Faculty of Computer and Information Science
Ain Shams University
2016-2017

Multi-View Data Mining Visualization Tool

Supervised by:

- Dr. Wedad Hussein
- T.A Amira Ali

Represented by:

- Mahmoud Nabawy Hassan Youssef
- Mohamed KordeyTahaElshafey
- Mina AeadMortagy
- Mostafa Ali Abd El-haleem
- Mostafa Amin AminShedeed

Acknowledgement

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them.

We are highly indebted to Dr. Wedad Hussein for her guidance and constant supervision as well as for providing necessary information regarding the project & also for her support in completing the project.

We would like to express our gratitude towards TA. Amira Ali for her kind cooperation and encouragement which help us in completion of this project.
Our thanks and appreciations also go to our team in developing the project and people who have willingly helped me out with their abilities.

Abstract

Since the amount of data for each organization is very large , the organizations are looking for ways to make use of that large amounts of data. This data is needed to generate or extract useful information as much as possible to help the organizations' managers to do some actions or take decisions to grow up without having to discover these large amounts of data on their own .

In our project we develop a desktop based software application that not only applies some processing techniques on that data to extract the useful information but also visualizes the results that come out from the processing phase. The visualization techniques are dependent on the problem that need to be solved and also the software can help the user to manipulate data on graphs.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation:

Organizations grow up and the amount of data that is available in these organizations is becoming very large. However, these organizations are still missing that part of information that enables the organizations' leaders to make quick , right decisions to improve the profit of their organization or enhance the costs of producing products ,…etc.

Data Mining can be defined as the process of discovering patterns in large data sets. It involves methods at the intersection of artificial intelligence, machine learning and statistics to extract information from a data set and transform it into an understandable structure for further use.

Data mining can be used to help the organizations with the decision making process. Also visualization techniques can be used to provide the organization with the big picture of the results through visualizing the result using some visualization techniques dependent on data mining algorithms used.

1.2 Problem Definition:

According to the increasing in the amount and the complexity of data, it became difficult for the user to search data to get the information he needs.

Even with the use of data mining techniques the interpretation of their results can be very difficult. Visualization techniques can help with the interpretation of the results and make them clearer to the user.

So we are going to apply not only some techniques of data mining to speed search and data retrieval but also some techniques of data visualization , to help users see data in different manners and to help updating/modifying data rapidly.

1.3 Objective:

The main objective of the project is to develop a data mining tool with built-in visualization capabilities , to help the user to apply data mining algorithms on his/her data set and then see the result visualized using a technique of his/her choice.

1.4 Document Organization

The following chapters of this document are organized as follows:

Chapter 2: Background
This chapter gives an overview of the scientific background behind the project . In this chapter we discuss the concept of data mining and the concept of visualization in more details and the different algorithms of data mining and different techniques of visualization used in the project.

Chapter 3: Analysis and Design
This chapter presents the outcomes of the analysis and design phases of the Project . It discusses the functional and non-functional requirements of the system ,this chapter also offers a detailed account of functionalities and processes offered by the project expressed as UML diagrams.

Chapter 4: Implementation
This chapter explains the technologies , tools , algorithms and the techniques used in the implementation of the system and the role of each of them in theproject.

Chapter 5: User Manual
The user manual offers a guideline on how to operate the system and make use of the different functionalities of the system.

Chapter 6: Conclusion and Future Work
This chapter presents the possible directions for the future work , also the final conclusions of our work is presented in this chapter.

Chapter 2

Background

In this chapter we introduce an overview of the scientific background behind the project ,basically our project is implemented using two fields

2.1-Data Mining:-

Data mining is the analysis step of the "knowledge discovery in databases" process, also it is the computing process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems . It is an interdisciplinary subfield of computer science , The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Data mining includes methods and techniques such as:-

2.1.1-Pre-Processing

Data pre-processing is an important step in the data mining process, Since data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time.

Data pre-processing includes:

- Data cleaning:a process used to determine inaccurate, incomplete or unreasonable data and then improve the quality through correcting of detected errors and omissions. Generally data cleaning reduces errors and improves the data quality.
- Data Integration: the merging of data from multiple data stores.
- Data Reduction:is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form. The basic concept is the reduction of multitudinous amounts of data down to the meaningful parts.

Also pre-processing can include other methods as:- Instance selection, normalization, transformation, feature extraction and selection, etc.

2.1.2-Clustering

Is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Common algorithms of clustering:

- Partitioning methods (K-Means , K-Medoids).

  →K-means is one of the simplest algorithm which uses unsupervised learning method to solve known clustering issues. It works really well with large datasets, also k-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean.

  →k-medoid is a classical partitioning technique of clustering that clusters thedata set of n objects into k clusters known a priori. A useful tool fordetermining k is the silhouette. It is more robust to noise and outliers ascompared to k-means because it minimizes a sum of pairwise dissimilaritiesinstead of a sum of squared Euclidean distances. A medoid can be defined as the object of a cluster whose average dissimilarity to all the objects in thecluster is minimal. i.e. it is a most centrally located point in the cluster.

- Hierarchical Methods (Agglomerative, Divisive).

Is a method of cluster analysis which seeks to build a hierarchy of clusters.

Strategies for hierarchical clustering generally fall into two types:

→ **Agglomerative**: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

→ **Divisive**: This is a "top down" approach: all observations start in one cluster,and splits are performed recursively as one moves down the hierarchy.

- Density-Based Methods (DBSCAN).
  It is a density-based clustering algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature

Why we chose K-Means to implement:

Three key features of *k*-means which make it efficient are often regarded as its biggest drawbacks:

- Euclidean distance is used as a metric and variance is used as a measure of cluster scatter.
- The number of clusters *k* is an input parameter: an inappropriate choice of *k* may yield poor results. That is why, when performing k-means, it is important to run diagnostic checks for determining the number of clusters in the data set.
- Convergence to a local minimum may produce counterintuitive ("wrong") results

K-Means Disadvantages :

1) Difficult to predict K-Value.

2) With global cluster, it didn't work well.

3) Different initial partitions can result in different final clusters.

4) It does not work well with clusters (in the original data) of Different size and Different density.

2.1.3-Classification

Is a data mining technique that assigns categories to a collection of data in order to aide in more accurate predictions and analysis. Also called sometimes called a Decision Tree, classification is one of several methods intended to make the analysis of very large data sets effective.

Common algorithms of clustering:

- Decision Tree.

  A decision tree is a simple representation for classifying examples, commonly used in data mining .The goal is to create a model that predicts the value of a target variable based on several input variables.

- Naive Bayesian.

  Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihoodtraining can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximationas used for many other types of classifiers

- KNN(K-nearest neighbor).

  Is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms.

Why we chose decision tree to implement:

- Decision trees implicitly perform variable screening or feature selection
- Decision trees require relatively little effort from users for data preparation
- Nonlinear relationships between parameters do not affect tree performance
- The best feature of using trees for analytics - easy to interpret and explain to executives

Disadvantages of decision trees.

1. Decision Trees do not work well if you have smooth boundaries. i.e they work best when you have discontinuous piece wise constant model. If you truly have a linear target function decision trees are not the best.
2. Decision Tree's do not work best if you have a lot of un-correlated variables. Decision tree's work by finding the interactions between variables. if you have a situation where there are no interactions between variables linear approaches might be the best.

2.1.4- Mining Frequent Patterns(Association Rules):

Frequent pattern mining searches for recurring relationships in a given data set. This section introduces the basic concepts of frequent pattern mining for the discovery of interestingassociationsandcorrelationsbetweenitemsetsintransactionalandrelational databases.

Common algorithms of mining frequent patterns:

- A priori

  is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

- FP Growth

  is an improvement of apriori designed to eliminate some of the heavy bottlenecks in apriori. The algorithm was planned with the bennefits of mapReduce taken into account, so it works well with any distributed system focused on mapReduce. FP-Growth simplifies all the problems present in apriori by using a structure called an FP-Tree. In an FP-Tree each node represents an item and it's current count, and each branch represents a different association.

Why we chose A priori to implement:

The Apriori Algorithm calculates more sets of frequent items.

Disadvantages of Apriori

- The candidate generation could be extremely slow (pairs, triplets, etc.).
- The candidate generation could generate duplicates depending on the implementation.
- The counting method iterates through all of the transactions each time.
- Constant items make the algorithm a lot heavier.
- Huge memory consumption

2.2-Visualization:-

The term visualization means representing data using graphs ,Datavisualization aims to communicate Data clearly and effectively through graphical representation . Data visualization has been used extensively in many applications for example, at work for reporting, managing business operations, and tracking progress of tasks. More popularly, we can take advantage of visualization techniques to discover data relationships that are otherwise not easily observable by looking at the raw data. Nowadays, people also use data visualization to create fun and interesting graphics.

Data visualization can include several techniques such as:-

- scatter-plot matrix and 3-D scatter-plot: can be used to visualize the results of clustering of data.
- parallel coordinates: can be used to visualize the results of both classification and clustering of data.
- Rule graph: can be used to visualize the results of mining frequent patterns.

Also there are other techniques and graphs of visualization can be used to visualize the data for example: pie chart, histogram ,column chart, line chart , tree-maps , circle segment technique , Pixel-oriented,.. etc.

2.3 - Existing Similar Projects:

1) **Weka(Waikato Environment for Knowledge Analysis) :**

Is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License.

Weka (pronounced to rhyme with Mecca) contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions.[1] The original non-Java version of Weka was a Tcl/Tk front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains,[2][3] but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research. Advantages of Weka include:

- Free availability under the GNU General Public License.
- Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces.

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka.[4] Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling.

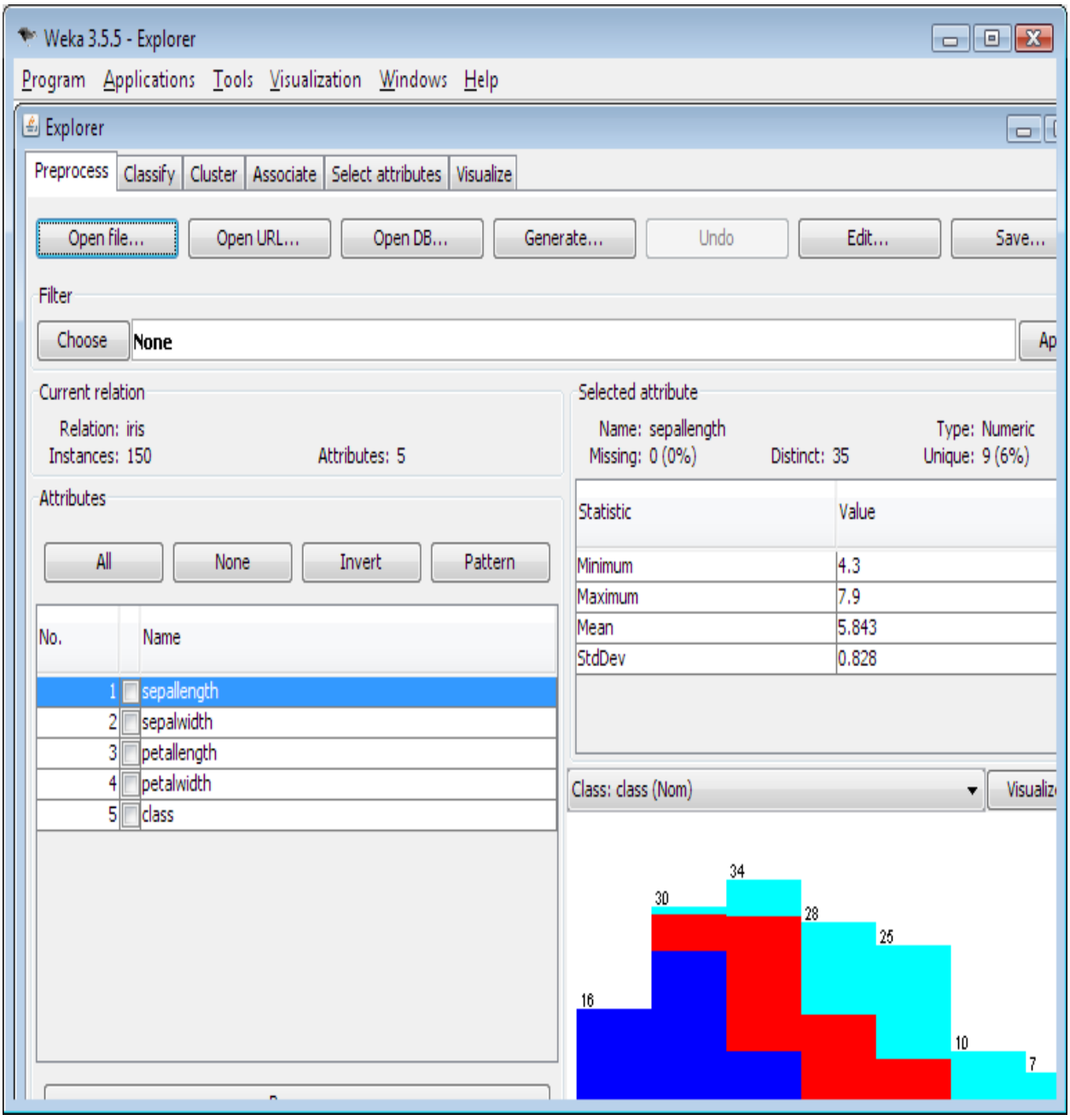


Fig.2.1 Weka 3.5.5

2) **Rattle GUI :**

is a free and open source software (GNU GPL v2) package providing a graphical user interface (GUI) for data mining using the R statistical programming language. Rattle is used in a variety of situations. Currently there are 15 different government departments in Australia, in addition to various other organizations around the world, which use Rattle in their data mining activities and as a statistical package.

