

Unveiling Shopping Patterns:Market-Basket Analysis Using Data Mining Techniques

Project Report

Submitted to the Faculty of Engineering of

**JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA,
KAKINADA**

In partial fulfillment of the requirements for the award of the Degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEER

CH.YAMINI NAGA SAI PRASANNA(21481A0549)

B.DHATRI SRI(21481A0536)

B.VENU NAYAK(21481A0530)

D.SIVA GANESH(21481A0557)

Under the Enviabale and Esteemed Guidance of

Dr.G.SRIDEVI, M. Tech, (Ph.D)

Associate Professor, Department of CSE



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE**

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

SESHADRIRAO KNOWLEDGE VILLAGE

GUDLAVALLERU – 521356

ANDHRA PRADESH

2022-23

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

SESHADRI RAO KNOWLEDGE VILLAGE, GUDLAVALLERU

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



CERTIFICATE

This is to certify that the project report entitled **“UNVEILING SHOPPING PATTERNS:MARKET BASKET ANALYSIS USING DATA MINING TECHNIQUES”** is a bonafide record of work carried out by **Ch.Yamini Naga Sai Prasanna(21481A0549),B.Dhatri Sri(21481A0536),B.Venu Nayak(21481A1A0530),D.Siva Ganesh(21481A0557)**, under the guidance and supervision of **Dr.G.SRIDEVI, Associate professor**, Computer Science and Engineering, in the partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of Jawaharlal Nehru Technological University Kakinada, Kakinada during the academic year 2022-23.

Project Guide
(Dr.G.Sridevi)

Head of the Department
(Dr. M. BABU RAO)

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mention of people who made it possible and whose constant guidance and encouragements crown all the efforts with success.

We would like to express our deep sense of gratitude and sincere thanks to **Dr.G.Sridevi, Associate Professor**, Computer Science and Engineering for his constant guidance, supervision and motivation in completing the project work.

We feel elated to express our floral gratitude and sincere thanks to **Dr. M. Babu Rao, Head of the Department**, Computer Science and Engineering for his encouragements all the way during analysis of the project. His annotations, insinuations and criticisms are the key behind the successful completion of the project work.

We would like to take this opportunity to thank our beloved principal **Dr. B.Karuna Kumar** for providing a great support for us in completing our project and giving us the opportunity for doing project.

Our Special thanks to the faculty of our department and programmers of our computer lab. Finally, we thank our family members, non-teaching staff and our friends, who had directly or indirectly helped and supported us in completing our project intime.

Team Members:

Ch.Yamini Naga Sai Prasanna(21481A0549)
B.Dhatri Sri(21481A0536)
B.Venu Nayak(21481A0530)
D.Siva Ganesh(21481A0557)

INDEX

TITLE	PAGENO
LIST OF TABLES	i
LIST OF FIGURES	ii
ABSTRACT	iii
CHAPTER 1: INTRODUCTION	1
Introduction	
Problem definition	
CHAPTER 2	5
Methodology	
Block Diagram	
Algorithm	
Data Preparation	12
Dataset Description	
Data Pre-processing	
CHAPTER 3: RESULTS	13
ORANGE tool description	
Screen shots	
CHAPTER 4: CONCLUSION AND FUTURE SCOPE	20
List of Program Outcomes and Program Specific Outcomes	
Mapping of Program Outcomes with graduated POs and PSOs	

LIST OF TABLES

1.1.1 Supervised learning Algorithms

1.1.2 Unsupervised Learning Algorithms

LIST OF FIGURES

- 1.1.3 block diagram for association technique
- 1.1.4 data table of market basket dataset
- 1.1.5 orange tool download and install
- 2.1.6 open new file
- 2.1.2 load the data set
- 2.1.3 Data Info of dataset
- 2.1.4 Data table before preprocessing
- 2.1.5 Install association rules
- 3.1.1 association rules before preprocessing
- 3.1.2 Frequent itemsets before preprocessing
- 3.1.3 Association rules to predictions
- 3.1.4 predictions before preprocessing
- 3.1.5 barplot
- 4.1.1 lineplot
- 4.1.2 preprocessing data
- 4.1.3 data table after preprocessing
- 4.1.4 frequent itemsets after preprocessing
- 4.1.5 Association rules after preprocessing
- 5.1.1 predictions after preprocessing
- 5.1.2 barplot after preprocessing
- 5.1.3 lineplot after preprocessing
- 5.1.4 final view of project

ABSTRACT

This study delves into the analysis of a market basket dataset to uncover patterns and associations between various products typically purchased together by customers. The dataset comprises transactions, with each row indicating the presence or absence of items such as Bread, Milk, Diapers, Beer, Eggs, and Cola. Through meticulously data cleaning and preprocessing, we convert the dataset into a binary matrix suitable for association rule mining. Initial exploratory data analysis reveals the frequency of individual items and common item pairs. Employing the Apriori algorithm, we generate frequent itemsets and derive association rules to identify significant to the product combinations. These insights offer valuable implications for retail and strategies, such as product placement, promotions, and inventory management. By understanding customer purchasing behavior, retailers can optimize their operations to enhance customer satisfaction and drive sales. This analysis demonstrates the potential of market basket analysis in transforming raw transactional data into actionable business intelligence.

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

MACHINE LEARNING

Machine learning is a subset of artificial intelligence that involves training algorithms to make predictions or decisions based on data. In the context of data mining, machine learning algorithms can be used to discover patterns and relationships in large datasets.

There are many different types of machine learning algorithms that can be used in data mining, including supervised learning, unsupervised learning, and reinforcement learning.

SUPERVISED LEARNING:

Supervised learning is a type of machine learning in which an algorithm is trained on a labeled dataset, meaning that each example in the dataset has a corresponding output label. The algorithm learns to make predictions or decisions based on these labeled examples, with the goal of being able to accurately predict the output for new, unseen examples.

In data mining, supervised learning algorithms are often used for classification or regression tasks. In classification tasks, the goal is to predict a categorical label or class for each example. For example, a supervised learning algorithm could be trained to predict whether an email is spam or not based on the contents of the email. In regression tasks, the goal is to predict a continuous value for each example. For example, a supervised learning algorithm could be trained to predict the price of a house based on its size and location.

There are many different supervised learning algorithms that can be used in data mining, including decision trees, support vector machines, and neural networks. The choice of algorithm depends on the specific task at hand and the characteristics of the dataset.

Algorithm	Description	Type
Linear regression	Finds a way to correlate each feature to the output to help predict future values	Regression
Logistic Regression	Extension of linear regression that's used for classification tasks. The output variable is binary (e.g., only black or white) rather than continuous (e.g., an infinite list of potential colors)	Classification rather than regression

Fig 1.1.1: Supervised Learning algorithms

Naïve Bayes	The Bayesian method is a classification method that makes use of the Bayesian theorem. The theorem updates the prior knowledge of an event with the independent probability of each feature that can affect the event.	Regression and Classification
Support Vector Machine	SVM, is typically, used for the classification task. SVM algorithm finds a hyperplane that optimally divided the classes. It is best used with a non-linear solver.	Regression and classification

UNSUPERVISED LEARNING:

Unsupervised learning is a type of machine learning in which an algorithm is trained on an unlabeled dataset, meaning that there are no output labels provided for each example in the dataset. The algorithm learns to discover patterns and relationships in the data without any preconceived notions about what those patterns might be.

In data mining, unsupervised learning algorithms are often used for tasks such as clustering and anomaly detection. Clustering involves grouping together similar examples in the dataset based on their features, while anomaly detection involves identifying examples that are significantly different from the rest of the dataset.

There are many different unsupervised learning algorithms that can be used in data mining, including k-means clustering, hierarchical clustering, and principal component analysis (PCA). The choice of algorithm depends on the specific task at hand and the characteristics of the dataset.

REINFORCEMENT LEARNING:

Reinforcement learning is a type of machine learning in which an agent learns to make decisions in a particular environment through trial and error. The agent interacts with the environment by taking actions, and it receives rewards or penalties based on the outcomes of those actions. The goal of the agent is to learn a policy, which is a mapping from states to actions, that maximizes its expected cumulative reward over time.

In data mining, reinforcement learning algorithms can be used for tasks such as game playing, robotics, and resource management. For example, a reinforcement learning algorithm could be used to train a robot to navigate a maze, or to train an autonomous vehicle to drive safely on the road.

There are many different reinforcement learning algorithms that can be used in data mining, including Q-learning, policy gradient methods, and actor-critic methods. The choice of algorithm depends on the specific task at hand and the characteristics of the environment.

Algorithm	Description	Type
K-Means clustering	Puts data into some groups (k) that each contains data with similar characteristics (as determined by the model, not in advance by humans)	Clustering
Gaussian mixture model	A generalization of k-means clustering that provides more flexibility in the size and shape of groups (clusters)	Clustering
Hierarchical clustering	Splits clusters along a hierarchical tree to form a classification system. Can be used for cluster loyalty-card customer	Clustering
Recommender system	Help to define the relevant data for making a recommendation	Clustering
PCA/T-SNE	Mostly used to decrease the dimensionality of	Dimension Reduction

Fig 1.1.2: Unsupervised learning algorithm

Exploring Data Mining Association Technique:

In data mining, machine learning algorithms can be used for a variety of tasks, including classification, clustering, and prediction. For example, a machine learning algorithm could be used to predict whether a customer is likely to purchase a particular product based on their past purchase history. Or, it could be used to cluster customers based on their purchasing behavior to identify different segments of the market.

Overall, machine learning is a powerful tool in data mining that can help businesses and organizations uncover valuable insights from their data. By using these algorithms to discover patterns and relationships in their data, businesses can make more informed decisions and improve their bottom line.

In recent years, the amount of data being generated by organizations has grown exponentially. This has led to an increasing demand for data mining techniques that can help organizations extract insights and knowledge from their data. Association analysis is one such technique that has gained widespread popularity due to its ability to identify interesting relationships between variables.

The objective of this project is to apply association analysis to a real-world dataset and identify significant patterns and relationships between variables. Specifically, we will be analyzing a dataset of customer transactions from an online retailer to identify frequent itemsets and association rules that can help the retailer understand customer behavior and optimize its product offerings.

To achieve this objective, we will be using the Apriori algorithm, which is a popular algorithm for frequent itemset mining. Apriori works by generating a set of candidate itemsets and pruning those that do not meet a minimum support threshold. The remaining frequent itemsets can then be used to generate association rules that provide insights into the relationships between different items.

In this project, we will also explore different measures of rule interestingness, such as support, confidence, and lift, to identify the most meaningful rules. We will then use these rules to make recommendations to the retailer on how to optimize its product offerings and improve customer satisfaction.

The rest of this project report is organized as follows. In the next section, we will provide an overview of the literature on association analysis and its applications. We will then describe the dataset used in this project and the preprocessing steps that were taken to prepare the data for analysis. Next, we will present the results of our analysis, including the frequent itemsets and association rules that we identified.

Problem Statement:

The project "Unraveling Consumer Behavior: Association Rule Mining on Market Basket Dataset" aims to uncover the hidden patterns and relationships between items that are frequently purchased together by customers in a retail store. By analyzing the market basket dataset, which contains information about the items purchased by customers in each transaction, the project seeks to identify the association rules that govern consumer behavior and to provide insights into how retailers can optimize their product offerings and marketing strategies.

