**Seminar Report**

On

**“Market Basket Optimization using Apriori Algorithm”**

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE

IN THE PARTIAL FULFILLMENT FOR THE AWARD OF THE DEGREE

OF

**BACHELER OF ENGINEERING**

in

**INFORMATION TECHNOLOGY**

Submitted by

**Aditya Avinash More**

Under the Guidance of

**Prof. Richa Agarwal**





**KJ’S EDUCATIONAL INSTITUTE**

**TRINITY COLLEGE OF ENGINEERING**

**AND RESEARCH**

**Academic Year 2023-2024**



***CERTIFICATE***

This is to certify that the Seminar Report entitled

“**Market Basket Optimization using Apriori** **Algorithm”**

Submitted by

## Aditya Avinash More (IT3051)

Is a record of bonafide work carried out by her/him under supervision and guidance of Prof. Richa Agarwal in partial fulfilment of the requirement for TE (Information Technology) 2019 course of Savitribai Phule Pune University, Pune in the academic year 2023-2024.

Prof. Richa Agarwal Dr. Vilas Gaikwad

Internal guide Head of

Department of IT. Department of IT.

TCOER, Pune.TCOER, Pune.

**ACKNOWLEDGEMENT**

Apart from individual efforts, the success of any seminar depends largely on the encouragement and guidelines of many others. We take this opportunity to express our gratitude to the people who have been instrumental throughout the Seminar work.

It is our privilege to express our gratitude towards our Seminar guide, **Prof. Richa Agarwal** for their valuable guidance, encouragement, inspiration and whole-heartedcooperation throughout the Seminar work. We thank him for being a motivation through all our highs and importantly, our lows.

We deeply express our sincere thanks to our Head of Department **Dr. Vilas Gaikwad** for encouraging and allowing us to present the Seminar on the topic “**Market Basket Optimization using Apriori Algorithm**” and providing us with the necessary facilities to enable us to fulfil our seminar requirements as best as possible. We take this opportunity to thank all faculty members and staff of Department of Information Technology, who have directly or indirectly helped our Seminar.

We pay our respects to honourable Principal **Dr. A. B. Auti** for their encouragement. Our thanks and appreciations also go to our family and friends, who have been a source of encouragement and inspiration throughout the duration of the Seminar.

**ABSTRACT**

Market Basket Optimization is a technique applied by retailers to understand customer’s shopping behaviour from their stores. The result of the effective analysis may improve the supplier’s profitability, quality of service, and customer satisfaction. The purpose of this seminar is to make use of anonymized data on customer’s transactional orders to focus on descriptive analysis on the customer purchase patterns, items which are bought together, and units that are highly purchased from the store to facilitate reordering and maintaining adequate product stock.

Nowadays, in various fields such as retail industries, real estate, banking sectors etc., massive quantities of data are stored in the databases. Therefore, it is very important to derive from the broad chunk of data the valuable piece of data. This method of extracting valuable data from data is referred to as data mining. This data Mining technique is useful for the Market Basket Analysis.

Analysis of customers purchase data can be done by analysing the available data in such a way that frequent itemset can be found and can be analysed to define an association rule. One of the algorithms which help in finding association rule for a frequent itemset and to identify the correlation is the Apriori algorithm. The apriori algorithm works on the principal that every subset of frequent itemset must be frequent. It will find the relationship between the two or more items. Minimum confidence and minimum support values used for mining rules are parameters of the foremost existence.

I

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| Figure | Figure Caption | Page |
| . No. |  | No. |
| 4.1 | Architecture of Market Basket Analysis Model | 8 |
|  |  |  |
| 4.2 | Work Flow of Apriori Algorithm | 9 |
|  |  |  |
| 4.3 | Use Case Diagram | 10 |
|  |  |  |
| 5.1 | The dataset used in seminar | 12 |
|  |  |  |
| 5.2 | Exploratory Data Analysis | 12 |
|  |  |  |
| 5.3 | Dataset after performing Data Cleaning | 13 |
|  |  |  |
| 5.4 | Bargraph for Top 10 Frequently sold products | 13 |
|  |  |  |
| 5.5 | Pie chart for Monthly Purchase of Products | 14 |
|  |  |  |
| 5.6 | Dataset after performing One Hot Encoding | 14 |
|  |  |  |
| 6.1 | Listing the frequent itemsets | 15 |
|  |  |  |
| 6.2 | Finding Association between the items | 15 |
|  |  |  |
| 6.3 | Final result diagram | 16 |
|  |  |  |

II

**ABBREVATIONS**

|  |  |
| --- | --- |
| Acronym | Definition |
|  |  |
| MBA | Market Basket Analysis |
|  |  |
| ARM | Association Rule Mining |
|  |  |
| ML | Machine Learning |
|  |  |

III

**TABLE OF CONTENTS**

Topics Page No.

* Abstract
* List of Figures
* Abbreviations

|  |  |
| --- | --- |
| Sr. No. | Table of Contents |
| Chapter 1 | Introduction……………………………………………... |
| 1.1 | Background |
| 1.2 | Importance of Seminar |
| 1.3 | Motivation |
| 1.4 | Scope |
| 1.5 | Expected Outcomes |

I

II

III

Page No.

1

1

1

1

2

2

|  |  |  |
| --- | --- | --- |
| Chapter 2 | Literature Survey………………………………………… |  |
| 2.1 | Market Basket Analysis using Apriori & FP growth algorithm | 3 |
| 2.2 | Comparative analysis of Market Basket Analysis through Data | 4 |
|  | Mining Techniques |  |

|  |  |  |
| --- | --- | --- |
| Chapter 3 Related Theory and Problem Definition…….…………… | |  |
| 3.1 | Problem Definition | 5 |
| 3.2 | Related Theory | 5 |
| Chapter 4 | Design Methodology………………………...................... |  |
| 4.1 | Proposed System Architecture | 8 |
| 4.2 | Internal Logic of System | 8 |
| 4.3 | Data Flow and UML diagrams | 9 |
| 4.4 | Technical Specifications | 10 |
| Chapter 5 | Implementation…………………………………………… |  |
| 5.1 | Implementation of proposed System | 11 |

|  |  |  |
| --- | --- | --- |
| 5.2 | System Flow | 11 |
| Chapter 6 Result and Discussion…………………………………… | |  |
| 6.1 | Result | 15 |
| 6.2 | Discussion |  |
|  |  | 16 |
| Chapter 7 Conclusion and Future Scope……………………………. | |  |
| 7.1 | Conclusion | 17 |
| 7.2 | Future Scope | 17 |
|  | References………………………………………………… | 18 |