Rattle provides considerable data mining functionality by exposing the power of the R Statistical Software through a graphical user interface. Rattle is also used as a teaching facility to learn the R software Language. There is a Log Code tab, which replicates the R code for any activity undertaken in the GUI, which can be copied and pasted. Rattle can be used for statistical analysis, or model generation. Rattle allows for the dataset to be partitioned into training, validation and testing. The dataset can be viewed and edited. There is also an option for scoring an external data file.

Fig. 2.2 Rattle GUI

Chapter 3
Analysis and Design

This chapter presents the outcomes of the analysis and design phases of the project. It discusses the functional and nonfunctional requirements of the system. This chapter also offers a detailed account of the functionalities and processes offered by the project expressed as UML diagrams.

3.1 System Overview

3.1.1 System Architecture

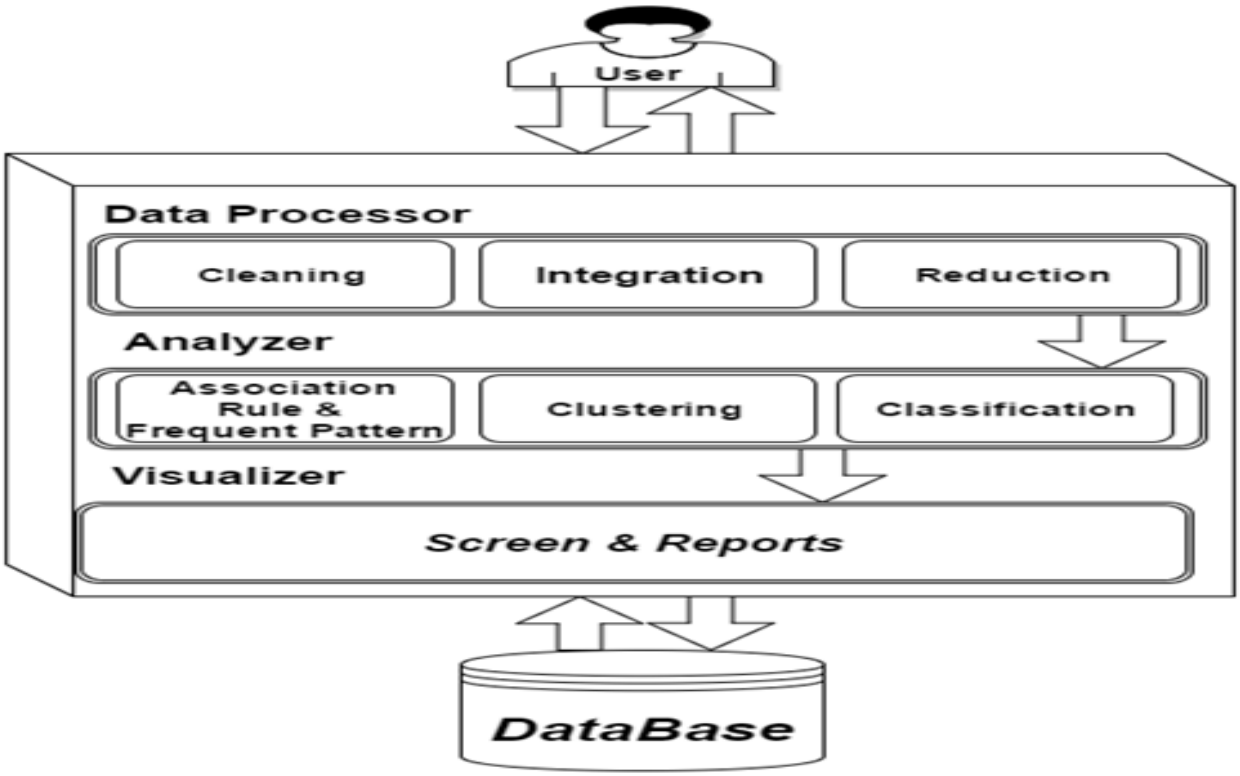


Fig. 3.1 System Architecture

1-Data Processor Module:-includes the preprocessing functions such as:-

a)Data Cleaning →Which is applied to fill the missing values in the data set.

b)Data Integration→Which is used to combine the data if there are multiple resources of it.

c)Data Reduction→Which is used to reduce the size of data when a inly sample of it is needed to be processed.

2-Analyzer Module:-Includes the algorithms for the data mining techniques such as:-

a)A priori →For Association Rules or Mining Frequent Patterns.
b)K-Means→For The Clustering Of Data.
c)Decision Tree(ID3)→For The Classification Of Data.

3-Visualizer Module:- Includes techniques applied and Screens for displaying / visualizing the results of the mined data.

Examples:-

a)Parallel Coordinates.
b)Scatter Plot Matrices.
c)Rule Graph.
d)General Visualization Techniques.

System Workflow:

→First , the user interacts with the interface and chooses one of two options:

- Data Mining and Visualization.
-General visualization.

→If General Visualization option is chosen the user then loads his data set and have only three general visualization techniques to deal with: (Histogram , Pie chart , Column chart).

→If Data Mining and Visualization option is chosen the user then loads his data set and then chooses from the data mining techniques:-

1-Preprocessing (Data Processor Block):- which includes functions such as
a-Data Cleaning.
b-Data Normalization.

2-Processing (Analyzer Block):-which deals with the data mining algorithms and techniques such as:-
a-Association Rules.
b-Clustering.
c-Classification.

→After applying the data mining algorithms the user can visualize the result using visualization techniques dependent on the data mining algorithms used (Visualizer Block), and also can manipulate the data on the graphs.

3.1.2Functional Requirements

1-The system should get input form database.
2-The system should apply pre-processing techniques on data :
-Data cleaning.
-Data integration.
-Data reduction.
3-The system should apply data mining techniques :
-Frequent patterns
-Classification
-Clustering
4-The system should visualize the output of data mining processes.
(Histogram , Column chart , pie chart , Scatter plots , parallel coordinates , rule graph).
5-The system should save the final results and outputs.
6-The system should give the ability to manipulate the output.

3.1.3 Nonfunctional Requirements

→Usability requirements: The system is very easily operatedand does not requires special skills from its users.

→ Performance requirements: Since we are developing a Desktop application, the speed of operation is an important factor.

→Efficiency requirements: Since the visualization of the data does not take so much time to change in real time or to draw the views upon the user's need.

3.1.4 System Users

A. Intended Users:

-The System is built for two types of users:-

1) Data Mining Specialist→ he/she can use the system to apply data mining techniques and algorithms dependent on the problem/situation and then can visualize the result using visualization techniques dependent on the used data mining algorithm .

2) Naïve User→he/she can use the system to only apply general visualization techniques on his/her data.

B. User Characteristics

1) Knowledge of computers

2)For Data Mining Specialist :- should be experienced in the field of the data mining ,aware of the data mining techniques , when to use each technique and the algorithms in each technique , and also should be aware of visualization techniques that can be used to visualize mined results.

3) For Naïve User :- he/ she will use general visualization techniques(Histogram , Pie Char , Column Chart), he/she should be aware of the differences between them and when/why to use each one of them.

3.2 System Analysis& Design

3.2.1 Use Case Diagram

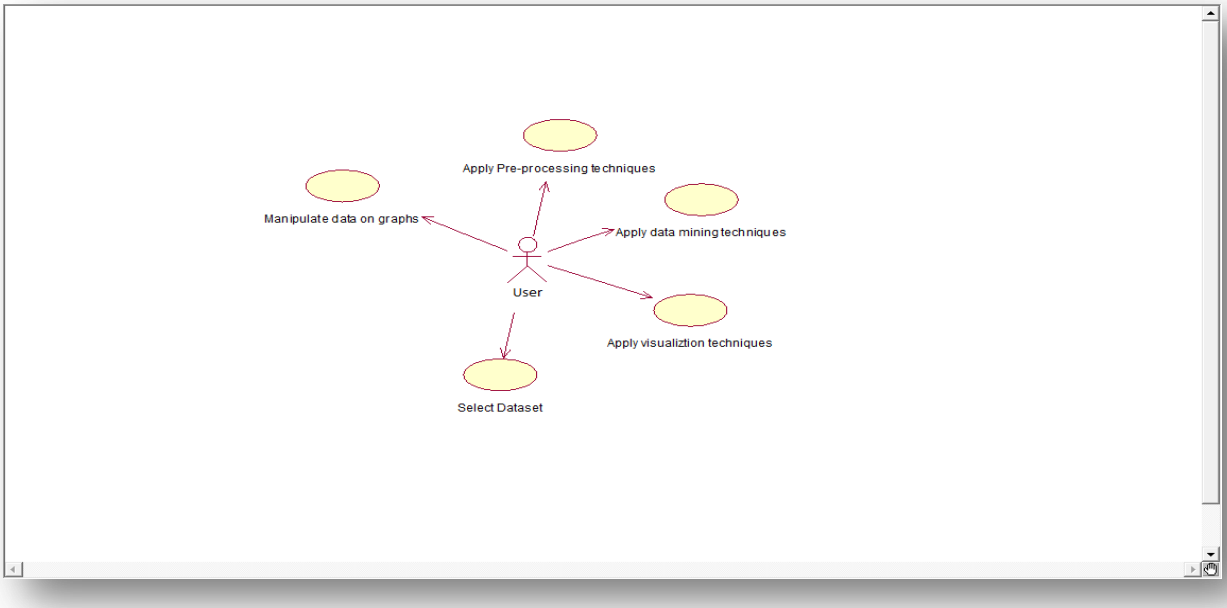


Fig. 3.2 Use Case Diagram

Use Case	Select Dataset
Brief Description	The user should load the dataset to work on it.
Post Condition	Load desired file
Flow Of Event	**Primary flow:-**the file is loaded successfully. **Alternate flows:-**file format is not correct**.** **Error flow:-**Light cut off

Use Case	Apply preprocessing techniques
Brief Description	User chooses the process to apply
Precondition	Loaded the desired file
Post Condition	The result is displayed in a data grid view.
Flow Of Event	**Primary flow:-**the expected result is displayed in a data grid view. **Alternate flow:-**file is not loaded **Error flows:-** -Electricity down.

	-Light cut off.

Use Case	Apply data mining techniques
Brief Description	User chooses the technique to apply dependent on the problem.
Precondition	In some techniques preprocessing should be applied first.
Post Condition	The result of the technique is displayed
Flow Of Event	**Primary flow:**-The output is displayed as expected. **Alternative flows:-** -Preprocessing is not applied. -Loading the file that is not suitable for the technique**.** **Error flows:-** -Wrong file format. -Light cut off.

Use Case	Apply visualization techniques
Brief Description	User chooses the visualization technique dependent on the data mining algorithm used.
Precondition	Data mining algorithms should be applied.
Post Condition	The view/graph is displayed.
Flow Of Event	**Primary flow:-**the desired graph is displayed and the user can understand it. **Alternative flow:-**incorrect visualization technique. -misunderstanding of the result. **Error flows:-** **-**Wrong file format. -Light cut off.

Use Case	Manipulate data on graphs.
Brief Description	User can zoom data on one graph or more dependent on the techniques used.
Precondition	Visualization should be applied.
Post Condition	Manipulation is done.
Flow Of Event	**Primary flow:-**the output is displayed

	as expected**.** **Alternative flow:-**incorrect visualization. **Error flows:-** -Wrong file format. -Light cut off.

