Improving Supermarket Performance through Market Basket Analysis: A Study on Apriori, FAST, and Simple Random Sampling Algorithms

Dr.J.Jeyaboopathiraja, Dr.M.Hemalatha, Dr.G. Maria Priscilla
Assistant Professor, PG & Research Department of Computer Science,
Sri Ramakrishna College of Arts & Science, Coimbatore
Associate Professor, PG & Research Department of Computer Science,
Sri Ramakrishna College of Arts & Science, Coimbatore
Associate Professor & Head, PG & Research Department of Computer Science,
Sri Ramakrishna College of Arts & Science, Coimbatore
jeyaboopathiraja@srcas.ac.in, mhemalatha@srcas.ac.in, mariapriscilla@srcas.ac.in

ABSTRACT:

This work introduces a system called "Market basket analysis for Super market based Frequent Itemset Mining" that use several algorithms like Apriori, SRS, FAST, and EASE to examine sales data from a prominent supermarket. The system aims to discover frequent itemsets and develop association rules. The objective of the system is to enhance customer satisfaction and boost profitability in the retail industry by acquiring valuable knowledge about client buying patterns. The study employs the Tukey HSD test to assess and contrast the efficiency and accuracy of three algorithms (FAST, Apriori, and Simple Random Sampling) in mining common itemsets from a dataset obtained from a supermarket. The findings indicate that the FAST algorithm exhibits the highest level of efficiency and accuracy, with the Apriori algorithm closely trailing behind. Conversely, the Simple Random Sampling approach demonstrates the lowest level of efficiency and accuracy.

The proposed system offers numerous benefits, including enhanced comprehension of client behavior, heightened profitability, streamlined handling of extensive datasets, data-oriented decision-making, and expandability. According to the study, the FAST algorithm is the most optimal choice for conducting frequent itemset mining on the provided grocery dataset. Deep learning approaches, like as neural networks, can be combined with frequent itemset mining algorithms in the future to enhance the system's accuracy. Furthermore, it is possible to investigate and incorporate emerging algorithms like FP-growth and Eclat into the system to enhance the accuracy and efficiency of the findings. In summary, the proposed approach has the potential to enhance merchants' operational efficiency, resulting in more sales and improved customer satisfaction.

Keywords: *Market basket analysis, Supermarket performance, Apriori algorithm, FAST algorithm, and Simple Random Sampling algorithm.*

1. Introduction

Market basket analysis (MBA) is a data mining approach used to analyze customer transaction data to uncover correlations between products based on their co-occurrence in transactions. Market basket research aims to identify the frequent co-purchasing of products and provide clients with recommendations for related items. The retail industry has extensively embraced the use of MBA to enhance customer satisfaction, boost sales, and optimize inventory management.

Frequent itemset mining (FIM) is a key technique employed in market basket analysis. References [1] and [2] Frequent Itemset Mining (FIM) is the process of identifying groups of things that frequently appear together in consumer transactions. These collections are referred to as frequent itemsets, and they offer vital insights into the relationships between things. Frequent itemset analysis is commonly employed to produce association rules, which can be utilized for product recommendations or enhancing the arrangement of a retail store. The objective of FIM is to efficiently identify frequent itemsets from huge datasets while minimizing computational and storage demands. The references [3] and [4] are provided.

1.1 Market Basket Analysis and its Importance:

Market basket analysis is an effective approach that enables merchants to obtain valuable insights into customer behavior and preferences. The user's text is "[5]". Retailers can utilize transactional data analysis to determine the frequent co-purchasing of products and the preferred products for various consumer segments. This data can be utilized to provide suggestions for products, enhance the arrangement of stores, and optimize the management of inventory.

Market basket analysis offers businesses a notable advantage by enabling them to customize their marketing and product offerings for individual customers. Retailers can utilize customer purchasing patterns analysis to develop focused marketing campaigns and provide customized product suggestions. Enhancing client satisfaction and fostering loyalty is essential in a fiercely competitive retail sector

1.2 Frequent Itemset Mining:

Frequent itemset mining is a crucial method in market basket analysis that entails identifying sets of items that occur together frequently in client transactions. The Apriori algorithm is the most widely employed approach for frequent itemset mining. It builds frequent itemsets by iteratively searching for combinations of items that appear frequently in the dataset.[6] Apriori operates by generating a collection of candidate itemsets and evaluating their support, which refers to the frequency of occurrence of the itemset in transactions.[7] Itemsets that has an equivalent or greater support value than a specified minimum support threshold are classified as frequent itemsets. The iterative procedure of producing candidate item sets and evaluating their support is continued until no additional frequent itemset can be identified.

