# Market Basket Analysis and Design of Chatbot for Recommendations to Startups in Tamil Region

Shruthee B[a], Shekinah Olive[a], Dr K Sundarakantham[a]

[a]*Department of Computer Science and Engineering, Thiagarajar College of Engineering, Tamil Nadu, 625015, India*

**Abstract**

Decision-makers of an organization can use Association Rule Mining (ARM), which is a data mining technique, to extract hidden information from databases and increase their overall profit. The Apriori algorithm is a traditional or classical method for association mining rule, but since it scans the entire database to generate the candidate item sets it takes a long period of time. The Apriori algorithm is being used on a Tamil dataset that is a dataset of grocery items which are needed in the Tamil household. This customized dataset is created by distributing a survey form to friends and families who come from Tamil backgrounds. The association rules that are formed need to cross the threshold value: minimum support:0.0045, minimum confidence: 0.2, and minimum lift: 3. For our database Apriori algorithm works better when compared to the FP-Growth algorithm. Finally, the acquired association rules are made available to the vendors via a chatbot.

*Keywords*: Apriori Algorithm; Association rule; Support; Lift; confidence; chatbot.

## 1. Introduction

Data mining is a popular method to obtain information from a large database. It is widely used in real-time data like textile showroom or supermarket data. Association rule mining is one of the popular techniques used in data mining that can generate association rules among the items in the dataset.

Association rules [4] involve the identification of unique patterns among the items in the given dataset and extract different association rules from the database to measure support and confidence; and is one of the basic Market Basket Analysis (MBA) that gives information to increase business sales [9] and also make purchases for customers easier.

### *1.1. Related Works*

R Suganya et al [1] presented a paper on frequent pattern mining using Apriori-based algorithms. The provided techniques for finding association rules in huge databases are not only effective but also quick. This method's crucial contribution is how much the I/O overhead associated with the Apriori algorithm is reduced. These techniques will be helpful in several actual boolean data storage database mining applications. These methods now only operate with boolean datasets, thus additional work is required to make them relevant to all types of datasets.

Wiwit Pura Nurmayanti et al [2] presented a paper on market basket analysis using an apriori algorithm on outdoor product sales data mainly focused on Indonesia. The Apriori algorithm provides transaction patterns in the sale of outdoor goods as the result. The paper uses a minimum support value of 0.296, a threshold confidence value of 0.774, and a threshold lift value of 1.49. It shows that consumer purchase patterns regarding the purchase of portable stove items have a potential to buy portable gas items. FP-Growth algorithm is used to find transaction patterns in the sale of outdoor goods in this paper.

Yusuf Kurina et al [3] expressed their idea of market basket analysis for knowing the sales pattern at the O! Fish restaurant using an Apriori algorithm. O! Fish restaurants may make more effective advertising tactics by mentioning goods (menus) that are frequently bought together using information about sales patterns. Data mining approaches utilizing Apriori algorithms may be coupled with knowledge of the goods that customers often buy together to determine their buying habits. To create association rules, an Apriori algorithm is utilized. By referring to a combination of goods that are frequently purchased concurrently, O! Fish restaurants may utilize information about the association's regulations in consumer menu purchases to develop more possible promotional techniques to increase sales.

Manpreet Kaur et al [4] published a paper on market basket analysis to identify the changing trends of market data using association rule mining Marketing, bioinformatics, education, nuclear science, and other industries can all benefit from using Market Basket Analysis, also known as Association Rule Learning or Affinity Analysis. The primary objective of an MBA is to help retailers gain knowledge about consumer purchasing patterns so that they may make sound decisions to help their business. For doing market basket analysis, a variety of algorithms are available. The current algorithms can only work with static data so, the changes that take

place over time are not included. This paper discusses association rule mining and provides a new algorithm which may help to examine customer behavior and assists in increasing sales.

Luis Cavique [5] designed a method on a scalable algorithm for market basket analysis. Market basket analysis is a useful technique for helping retailers connect different item sets with each other to help them flourish their business. Finding big baskets is crucial in any industry that deals with a huge number of things, but it's especially important in retail. Even though certain methods can locate big item sets, they may not be computationally efficient. This work aims to offer a big itemset pattern discovery technique for market basket analysis. The market basket aim is converted into a maximum-weighted clique aim in order to acquire the condensed data, which are then employed in this method.