3.2.2 Class Diagram

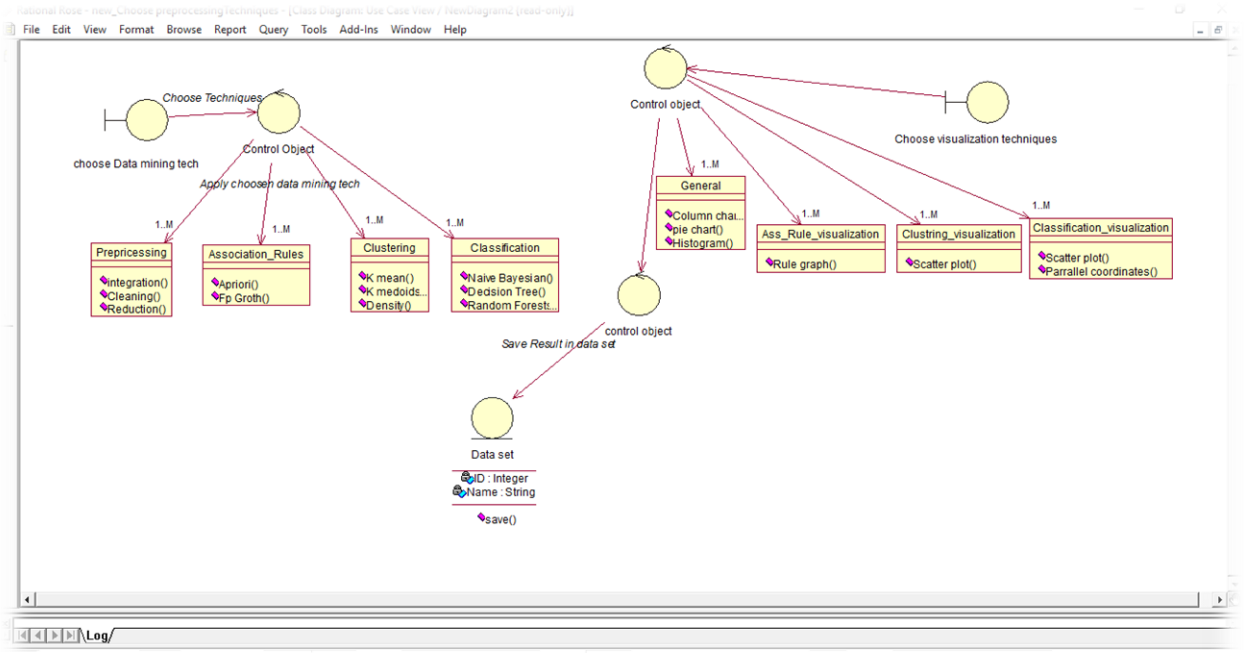


Fig. 3.3 Class Diagram

3.2.3Sequence Diagrams

a)Sequence diagram for pre-processing

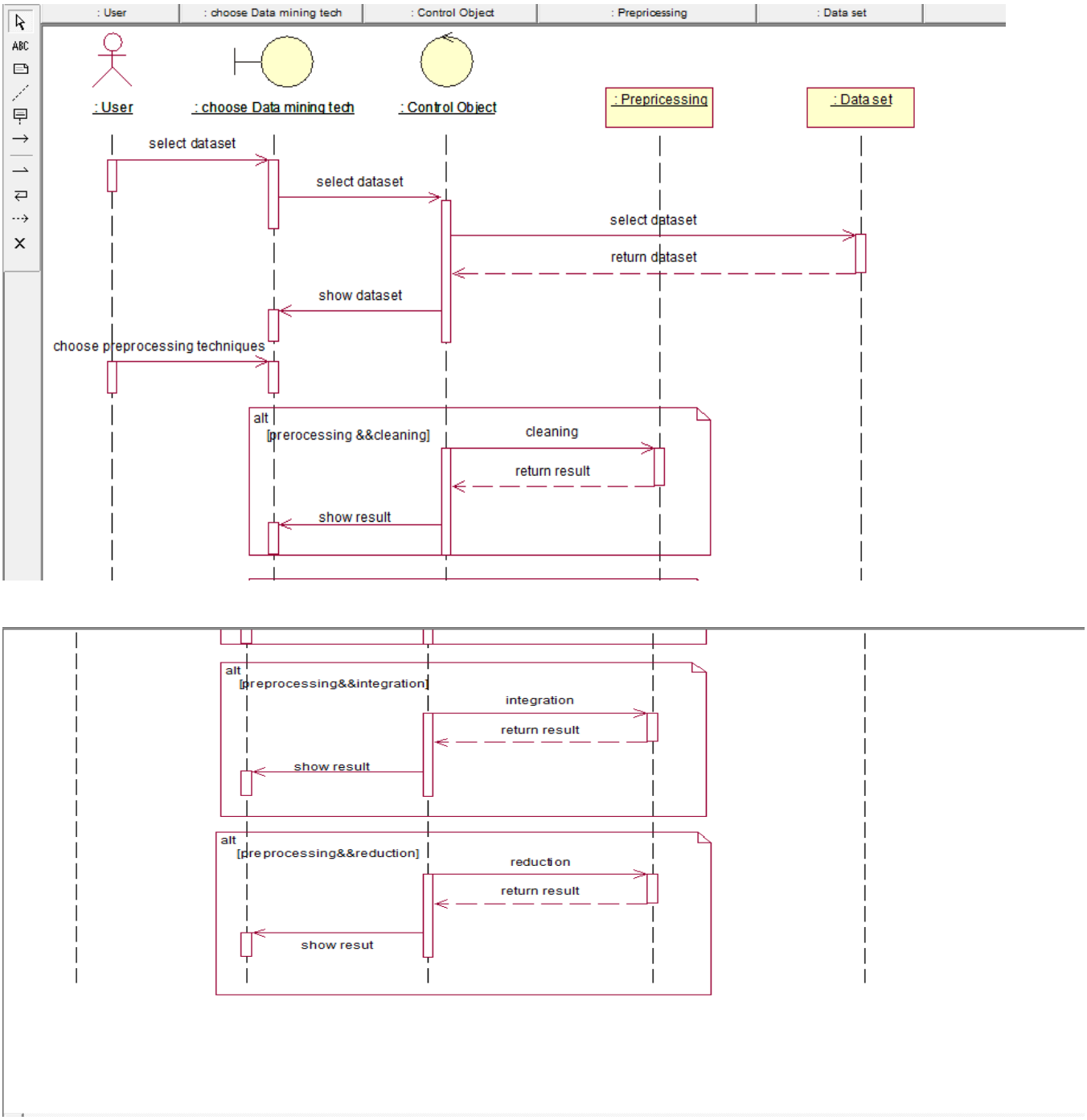
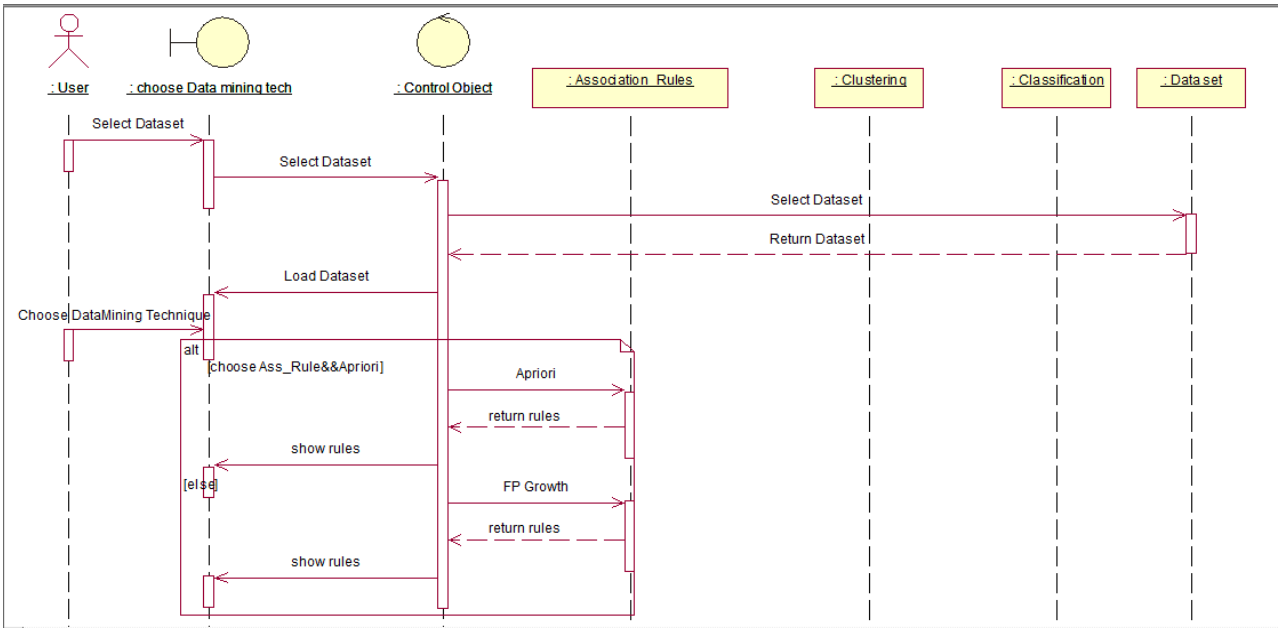
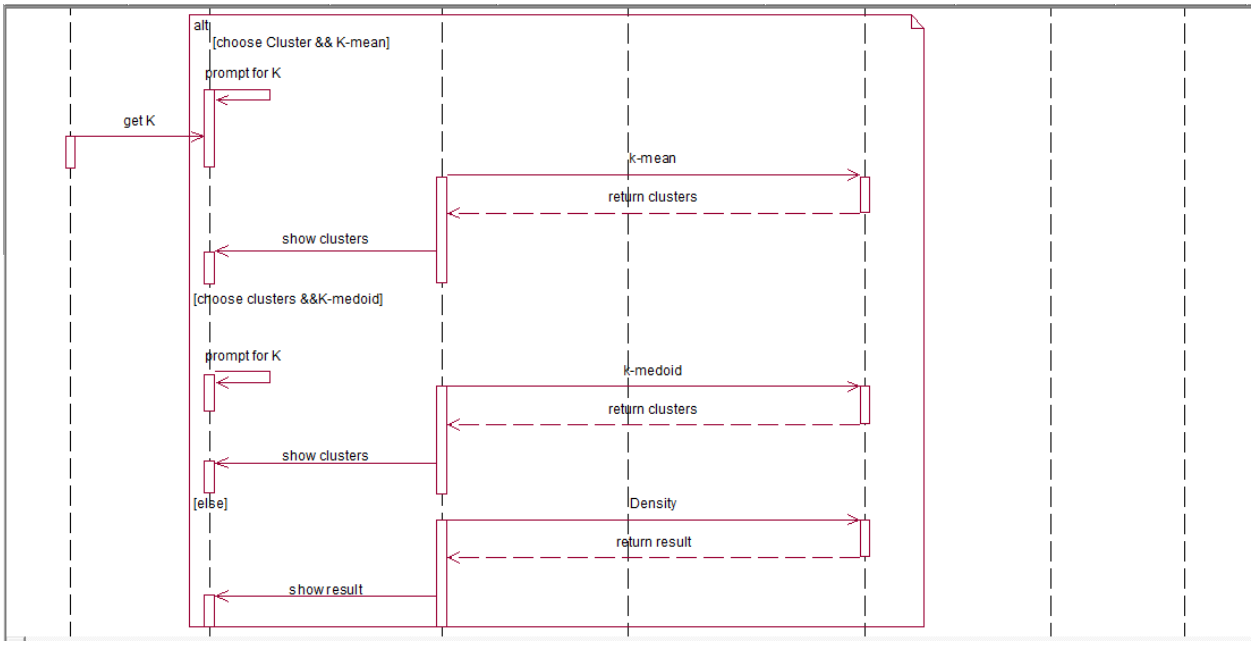


Fig. 3.4 Sequence diagram(pre-processing)

b)Sequence diagram for choosing data mining technique.





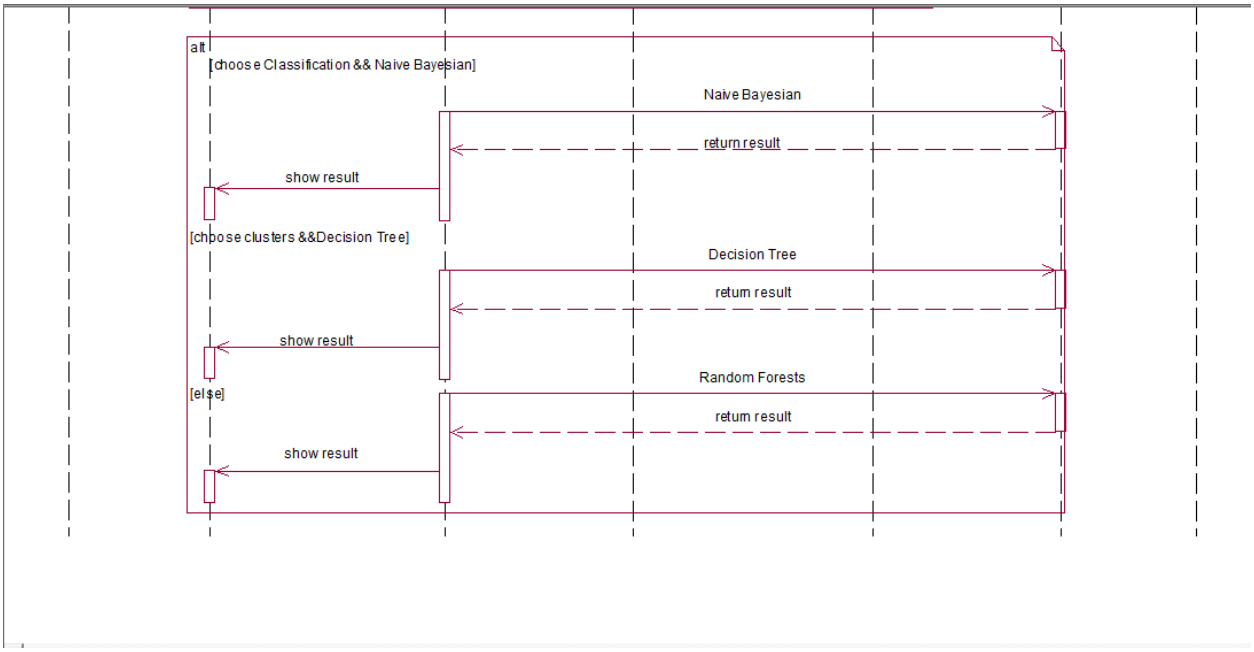
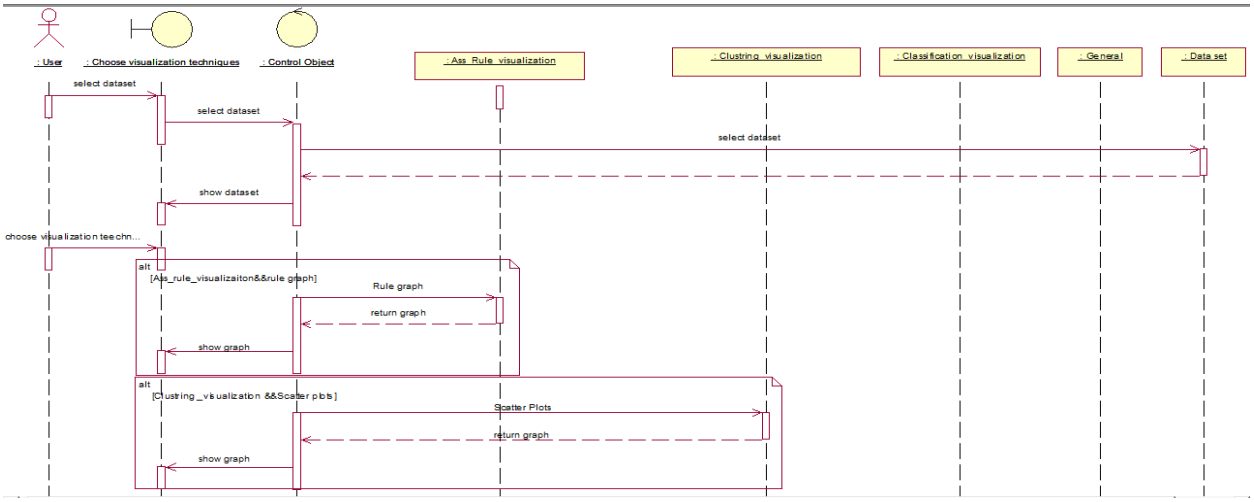


Fig. 3.5 Sequence diagram(choosing data mining technique)

c)Sequence diagram for choosing visualization technique.