The specific problem statement for the project is:

Given a large market basket dataset containing information about the items purchased by customers in each transaction, the goal of this project is to use association rule mining techniques to identify the frequent itemsets and association rules that govern consumer behavior. These rules should provide insights into which products are frequently purchased together, which products are often purchased as substitutes or complements, and which products are likely to be purchased by specific customer segments. The project should also evaluate the effectiveness of different association rule mining algorithms and provide recommendations for how retailers can use these insights to optimize their product offerings and marketing strategies.

CHAPTER 2

METHODOLOGY:

The Association technique is a popular data mining technique used to find interesting relationships between variables in a dataset. Here is a high-level overview of the methodology for association analysis:

1. **Data Collection:** The first step in any data analysis project is to collect and prepare the data. In the case of association analysis, this involves collecting transaction data from a retail store or online retailer. This transaction data typically includes information on the products purchased by customers and the time and date of each transaction.
2. **Data Preparation:** Once the data has been collected, it needs to be cleaned and prepared for analysis. This may involve removing duplicate transactions, identifying and removing outliers, and transforming the data into a format suitable for association analysis.
3. **Data Analysis:** The next step is to analyze the data using association analysis techniques. This involves identifying frequent itemsets, which are groups of items that are frequently purchased together, and generating association rules, which are relationships between different items in the dataset.
4. **Rule Evaluation:** Once the association rules have been generated, they need to be evaluated to determine their usefulness and relevance. This may involve calculating metrics such as support, confidence, and lift to assess the strength of the relationships between different items.
5. **Rule Deployment:** The final step is to deploy the association rules in a real-world setting. This may involve using the rules to identify product bundles or promotions that can be offered to customers to increase sales and customer satisfaction.

Notation and Basic Concepts:

Let $\Omega = \{i_1, i_2 \dots i_m\}$ be a universe of items. Also, let $T = \{t_1, t_2 \dots t_n\}$ be a set of all transactions collected over a given period of time. To simplify a problem, we will assume that every item i can be purchased only once in any given transaction t . Thus $t \subseteq \Omega$ (“ t is a subset of Ω ”). In reality, each transaction t is assigned a number, for example a transaction id (TID). Let now A be a set of items (or an itemset).

transaction t is said to contain A if and only if $A \subseteq t$. Now, mathematically, an association rule will be an implication of the form

$$A \Rightarrow B$$

Where both A and B are subsets of Ω and $A \cap B = \emptyset$ (“the intersection of sets A and B is an empty set”).

Support :

The support of an itemset is the fraction of the rows of the database that contain all of the items in the itemset. Support indicates the frequencies of the occurring patterns. Sometimes it is called frequency. Support is simply a probability that a randomly chosen transaction t contains both itemsets A and B . Mathematically,

$$\text{Support}(A \Rightarrow B) = P(A \subseteq t \wedge B \subseteq t)$$

We will use a simplified notation that

$$\text{Support}(A \Rightarrow B) = P(A \wedge B)$$

Confidence:

Confidence denotes the strength of implication in the rule. Sometimes it is called accuracy. Confidence is simply a probability that an itemset B is purchased in a randomly chosen transaction t given that the itemset A is purchased. Mathematically,

$$\text{Confidence}(A \Rightarrow B) = P(B \subseteq t \mid A \subseteq t)$$

We will use a simplified notation that

$$\text{Confidence}(A \Rightarrow B) = P(B|A)$$

In general, a set of items (such as the antecedent or the consequent of a rule) is called an itemset. The number of items in an itemset is called the length of an itemset. Itemsets of some length k are referred to as k -itemsets. Generally, an association rules mining algorithm contains the following steps:

- The set of candidate k-itemsets is generated by 1-extensions of the large (k -1)- itemsets generated in the previous iteration

Supports for the candidate k-itemsets are generated by a pass over the database.

- Itemsets that do not have the minimum support are discarded and the remaining itemsets are called large item sets
This process is repeated until no more large itemsets are found.

Association rules:

We can define association rule as follows:

Let $I = \{i_1, i_2, \dots, i_m\}$ be a set of literals or items, $D = \{t_1, t_2, \dots, t_n\}$ be a set of transactions, where each transaction t_i is an itemset such that $t_i \subseteq I$. Each transaction, t , has a transaction-id ($t.id$) and an itemset ($t.Itemset$), i.e., $t = (t.id, t.Itemset)$. A transaction t contains an itemset X if X is a subset of $t.Itemset$. An Association rule, R , denoted by $R: X \rightarrow Y$, where X and Y are itemsets that don't intersect. Each rule R has two value measures, support and confidence, denoted by $sup(R)$ and $conf(R)$ respectively. The support of an item set, X , has support, s , in transaction set, D , if $s\%$ of transaction in D contain X . Then, $sup(R: X \rightarrow Y) = sup(X \rightarrow Y)$, $conf(R: X \rightarrow Y) = sup(X \rightarrow Y) / sup(X)$. Different transactions may contain same itemset, especially for remote sensed imagery. This suggests a way to eliminate duplicate calculation. Some concepts are given below. Let $U = \{t \mid t \text{ is any possible transaction}\}$, while $D = \{t \mid t \text{ is a transaction already happened}\}$.

Frequent Itemsets:

In many (but not all) situations, we only care about association rules or causalities involving sets of items that appear frequently in baskets. For example, we cannot run a good marketing strategy involving items that no one buys anyway. Thus, much data mining starts with the assumption that we only care about sets of items with high support; i.e., they appear together in many baskets. We then find association rules or causalities only involving a high-support set of items i.e., $\{X_1, \dots, X_m\}$. Y must appear in at least a certain percent of the baskets, called the support threshold.

The AIS algorithm was the first algorithm proposed for mining association rule. In this algorithm only one item consequent association rules are generated, which means that the consequent of those rules only contain one item, for example we only generate rules like $X \cap Y \rightarrow Z$ but not those rules as $X \rightarrow Y \cap Z$. The main drawback of the AIS algorithm is too many candidate itemsets that finally turned out to be small are generated, which requires more space and wastes much effort that turned out to be useless. At the same time this algorithm requires too many passes over the whole database. Apriori is more efficient during the candidate generation process. Apriori uses pruning techniques to avoid measuring certain itemsets, while guaranteeing completeness. These are the itemsets that the algorithm can prove will not turn out to be large. However there are two bottlenecks of the Apriori algorithm. One is the complex candidate generation process that uses most of the time, space and memory. Another bottleneck is the multiple scan of the database. Based on Apriori algorithm, many new algorithms were designed with some modifications or improvements.

Algorithm for Finding Frequent Itemsets

1. Given support threshold s , in the first pass we find the items that appear in at least fraction s of the baskets. This set is called L_1 , the frequent items.
2. Pairs of items in L_1 become the candidate pairs C_2 for the second pass. We hope that the size of C_2 is not so large that there is not room for an integer count per candidate pair. The pairs in C_2 whose count reaches s are the frequent pairs, L_2 .
3. The candidate triples, C_3 are those sets $\{A; B; C\}$ such that all of $\{A; B\}$, $\{A; C\}$, and $\{B; C\}$ are in L_2 . On the third pass, count the occurrences of triples in C_3 ; those with a count of at least s are the frequent triples, L_3 .
4. Proceed as far as you like (or the sets become empty). L_i is the frequent sets of size i ; C_{i+1} is the set of sets of size $i + 1$ such that each subset of size i is in L_i .

BLOCK DIAGRAM:

The block diagram for Association Technique typically involves the following components:

1.Data Preparation: This component involves cleaning and preparing the data for analysis. The data may be collected from various sources, such as transactional databases, web logs, or customer feedback forms. The data may need to be transformed or aggregated to facilitate analysis.

2.Frequent Itemset Generation: This component involves identifying groups of items that are frequently purchased together by customers. This is typically done using algorithms such as the Apriori algorithm or the FP-Growth algorithm.

3.Rule Generation: Once the frequent itemsets have been identified, the next step is to generate association rules. Association rules are relationships between different items in the dataset that occur with a certain frequency. The strength of these rules is typically measured using metrics such as support, confidence, and lift.

4.Rule Evaluation: This component involves evaluating the association rules to determine their usefulness and relevance. This may involve filtering out rules that are not interesting or useful and selecting the most relevant rules for deployment.

5.Rule Deployment: The final component involves deploying the association rules in a real-world setting. This may involve using the rules to identify product bundles or promotions that can be offered to customers to increase sales and customer satisfaction.

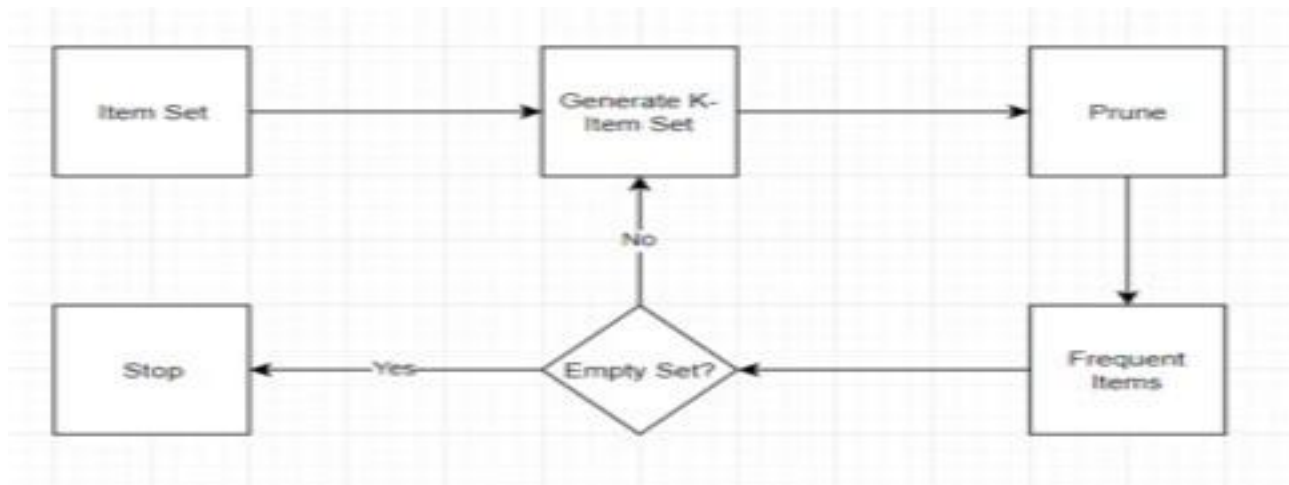


Fig 1.1.3:block diagram for assosiation technique

ALGORITHM:

Apriori Algorithm:

Find frequent itemsets using an iterative level-wise approach based on candidate generation.

Input:

- D , a database of transactions;
- min_sup , the minimum support count threshold.