Kwei Tang et al [6] published a paper on context-based market basket analysis in a multi-store environment. A novel method was provided to conduct market basket analysis across various stores and time periods. According to the user's application and demands, the user first establishes a time concept hierarchy and a location hierarchy. By merging the levels of the two hierarchies, a collection of contexts is methodically obtained. Then an effective approach for deriving association rules that satisfies all context-specific support and confidence requirements was created. This method enables a decision-maker to examine purchasing trends at various concepts like time and location for a combination of days and stores, a combination of quarters and states, or a combination of days and regions. The association rules are highly ordered in addition to being flexible since they are acquired from the time and place hierarchies. The algorithm may produce more detailed and substantial information compared to the store-chain rules and the conventional rules, according to a numerical evaluation of its performance.

Abhishek B Rao et al [7] published a paper on the application of market analysis in healthcare. In the modern day, data analysis is essential since it enables us to comprehend patterns via thoughtful exploration. The Apriori technique is extensively used by academics to locate frequently recurring objects in transactional databases, making it one of the primary strategies. In this essay, market basket analysis' applicability to the healthcare sector is discussed. The Apriori method is used in the current study to identify common illnesses that co-occur in a region. This could encourage locals to be more aware of common illnesses and to take every safety action at their disposal to protect their health.

Dr Sandeep A. Thorat et al [8] published a paper on computer–human interaction. The popular approach is through the use of chatbots, which are computer programs designed to make this process easy and engaging. Current artificial intelligence techniques often struggle to provide the most appropriate response to user queries, leading to the predominance of rule-based chatbot systems in industry. This paper presents a comprehensive study on the implementation of rule-based chatbot systems, including discussions on performance measurement parameters for such systems. Furthermore, a comparison is made between two of the most widely used rule-based chatbot implementation frameworks, Google Dialogflow and IBM Watson. Finally, the paper concludes by listing expectations for future chatbot systems.

Robert C. Blattberg et al [9] published a chapter on the examination of the items that customers commonly purchase together. This information is to determine which products should be promoted or cross-sold together. This term is derived from the shopping carts used by customers in supermarkets during their shopping trips. With the advent of the internet, there are now new opportunities for collecting and analyzing such data. This chapter provides an overview of the fundamental concepts of "confidence," "support," and "lift" in relation to market basket analysis, and explains how these concepts can be translated into practical metrics that can be expanded upon.

Jagdish Singh et al [10] published a paper to present the implementation of a rule-based enquiry chatbot that is designed exclusively for students of Asia Pacific University (APU). This chatbot is called ‘APU Admin Bot’. It can help students with a fast solution in a methodical way to resolve their questions and doubts instead of seeking help from the administrative staff. This chatbot makes use of a rule-based approach in the area of pattern recognition. It uses certain words or phrases and sometimes even actions to fetch and display the whole set of responses from the chatbot. It is built entirely from the Chatfuel platform. The chatbot is a messaging platform and is more reliant on a code-less platform.
Bulleted lists may be included and should look like this:

- First point
- Second point
- And so, on

*1.2. Our Contributions*

Our contributions to this paper include,

- Creation of customized regional language dataset.

- Creation of association rule for the Tamil language customized grocery dataset.
- Creation of a chatbot, to use the association rules generated for the retailer's purpose.

*1.3. Overview of our technique*

This work makes use of the Apriori algorithm. It is a data mining technique which is widely utilized to do market basket analysis.

The novelty of the proposed work is that it has a regional language dataset created from scratch and a chatbot which will give suggestions to the retailers based on the association rule; backed by the Apriori algorithm.

Market Basket Analysis (MBA) is a technique of determining what product is purchased when a specific product is purchased by a customer. This is determined by analyzing the customer's purchase history.

So, the purchase history of customers is collected in order to determine the products which are most likely to be purchased with a product.

The products purchased by the customers are referred to as itemsets. Market Basket Analysis is done on the customer purchase history to find relationships between the different item sets.

Since MBA is useful in finding the association or relationships among the items purchased by the customer, it can be made use of by retailers to help their business. Retailers can gather customer purchase data and analyze that data using Market Basket Analysis and extract useful information regarding the relationship between the different item sets.