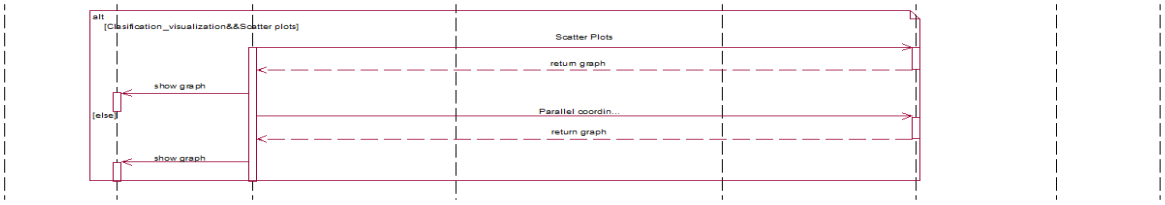


Fig. 3.6 Sequence diagram(choosing Visualization technique)

Chapter 4
Implementation

This chapter explains the algorithms and techniques used in the implementation of the system and the role of each of them in the project.

-First we load the required data set to work on it.

4.1- Pre-Processing

4.1.1 Data Cleaning (Remove Noise).

4.1.1.1-Steps:

- Move onto Pre-processing tap to choose the method from combo box
- If the data cleaning is chosen , by clicking on apply method button the whole data set will be looped to replace missing values

→String values replaced with NULL & Numeric values replaced with 0s.

4.1.2 – Data Normalization:

4.1.2.1 – Steps:

- If data normalization is chosen ,the user should choose the method.
- If Min-Max Normalization is chosen the user should enter minimum and the maximum values for data normalization and by clicking the apply algorithm button the equation of the normalization will be applied on the whole numeric values within the data set to be within the range specified by the user.

4.2-Association Rules

4.2.1 – A priori:

4.2.1.1- Pseudo Code:

```
procedureApriori (T, minSupport){
        //T is the database and minSupport is the minimum support
        L1= {frequent items};
            for (k= 2; Lk-1 !=∅; k++) {
            Ck= candidates generated from Lk-1
            //that is cartesian product Lk-1 x Lk-1 and eliminating any k-1 size
            itemset that is not frequent
for each transaction t in database do{
        #increment the count of all candidates in Ck that are contained in t
            Lk = candidates in Ck with minSupport
              }//end for each
        }//end for
        returnU_k L_k;
```

}

4.2.1.2-Steps:

- Move onto Association Rules tap,choose the algorithm and specify minimum support and minimum confidence and then click on find frequent button.
- By clicking on the button the FindFrequent method will be called which in its turn calls the ExtractSupported method to find the items that satisfy minimum support .
- Then by clicking CheckConfidence button the GetConfidence method will be fired to calculate the confidence of the item and diplay only those that satisfy the mininmum confidence.
- By clicking on visualize button the graph will be display on the visualize tap.

4.2.2 - Association Rule Visualization(Rule Graph)

Visualize association rule using vertices and edges where vertices typically represent items or item sets and edges indicate relationships and rules, where the size of vertices is support, and the color of vertices is the value of lift.

4.2.2.1 - Steps:-

- rule graph are u
- sing to visualize frequent association rule, containing a member functions:-
  1) Get unique lift values from List 3
  2) Get unique support values from List 4
  3) Generate randomly sizes base on support values
  4) generate colors which generate random colors for each uniquely lift value
  5) Generate vertices and edges for each rule in Frequent Item Set (List 4) based on its assigning color and it’s assigning size

4.3- Clustering

4.3.1 - K-Means:

4.3.1.1- Pseudo Code:

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

- (1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
- (2) repeat
- (3) (re)assign each object to the cluster to which the object is the most similar,
- based on the mean value of the objects in the cluster;
- (4) update the cluster means, that is, calculate the mean value of the objects for
- each cluster;
- (5) until no change;

4.3.1.2 - Steps :

- Move onto clustering tap , enter the number of clusters .
- By clicking apply algorithm button an event to check if number of clusters is not null will be fired.
- If number of clusters equal null a message will be fired to warn the user and ask him to enter the number of clusters.
- If number of clusters not equal null the showData function will be called which in its turn it calls the Cluster function which randomly choose the clusters.
- Then initclustering to assign the rest of data objects/items to the chosen clusters , and also Allocate function is called to hold the positions of the clusters in a matrix.
- Then updateMeans function will be called to recalculate the new centroids and the distance between the objects and centroids using functions Distance and MinIndex to assign objects to the closest centroid.
- If there is a change in the centroids the UpdateClustering function will be called to update clusters with new centroids.

4.3.2 - Clustering Visualization(Scatter plot matrix)

Scatter plot matrix are a great way to roughly determine if you have a linear correlation between multiple variables.

4.3.2.1 - Steps :

- Scatter plot matrix are using to visualize data mining clustering result .
- Class scatter plot matrix is used to generate scatter plot matrix containing an **member variables** :

  1. data table.
  2. 2D array of charts controls.
  3. array of integers to represent cluster number for each data table row.

  and**member functions :**

  1. constructor to declare data table and declare 2D array of charts.
  2. Generate colors which generate random colors for each cluster.
  3. Generate matrix which generate number of chart and assign each one of them to corresponding element in 2D array and add points to each chart according to its position on matrix .
  4. Chart mouse move which is an event for each chart to detect if cursor position is on point or not if change point color and corresponding points of other charts.

4.4- Classification

4.4.1 - Decision Tree

4.4.1.1 Pseudo Code

- (1) create a node *N*;
- (2) **if** tuples in *D* are all of the same class, *C*, **then**
- (3) return *N* as a leaf node labeled with the class *C*;
- (4) **if** *attribute list* is empty **then**
- (5) return *N* as a leaf node labeled with the majority class in *D*; // majority voting
- (6) apply **Attribute selection method**(*D*, *attribute list*) to **find** the "best" *splitting criterion*;
- (7) label node *N* with *splitting criterion*;

- (8) **if** *splitting attribute* is discrete-valued **and** multiway splits allowed **then**
- (9) *attribute list attribute list splitting attribute*; // remove *splitting attribute*
- (10) **for each** outcome *j* of *splitting criterion*
- (11) let *Dj*be the set of data tuples in *D* satisfying outcome *j*; // a partition
- (12) **if** *Dj*is empty **then**
- (13) attach a leaf labeled with the majority class in *D* to node *N*;
- (14) **else** attach the node returned by **Generate decision tree**(*Dj*, *attribute list*) to node *N*;

endfor

- (15) return *N*;

4.4.1.2 - Steps

- First we will get copy of the entered data
- By using convert to ARFF we will take the entered data and convert to the ARFF components like the attributes and there unique values and the target of classification "the last attribute in data"
- Then we will built the classifier "Decision Tree" by initialize new instance of class Classifier
- The class Classifier take the ARFF component and build the tree by DecisionBuilder class
- The tree components leaf and nodes it's built by know the importance of attribute by calculate the Entropy
- The we draw it and save the result for make visualization on the classifier

4.4.2 - Parallel coordinate

4.4.2.1 - Steps

- First we will get copy of the data
- Then make series of the graph equal the number of rows of the data

- Then normalize and convert the nominal data to numeric data to make easy to draw
- For every row in the data we get the value of attributes and draw it with random color

4.5 - General Visualization

4.5.1- Steps

Load Data :

- First we have to load dataset.
- Open file dialog will help us to get database path.
- Stream reader help us to read database file and fill data table.

Choose Techniques :

- After loading database we have to choose one technique of general techniques :-

1. **Histogram**

   A **histogram** is a graphical representation of the distribution of numerical data.

   i) Bin the range of values that is, divide the entire range of values into a series of intervals.

   ii) Then count how many values fall into each interval .

2. **Pie Chart.**

   A **pie chart** is a circular statistical graphic which is divided into slices to illustrate numerical proportion.

3. **Column Chart.**

   Column chart are used to compare values across categories by using vertical bars.

Visualize Data :

- Select column name you want to visualize .
- If selected technique is **Histogram** :

I. By using **generate histogram** function histogram chart are generated and positioned on panel control .

II. Declare an object from **class Histogram** which contain **member variable** of data type **data table** which declared on class **constructor**.

III. Then by using class **member function get bins** which takes column name as parameter and return an array of objects of

Bin data class every element have **three member variable** that define each interval (**minimum**–**maximum**–**frequency**),first we find minimum and maximum value of this column and then calculate width of each bin(interval) by :

1. Count number of rows in this column then take an square root .
2. Width equals maximum- minimum divide to square root.

then define each interval by it's minimum and maximum value , finally looped on column values to define frequency for each bin(interval).

IV. Generate random colors by using **generate random colors function** which return an array of unique colors assigned for each bin(interval).

- Else if selected technique is **pie chart** :
  - I. By using **generate Piechart** function Pie chart are generated and positioned on panel control .
  - II. Declare an object from **class PieChart**which contain **member variable** of data type **data table** which declared on class **constructor**.
  - III. Then by using class **member function count data** which takes column name as parameter and return an array of objects of **Pie Chart data** class every element have **two member variable** (bin name-frequency), which contain unique values of this column and frequency for each unique value .

Chapter 5
User Manual

5.1.Overview:

The product is a Multi-View Data Mining Visualization .The User can apply some pre-process (Remove Noisy - Normalization) on data before apply Data Mining technique

(Association Rule – Clustering – Classification) on data before visualize it Or Visualize the data directly not need to apply the technique of data mining .

5.2.Operating the System:

5.2.1.Selectingdata mining and visualization or general visualization

The user may select data mining and visualization to apply data mining technique before visualization or select general visualization to visualize the data directly .

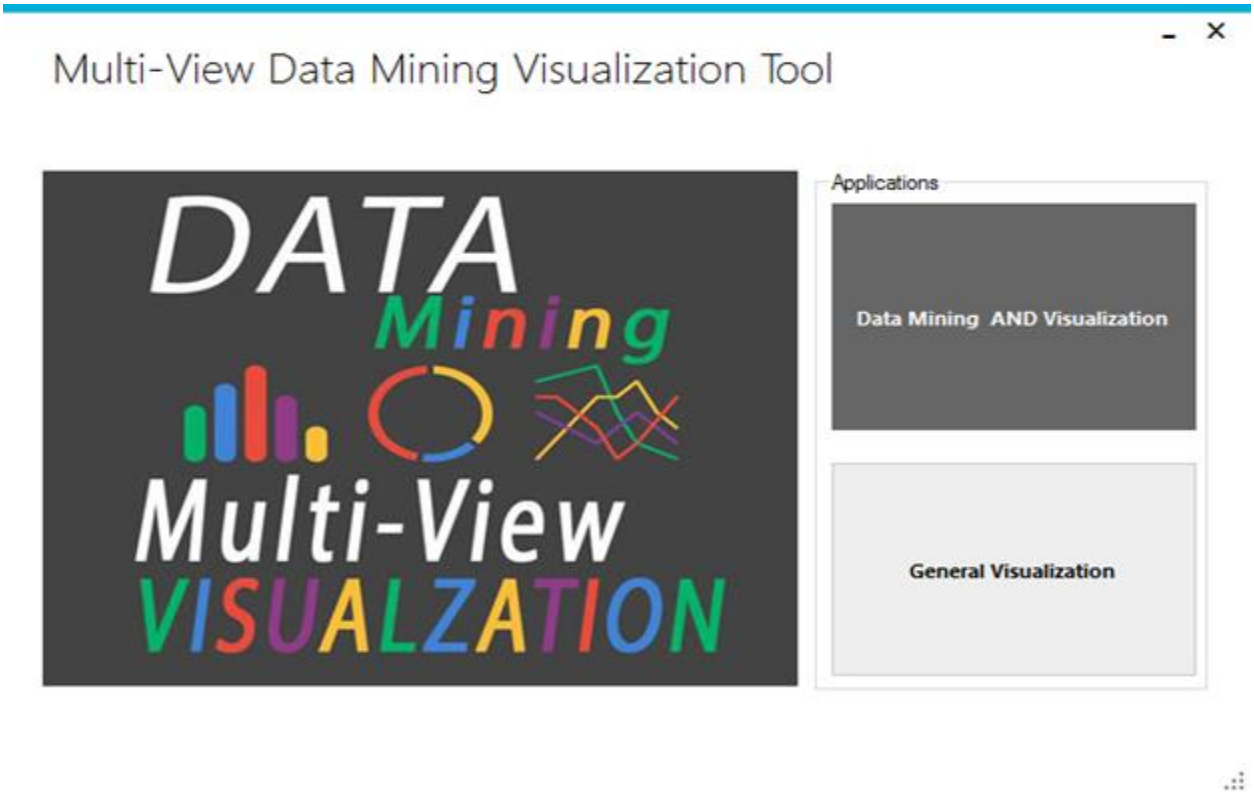


Fig. 5.1 Data Mining and Visualization / General visualization

5.2.2.SelectingData Mining and Visualization

The user click on Data Mining And Visualization
to open the form of Data Mining And Visualization