Output:

- L , frequent itemsets in D . Method:

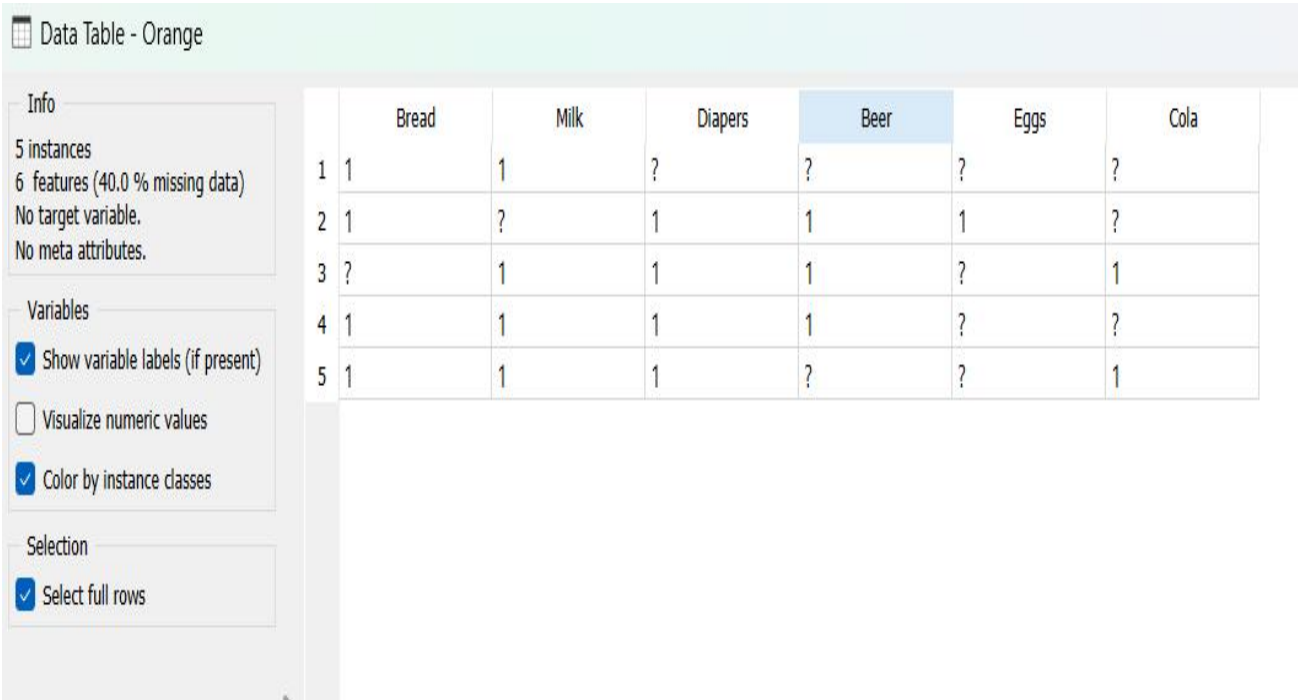
- (1) $L_1 = \text{find_frequent_1-itemsets}(D)$;
- (2) for ($k=2$; $L_{k-1} \neq \emptyset$; $k++$) {
- (3) $C_k = \text{apriori_gen}(L_{k-1})$;
- (4) for each transaction $t \in D$ { // scan D for counts
- (5) $C_t = \text{subset}(C_k, t)$; // get the subsets of t that are candidates
- (6) for each candidate $c \in C_t$
- (7) $c.\text{count}++$;
- (8) }
- (9) $L_k = \{c \in C_k \mid c.\text{count} > \text{min_sup}\}$
- (10) }
- (11) return $L = \bigcup_k L_k$;

Data Preparation

DataSet Description

A dataset containing transaction ID and items typically includes data on customer transactions in a retail store or online retailer. The dataset includes a unique identifier for each transaction (transaction ID) and the items that were purchased in the transaction.

The item data may include the name or code of the item purchased, along with other information such as the price, quantity, or category of the item. For example, a transaction may include the purchase of items such as “bread” , “milk” , “diapers” , “beer” , “eggs” , “cola” .



The screenshot shows the Orange Data Table widget interface. On the left, a sidebar contains settings for the widget. The main area displays a table with 5 rows and 6 columns. The columns are labeled 'Bread', 'Milk', 'Diapers', 'Beer', 'Eggs', and 'Cola'. The rows are numbered 1 to 5. The data is as follows:

	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	?	?	?	?
2	1	?	1	1	1	?
3	?	1	1	1	?	1
4	1	1	1	1	?	?
5	1	1	1	?	?	1

The sidebar settings are as follows:

- Info:** 5 instances, 6 features (40.0 % missing data), No target variable, No meta attributes.
- Variables:** ☒ Show variable labels (if present), ☐ Visualize numeric values, ☒ Color by instance classes.
- Selection:** ☒ Select full rows.

Fig 1.1.4:Data table of market-basket dataset

CHAPTER 3

RESULTS

ORANGE tool description:

Orange is an open-source data visualization and analysis tool designed for users seeking intuitive yet powerful solutions in machine learning and data mining. Its hallmark feature is a visual programming interface, facilitating the construction of data analysis workflows through interconnected components (widgets). With this approach, users can perform various tasks seamlessly, including data preprocessing, exploratory data analysis, predictive modeling, and visualization. Orange offers an array of preprocessing techniques, allowing users to handle missing values, scale features, encode categorical variables, and select relevant features effortlessly. Moreover, its extensive collection of visualization tools enables users to explore datasets visually, uncovering relationships, distributions, and patterns. Through integration with machine learning algorithms and ensemble learning methods, Orange empowers users to train models for classification, regression, clustering, and association rule mining. Model evaluation tools further aid in assessing model performance, ensuring robust and reliable results. With its blend of usability and versatility, Orange serves as a valuable asset for data scientists, researchers, and analysts across various domains, fostering innovation and insight discovery.