**Chapter 1**

**INTRODUCTION**

**1.1 Background:**

Market Basket analysis is a data mining method focusing on discovering purchase patterns of the customers by extracting association or co-occurrences from a store’s transactional data. For example, when the person checkout items in a supermarket all the details about their purchase goes into the transaction database. Later, this huge data of many customers is analysed to determine the purchasing pattern of customers. Also, decisions like which item to stock more, cross selling, up selling, store shelf arrangement are determined. Association rule mining (ARM) identifies the association or relationship between a large set of data items and forms the base for market basket analysis. Association rule mining has been widely used in various industries besides supermarkets, such as mail order, telemarketing production, fraud detection of credit card and e-commerce. One of the challenges for companies that have invested heavily in customer data collection is how to extract important information from their vast customer databases and product feature databases, in order to gain competitive advantage. Market basket analysis has been intensively used in many companies as a means to discover product associations. A retailer must know the needs of customers and adapt to them. Market basket analysis is one possible way to find out which items can be put together.

**1.2 Importance of the seminar:**

The recent advancements in the field of data analytics have opened up a world of new opportunities for players in the food and beverage sector. Businesses capable of identifying these opportunities have succeeded in improving sales and operational efficiency. Market basket analysis can help food industry players to drive profitable growth by analysing the purchase patterns of its customers.

It can be used effectively to increase the overall spending from the customer by bundling frequently purchased items at a discounted price. we believe that leveraging market basket analysis can help food and beverage companies to gain a front-liner advantage in today’s complex business world

**1.3 Motivation:**

* Market basket analysis is a technique that allow to discover the relationships between products.
* It is based on searching for the most common combination of products in each transaction.
* In our analysis we use the Apriori algorithm and the main goal is to sorting the products and creating the best relationship.

1

**1.4 Scope:**

* Cross Selling
* Customer Behaviour
* Fraud Detection
* Product Placement
* Deliver online marketing Campaigns

**1.5 Expected Outcome:**

* Right store identification: Identification of the right point-of-sale or store to drive increased footfalls, run personalized promotions, boost sales and maximize profits.
* Focused market strategies: Creation of customer segment based on their buying behaviors so as to improve ROI by precisely hyper-targeting the market.
* Improved selling opportunities: Prediction of sales of products at right time and place to the right customer and driving campaigns to increase cross-sell and upsell possibilities.
* Better SKU placements: Strategic placement of a product SKU next to the associated popular selling product SKUs to improve visibility and sales through products ‘bundling’.

2

**Chapter 2**

**LITERATURE SURVEY**

**2.1** **JI HYUNG YOON AND BEAKCHEOL JANG, (Member, IEEE), “Evolution of Deep Learning-Based Sequential Recommender Systems: From Current Trends to New Perspectives”, National Research Foundation of Korea Fund under Grant NRF2022R1F1A1063961, 1-7 June 2023.**

The literature survey identifies the limitations faced by earlier recommender systems in providing personalized recommendations. It introduces the evolution of deep learning-based sequential recommender systems, emphasizing the role of Graph Neural Networks (GNNs) and Generative Adversarial Networks (GANs) in enhancing recommendation models. The paper examines various techniques, such as Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), Graph Neural Networks, Generative Adversarial Networks, and Self-Supervised Learning (SSL), highlighting their effectiveness in capturing complex user-item interactions and improving recommendation diversity.

The implementation section outlines the data processing process in deep learning-based recommendation systems, emphasizing the importance of optimizing the model update process through techniques like checkpoint broadcasts and localized machine learning. It also presents a comprehensive system diagram showcasing the evolution of recommender models and the significance of considering temporal factors in user-item interactions.

The results section discusses the algorithms and features of popular deep learning-based models, emphasizing their ability to capture dependency patterns and their integration of various data forms for model learning. The survey also mentions the growing integration of natural language processing techniques in recommendation systems, indicating the potential of deep learning-based sequential recommender systems.

The conclusion of the paper, however, is not explicitly stated in the provided content. It points to the need for further research and development in the field without specifying particular conclusions or future directions.

3

**2.2 RAKESH GUPTA AND MANOJ KUMAR TRIVEDI, “AEHO: Apriori-Based Optimized Model for Building Construction to Time-Cost Tradeoff Modeling”, Department of Civil Engineering, MITS, Gwalior, Madhya Pradesh 474005, India, 14-22 Sep, 2021**

This literature review focuses on the AEHO model, an Apriori-based optimized approach for addressing time-cost trade-off issues in building construction projects. The paper highlights the challenges faced in balancing project duration and cost in the construction industry and explores various optimization techniques and models used for informed decision-making by construction managers.

The study aims to develop a model that can assist construction managers in making optimal decisions to complete projects within the designated timeline and budget. The AEHO model utilizes a combination of the Apriori algorithm and swarm intelligence to generate non-dominated solutions for the time-cost tradeoff. It incorporates techniques such as Ant Colony Optimization Algorithm, Support Vector Machines, Simulated Annealing with Fuzzy Logic, Genetic Algorithm, Integer Linear Programming, and Multiple Attribute Utility Function Theory to optimize decision-making processes.

The implementation of the AEHO model involves the integration of discrete-event simulation with optimization techniques to minimize computational costs and achieve optimal project outcomes. The results demonstrate the effectiveness of the AEHO methodology in evaluating project management issues and providing better solutions for time-cost tradeoff in construction projects. The paper concludes by emphasizing the significance of optimization strategies and methodologies in achieving desirable results in construction projects and highlights the potential of other techniques such as Support Vector Machines and Particle Swarm Optimization in addressing time-cost trade-off analysis. 4

**2.3 Qiuqiang Lin and Chuanhou Gao, Senior Member, IEEE, “Discovering Categorical Main and Interaction Effects Based on Association Rule Mining”, The School of Mathematical Sciences, Zhejiang University, Hangzhou 310027, China. E-mail: {11735034, gaochou}@zju.edu.cn, 2 Feb2023.**

This literature survey delves into the challenge of efficiently discovering categorical main and interaction effects in association rule mining. It addresses several critical problems faced in traditional methods, such as efficiency and complexity, computational burden, and interpretability. The primary problem statement revolves around the computationally expensive nature of fitting models with numerous features and interactions, leading to the need for feature selection.