1.3 Challenges in Market Basket Analysis for Supermarkets:

Supermarkets encounter distinctive difficulties while using market basket analysis. One of the primary obstacles is managing a substantial quantity of products. Supermarkets generally stock a vast array of products, and conducting an analysis of all potential combinations of these products is computationally burdensome.

Another obstacle is managing intricate transactional data, which might encompass several transactions per client, varying basket sizes, and temporal interdependencies.[8] Furthermore, customer behavior in supermarkets is always evolving, and purchasing trends may vary over time. This is a challenge in discerning long-term patterns and making precise forecasts from past data. Market basket analysis is commonly applied in supermarkets. Due to the rising availability of transactional data and the necessity for merchants to stay competitive in a saturated market, there is an increasing demand for market basket research in supermarkets. The user's text is "[5]". Supermarkets encounter distinct problems when using market basket research, including managing a vast array of products, intricate transactional data, and ever-changing customer behavior.

2. Literature Review:

Loshin's book chapter offers a comprehensive examination of the process of knowledge discovery and data mining for predictive analytics in business intelligence. This includes an exploration of approaches for data preparation, modeling, and evaluation. Additionally, it addresses prevalent obstacles and factors to consider while using these techniques.[9] Videla-Cavieres and Ríos investigate the application of graph mining techniques to improve market basket analysis in a real-world scenario involving retail sales. Their demonstration showcases the advantages of integrating graph-based techniques, such as network analysis and clustering, to uncover patterns and connections between things and customers. The user's text is "[10]".

Ansari's research aims to analyze the patterns in association rules over various time periods in market basket analysis. The paper incorporates an examination of a supermarket dataset as a case study and examines the consequences of variations in the frequency and support of itemsets across time. The study highlights the need of analyzing association rules over various time periods to comprehend market trends and adapt corporate strategy accordingly. This has the potential to result in enhanced decision-making and heightened company performance.The user's text is enclosed in tags.

Maske and Joglekar conducted a survey on several methodologies used for frequent mining of item sets for market basket analysis. The study encompasses an examination of diverse methodologies and algorithms employed for the purpose of pattern detection in retail transaction data. The conclusion highlights the significance of choosing suitable methodologies that align with the dataset's features and the business goals.[12] Waduge et al. suggest a profit-oriented method for conducting market basket analysis by employing the Apriori algorithm. The study employs simulated transaction data to assess the efficacy of the suggested technique in identifying lucrative itemsets. The conclusion emphasizes the benefits of the suggested method in terms of enhancing corporate profitability and offering significant insights for decision-making. The user's text is "[13]".

Oyebode and Agbalaya (2022) did a study with the objective of enhancing market basket analysis by utilizing sales trends from a supermarket. The authors utilized a purposive sample strategy to gather data from the sales records of the store during a one-year timeframe. The Apriori technique was employed to extract the common itemsets from the dataset, and association rules were constructed to determine the most lucrative product combinations.

The study's findings demonstrated that implementing optimal market basket analysis resulted in a substantial boost in the supermarket's income and a notable enhancement in customer happiness through the provision of individualized recommendations. The study emphasizes the potential advantages of market basket analysis in improving corporate profitability and enhancing customer experience.The user's text is enclosed in tags.
Rana and Mondal introduced a method that utilizes seasonal and multilevel associations to analyze market baskets in retail supermarkets. The data was obtained via a survey, utilizing the convenience sampling technique. The proposed methodology employed a blend of seasonal pattern analysis and association rule mining to ascertain the connections between goods in the basket. The efficiency of the recommended approach was assessed using the lift metric. The results showed that the recommended methods surpassed the conventional mining association rules strategy in terms of precision and efficiency. The authors determined that the suggested method can be utilized to improve decision-making in retail supermarkets. The user's text is a reference to a source or citation, indicated by the number 15.