Screen shots

Step 1: Download and install ORANGE

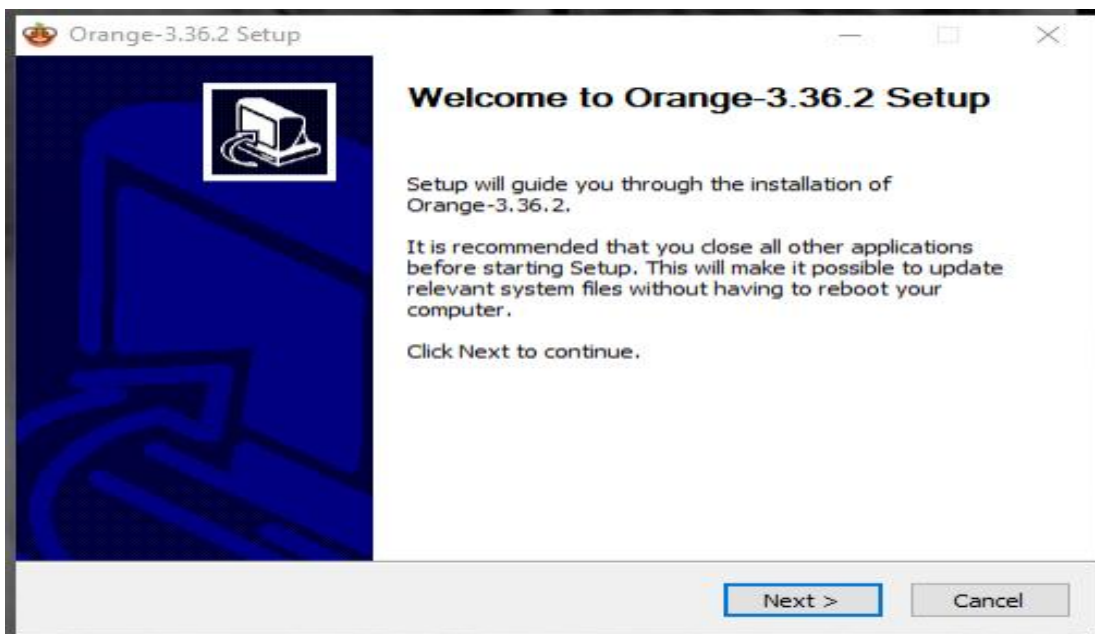


Fig 1.1.5: Orange tool download and install

Step 2: Open Orange and Select new to start a new project

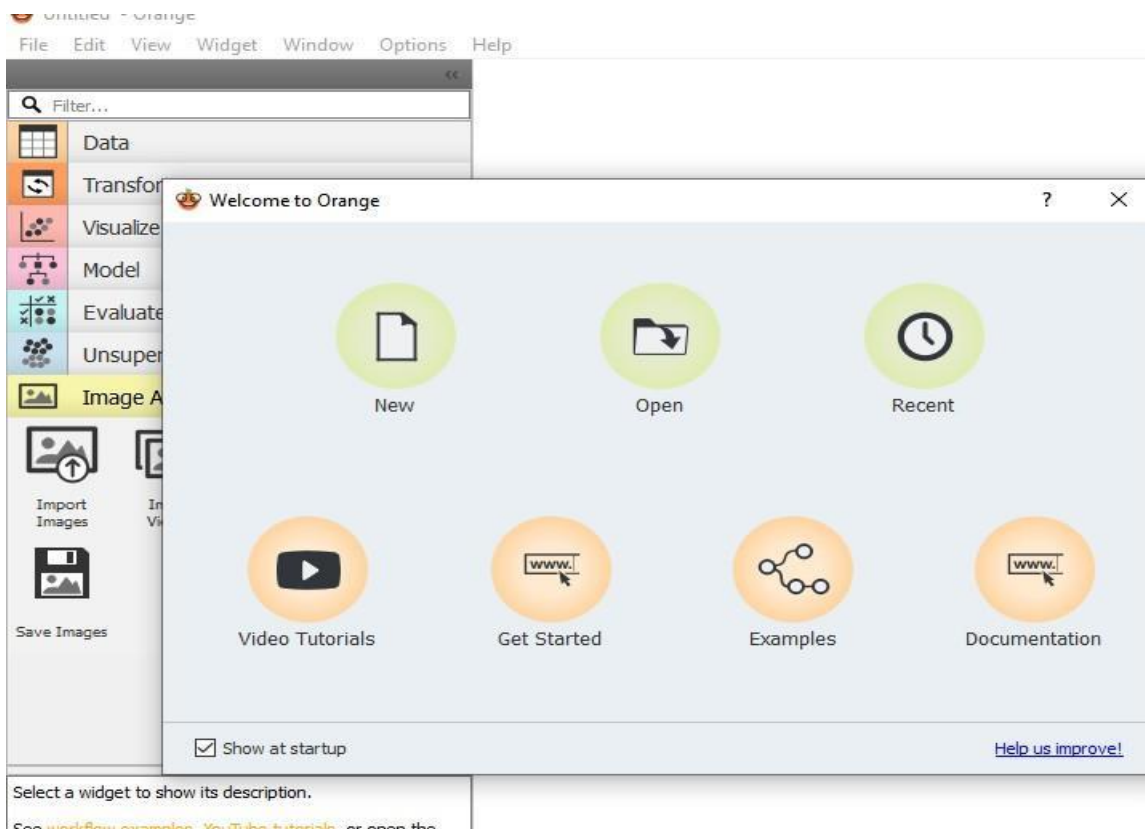


Fig 2.1.1:Open new file

Step 3: From the Data, select file. Double click on it and Load the dataset

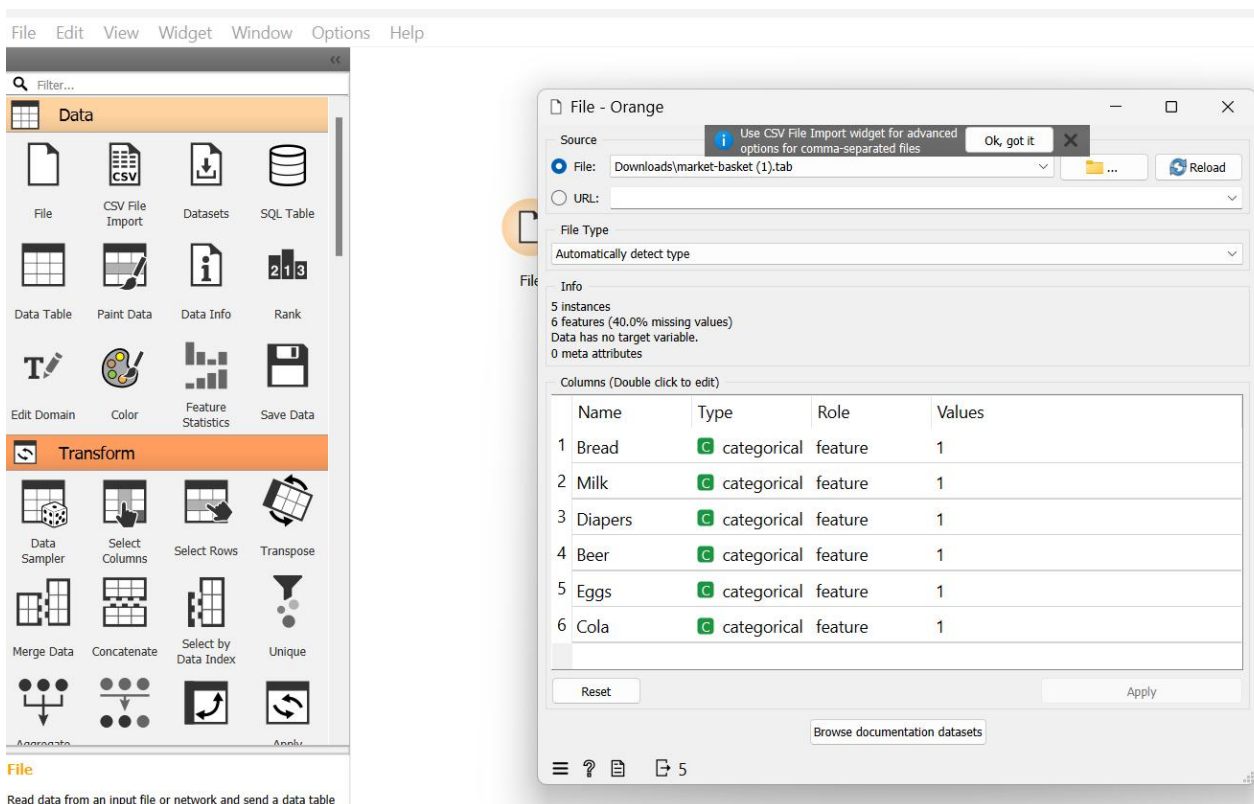


Fig 2.1.2:Load the dataset

Step 4: The dataset can be viewed with a Data Table and its information with Data Info

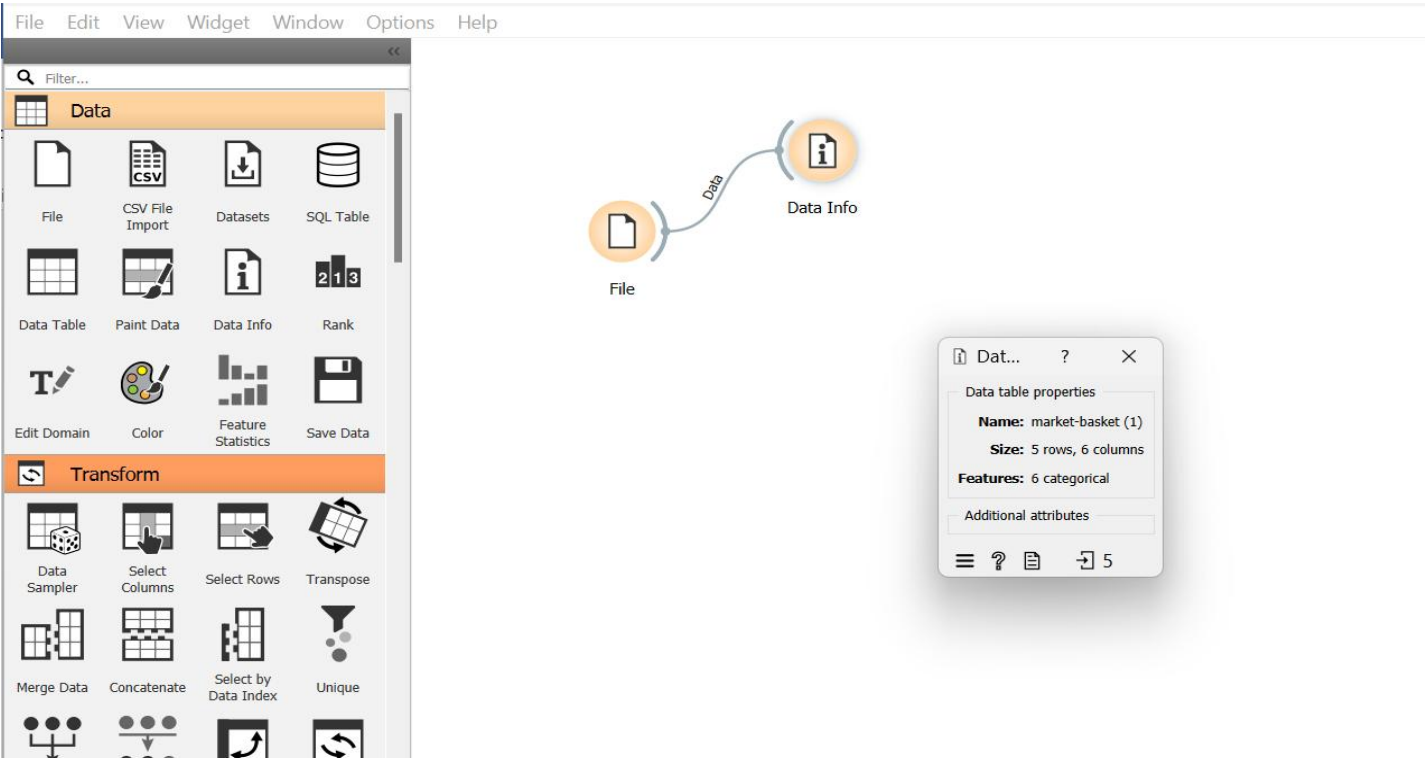


Fig 2.1.3:Data info of dataset

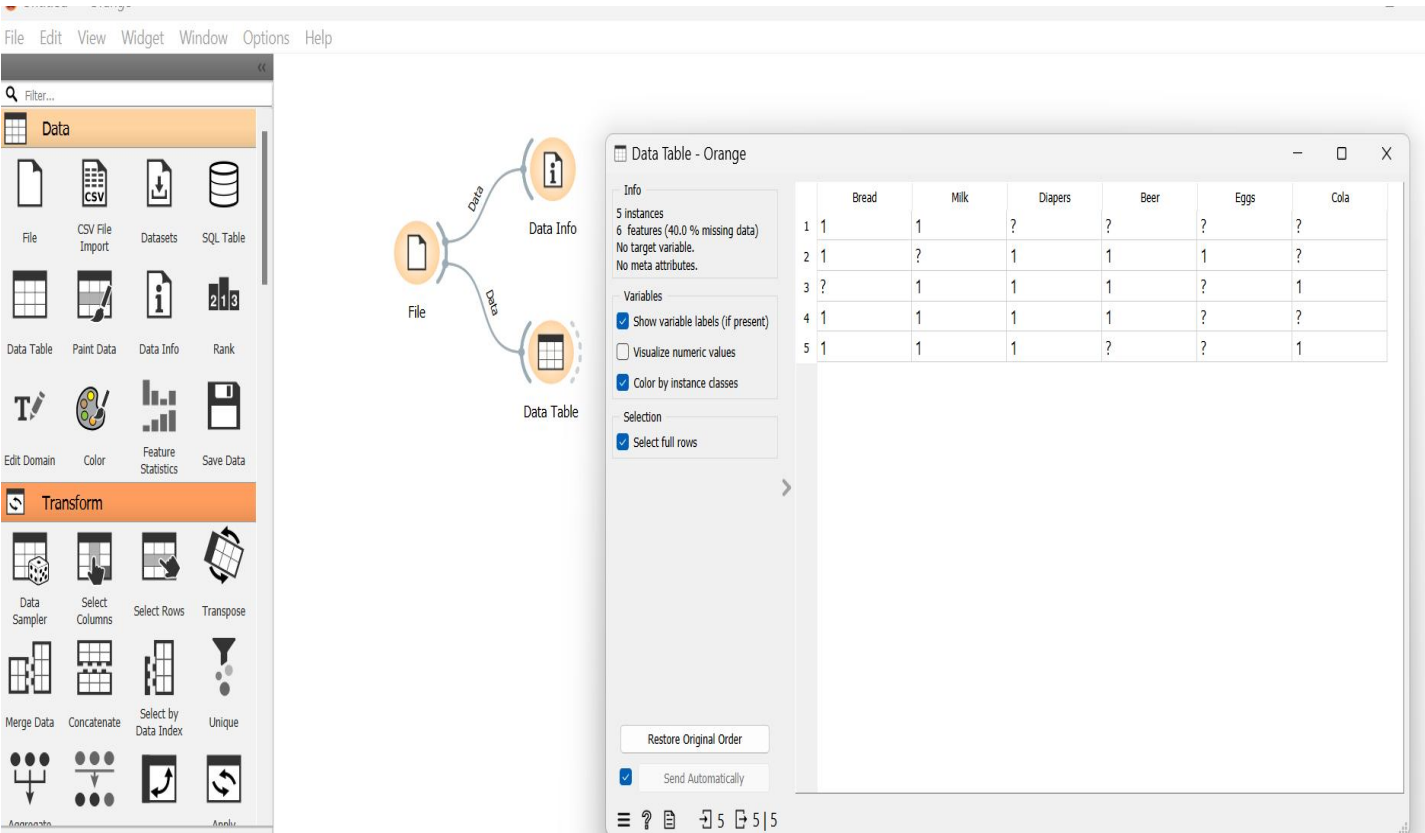


Fig 2.1.4:Data table before preprocessing

Step 5:Goto options->addons to insall associate

- From the associate, association rules. Double click on it and adjust min. Supp ,min, conf

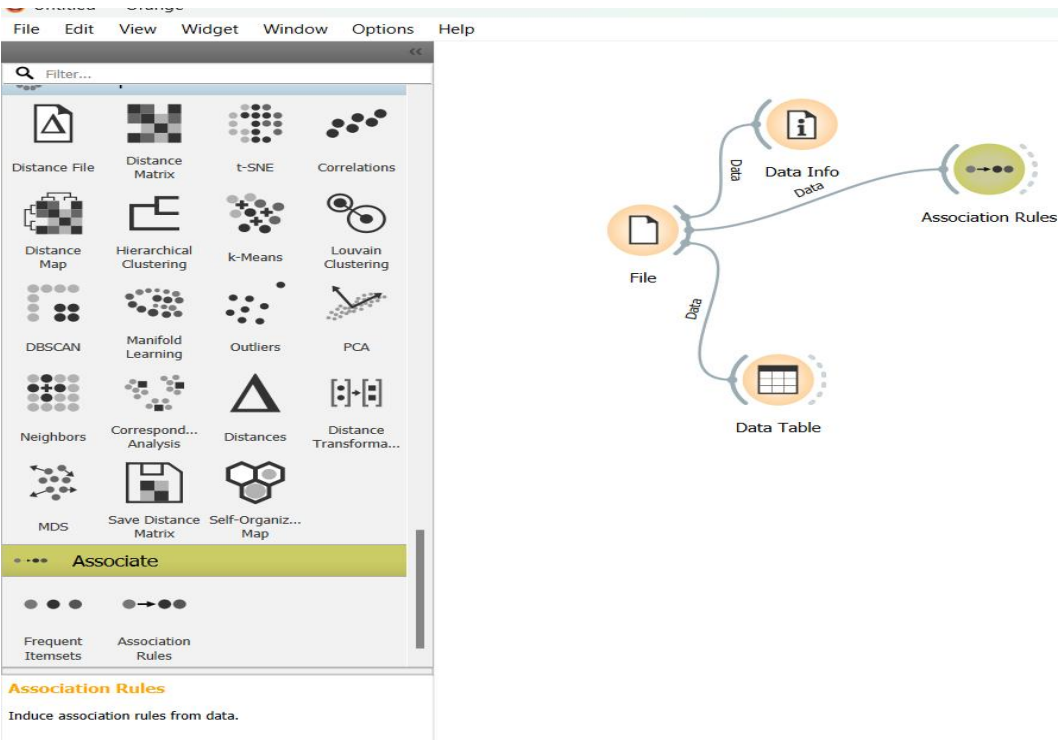


Fig 2.1.5:Install association rules

The screenshot shows the 'Association Rules - Orange' window. It displays the results of association rule mining. The window has a left sidebar with filters and a main table of results.