Fig. 5.2 Choose Data Mining and Visualization

5.2.2.1 Load Dataset

The user load a dataset to from open in file menu

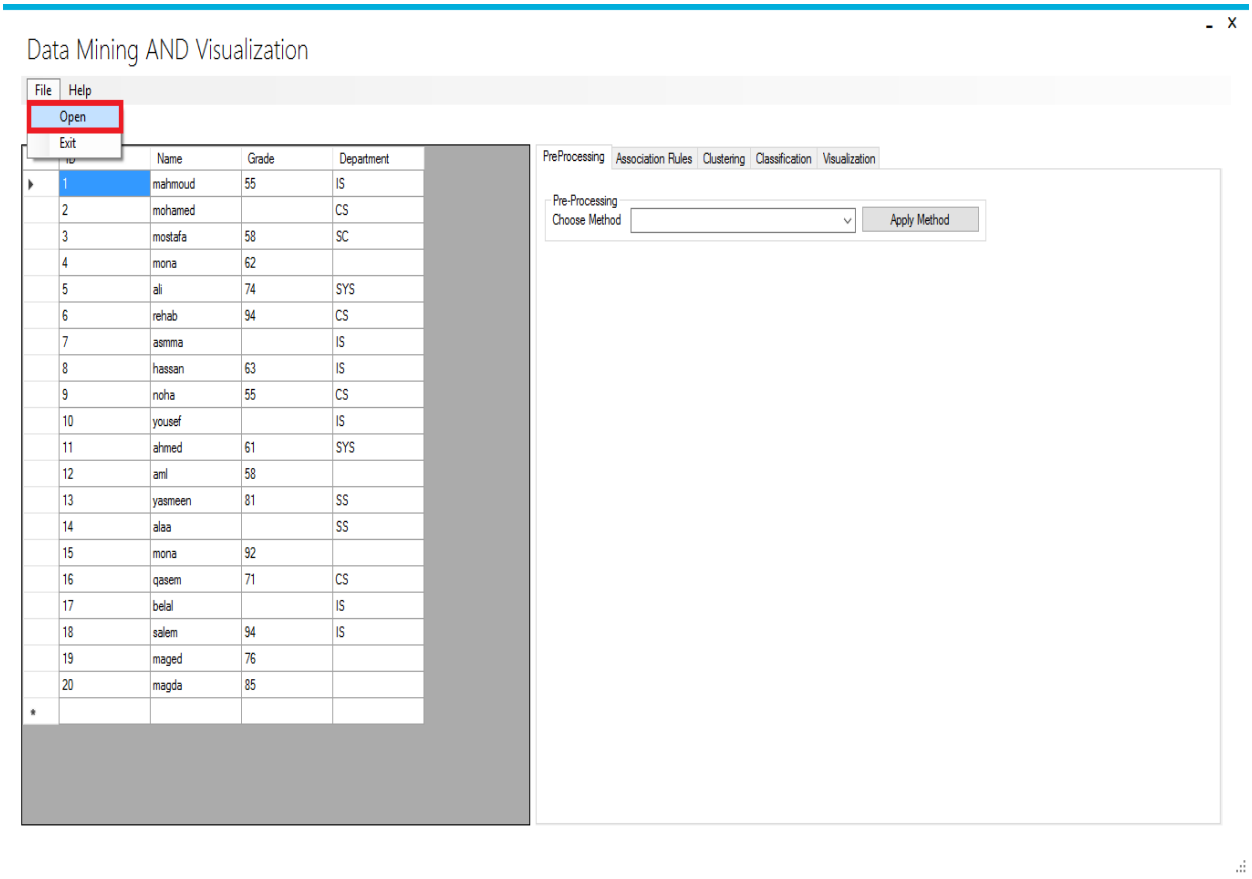


Fig. 5.3 Load Dataset in Data Mining And Visualization

5.2.2.2 Apply Pre-Processing (Data Cleaning(Remove Noise))

The user choose pre-processing -> Data Cleaning(Remove Noise) to remove noise in dataset

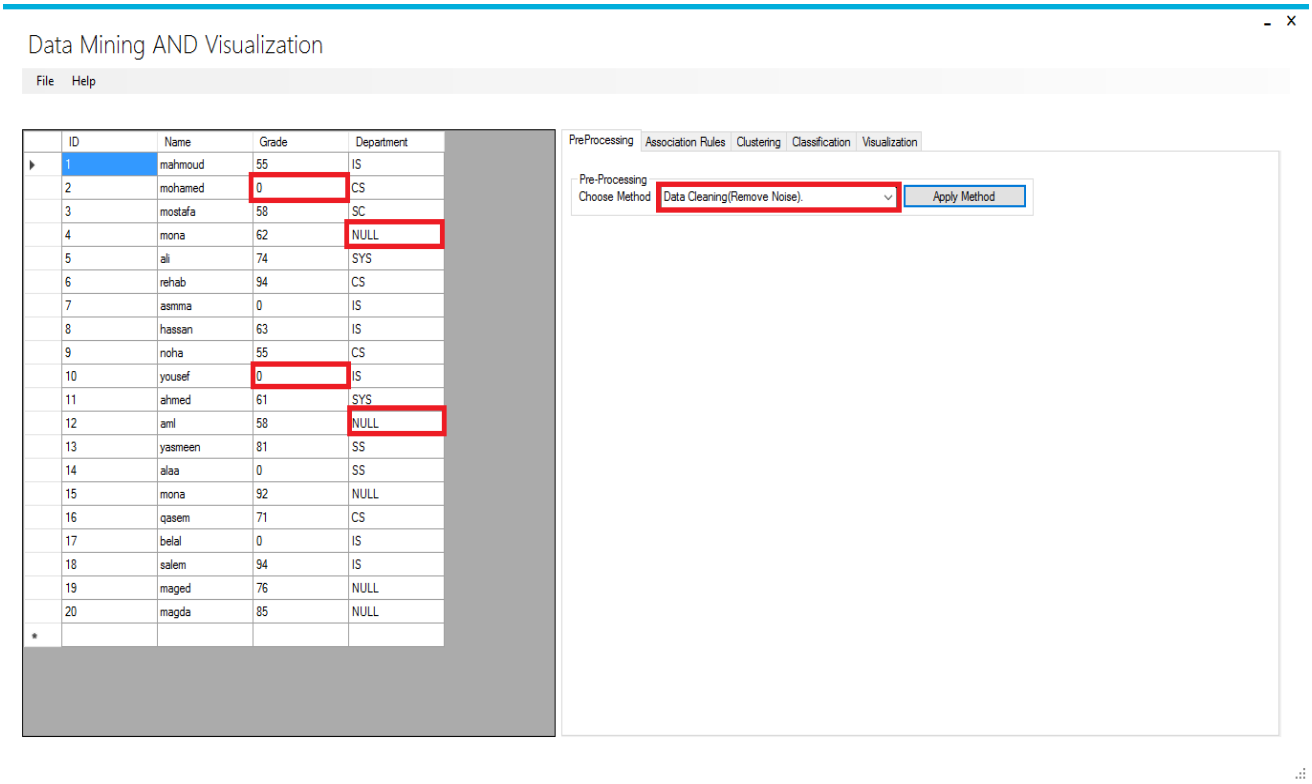


Fig. 5.4 Apply Data Cleaning(Remove Noise)

5.2.2.3 Apply Pre-Processing (Data Normalization)

The user choose pre-processing -> Data Normalization
then enter Minimum and Maximum value in two text box to apply normalization in data set

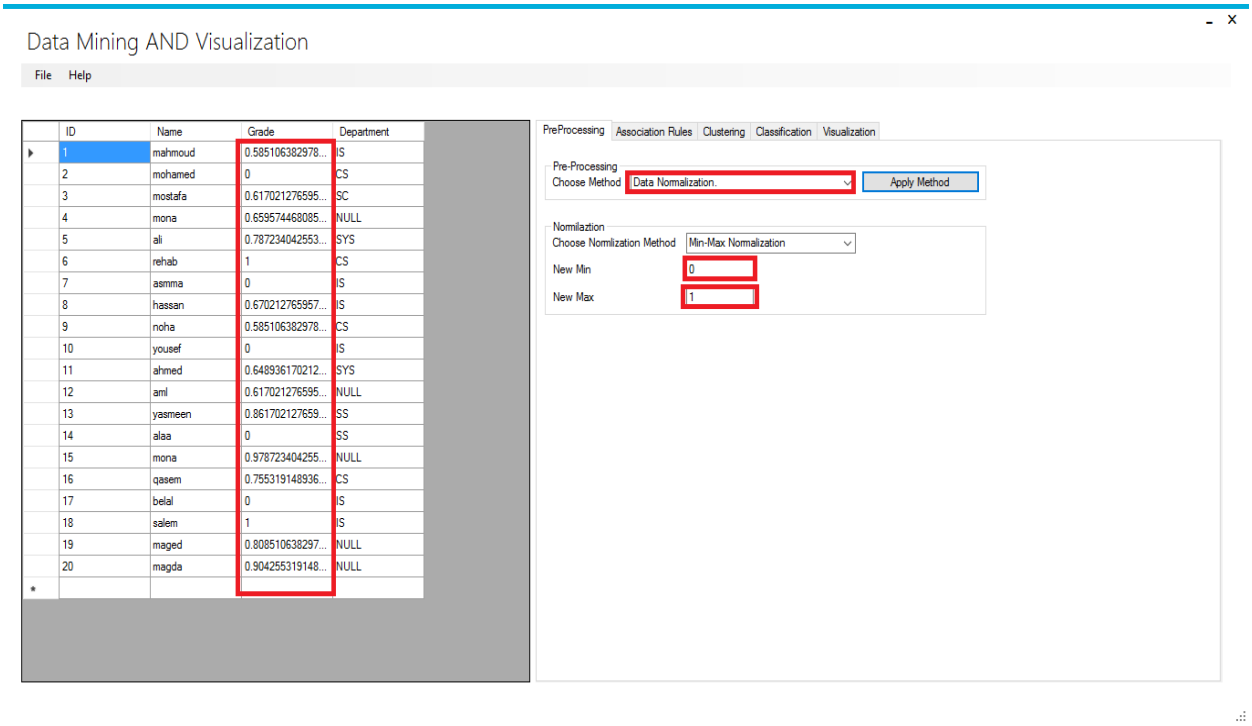


Fig. 5.5 Apply (Data Normalization)

5.2.2.4 Apply Data Mining Techniques (Association Rule (Apriori))

The user choose Association Rule ->Apriori then enter Minimum support and Minimum confidence then click on find frequent to find frequent item and click on check confidence to check confidence between the frequent item . After apply algorithm click on Visualize to visualize the result using (Rule Graph) .

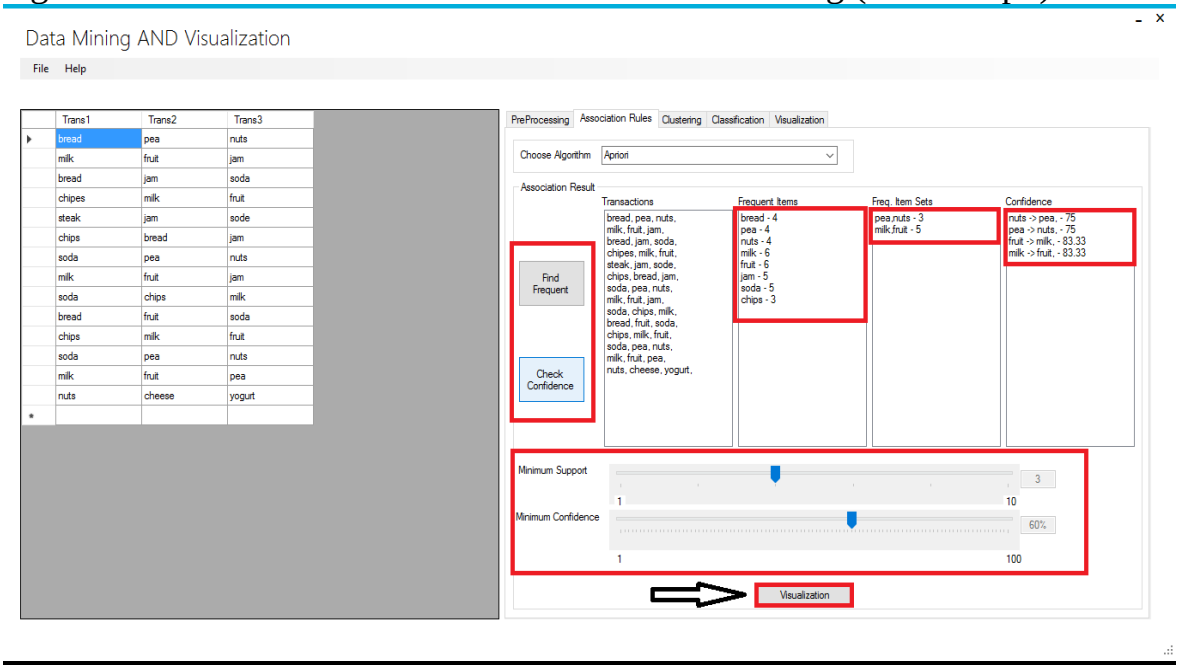


Fig. 5.6 Apply Apriori

Visualize the result using (Rule Graph)

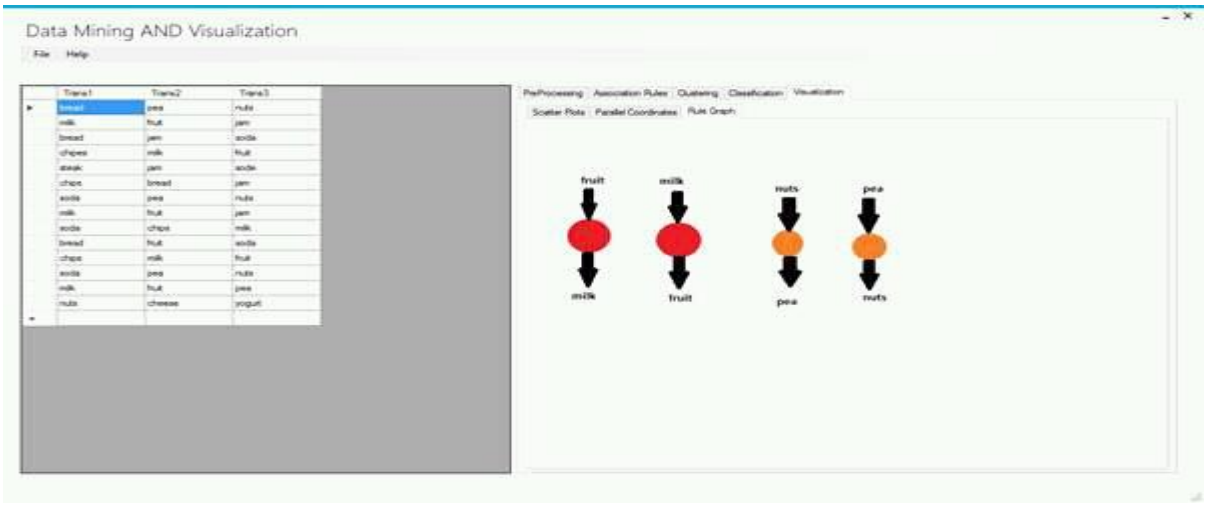


Fig. 5.7 Visualize Apriori (Rule Graph)