This relationship between different itemsets is known as the association rule. Each rule is determined by its:

- Confidence
- Support
- Lift

Key differences between the proposed work and existing work,

(i) While there are a number of works available for English language datasets, works on regional language datasets are very low.

(ii) The proposed work applies the Apriori algorithm for low resource Tamil language, which is used by the chatbot application.

**2. Proposed Methodology**

We propose the use of the Apriori algorithm to find association rules for the regional language dataset. The overall architecture of the methodology that is put forward is shown in Fig 1. The following sections explain the methodology in detail.

*2.1. Association Rule Mining*

We have used association rule mining to perform market basket analysis. An approach to machine learning, based on rules is association rule mining. This helps us to find significant connections between various items as they occur together in the data collection.

Three core measures that are used in the association rule are Support, Confidence and Lift.

- Support
- Confidence
- Lift

Support:
The percentage of groups that include all the items stated in an association rule is known as the rule's support. This is calculated from all the groupings that were taken into consideration as the percentage value.

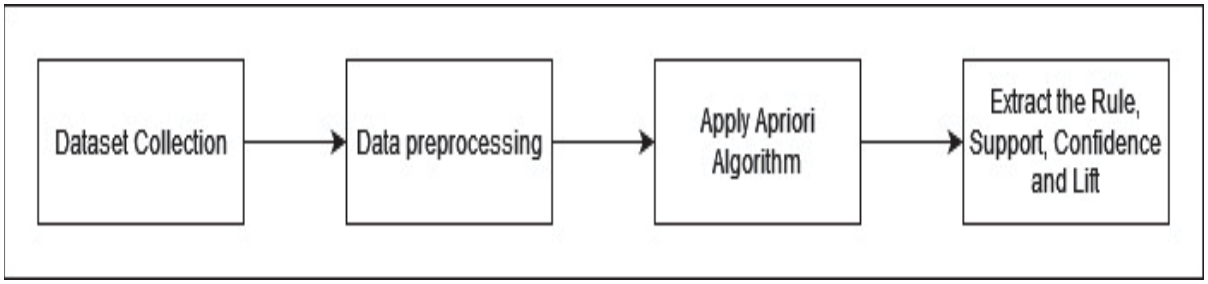



Fig. 1. Workflow of the Market Basket Analysis

Support (X ⇒ Y) = Count of X ⇒ Y/ Total number of Transactions

Confidence:

Confidence is defined as the percentage of how frequently the rule head appears across all the groups that also include the ruling body This rule's reliability is indicated by the confidence value.

Confidence (X ⇒Y) = Support (X ⇒Y) / Support (X)

The proposed methodology is as follows:

- Dataset collection: The data for the dataset is collected from local retail stores and acquaintances which makes it more reliable and accurate for real-time use.
- Data preprocessing: The collected data can have null values. These if left in the dataset will give unreliable results, therefore we removed them to only take the required text data into consideration.
- Apply Apriori algorithm: Apriori algorithm makes use of prior knowledge of frequent itemset properties, which helps to generate the association rules. The frequent itemsets are generated in a level-wise manner. The Apriori property that, all the nonempty subsets of a frequent itemset must also be frequent, is followed over here. The algorithm is applied with minimum support of 0.0045, minimum confidence of 0.2, a minimum lift of 3, minimum length of 2.
- Extraction of Rule, Support, Confidence and Lift: After applying the Apriori algorithm to the records of the dataset we obtained the association results. From which the rules and their necessary parameters can be extracted.

Table 1. An example of a table

| An example of a column heading | Column A (*t*) | Column B (*T*) |
|---|---|---|
| And an entry | 1 | 2 |
| And another entry | 3 | 4 |
| And another entry | 5 | 6 |

If table footnotes should be used, place footnotes to tables below the table body and indicate them with superscript lowercase letters. Be sparing in the use of tables and ensure that the data presented in tables do not duplicate results described elsewhere in the article.

*2.2. Apriori Algorithm*

The apriori algorithm is a popularly used algorithm in association rule learning. It is used to identify the items in a data set, and create itemsets out of it to create association rules.