To address these issues, the survey introduces the EARAF (Efficient Association Rule-based Feature Selection) algorithm, inspired by association rule mining techniques. EARAF efficiently mines relevant main and interaction effects from data and enhances computational complexity through methods like random sampling and min-heap data structures. The algorithm is applicable to both categorical and numerical features and has demonstrated effectiveness in various experiments.

The implementation sections provide a detailed overview of the algorithm and its computational complexities, suggesting parameter selection strategies based on theoretical analysis. The paper also discusses possible extensions of the proposed algorithm and presents the results of experiments, showcasing the algorithm's effectiveness.

Furthermore, the survey highlights key findings from using EARAF in real-world scenarios. It effectively identifies common patterns in datasets, boosts the performance of existing models, and overcomes tricky data challenges with different approaches for various datasets.

In the context of related work, the survey references other algorithms and models that have achieved state-of-the-art results in sequential recommendation, including SAS rec, BERT4Rec, Transformers4Rec, SSE-PT, RecGAN, and Self-attentive multi-adversarial network.

In conclusion, the EARAF algorithm offers a promising solution to the challenges associated with feature selection and interaction effects in association rule mining. Its principles and experimental results demonstrate its effectiveness in improving classification tasks.

5

**2.4 Maggi Bansal, Inderveer Chana, and Siobhan Clarke, “UrbanEnQoSPlace: A Deep Reinforcement Learning Model for Service Placement of Real-Time Smart City IoT Applications”, Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, Punjab 147004, India, 4 July, 2023.**

This literature survey explores the development of the UrbanEnQoSPlace model, which addresses the challenge of optimizing the placement of real-time smart city IoT applications. The initial problem revolved around the poor performance of Policy Gradient (PG) algorithms in handling the stochastic optimization decision problem with discrete action-space in the UrbanEnQoSMDP. Consequently, the chosen PG algorithms underperformed in comparison to Value-Based (VB) algorithms.

The problem statement emphasizes the limitations of PG algorithms and their unsuitability for the complex UrbanEnQoSMDP, prompting the need for an alternative approach.

The survey discusses the novel solution employed, utilizing Deep Reinforcement Learning (DRL), specifically the Dueling Deep Q-Network (DQN) algorithm. The DQN algorithm was customized with an "ε-greedy with mask" policy to handle discrete placement actions for services in the placement problem efficiently. This approach enabled the model to optimize the placement of all application services simultaneously within a single decision step. The UrbanEnQoSPlace model outperformed other DRL algorithms, achieving higher rewards, stable training, and reduced exploration.

In terms of implementation, the survey details the utilization of the UrbanEnQoSPlace model as a multi-action DRL model based on Dueling DQN. It emphasizes the model's role in optimizing service placement within the 'Urban IoT-Federated MEC-Cloud' architecture, factoring in energy and Quality-of-Service (QoS) requirements. The model's ability to generate multiple actions for each service within a single decision step is highlighted, aiming to achieve the highest reward while minimizing constraint violations.

The survey presents compelling results, showcasing the UrbanEnQoSPlace model's superiority over other DRL models, achieving optimal placement of application services with the highest rewards and minimal constraint violations. The model outperforms DQN and Double DQN models, demonstrating efficiency in its policy and constraint violation management.

The conclusion section highlights key insights:

1. The complexity of applications and the associated latency and energy consumption issues were addressed through a dueling approach, enhancing performance and stability.
2. The impact of varying the number of placement servers on rewards, with an emphasis on resource-intensive applications.
3. The benefits of the model when dealing with an increased number of applications due to the state-value separation mechanism, reducing unnecessary exploration for less valuable states.

6

**2.5 Monerah M. Alawadh and Ahmad M. Barnawi, “A Survey on Methods and Applications of Intelligent Market Basket Analysis Based on Association Rule”, King Abdulaziz University, Jeddhah, 21589, The Kingdom of Saudi Arabia corresponding Author: Monerah M. Alawadh, 23 Nov, 2021.**

This literature survey explores the challenges and solutions in the domain of market basket analysis (MBA) based on association rule learning. It addresses various issues encountered in this field, including the exponential growth of data, data transformation and management, the need for real-time analysis, and the expanding application of MBA beyond traditional supermarket stores.

The problem statement emphasizes the need for efficient algorithms for MBA across different computing environments. The document discusses and evaluates various algorithms, such as Apriori, FP-Growth, Eclat, Clique, genetic algorithms, and artificial neural networks. It explores their strengths and weaknesses in diverse computing settings, providing insights into the role of these algorithms in predictive modeling and association rule learning.

The survey highlights the following methods and techniques used to tackle these challenges:

1. Apriori Algorithm: Widely used in MBA, it identifies frequent itemsets and generates association rules based on support and confidence.
2. Swarm Intelligence: Utilizes techniques like particle swarm optimization and ant colony optimization for clustering big data and post-processing capabilities.
3. Neural Networks: Increasingly employed for handling large datasets and gaining insights into consumer behavior.
4. Clustering-based Approaches: Focus on generating appropriate intervals using density variance or dense regions to identify meaningful associations.
5. Fuzzy-based Approaches: Utilize fuzzy logic and linguistic terminologies to define partitions for association rule learning, addressing optimization challenges.

The survey delves into practical implementations of these techniques, including predicting customer purchasing behavior, conducting market-based surveys, analyzing consumer demand, optimizing product positioning, and implementing successful promotions. It also highlights the potential for consumer segmentation based on purchasing behavior.

Key findings in the results section encompass the benefits of association rule learning techniques in MBA for retailers and supermarket operators. These techniques assist in predicting customer purchase behavior, conducting surveys, understanding consumer demand, optimizing product placement, identifying effective promotions, and segmenting consumers. Additionally, association rule learning can optimize sales by determining the right placement of products on store shelves, resulting in improved customer satisfaction.