5.2.2.5 Apply Data Mining Techniques (Clustering (K-Means))

The user choose Clustering -> K-Means then enter No. of Clusters to define the number of clusters then click on Apply Algorithm .After apply algorithm click on Visualize to visualize the result using (Scatter Plot)

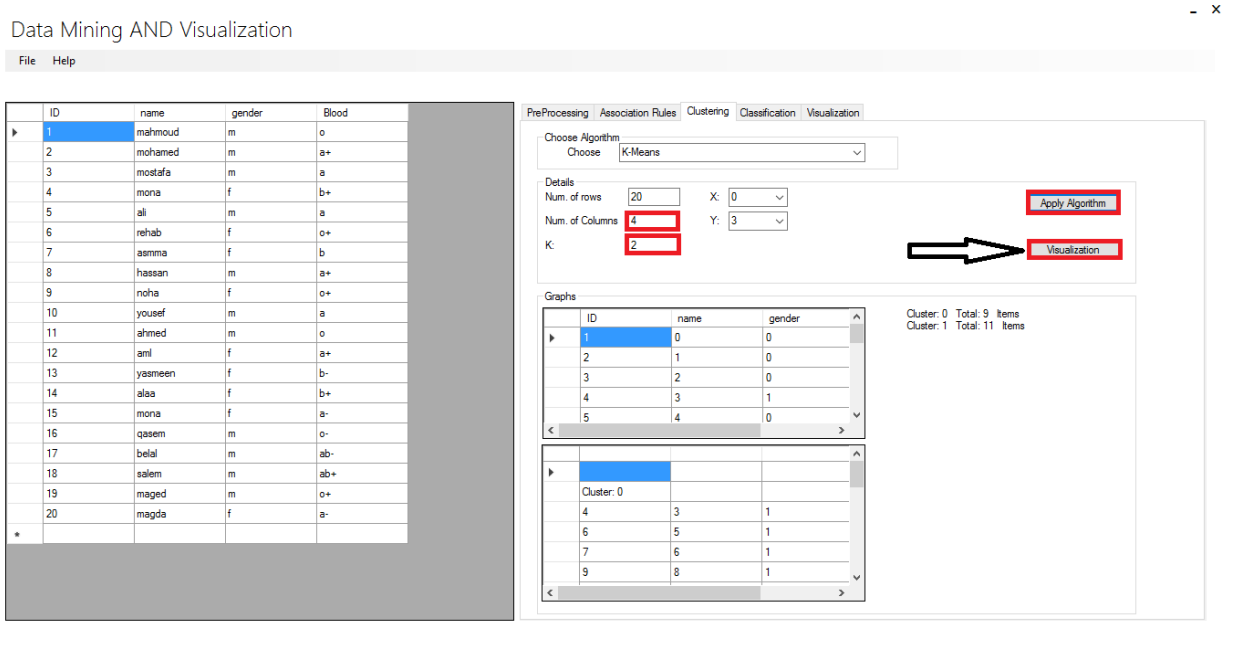


Fig. 5.8 Apply K-Means

Visualize the result using (Scatter Plot)

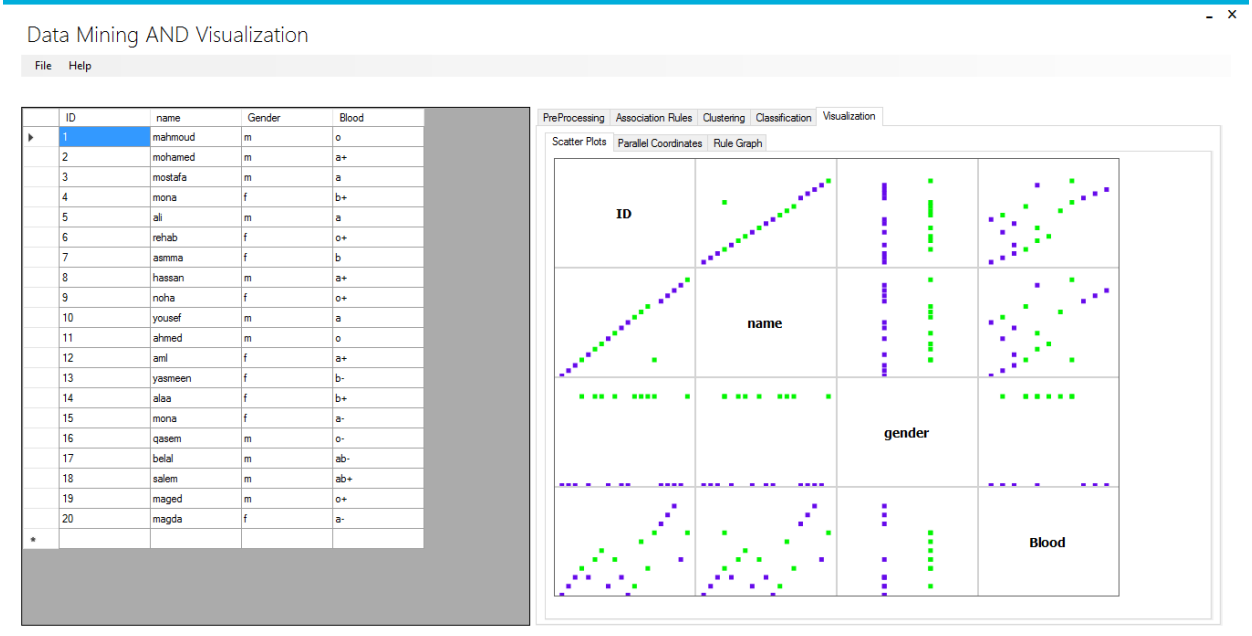


Fig. 5.9 Visualize K-Means (Scatter Plot)

5.2.2.6 Apply Data Mining Techniques (Classification (ID3))

The user choose Classification -> ID3 then click on Apply Algorithm.
After apply algorithm click on Visualize the result using (Parallel Coordinate)

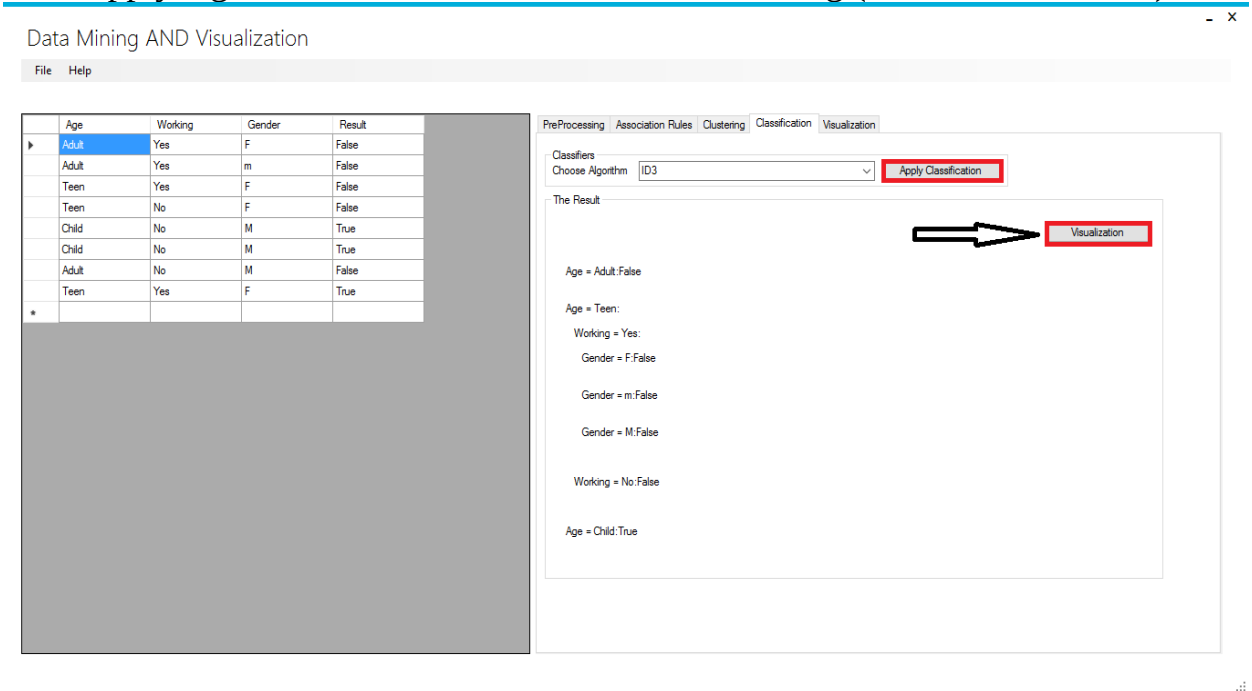


Fig. 5.10 Apply ID3

Visualize the result using (Parallel Coordinate)

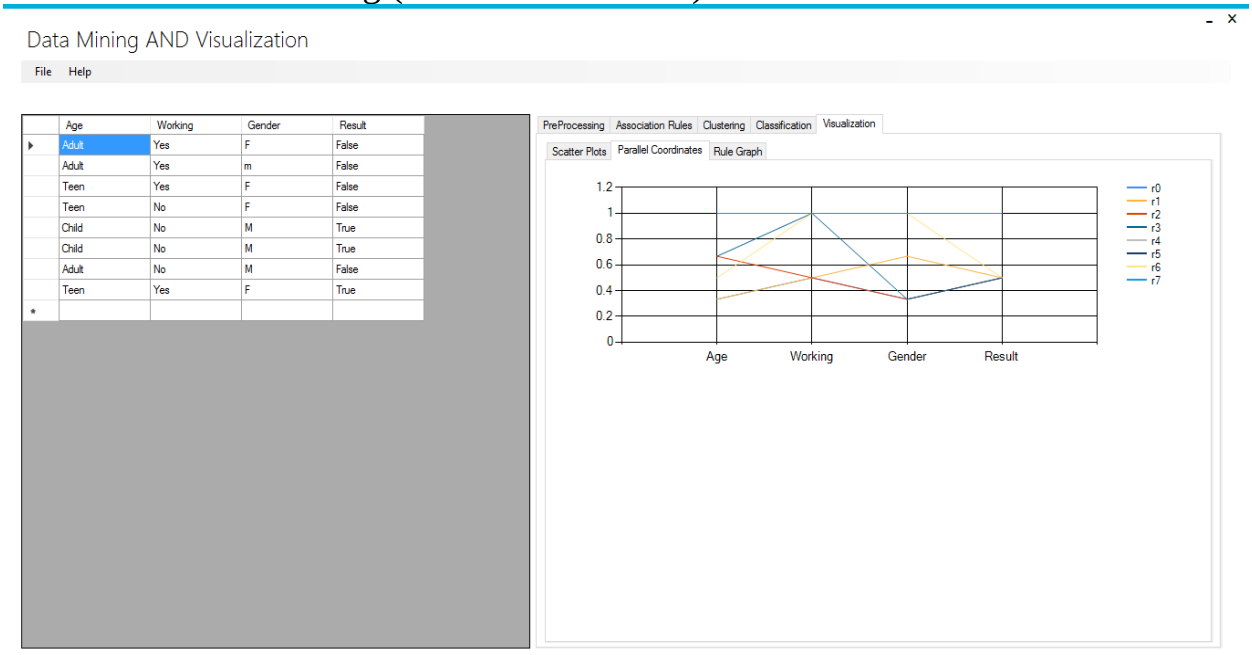


Fig. 5.11 Visualize ID3 (Parallel Coordinate)