Info
Rules: 8 (shown 8)

Find association rules
Min. supp.: 40 %
Min. conf.: 80 %
Max. rules: 10k
☐ Induce only classification rules
☐ Restrict search by below filters
Find Rules

Filter by Antecedent
Contains:
Items, min: 1 max: 999

Filter by Consequent
Contains:
Items, min: 1 max: 99

☒ Send selection

Supp	Conf	Covr	Strg	Lift	Levr	Antecedent	Consequent
0.600	1.000	0.600	1.333	1.250	0.120	Beer=1	Diapers=1
0.400	1.000	0.400	2.000	1.250	0.080	Bread=1, Beer=1	Diapers=1
0.400	1.000	0.400	2.000	1.250	0.080	Milk=1, Beer=1	Diapers=1
0.400	1.000	0.400	2.000	1.250	0.080	Cola=1	Milk=1
0.400	1.000	0.400	2.000	1.250	0.080	Cola=1	Diapers=1
0.400	1.000	0.400	2.000	1.250	0.080	Diapers=1, Cola=1	Milk=1
0.400	1.000	0.400	2.000	1.250	0.080	Milk=1, Cola=1	Diapers=1
0.400	1.000	0.400	1.500	1.667	0.160	Cola=1	Milk=1, Diapers=1

Fig 3.1.1:Association rules before preprocessing

From the associate, frequent itemsets. Double click on it and adjust min. Supp

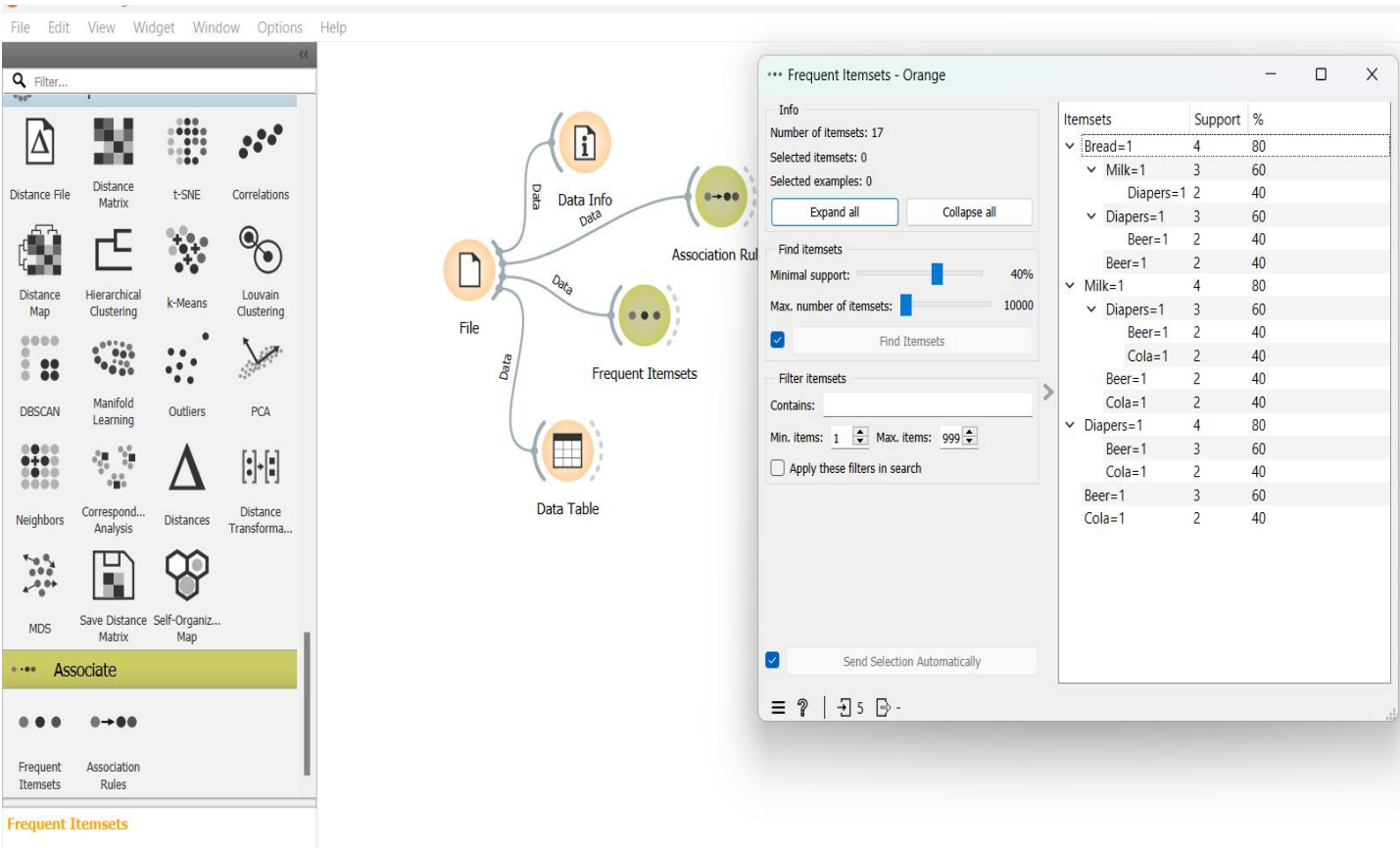


Fig 3.1.2:frequent itemsets before preprocessing

Step 6: From Evaluate, select predictions and connect it with association rules.