The effectiveness of clustering-based and fuzzy-based approaches is highlighted, along with the increasing popularity of MBA research across various industries such as healthcare, retail, hospitality, and sports. Furthermore, the utilization of neural networks and intelligent data mining techniques marks a potential paradigm shift in market basket analysis. The document suggests that association rule learning has a promising future, as evidenced by the growing number of research papers dedicated to this field.Top of Form

7

**Chapter 3**

**RELATED THEORY AND PROBLEM DEFINITION**

**3.1 Problem Definition:**

Nowadays people buy daily goods from online e-commerce website where we can purchase products from nearby grocery shops. There are many supermarkets that provide goods to their customer. The problem many retailers face is the placement of the items. They are unaware of the purchasing habits of the customer so they don’t know which items should be placed together in their store. With the help of this analysis shop managers can determine the strong relationships between the items which ultimately helps them to put products that co-occur together close to one another. Also, decisions like which item to stock more, cross selling, up selling, store shelf arrangement are determined

**3.2 Related Theory:**

* **Machine learning:** Machine learning is a method of teaching prediction based on some data. it is a branch of artificial intelligence. Which numerically improves on data over as more data as add in algorithm the performance of the system is improved. These are the three types of machine learning:

1. Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
2. Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).
3. Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). As it navigates its problem space, the program is provided feedback that’s analogous to rewards, which it tries to maximize.

* **Why Python?:**

Machine learning is the study area that helps computers to learn without being specifically programmed. Now a days, python libraries, frameworks and modules are very simple and powerful as compared with the old days. Python has replaced many of the industry's languages, one of the reasons for this is its vast collection of libraries, and is one of the most popular programming languages for this task today.

8

* **Python libraries used in the seminar are:**

**Pandas:** Pandas is an open source, Python licensed library that offers high-performance, easy-to-use data structures, and data analysis tools to the Python programming language. The Data Frame is the core data structure. Data frame allows tabular data to be stored and manipulated in observation rows and variable columns. A broad variety of stored data types is available, such as CSVs, TSV's (Tab separated values), JSONs (Hypertext Mark-Up Language), and more. Pandas can read different types. A Data Frame consists of both a row and a column index, a two-dimensional set of values. A series is a special collection of index values.

**NumPy**: NumPy is an array-processing application for general purposes. It stands for 'Numerical

Python'. It is a library of multidimensional array objects, and a set of array processing routines.

NumPy has functions built in for linear algebra and the generation of random numbers.

**Matplotlib:** Is the art of displaying data through charts, icons, presentations and more. It is most common to translate complex data for a non-technical audience into comprehensible insights. Matplotlib is one of the most powerful Data Visualization Python packages used. This is a cross

* platform framework designed to make Two dimensional graphs from records in arrays. This also provides an object-oriented API which helps, for example, to embed plots into implementations using Python GUI toolkits such as PyQt.

**Seaborn:** Seaborn is an enhancement to matplotlib and not a substitution for it. The reason for this is that it is placed on top of matplotlib and you will often explicitly invoke matplotlib functions to draw simpler plots already available through the namespace pyplot. Matplotlib is completely scalable but it can be difficult to know what settings to change to achieve an appealing plot. Seaborn comes with a number of custom themes to track the matplotlib look and a high-level user interface. It is closely integrated with the PyData0 stack, including support of SciPy and stats models data structures for NumPy and Pandas, and statistical routines.

* **One Hot Encoding:**

What is categorical Data?

Categorical data are variables that contain label values rather than numeric values. why should we convert it to numerical values in the first place? Many machine learning algorithms cannot operate on label data directly. They require all input variables and output variables to be numeric.

How to convert categorical data to numerical data?

One way to do so is One-Hot-Encoding. One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do prediction. One hot encoding makes our training data more useful and expressive, and it can be rescaled easily. By using numeric values, we more easily determine a probability for our values.

One Hot Encoding:

One hot encoding is one method of converting data to prepare it for an algorithm and get a better prediction. With one-hot, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns.

9

* **Association rule:** Association rule mining is a technique to identify frequent patterns and associations among a set of items. For instance, if item A and B are bought together more frequently then several steps can be taken to increase the profit. For example:

1. A and B can be placed together so that when a customer buy one of the product, he doesn't have to go far away to buy the other product.
2. People who buy one of the products can be targeted through an advertisement campaign to buy the other.
3. Collective discounts can be offered on these products if the customer buys both of them.
   1. Both A and B can be packaged together

* **Apriori Algorithm:** To implement association rule mining, many algorithms have been developed. Apriori algorithm is one of the most popular and arguably the most efficient algorithms among them. The key concept in the Apriori algorithm is that it assumes all subsets of a frequent itemset to be frequent.

Basic idea behind this algorithm is,

1. Only a large item set can be an item set if all its subsets are large item sets.
2. It is possible to accept collections of products that have minimal support.
3. From frequent item sets, association rules can be created

There are three major components of Apriori algorithm:

1. **Support:**

Support refers to the default popularity of an item and can be calculated by finding number of transactions containing a particular item divided by total number of transactions. Suppose we want to find support for item B.

This can be calculated as: Support(B) = (Transactions containing (B))/ (Total Transactions)

1. **Confidence:**

Confidence refers to the likelihood that an item B is also bought if item A is bought. It can be calculated by finding the number of transactions where A and B are bought together, divided by total number of transactions where A is bought. Mathematically, it can be represented as:

Confidence(A→B) = (Transactions containing both (A andB)) / (Transactions containing A)

1. **Lift:**

Lift (A -> B) refers to the increase in the ratio of sale of B when A is sold. Lift (A –> B) can be calculated by dividing Confidence (A -> B) divided by Support(B).