5.2.2.7 Help

The user choose Help -> How To Use to open form Help

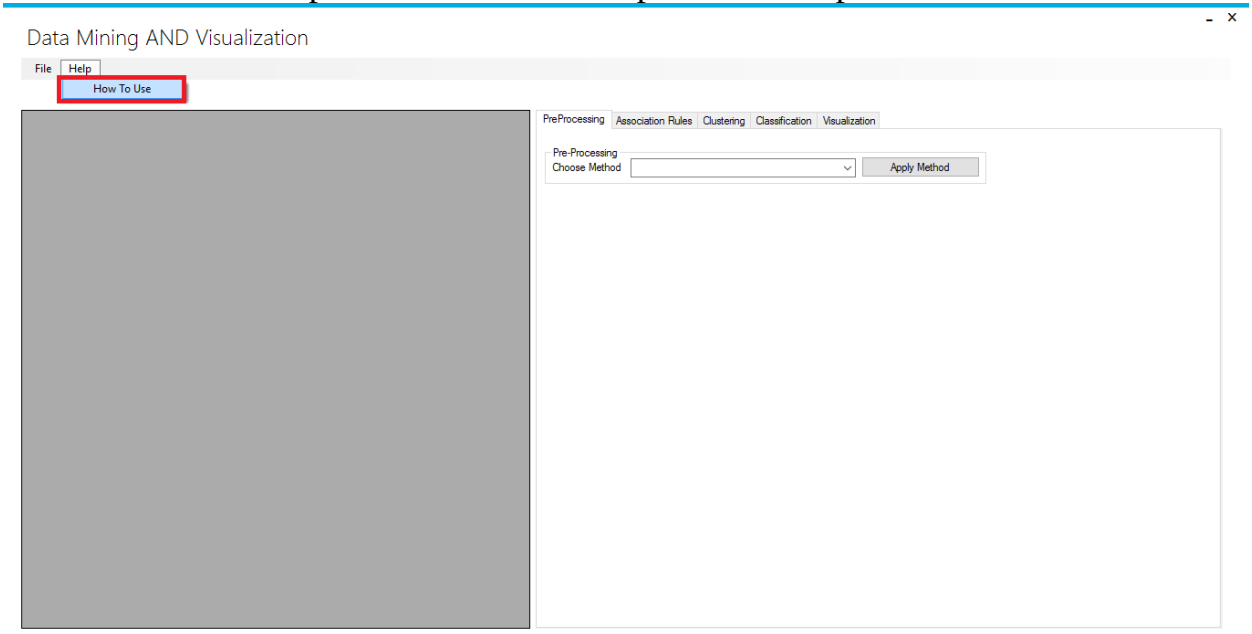


Fig. 5.12 How To Use in Data Mining And Visualization

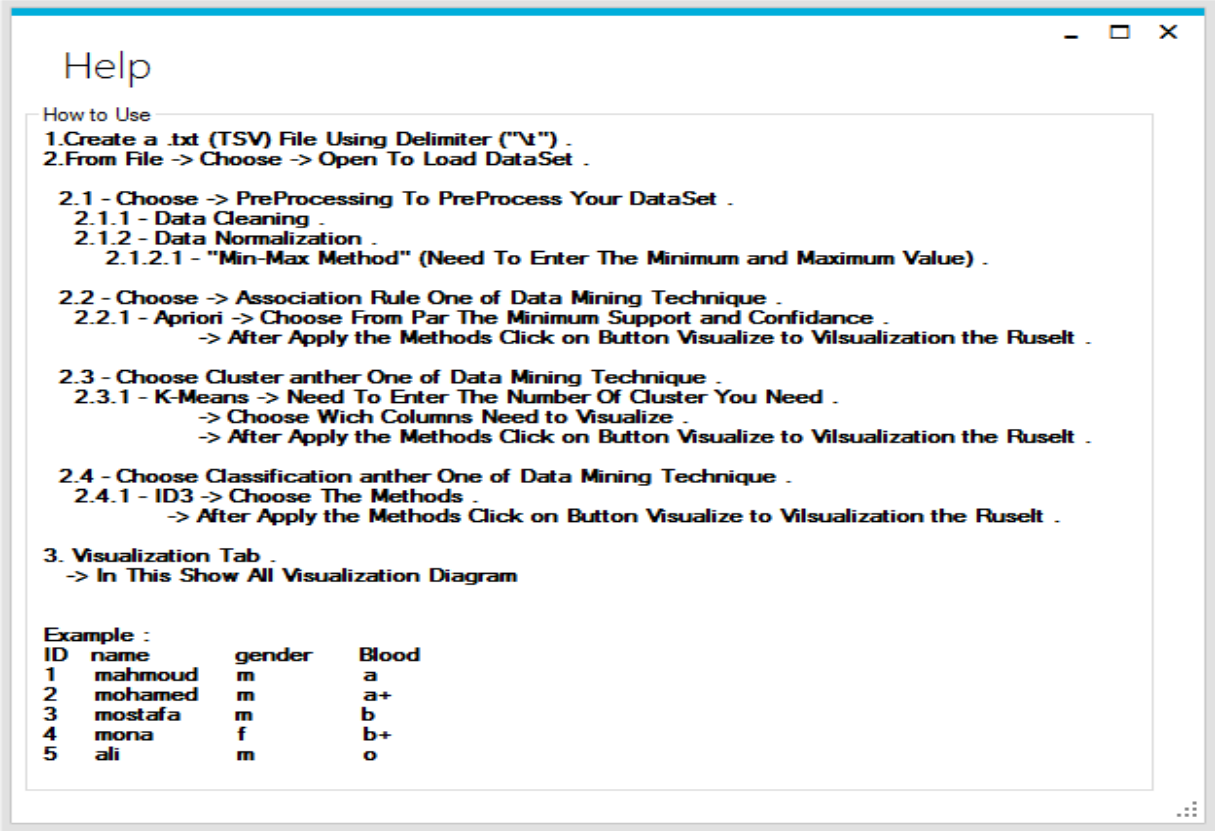


Fig. 5.13Help in Data Mining And Visualization

5.2.3.SelectingGeneral Visualization

The user click on General Visualization to open the form of General Visualization

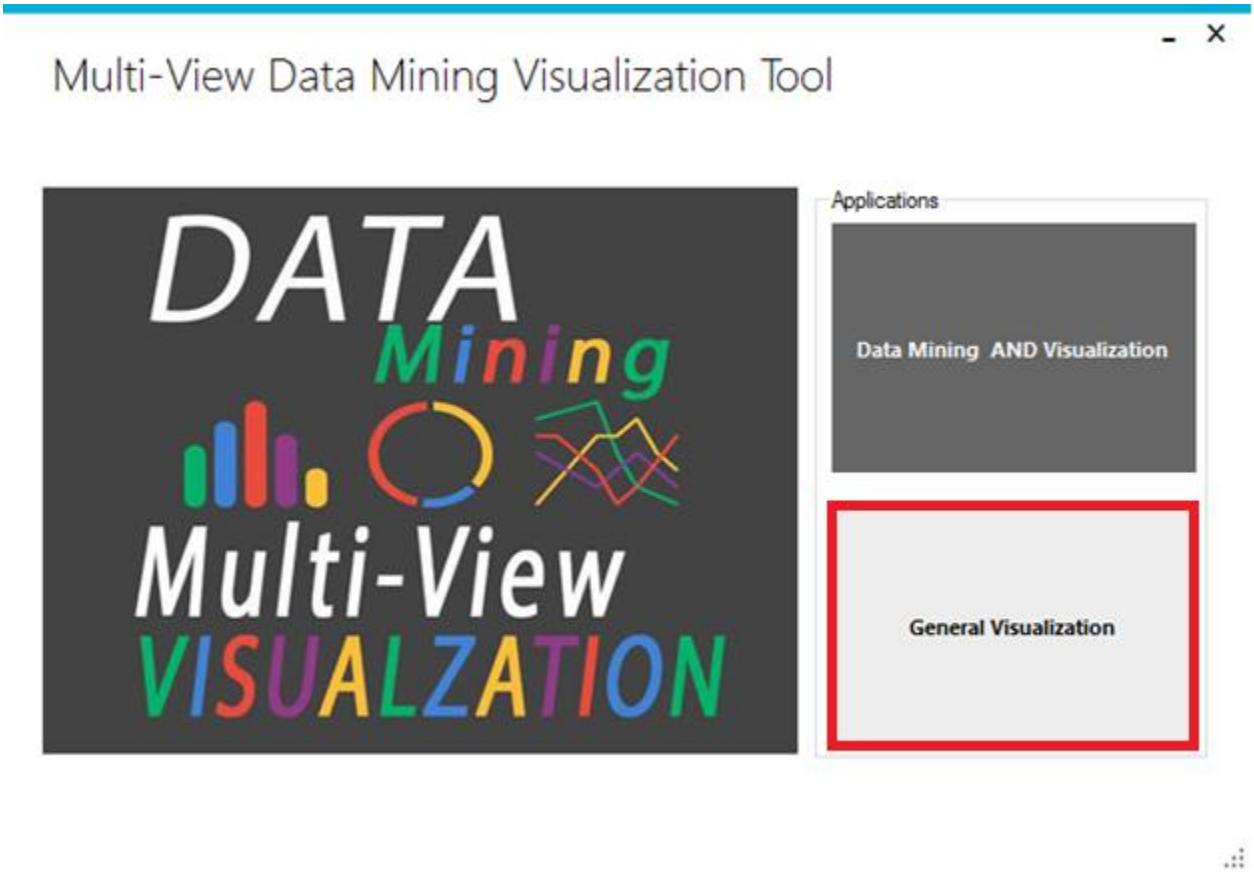


Fig. 5.14 Choose General Visualization

5.2.3.1 Load Dataset

The user load a dataset to from open in file menu

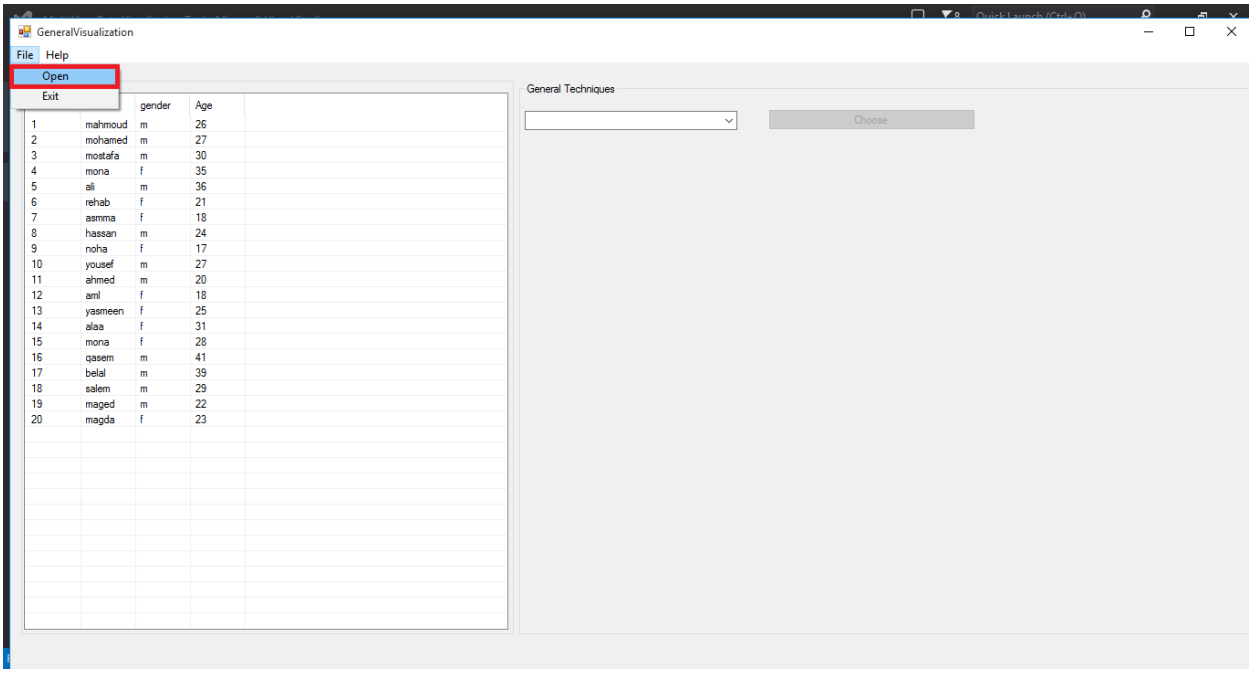


Fig. 5.15 Load Dataset in General Visualization

5.2.3.2.Histogram

The user choose Technique -> Histogram then choose which column need to visualize then Click on visualize

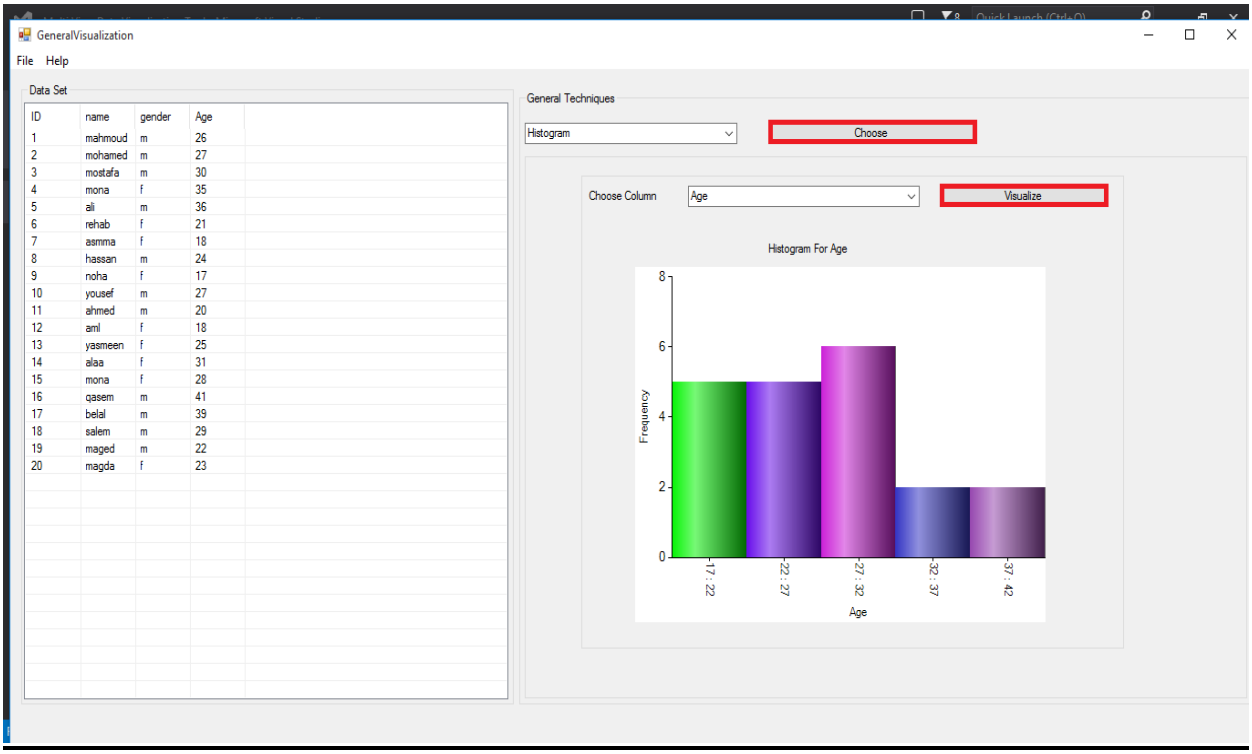


Fig. 5.16 Visualize Histogram

5.2.3.3.Pie Chart

The user choose Technique -> Pie Chart then choose which column need to visualize then Click on visualize