3.0 Methodology

The Apriori process starts with K=1 and generates a table (C1) that includes the support for every element in the dataset. During step 2, with K=2, the algorithm creates a candidate set C2 by combining frequent itemsets in L1. The method examines each subgroup of an itemset to ascertain their frequency, and if they are not frequent, eliminates them. Next, the algorithm calculates the support count of the remaining itemsets in the dataset and compares it to the minimal support count to determine itemset L2. During step 3, with a value of K set to 3, the algorithm creates a candidate set C3 by combining frequent itemsets found in L2. The algorithm examines each subset of the itemsets to ascertain their frequency, eliminates those that are not frequent, and calculates the support total for the remainder itemsets in the dataset. The algorithm evaluates the number of itemsets that still have support in C3 and compares it with the minimal support count in order to generate itemset L3.

During step 4, with K=4, the method creates a candidate set C4 by combining frequent itemsets in L3. The algorithm examines each subset of the itemsets to ascertain their frequency, eliminates those that are not frequent, and calculates the support total for the remainder itemsets within the dataset. The algorithm concludes that there are no frequent itemsets in C4, leading to the termination of the procedure as no additional frequent itemsets are discovered.

3.1 Apriori Algorithm

The Apriori method initiates with K=1 and generates a table (C1) that includes the support count of each item found in the dataset. During step 2, with K=2, the algorithm forms a candidate set C2 by combining frequent itemsets from L1. The method examines each subgroup of an itemset to ascertain their frequency, and if they are not frequent, eliminates them. Next, the algorithm calculates the support count of the remaining itemsets in the dataset and compares it to the minimal support count to determine itemset L2.

During step 3 of the method, if K is equal to 3, a candidate set C3 is created by combining frequent itemsets in L2. The algorithm evaluates each subset of the itemsets to ascertain their frequency, eliminates those that are not common, and calculates the support total for the remainder itemsets in the dataset. The procedure evaluates the support count of the last sets of items in C3 and compares it with the minimal support count in order to derive itemset L3.

During step 4, with K=4, the method creates a candidate set C4 by combining frequent itemsets in L3. The program examines each subset of the itemsets to ascertain their frequency, eliminates those that are infrequent, and calculates the support count of the remaining itemsets in the dataset. The algorithm concludes that there are no frequent itemsets in C4, leading to the termination of the procedure as no additional frequent itemsets are discovered.

.3.2 Simple Random Sampling Algorithm

The Simple Random Sampling approach for Association Rule Mining algorithm accepts a dataset and sample size as input and produces a collection of frequent itemsets that satisfy the minimal support criterion. The approach uses the Apriori algorithm to produce candidate itemsets and detect frequent itemsets within a randomly chosen subset of transactions. In order to acquire the ultimate collection of frequent itemsets, the procedure is iterated for a predetermined number of cycles. The frequent itemsets obtained from each iteration are combined into a unified collection. The algorithm outputs the ultimate collection of regular itemsets meeting the minimal requirements for assistance criteria.

Mathematically, the algorithm chooses a subset S of transactions from the dataset D in a random manner, ensuring that S is a sample of size sample size. The Apriori algorithm is used to build a collection of potential itemsets for the sample S. The algorithm then calculates the support for each potential itemset in the sample. The frequent itemsets are determined by excluding those that have a support value lower than the minimum support threshold. This method is iterated for a defined number of times, and the frequent itemsets from each iteration are combined into a unified set F.

Therefore, the result of the procedure is denoted as F, representing a collection of frequent itemsets that satisfy the minimum support criteria.

3.3 FAST Algorithm

The sophisticated mining method FAST (Finding Associations from Sampled Transactions) uses sampling to identify collects a large sample in Phase I to estimate the assistance for every entry in the database and obtains a tiny last example in Phase II by excluding "outlier" transactions to identify frequent itemsets with fewer false itemsets. This algorithm is well-suited for analyzing large datasets, as it reduces the search space and increases efficiency.