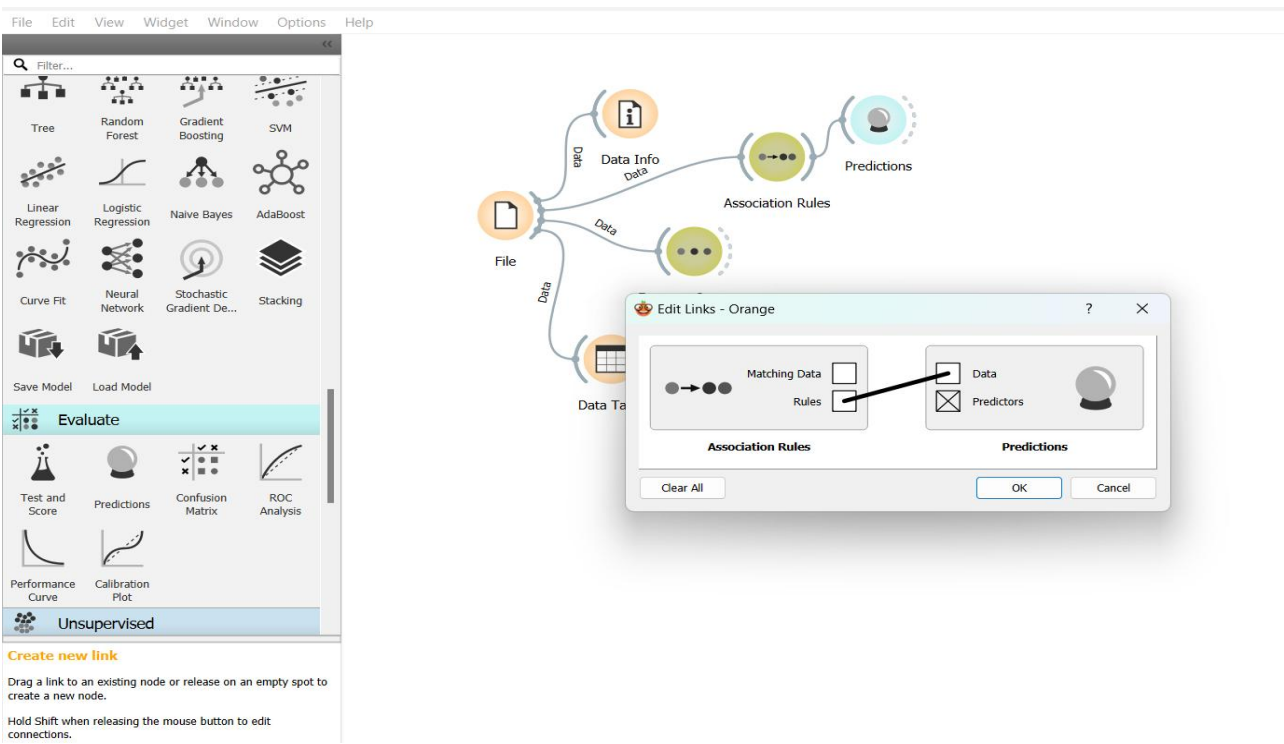


Fig 3.1.3:Association rules to predictions

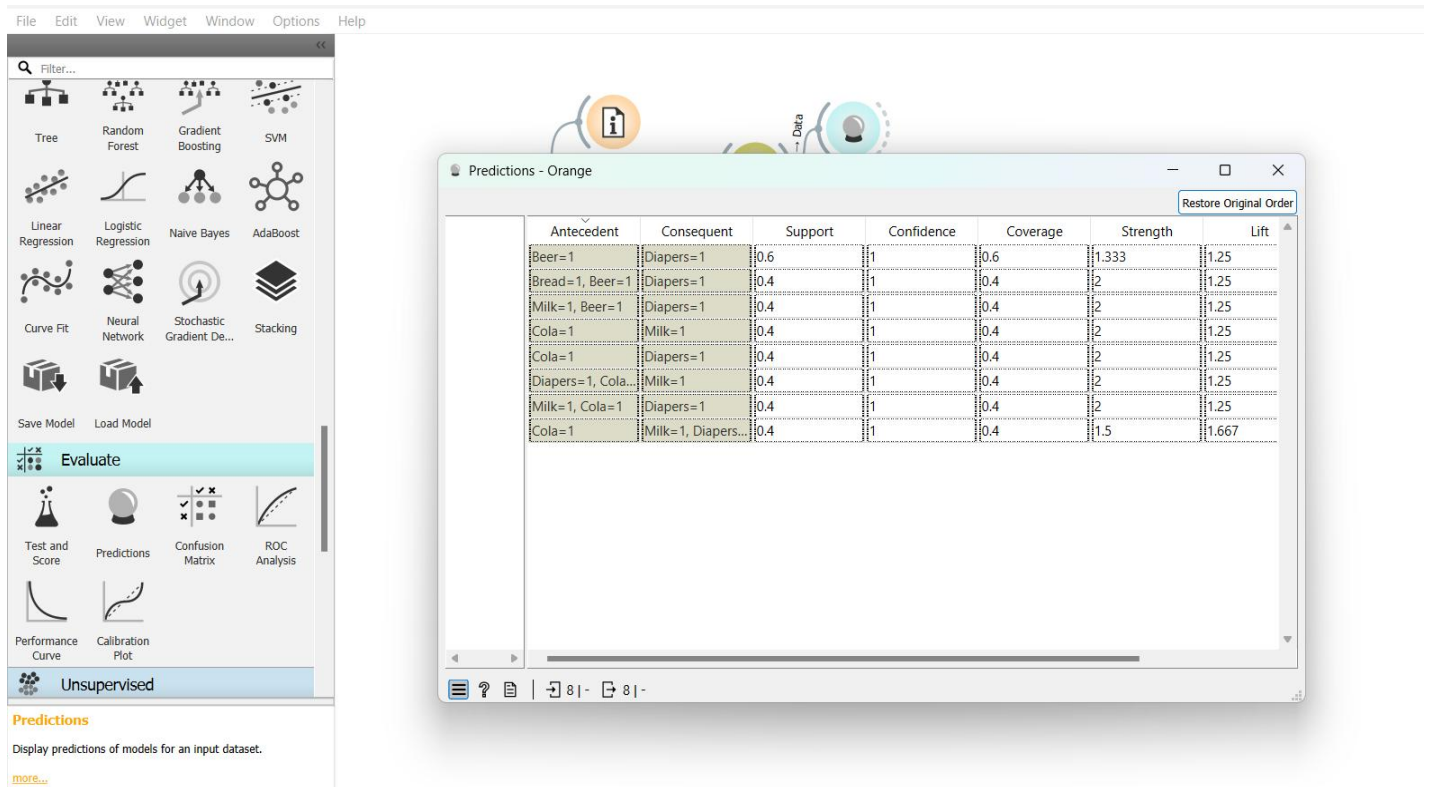


Fig.3.1.4 :predictions before preprocessing

Step 7: From vizualize, select -Barplot



Fig 3.1.5:Barplot

-Line plot

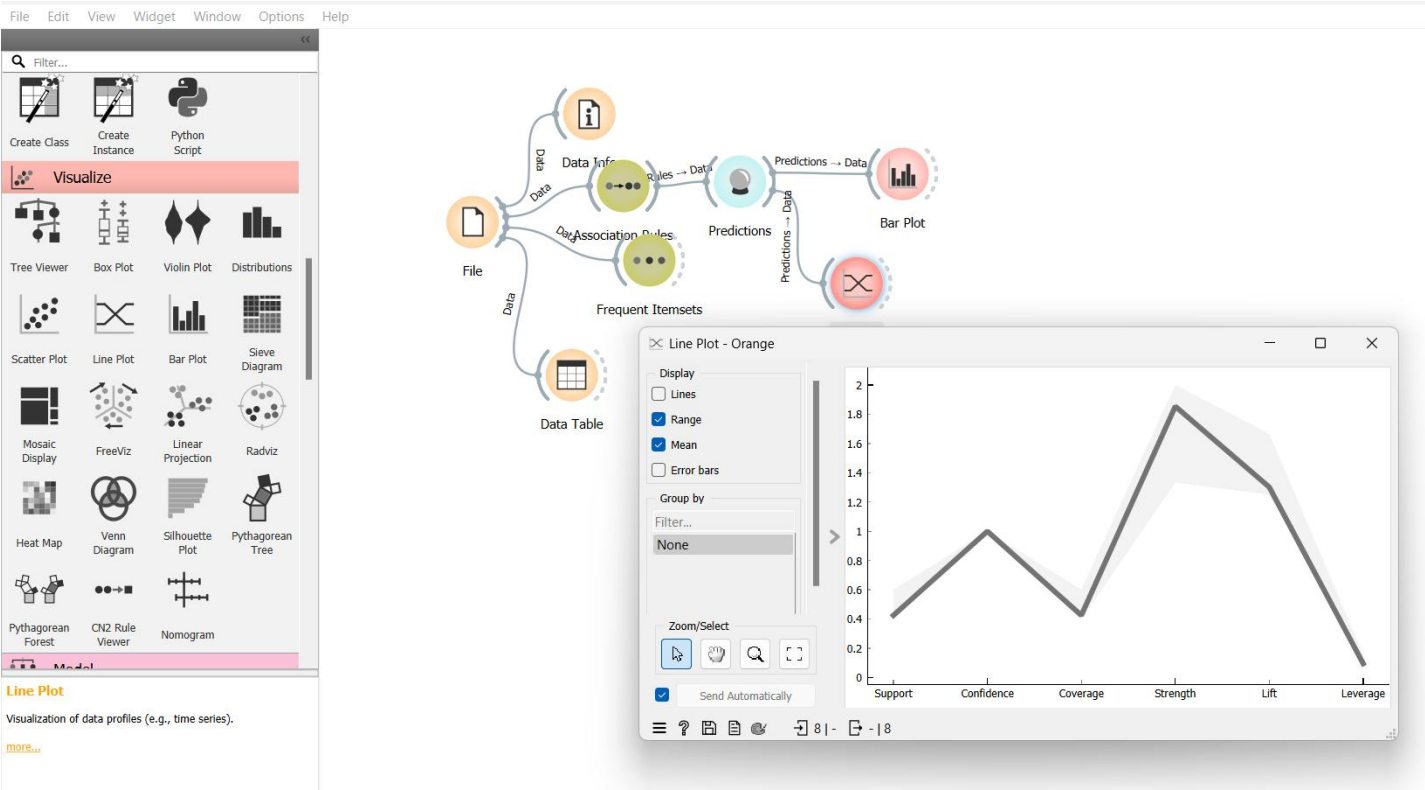


Fig 4.1.1:lineplot

Step 8:As we have missing values we need to preprocess the data.