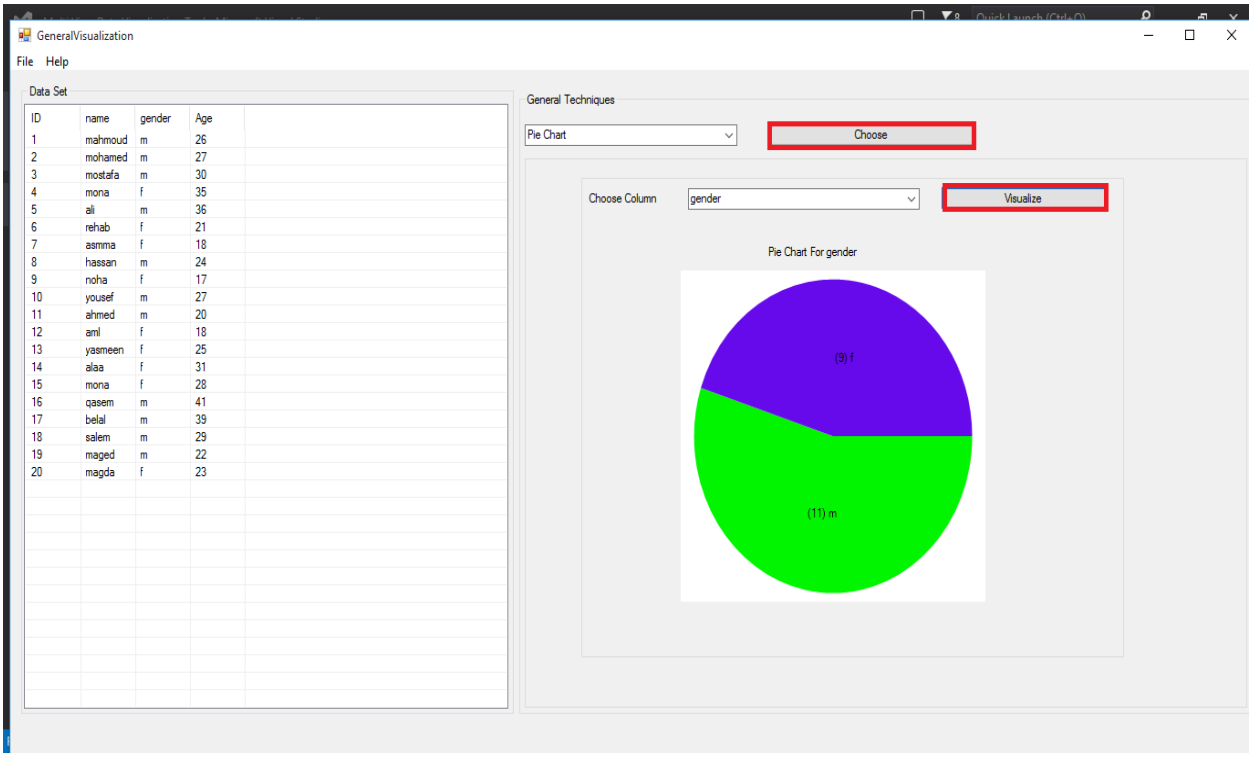


Fig. 5.17 Visualize Pie Chart

5.2.3.4.Column Chart

The user choose Technique -> Column Chart then choose which column need to visualize then Click on visualize

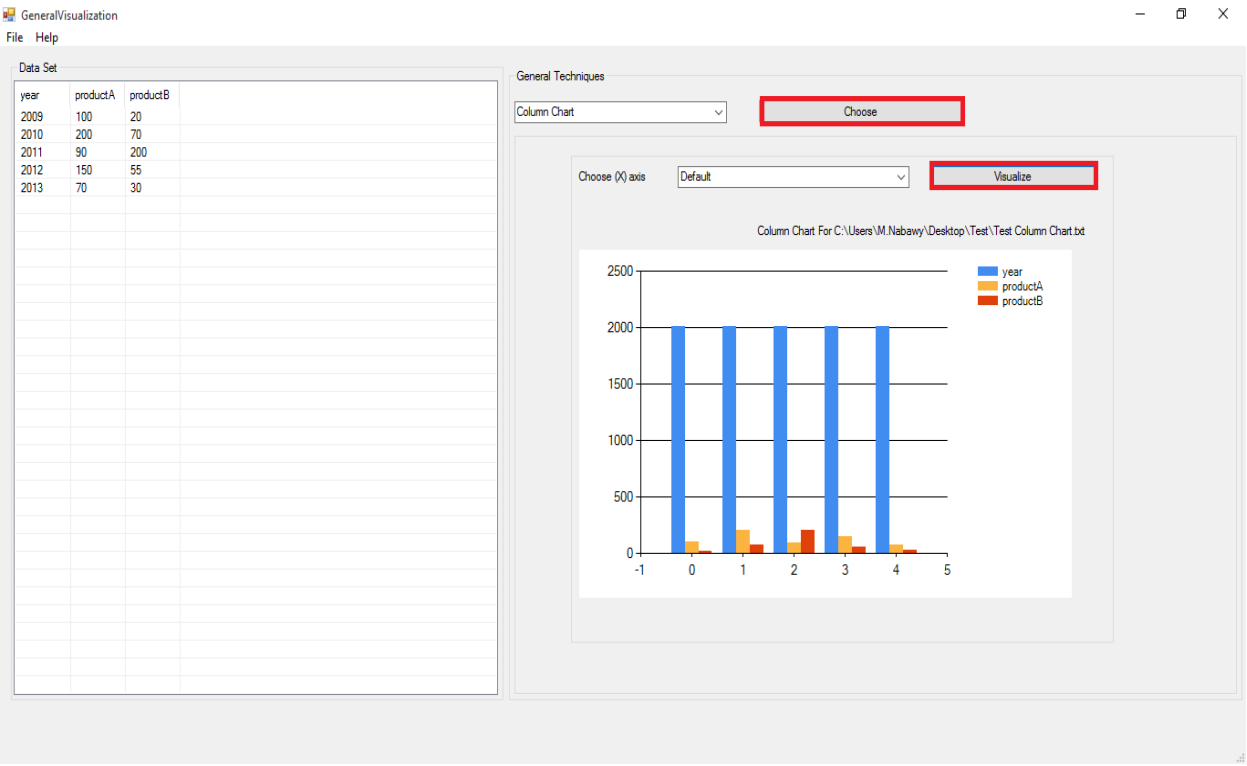


Fig. 5.18 Visualize Column Chart

5.2.3.5 Help

The user choose Help -> How To Use to open form Help

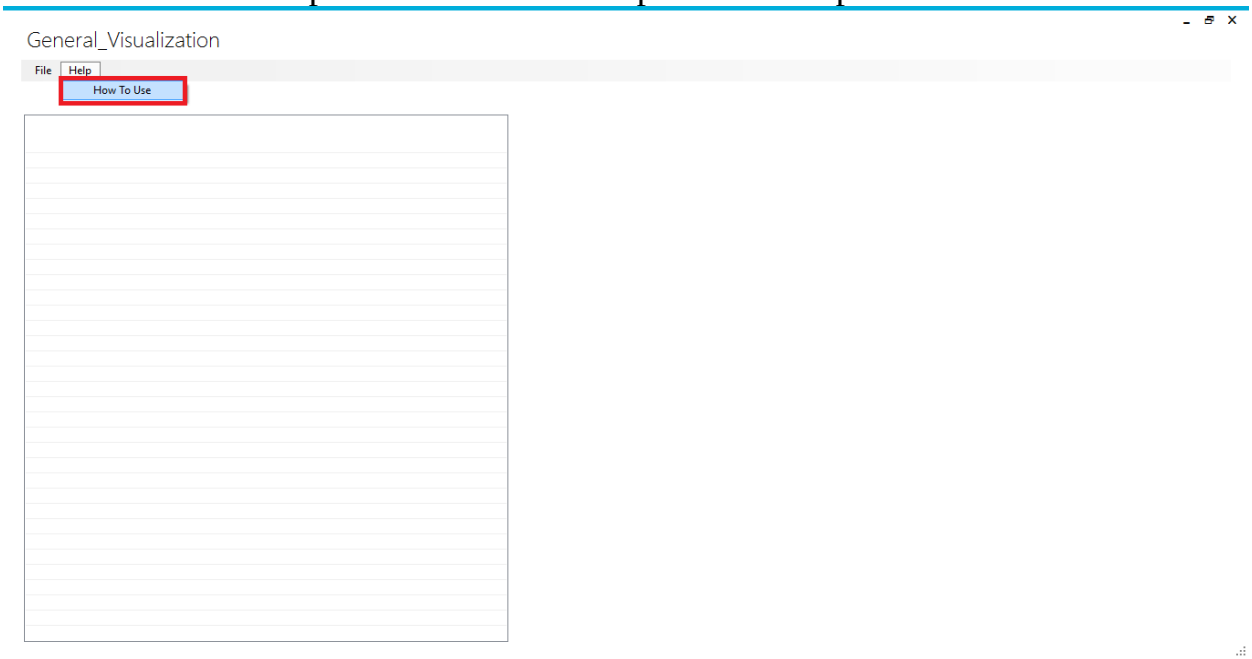


Fig. 5.19 How To Use in General Visualization

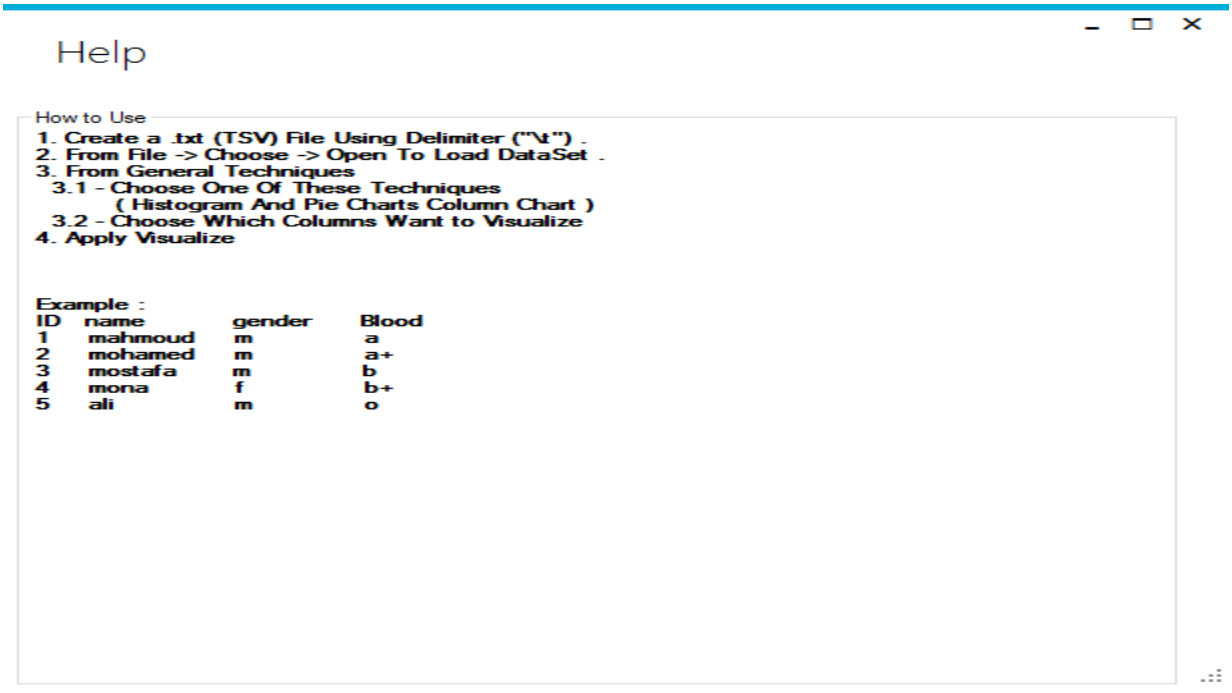


Fig. 5.20 Help in General Visualization

Chapter 6
Conclusions and Future Work

6.1: Conclusions

The final conclusions of our work are presented in this chapter. Due to the huge amount of data for the organizations are looking for easier ways to access, make some operation on it and visualize their data.
the processes like “cluster , classification “and visualization is useful for successful decision making.
Since data is huge, it will be very difficult to understand what is useful of this huge data
Our system is a disk top program for the organization’s data to filter the data ,improve of their quality , get some information of databyimplementation some technics of data mining , and get multiple views in data which makes it easier for the user to monitor his data and performance in real time and reflect on it and also facilitates better decision-making.

6.2: Future Work

Also, this chapter presents the possible directions for future work. We can enable the user to work on different types of files , get a report of all the algorithms and techniques used.
We can allow the user to use more visualization than the present ones.

References

1. https://www.analyticsvidhya.com/blog/2015/07/guide-data-visualization-r
2. https://github.com/csuldw/MachineLearning
3. http://www.ecs.syr.edu/faculty/chin/cse774/readings/rbac/role_graph_model_and_conflict_of_interest.pdf
4. https://www.techwalla.com/articles/advantages-disadvantages-of-decision-trees
5. https://www.singularities.com/blog/2015/08/apriori-vs-fpgrowth-for-frequent-item-set-mining
6. http://playwidtech.blogspot.com.eg/2013/02/k-means-clustering-advantages-and.html
7. https://www.quora.com/What-are-the-disadvantages-of-using-a-decision-tree-for-classification
8. https://en.wikipedia.org/wiki/Apriori_algorithm
9. https://en.wikipedia.org/wiki/K-medoids
10. https://en.wikipedia.org/wiki/DBSCAN
11. https://en.wikipedia.org/wiki/Hierarchical_clustering
12. https://en.wikipedia.org/wiki/Naive_Bayes_classifier
13. http://www.simafore.com/blog/bid/62333/4-key-advantages-of-using-decision-trees-for-predictive-analytics