An itemset has a low likelihood of occurring if:

- Minimum support threshold is P(I), where I is any itemset that isn't empty.
- Any subset in the itemset's value that is below the minimum support level is ignored.

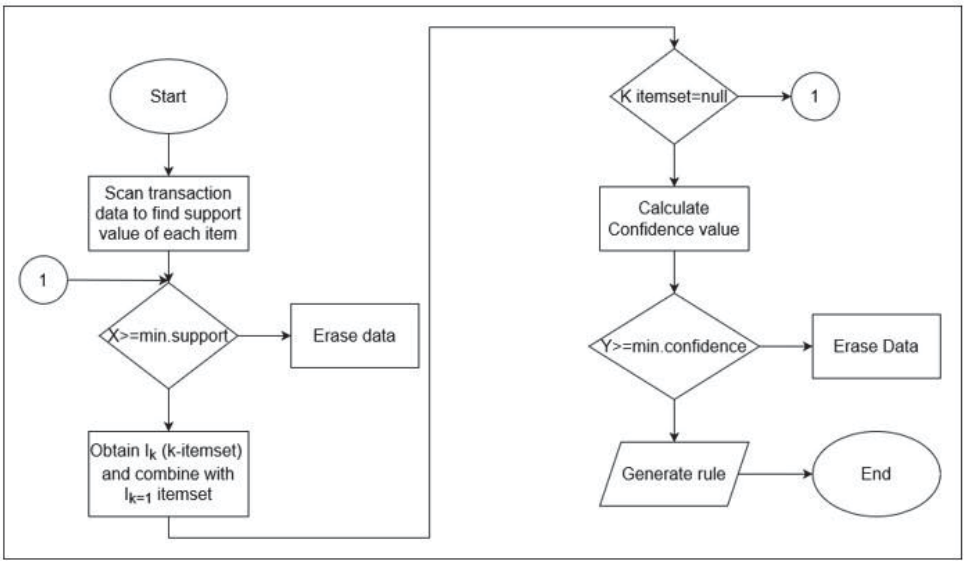



Fig. 2. Flowchart of Apriori algorithm

The Anti-monotone Property is the second feature. An illustration for it would be that if the chance of buying a burger is below the minimum support, the possibility of buying a burger and fries will be lower as well.

Steps in the apriori algorithm:

- Computing the support, also known as frequency, for each item individually.
- Set a threshold value for support.
- Selecting the frequent items.
- Find the frequent itemsets support.
- Repeat for the dataset.
- Construct the association rules and calculate their confidence.
- Check the list of rules.

The implementation for the dataset is as follows:

| Algorithm 1 (Apriori algorithm for MBA to find association items) |
|---|
| 1. Download Apriori on your environment |
| 2. Import necessary packages and dataset creation |
| 3. Give the custom dataset as input |
| 4. Do data preprocessing and perform descriptive analysis |
| 5. Set minimum support, confidence, and lift. |
| 6. For each item in the association rule results: |
| 6.1. Extract the rule, support, confidence and lift. |
| 6.2. Print association rules, which is expected. |

*2.3. Chatbot*

A chatbot is a computer program used to understand questions a user asks and automate the response process to simulate human conversation. The process of finding information is made easier through Chatbots by giving a direct reply to the requested queries. The queries can be of any format like text, audio, or both.

The chatbot we have developed is a simple chatbot which is text-based. It is pre-programmed to give replies to a limited set of inputs with answers that had been written when the chatbot was developed. It operates much similarly to the ones for an interactive FAQ chatbot. The chatbot does not give answers to untrained inputs but just gives an unavailable response.