Mathematically it can be represented as: Lift(A→B) = (Confidence (A→B)) / (Support (B))

10

**Chapter 4**

**DESIGN METHODOLOGY**

**4.1 Proposed System Architecture**

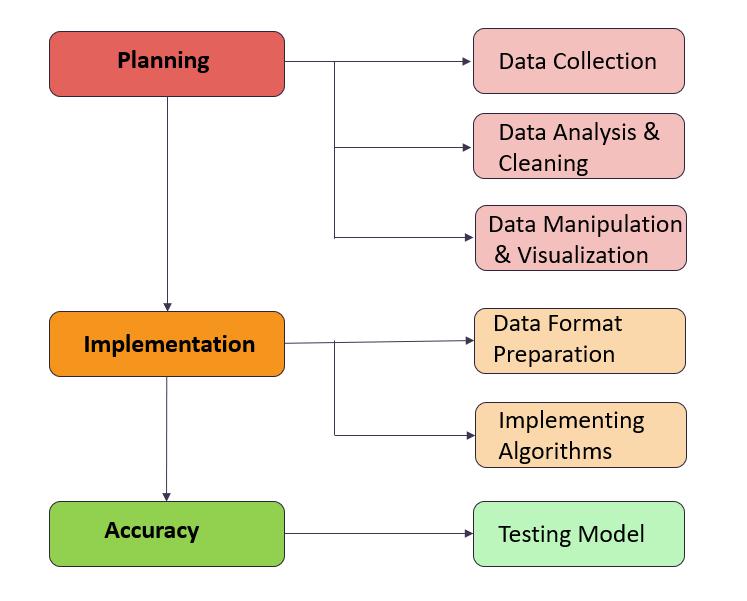


Figure 4.1 Architecture of Market Basket Analysis Model

**4.2 Internal Logic of System**

Market Basket Analysis using machine learning Model can be divided into following modules:

* Collection of data from the grocery store.
* Performing the data analysis and cleaning the data.
* Finding the association between the different items.
* Implementation of Apriori Algorithm and finding the frequent products.

First user has to provide some data as input. After that using Association rule our model will find the frequent itemsets. This prediction is based on various support, confidence and lift count. User will get the frequent products as an output.

11

**4.3 Data Flow and UML Diagrams**

* **How the Apriori Algorithm Works?**

Apriori is an [algorithm](https://en.wikipedia.org/wiki/Algorithm) for frequent item set mining and [association rule learning](https://en.wikipedia.org/wiki/Association_rule_learning) over relational [databases.](https://en.wikipedia.org/wiki/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. The frequent item sets determined by Apriori can be used to determine [association](https://en.wikipedia.org/wiki/Association_rules) [rules](https://en.wikipedia.org/wiki/Association_rules) which highlight general trends in the [database:](https://en.wikipedia.org/wiki/Database) this has applications in domains such as [market basket analysis.](https://en.wikipedia.org/wiki/Market_basket_analysis)

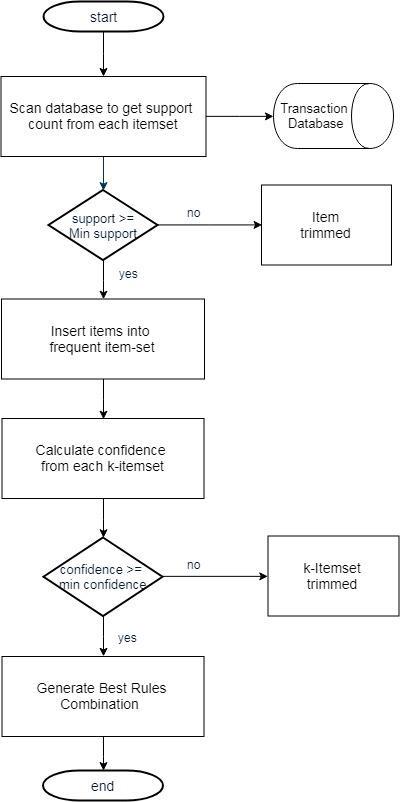


Figure 4.2 Work Flow of Apriori Algorithm

12

* **UML Diagram:**

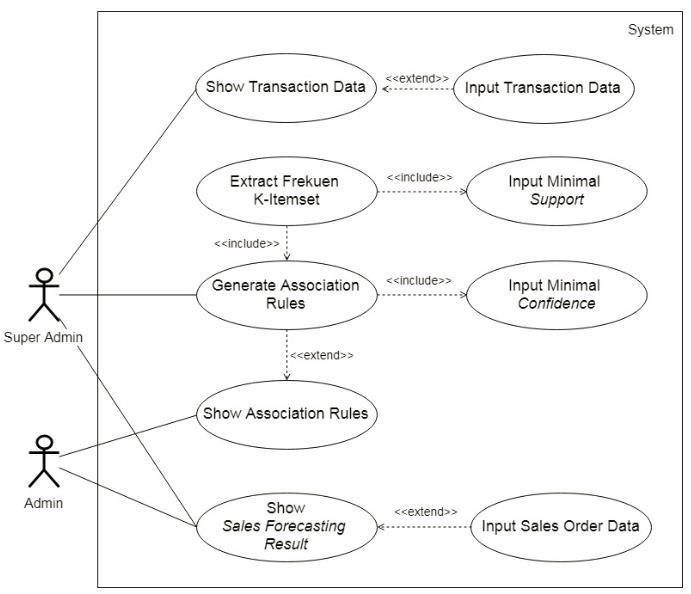


Figure 4.3 Use Case Diagram

**4.4 Technical Specifications**

**Hardware Requirement:**

* + PC Version: Intel® Core™ i5 x64 based
* RAM: 4GB and above
* HDD: 500 GB
* Processor: Intel® Core™ i5

**Software Requirement:**

* Python
* Windows 10
* Browser (Google Chrome)
* Code editor (Google colab Notebook)

13

**Chapter 5**

**IMPLEMENTATION**

**5.1 Implementation of Proposed System:**

* **Python:**

Python is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale seminars.

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library

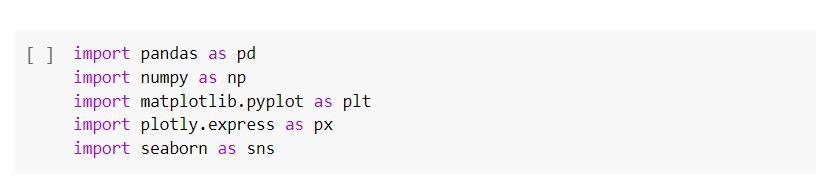
* **Association Rule Mining:**

Association rule learning is a [rule-based machine learning](https://en.wikipedia.org/wiki/Rule-based_machine_learning) method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.

