**Disadvantages of FP-Growth Algorithm**

This algorithm also has some disadvantages, such as:

* FP Tree is more cumbersome and difficult to build than Apriori.
* It may be expensive.
* The algorithm may not fit in the shared memory when the database is large.

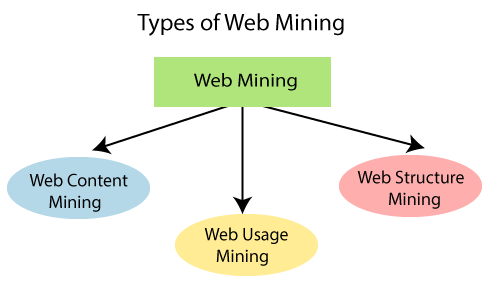
**Difference between Apriori and FP Growth Algorithm**

* Apriori and FP-Growth algorithms are the most basic FIM algorithms. There are some basic differences between these algorithms

**Web mining**

**Web Mining** is the process of Data Mining techniques to automatically discover and extract information from Web documents and services. The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns.

**There are three types of data mining:**



**Web Content Mining:** Web content mining is the application of extracting useful information from the content of the web documents. Web content consist of several types of data – text, image, audio, video etc. Content data is the group of facts that a web page is designed.

**Web Structure Mining:** Web structure mining is the application of discovering structure information from the web. The structure of the web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages.

**Web Usage Mining:** Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets. And these patterns enable you to understand the user behaviours or something like that.

**Application of Web Mining:**

Web mining has an extensive application because of various uses of the web. The list of some applications of web mining is given below.

* Marketing and conversion tool
* Data analysis on website and application accomplishment.
* Audience behavior analysis
* Advertising and campaign accomplishment analysis.
* Testing and analysis of a site.

Web structure mining technique:

* Page rank
* Clever

PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages. According to Google:

*PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.*

Extra

Mention Some Of The Data Mining Techniques?

* ○ Statistics
* ○ Machine learning ○
* Decision Tree
* ○ Hidden markov models
* ○ Artificial Intelligence
* ○ Genetic Algorithm ○ Meta learning