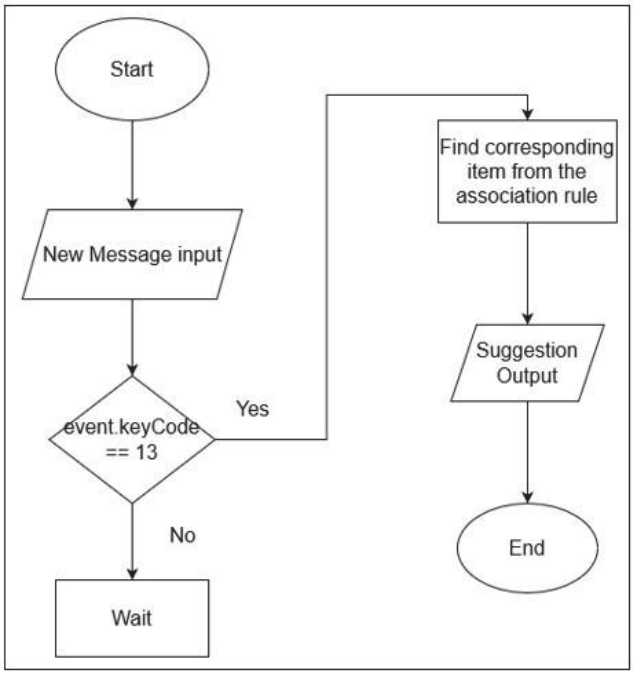



Fig. 3. Flowchart of Apriori algorithm

This chatbot can be used by people who would need a recommendation tool to give them suggestions regarding their grocery business, whether it be a new or existing business. A user would not need to rely on searching the internet to find the related items for their given inputs.

## 3. Experimental results and analysis

### *3.1. Dataset collection*

The data set is created by gathering information from our acquaintances and nearby stores. The custom dataset contains transaction data of about 10000 different transactions, all in the Tamil language.

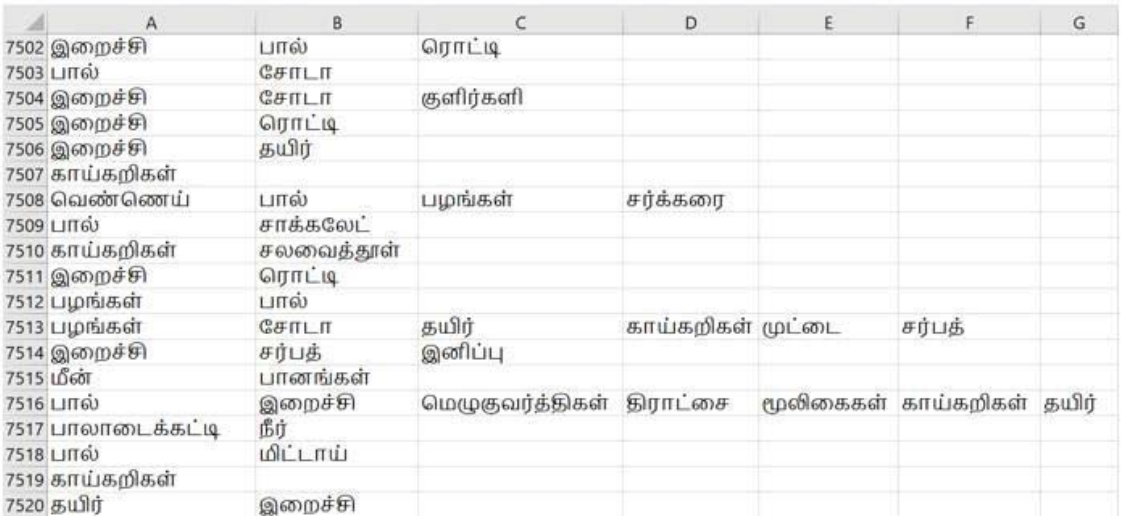
The dataset has a large variety of items and there are transactions of different lengths ranging from 1 item in a transaction to 20 items in a transaction.

Most of the items are checked for spelling mistakes. The dataset itself is random and is not ordered in any specific pattern.

Table 1. Parameters of the Apriori algorithm used

| Parameters | Values |
|---|---|
| minSupport | 0.0045 |
| minConfidence | 0.2 |
| minLift | 3 |
| minLength | 2 |

The dataset is stored in Excel with a variety of items in different transactions.



| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 7502 | இறைச்சி | பால் | ரொட்டி | | | | |
| 7503 | பால் | சோடா | | | | | |
| 7504 | இறைச்சி | சோடா | குளிர்களி | | | | |
| 7505 | இறைச்சி | ரொட்டி | | | | | |
| 7506 | இறைச்சி | தயிர் | | | | | |
| 7507 | காய்கறிகள் | | | | | | |
| 7508 | வெண்ணெய் | பால் | பழங்கள் | சர்க்கரை | | | |
| 7509 | பால் | சாக்கலேட் | | | | | |
| 7510 | காய்கறிகள் | சலவைத்தூள் | | | | | |
| 7511 | இறைச்சி | ரொட்டி | | | | | |
| 7512 | பழங்கள் | பால் | | | | | |
| 7513 | பழங்கள் | சோடா | தயிர் | காய்கறிகள் | முட்டை | சர்பத் | |
| 7514 | இறைச்சி | சர்பத் | இனிப்பு | | | | |
| 7515 | மீன் | பானங்கள் | | | | | |
| 7516 | பால் | இறைச்சி | மெழுகுவர்த்திகள் | திராட்சை | மூலிகைகள் | காய்கறிகள் | தயிர் |
| 7517 | பாலாடைக்கட்டி | நீர் | | | | | |
| 7518 | பால் | மிட்டாய் | | | | | |
| 7519 | காய்கறிகள் | | | | | | |
| 7520 | தயிர் | இறைச்சி | | | | | |

Fig. 4. Customized Dataset Screenshot

*3.2. Analysis of the Apriori Algorithm*

The advantages of using the Apriori algorithm for extracting association rules:

(i) One of the simplest and easily understandable algorithms is the Apriori algorithm.
(ii) The derived associations are understandable and easy to explain to the client.
(iii) There is no need to use labelled data because the algorithm is a fully unsupervised algorithm. So, it can be used in various circumstances as unlabeled data is more accessible.
(iv) It is a comprehensive algorithm, so it detects each rule for a specified level of support and confidence.
(v) Apriori algorithm is the most suitable for the work because it uses join functions to check all possible combinations that fits above the given threshold.
(vi) Join and Prune steps are easy to implement on large itemsets in large databases.

Table 2. Comparative analysis of results

| Paper | Description |
|---|---|
| Frequent Pattern Mining Using Apriori-Based Algorithm [1005D | They presented a paper on frequent pattern mining using apriori-based algorithms. The provided techniques for finding association rules in huge databases are not only effective but also quick. |
| Market Basket Analysis with Apriori Algorithm and Frequent Pattern Growth (FP-Growth) on Outdoor Product Sales Dat [2] | They presented a paper on market basket analysis using an apriori algorithm on outdoor product sales data mainly focused on Indonesia. The Apriori algorithm generates transaction patterns in the sale of outdoor goods. The level of strength of the rules at minimum support is 0.296, the confidence is 0.774, and the lift value is 1.49. This shows that the consumers who buy portable stove items also have the potential to buy portable gas items. |
| Proposed method | We applied the Apriori algorithm to the Tamil Language dataset, which mostly focuses on locally available groceries. It helped us to understand the purchase patterns of people living here. |

The output obtained is as follows.

| | LHS | RHS | Support | Confidence | Lift | Frequency |
|---|---|---|---|---|---|---|
| 0 | முழு கோதுமை பாஸ்தா | இடலை எண்ணெய் | 0.007961 | 0.271493 | 4.142195 | 0.191057 |
| 1 | பாஸ்தா | இறால் | 0.005042 | 0.322034 | 4.528301 | 0.121003 |
| 2 | பாஸ்தா | இறைச்சி | 0.006103 | 0.389831 | 3.460721 | 0.146477 |
| 3 | தக்காளி சட்னி | மாட்டிறைச்சி | 0.005307 | 0.377358 | 3.859092 | 0.127372 |
| 4 | மூலிகை & மிளகு | மாட்டிறைச்சி | 0.015921 | 0.323450 | 3.307793 | 0.382115 |
| 5 | ஸ்பாகெட்டி - அரிசி மாவு | இடலை எண்ணெய் | 0.005042 | 0.201058 | 3.067562 | 0.121003 |
| 6 | காய்கறிகள் - இடலை எண்ணெய் | பால் | 0.005042 | 0.395833 | 3.034991 | 0.121003 |
| 7 | காய்கறிகள் - ஸ்பாகெட்டி | இடலை எண்ணெய் | 0.006369 | 0.200837 | 3.064184 | 0.152846 |
| 8 | கோழிக் கறி - இடலை எண்ணெய் | பால் | 0.004909 | 0.411111 | 3.152131 | 0.117819 |
| 9 | சூப் - நீர் | இடலை எண்ணெய் | 0.005174 | 0.222857 | 3.400150 | 0.124187 |
| 10 | மாட்டிறைச்சி - பால் | இடலை எண்ணெய் | 0.004909 | 0.224242 | 3.421286 | 0.117819 |