**5.2 System Flow**

Market Basket Optimization using Machine Learning Model can be divided into following steps:

**Step 1: Importing the required python libraries**



**Step 2: The dataset**

The dataset used in this seminar was downloaded from Kaggle.com

14

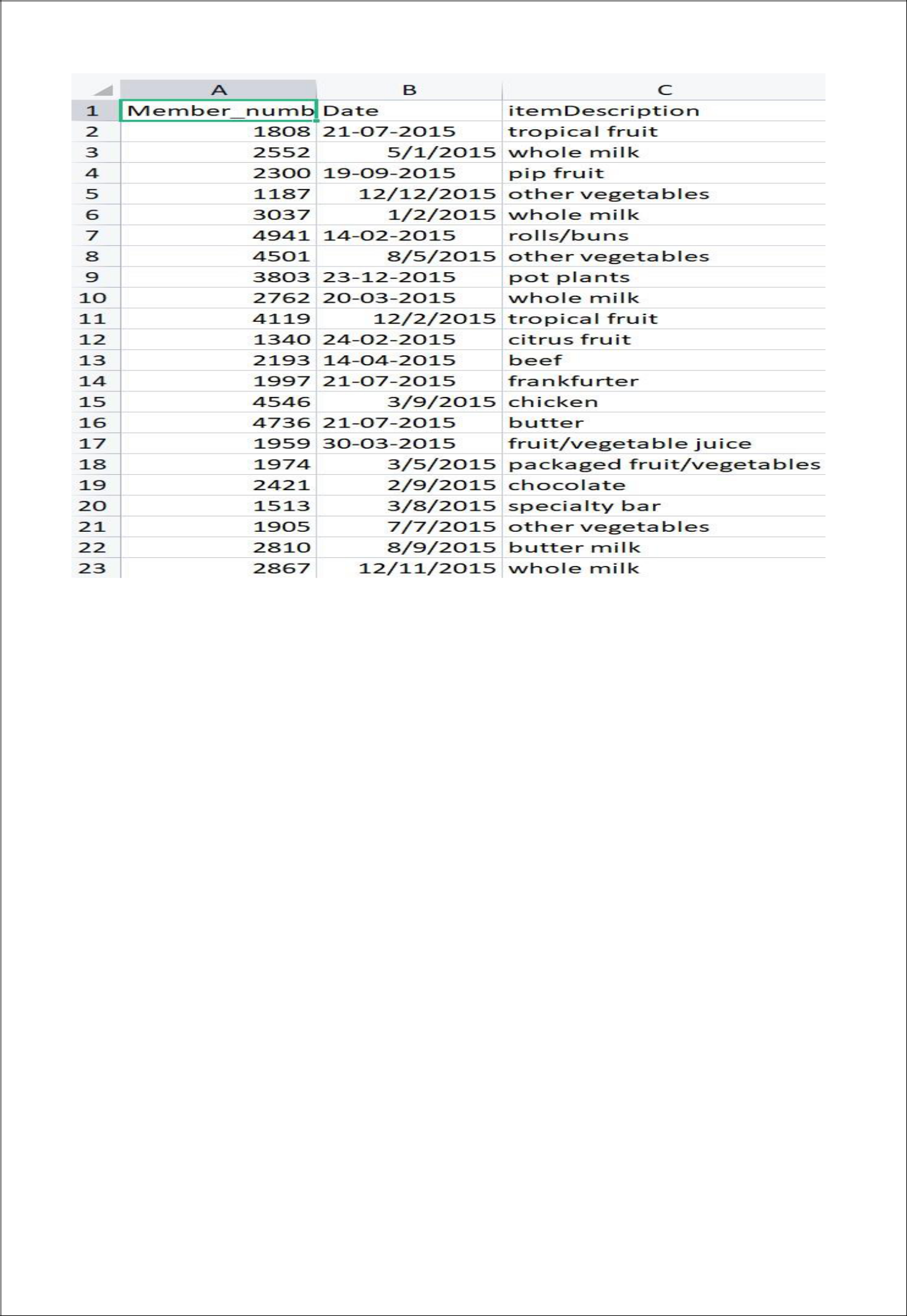


Figure 5.1The dataset used in seminar

**Step 3: Data Analysis**

**3.1) Exploratory Data Analysis:** From the Exploratory Data Analysis we have analysed:

1. The dataset is 2 Dimensional having 3 columns & 38765 Rows.
2. Total count of elements is 38765
3. There are no any NAN values in the dataset
4. There is no duplicated transaction of products