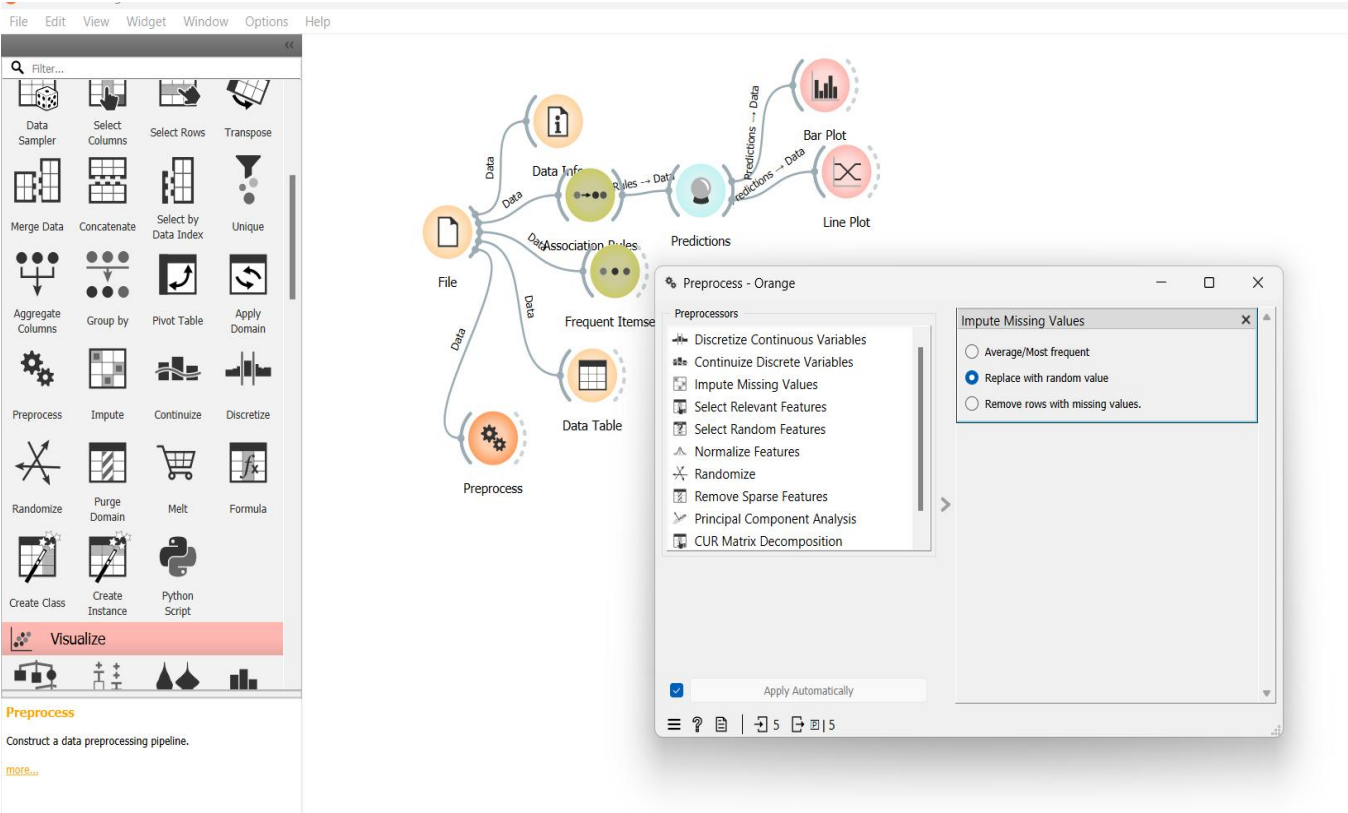


Fig 4.1.2:Preprocessing data

Step 9: Adding Datable to preprocess to view the data set

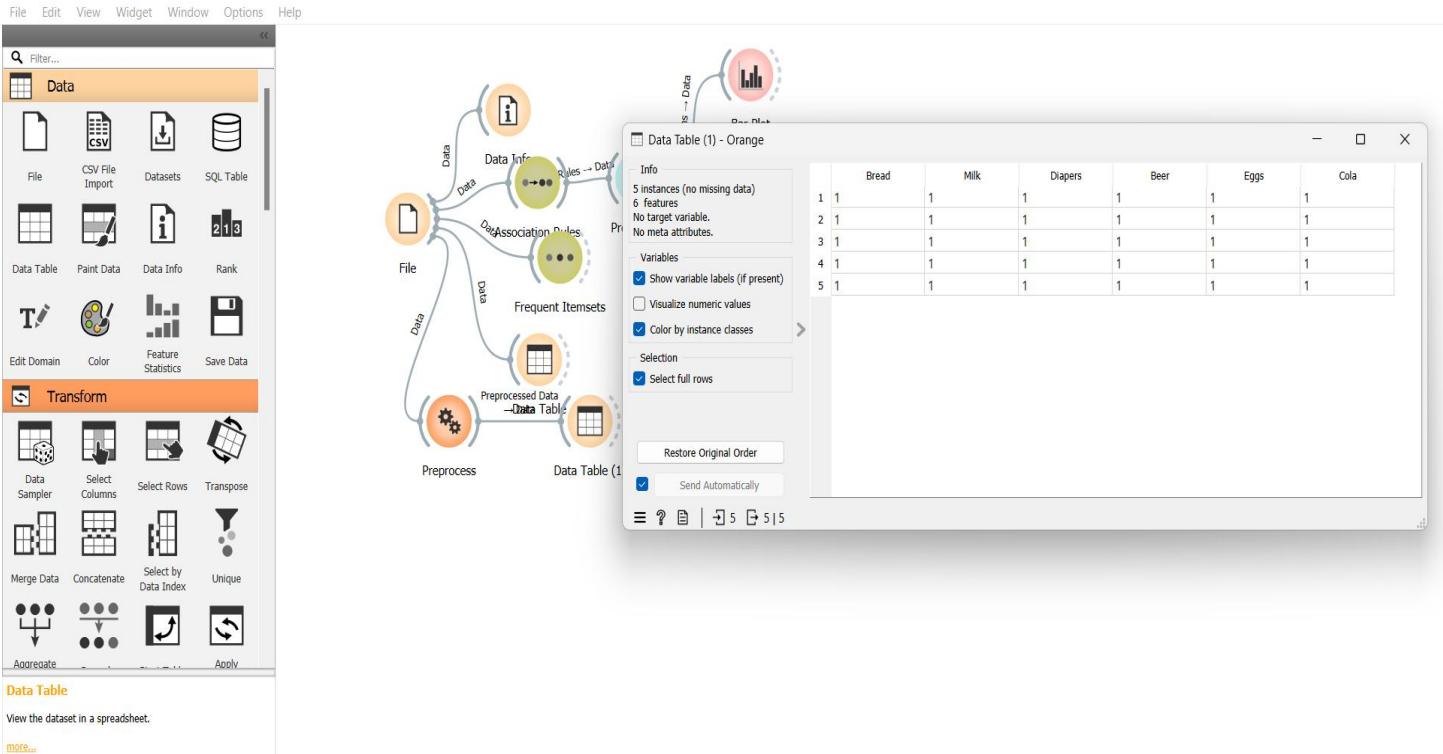


Fig 4.1.3: data table after preprocessing

Step 10: Select frequent itemsets from Associate and connect it with preprocess to generate frequent itemsets for preprocessed data

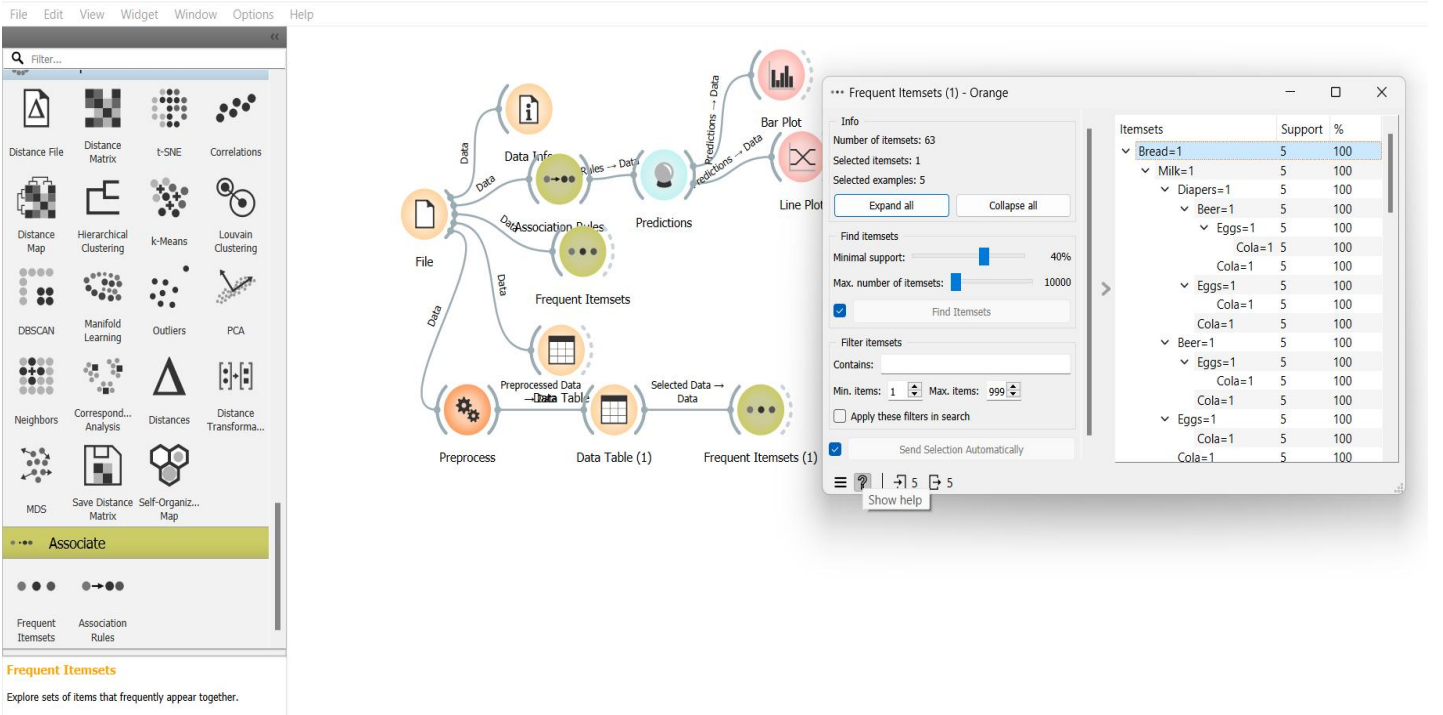


Fig 4.1.4: frequent itemsets after preprocessing

The screenshot displays the Orange3 data mining software interface. The main window shows a workflow canvas with the following nodes: File, Data, Data Table, Preprocess, Data Table (1), Frequent Itemsets, Association Rules (1), and Data Table (2). The 'Association Rules (1)' widget is selected, showing its configuration panel. The configuration panel includes settings for 'Find association rules' (Rules: 602 shown, 602 total), 'Min. supp.: 40%', 'Min. conf.: 80%', 'Max. rules: 10k', and 'Filter by Antecedent' (Contains:). The 'Association Rules (1)' widget displays a table of rules with columns: Supp, Conf, Covr, Strg, Lift, Lev, and Antecedent. The table shows 10 rules, including 'Milk=1 → Bread=1', 'Bread=1 → Milk=1', 'Diapers=1 → Bread=1', 'Bread=1 → Diapers=1', 'Diapers=1 → Milk=1', 'Milk=1 → Diapers=1', 'Milk=1, Diapers=1 → Bread=1', 'Bread=1, Diapers=1 → Milk=1', 'Diapers=1 → Bread=1, Milk=1', and 'Bread=1, Milk=1 → Diapers=1'.

The screenshot shows the Orange3 data mining software interface. In the background, a workflow is visible with nodes for 'Data', 'Data Info', 'Rules', 'Data', 'Predictions', 'Bar Plot', and 'Line Plot'. The 'Predictions' node is highlighted, and its output is displayed in the foreground window titled 'Predictions (1) - Orange'.

The 'Predictions (1) - Orange' window displays a table with 16 rows and 7 columns. The columns are: Antecedent, Consequent, Support, Confidence, Coverage, Strength, and Lift. The table contains 16 rules, each with an antecedent, a consequent, and numerical values for Support, Confidence, Coverage, Strength, and Lift. The rules are as follows:

Antecedent	Consequent	Support	Confidence	Coverage	Strength	Lift
Milk=1	Bread=1	1	1	1	1	1
Bread=1	Milk=1	1	1	1	1	1
Diapers=1	Bread=1	1	1	1	1	1
Bread=1	Diapers=1	1	1	1	1	1
Diapers=1	Milk=1	1	1	1	1	1
Milk=1	Diapers=1	1	1	1	1	1
Milk=1, Diapers=1	Bread=1	1	1	1	1	1
Bread=1, Diapers=1	Milk=1	1	1	1	1	1
Diapers=1	Bread=1, Milk=1	1	1	1	1	1
Bread=1, Milk=1	Diapers=1	1	1	1	1	1
Milk=1	Bread=1, Diapers=1	1	1	1	1	1
Bread=1	Milk=1, Diapers=1	1	1	1	1	1
Beer=1	Bread=1	1	1	1	1	1
Bread=1	Beer=1	1	1	1	1	1
Beer=1	Milk=1	1	1	1	1	1
Milk=1	Beer=1	1	1	1	1	1

The table is displayed in a window titled 'Predictions (1) - Orange' with a 'Restore Original Order' button. The window is part of the Orange3 interface, which also shows a workflow in the background and a 'more...' link at the bottom left.