Fig. 5. Output of Apriori algorithm


Fig. 6. Output of Chatbot

Some of them with their obtained parameters are:

Table 3. Results obtained

| Antecedents | Consequents | antecedent support | consequent support | Support | Confidence | Lift | Leverage |
|---|---|---|---|---|---|---|---|
| (உழுத்தம் பருப்பு, கோதுமை) | (கடலை பருப்பு, திணை) | 0.182857 | 0.182857 | 0.182857 | 1.0 | 5.468750 | 0.149420 |
| (கடலை பருப்பு, திணை) | (உழுத்தம் பருப்பு, கோதுமை) | 0.182857 | 0.182857 | 0.182857 | 1.0 | 5.468750 | 0.149420 |
| (உழுத்தம் பருப்பு, கோதுமை) | (திணை, பயத்தம் பருப்பு) | 0.182857 | 0.182857 | 0.182857 | 1.0 | 5.468750 | 0.149420 |
| (திணை, பயத்தம் பருப்பு) | (உழுத்தம் பருப்பு, கோதுமை) | 0.182857 | 0.182857 | 0.182857 | 1.0 | 5.468750 | 0.149420 |
| (உளுந்து, திணை) | (உழுத்தம் பருப்பு, கோதுமை) | 0.182857 | 0.182857 | 0.182857 | 1.0 | 5.468750 | 0.149420 |

## 4. Conclusion

We applied the Apriori algorithm to the Tamil Language dataset, which mostly focuses on locally available groceries. It helped us to understand the purchase patterns of people living here. It can be used to assist retail stores with their stocking schedule and stock maintenance. With the help of this study, the stores can make offers and discounts to increase their sales. This can also be used by the producers to understand customer demand and purchase patterns. We obtained 22 association rules for the dataset.

## References

1. R. Suganya, R. Tamil Selvi, 2014. Frequent Pattern Mining Using Apriori Based Algorithm; IJERT.
2. Wiwit Pura Nurmayanti, Hanipar Mahyulis Sastriana, Abdul Rahim, Muhammad Gaz-ali. Ristu Haiban Hirzi, Zuhut Ramdani, Muhammad Malthuf, 2021. Market Basket Analysis with Apriori Algorithm and Frequent Pattern Growth (Fp-Growth) on Outdoor Product Sales Data; International Journal of Educational Research and Social Sciences.
3. Yusuf Kurnia, Yohanes Isharianto, Yo Ceng Giap, Aditiya Hermawan & Riki, 2019. Study of application of data mining market basket analysis for knowing sales pattern (association of items) at the O! Fish restaurant using apriori algorithm; Journal of Physica: Conference Series.
4. Manpreet Kaur & Shivani Kang; 2016. Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining; Procedia Computer Science 85, p. 78-85.
5. Luis Cavique; A scalable algorithm for the market basket analysis, 2007. Journal of Retailing and Consumer Services 14, p. 400-407.
6. Kwei Tang, Yen-Liang Chen, Hsiao-Wei Hu, 2008. Context-based market basket analysis in a multiple-store environment; Decision Support Systems 45, p. 150-163.
7. Abishek B. Rao, Jammula Surya Kiran, Poornalatha G, 2021. Application of market–basket analysis on healthcare; International Journal of System Assurance Engineering and Management.
8. Dr Sandeep A. Thorat, Vishakha D. Jadhav, 2020. A Review on Implementation Issues of Rule-based Chatbot Systems; International Conference on Innovative Computing and Communication.
9. Robert C. Blattberg, Byung-Do Kim, Scott A. Neslin, 2008. Market Basket Analysis; Database Marketing; Springer; p. 339–351.
10. Jagdish Singh, Minnu Helen Joesph and Khurshid Begum Abdul Jabbar, 2019. Rule-based chabot for student enquiries; International conference on computer vision and machine learning.