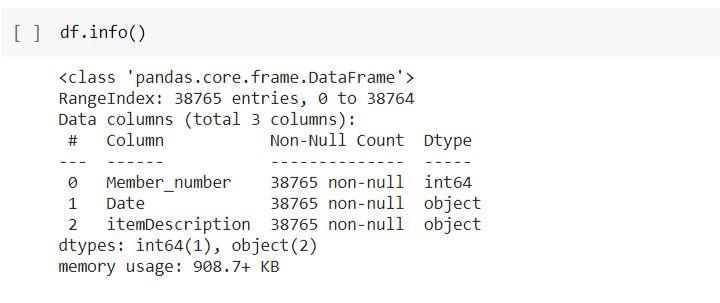


Figure 5.2 Exploratory Data Analysis

15

**3.2) Data Cleaning:**



Figure 5.3 Dataset after performing Data Cleaning

**3.3) Data Visualisation:** Visualisation is used for the better understanding of dataset

* Frequently Bought Products

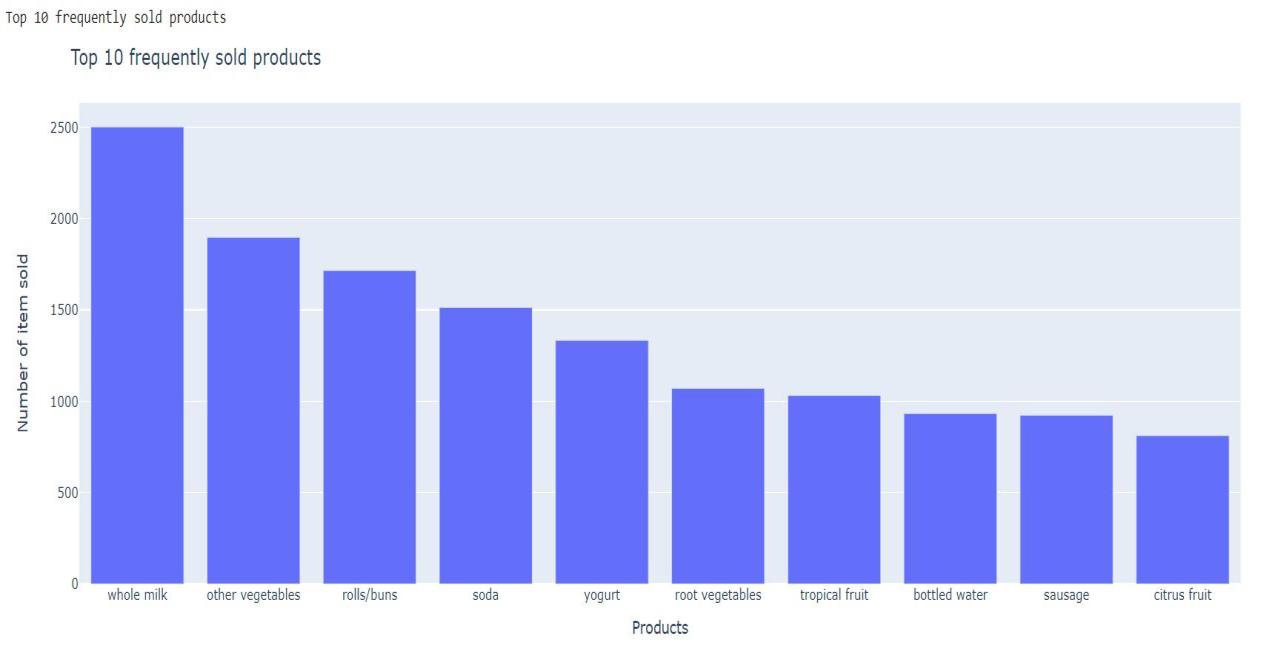


Figure 5.4 Bargraph for Top 10 Frequently sold products

16

* Monthly purchase of Products:

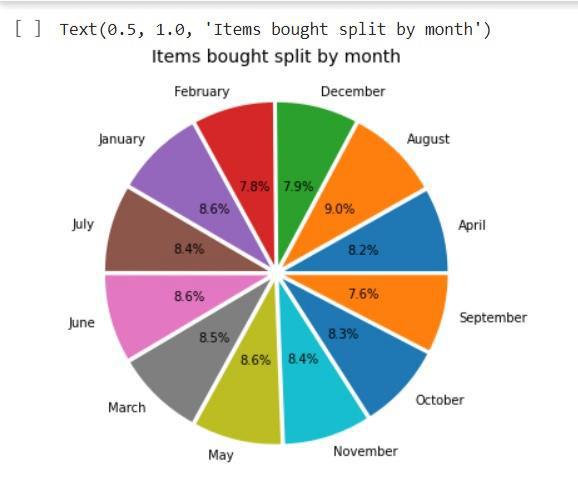


Figure 5.5 pie chart for Monthly Purchase of Products

**Step 4: Implementation of Apriori Algorithm**

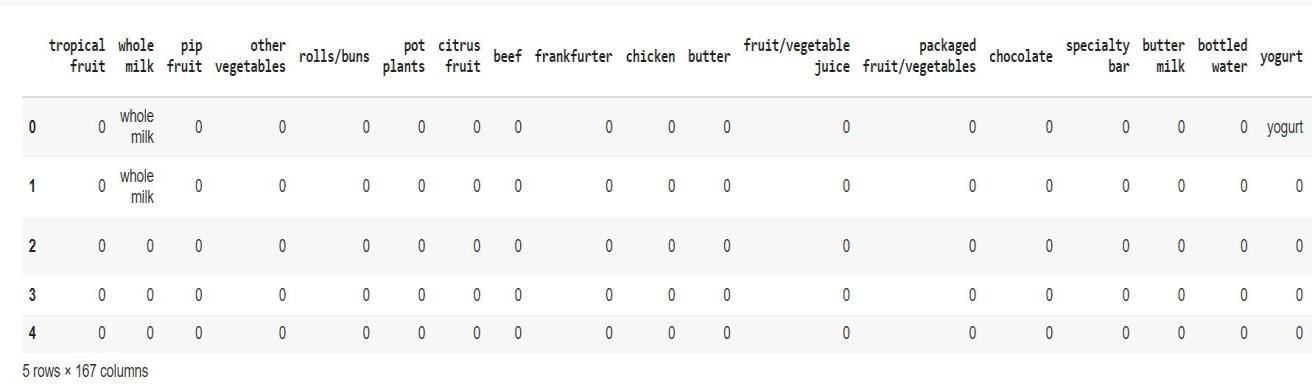


Figure 5.6 Dataset after performing One Hot Encoding

17

**Chapter 6**

**RESULT AND DISCUSSION**

**6.1 Result**

**The proposed system produces following outputs:**

* Creating the Association rules between the products
* Finding Frequent Items in the dataset