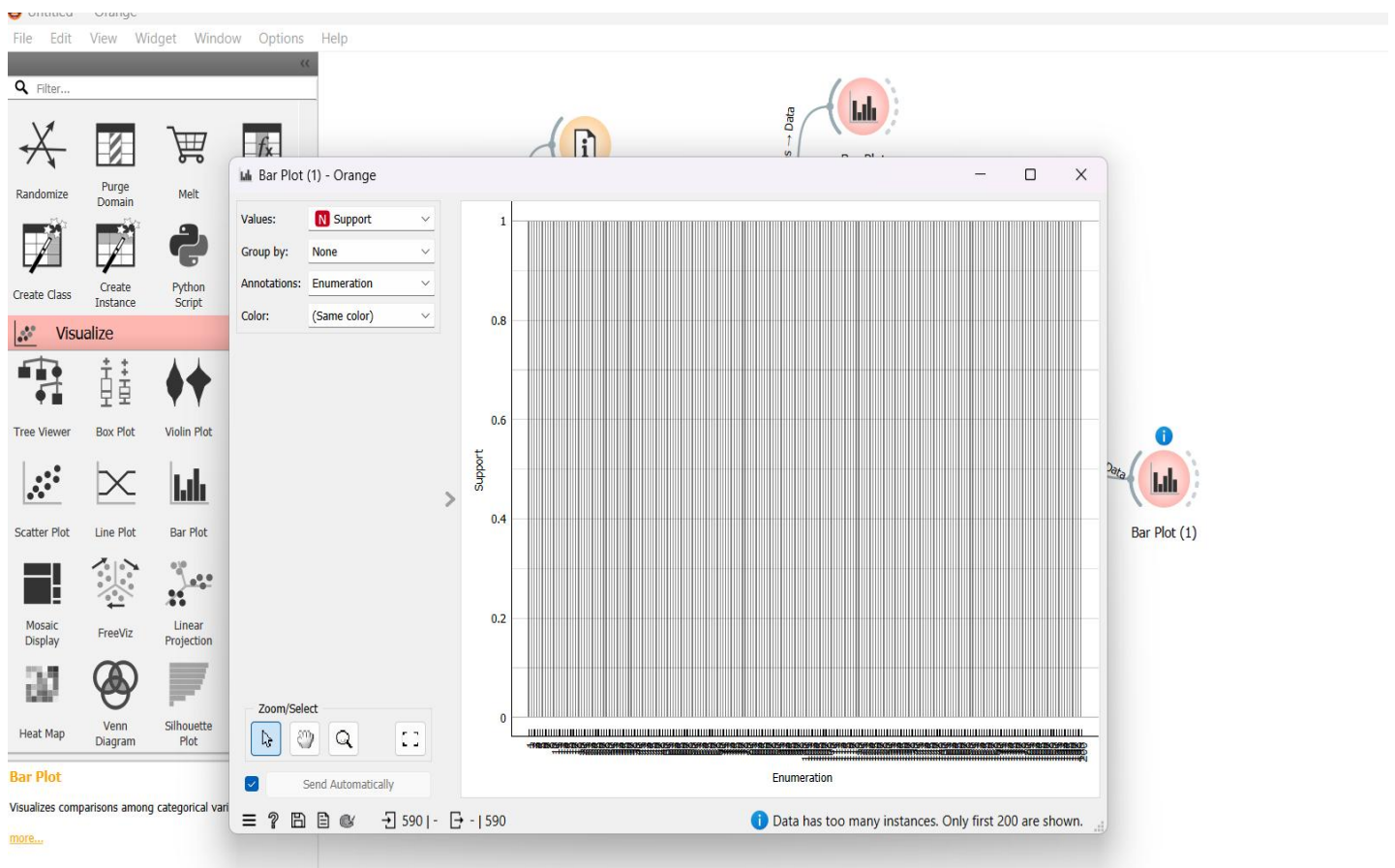


Fig 5.1.2: barplot after preprocessing

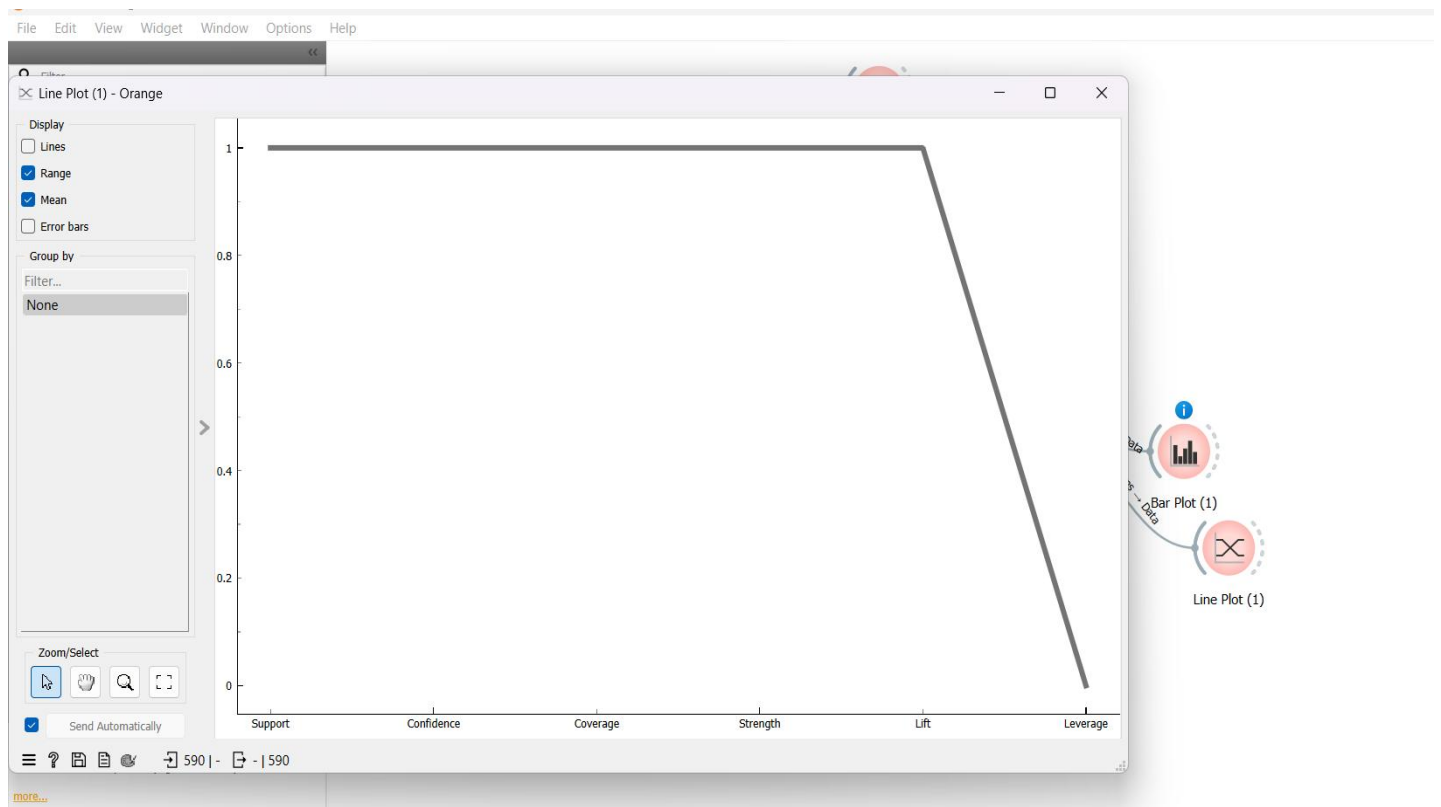


Fig 5.1.3: line plot after preprocessing

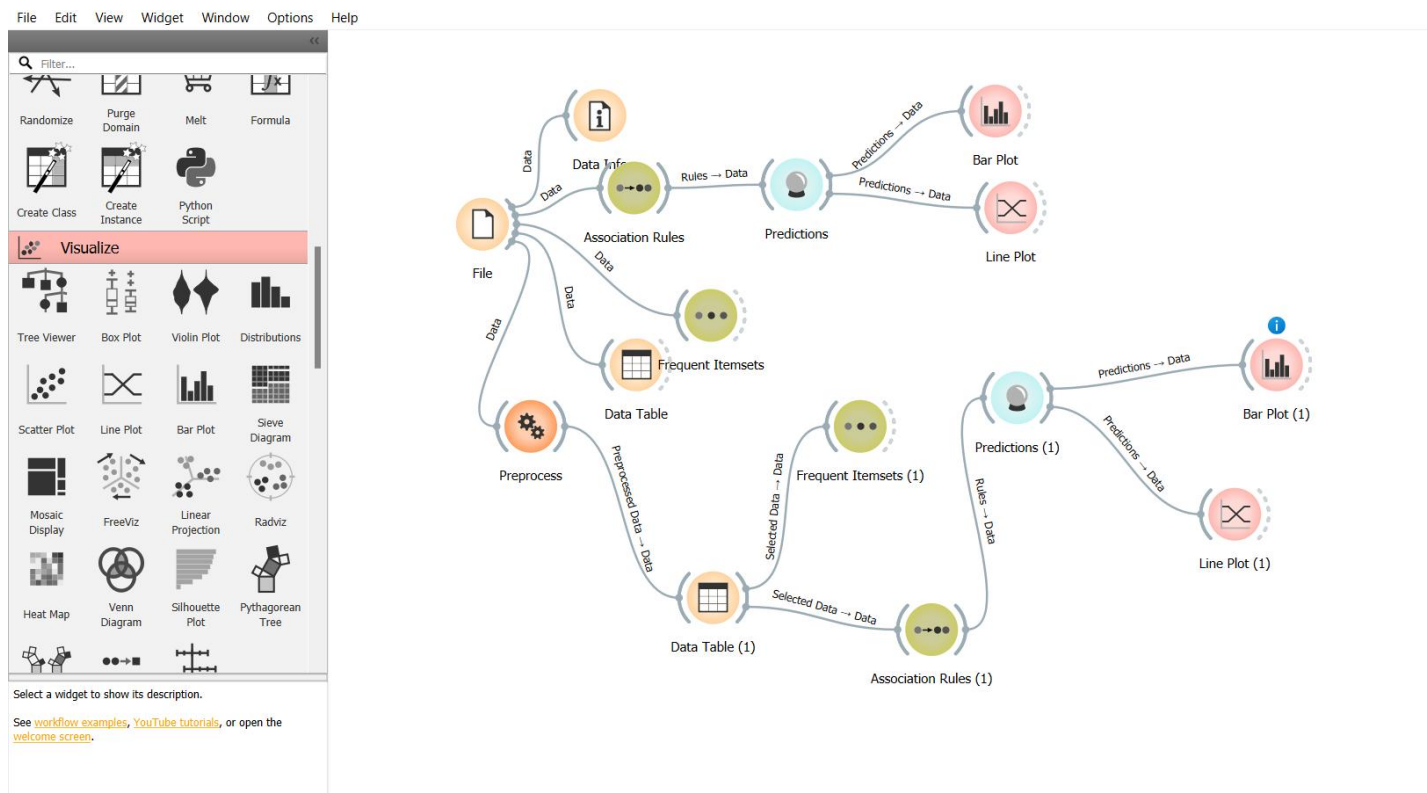


Fig 5.1.4:final view of project

By applying association rules like the Apriori algorithm, you can discover frequent item sets and association rules that indicate which items are often found together.

CHAPTER 4

CONCLUSION

In conclusion, market basket data set using association rules can provide valuable insights into customer purchase patterns and product relationships. By analyzing transaction data and identifying frequent itemsets and association rules, businesses can gain a better understanding of which products are often purchased together, and use this information to optimize their marketing strategies and drive sales.

Market basket analysis using association rules can also help businesses to identify cross-selling opportunities and improve their product offerings. For example, if a particular item is frequently purchased with another item, businesses can bundle these products together or offer targeted promotions to encourage customers to purchase both items.

Overall, market basket data set using association rules is a powerful tool for businesses looking to gain a competitive edge by understanding their customers' buying behavior and preferences. By leveraging the insights provided by association rules, businesses can develop more effective marketing strategies, drive revenue growth, and enhance the overall customer experience.

FUTURE SCOPE:

The current project analyzes the market basket dataset to identify association rules between products. However, future work can focus on incorporating customer demographics, such as age, gender, income, etc., to identify the association rules that are specific to different customer segments.

SESHADRI RAO GUDLAVALLERU ENGINEERING COLLEGE

vdfid(An Autonomous Institute with Permanent Affiliation to JNTUK, Kakinada)

Seshadri Rao Knowledge Village, Gudlavalleru

Department of Computer Science and Engineering

Program Outcomes (POs)

Engineering Graduates will be able to:

- 1. Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
- 2. Problem analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
- 3. Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
- 4. Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions., component, or software to meet the desired needs.
- 5. Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
- 6. The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
- 7. Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
- 8. Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
- 9. Individual and team work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
- 10. Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write

effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Program Specific Outcomes (PSOs)

PSO1 : Design, develop, test and maintain reliable software systems and intelligent systems.

PSO2 : Design and develop web sites, web apps and mobile apps.

PROJECT PROFORMA

Classification of Project	Application	Product	Research	Review
	√			

Note: Tick Appropriate category

Data Mining Outcomes	
Course Outcome (CO1)	Describe fundamentals, and functionalities of data mining system and data preprocessing techniques.
Course Outcome (CO2)	Illustrate the major concepts and operations of multi dimensional data models.
Course Outcome (CO3)	Analyze the performance of association rule mining algorithms for finding frequent item sets from the large databases.
Course Outcome (CO4)	Apply classification algorithms to solve classification problems.
Course Outcome (CO5)	Use clustering methods to create clusters for the given data set.

Mapping Table

CS3509 : DATA MINING															
Course Outcomes	Program Outcomes and Program Specific Outcome														
	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12		PSO 1	PSO 2
CO1	1	1										1			
CO2	1											1			
CO3	2	3	2									2		1	
CO4	2	2	3	2								2		2	
CO5	1	2	3	1								2		1	

Note: Map each Data Mining outcomes with POs and PSOs with either 1 or 2 or 3 based on level of mapping as follows:

1-Slightly (Low) mapped 2-Moderately (Medium) mapped 3-Substantially (High) mapped