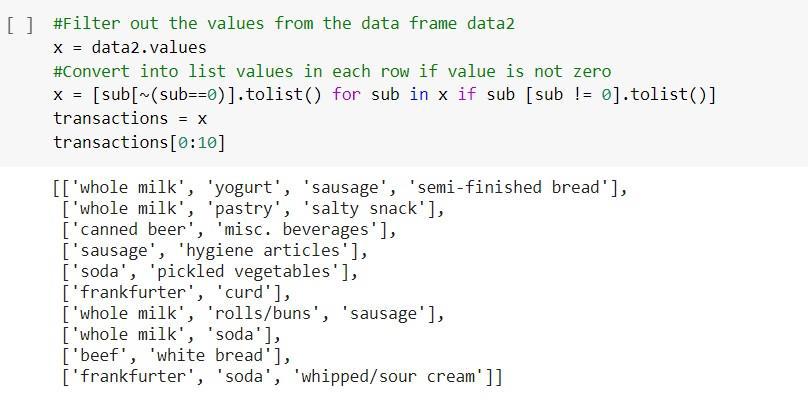


Figure 6.1 Listing the frequent items

* Creating the Association rules between the products

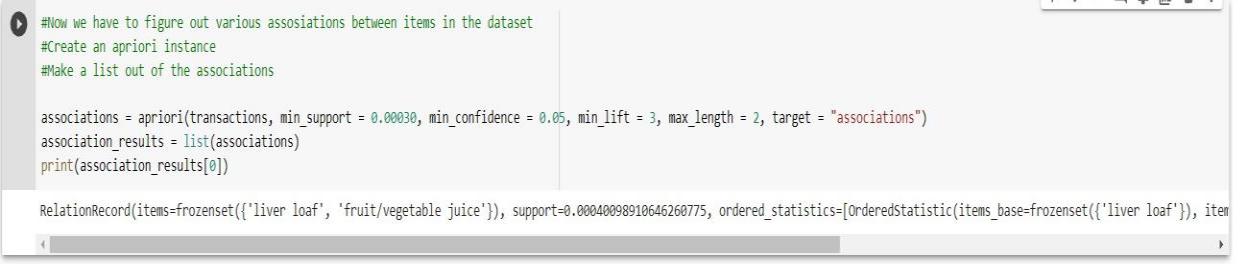
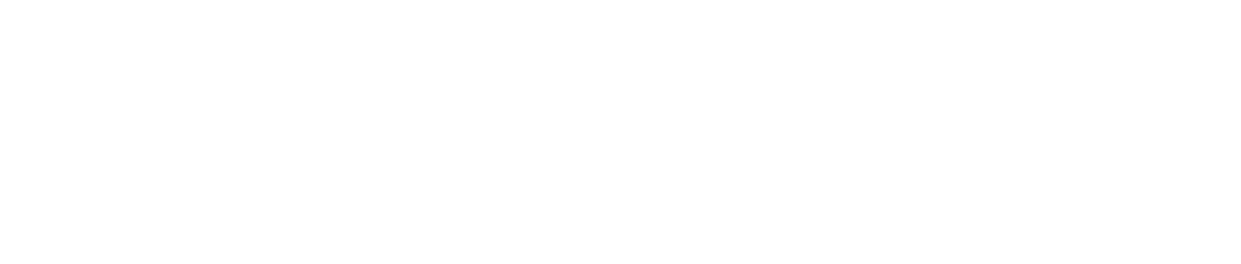


Figure 6.2 Finding Association between the items

18

Final Result:



Figure 6.3 Final result

**6.2 Discussion:**

This model is based on the machine learning algorithms and we were trying to find the relation between two or more products which are provided in the dataset at Kaggle. To predict the frequent itemset we used machine learning algorithms i.e., Apriori Algorithm. The Apriori algorithm effectively generates highly informative frequent itemsets and association rules for the data of the Grocery store with the help of support, and confidence. The frequent data items are generated from the given input data and based on the frequent item sets strong association rules were generated. From this we can conclude that the Apriori Algorithm is efficient algorithm for data mining irrespective of the size of data.

19

**Chapter 7**

**CONCLUSION AND FUTURE SCOPE**

**7.1 Conclusion:**

Market Basket Analysis is a conceptual framework that originates in the marketing field and has been used effectively in fields such as bioinformatics, nuclear science, immunology, and geophysics more recently. One reason for MBA's increasing adoption across scientific fields is that by using an inductive approach to theorizing, researchers are able to evaluate the existence of association rules. Considering all, by summing up the whole, we agree that recommendation system can have an efficient impact on marketing and sales research that can be used to make strategic business decisions.

The Apriori algorithm effectively generates highly informative frequent itemsets and association rules for the data of the Grocery store. The frequent data items are generated from the given input data and based on the frequent item sets strong association rules were generated.

**7.2 Future Scope:**

Seminar can be improved by implementing new and advanced mining algorithms along with apriori algorithm for better performance and fast results for sparse dataset. In the current approach, we only use association rules to exploit the collective information i.e., building a model by finding similarity between customers’ products associations. In future work, association rules can also be used to exploit the content-based information i.e., finding a similarity between products, and recommending a product based on interest of a similar products. Content based recommendation system is not based on a lot of user data since the calculation of similarities takes place at the product level. Perhaps we can build recommendation system in future work, incorporating the two approaches into a hybrid approach that can benefit from the strengths of both item-based and customer-based approaches.

This application can be extended to other areas such as: sales tracking, product tracking, discount and calculation of prices etc. This method can be applied in future to very large databases where memory space is valuable and needs enhancement. It can be further tuned for improved efficiency and performance.

20

**REFERENCES**

[1] JI HYUNG YOON AND BEAKCHEOL JANG, (Member, IEEE), “Evolution of Deep Learning-Based Sequential Recommender Systems: From Current Trends to New Perspectives”, National Research Foundation of Korea Fund under Grant NRF2022R1F1A1063961, 1-7 June 2023.

[2] RAKESH GUPTA AND MANOJ KUMAR TRIVEDI, “AEHO: Apriori-Based Optimized Model for Building Construction to Time-Cost Tradeoff Modeling”, Department of Civil Engineering, MITS, Gwalior, Madhya Pradesh 474005, India, 14-22 Sep, 2021

[3] Qiuqiang Lin and Chuanhou Gao, Senior Member, IEEE, “Discovering Categorical Main and Interaction Effects Based on Association Rule Mining”, The School of Mathematical Sciences, Zhejiang University, Hangzhou 310027, China. E-mail: {11735034, gaochou}@zju.edu.cn, 2 Feb2023.

[4] Maggi Bansal, Inderveer Chana, and Siobhan Clarke, “UrbanEnQoSPlace: A Deep Reinforcement Learning Model for Service Placement of Real-Time Smart City IoT Applications”, Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, Punjab 147004, India, 4 July, 2023.

[5] Monerah M. Alawadh and Ahmad M. Barnawi, “A Survey on Methods and Applications of Intelligent Market Basket Analysis Based on Association Rule”, King Abdulaziz University, Jeddhah, 21589, The Kingdom of Saudi Arabia corresponding Author: Monerah M. Alawadh, 23 Nov, 2021.

[6] Maliha Hossain, A H M Sarowar Sattar, Mahit Kumar Paul, “Market Basket Analysis Using Apriori and FP Growth Algorithm”, International Conference on Computer and Information Technology (ICCIT), 18-20 December, 2019

[7] Shish Kumar Dubey, Sonu Mittal, Seema Chattani, Vinod Kumar Shukla, “Comparative Analysis Of Marley Basket Analysis through Data Mining Techniques”, International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) ,17–18 March, 2021

21