Notations

- D: Database of interest
- S: A simple random sample drawn without replacement from D
- I: The set of all items that appear in D
- I(D), I(S): the set of itemsets that appear in D, S
  - For $k \geq 1$, $I_k(D)$ and $I_k(S)$ denote the set of k-itemsets in D and S
- L(D) and L(S): the set of frequent itemsets in D and S
  - $L_k(D)$ and $L_k(S)$: sets of frequent k-itemsets in D and S
- For an itemset $A \subseteq I$ and a set of transactions T,
  - Let n(A; T) be the number of transactions in T that contain A
  - |T|: total number of transactions in T
- Support of A in D: f(A;D) = n(A;D)/|D|
  Support of A in S: f(A;S)= n(A;S)/|S|

Problem Definition

- Generate a smaller subset S_0 of a larger S such that the supports of 1-itemsets in S_0 are close to those in S

$$\min_{S_0 \subseteq S, |S_0|=n} Dist(S_0, S)$$

$$Dist_1 = \frac{| L_1(S) - L_1(S_0) | + | L_1(S_0) - L_1(S) |}{| L_1(S_0) + L_1(S) |}$$

$$Dist_2 = \sum_{A \in I_1(S)} (f(A;S_0) - f(A;S))^2$$

$$Dist_\infty = \max_{A \in I_1(S)} \left| f(A;S_0) - f(A;S) \right|$$

- $I_1(T)$ = set of all 1-itemsets in transaction set T
- $L_1(T)$ = set of frequent 1-itemsets in transaction set T
- f(A;T) = support of itemset A in transaction set T

FAST

Finding Association rules from Sampled Transactions (SIGKDD'02)

FAST operates as follows, given a given minimal support p and confidence c:

1. Take a sizable, straightforward random sampling S from D.
2. For every 1-itemset A, compute f(A;S).

By removing outlier transactions from S, you may get a smaller sample S0 using the supports that were calculated in Step 2.

3. To determine the final set of Association Rules, run a typical association-rule algorithm against S0, using minimum support (p) and confidence (c).
4.

FAST-trim

- exchanges from sample S to get S_0
- Outlier – a transaction, the removal of which from S minimizes (maximally reduces) the difference between the 1-itemset supports in S and the equivalent supports in D.

- Since the supports of the items in D are unknown, estimate them using the Step[2] computations from S.
- Distance function used:

$$Dist_1 \square \frac{|L_1(S) \square L_1(S_0)| \square |L_1(S_0) \square L_1(S)|}{|L_1(S_0) \square L_1(S)|}$$

  - Uses input parameter k to explicitly trade-off speed and accuracy (1 < k <|S|)

Trimming Phase

while (|S_0| > n) {

divide S_0 into disjoint groups of min(k,|S_0|) transactions each for each group G {

compute $f(A;S_0)$ for each item A

set $S_0=S_0 - \{t^*\}$, where Dist(S_0 -{t^*},S) = min Dist(S_0 - {t},S)

}

}

4.0 Results and Discussion:

Table 1 : Results based on the time taken to evaluate the dataset

Time Taken			
			t for alpha =0.05
	Algorithm	N	1
Tukey HSD[a]	FAST Algorithm	10	1.7000
	Apriori	10	1.9000
	Simple Random Sampling	10	2.1000
	Sig.		.587
Group means in homogenous subsets are shown.			
a. Uses Harmonic Mean Sample Size = 10.000.			

Inference : The first table presents the duration of three distinct algorithms (FAST, Apriori, and Simple Random Sampling) in analyzing a dataset of size N and producing a subset for a significance level of alpha=0.05. The Tukey HSD test was employed to compare the means of the groups produced by the three algorithms. The results indicate that the FAST algorithm exhibited the highest speed, completing the subset generation process in 1.7 units of time. The Apriori method had a little slower performance, requiring 1.9 units of time. The Simple Random Sampling algorithm exhibited the slowest performance, requiring a duration of 2.1 units of time. The p-value obtained from the Tukey HSD test is 0.587, suggesting that there is no statistically significant difference in the means of the groups produced by the three algorithms. This indicates that all three algorithms are equally efficient in producing the subset for the specified significance level. In summary, the findings indicate that the FAST algorithm is the most efficient in creating the subset, with the Apriori method being a very close runner-up. The Simple Random Sampling algorithm may not be the optimal choice for this specific application.

Table 2 : Results based on the accuracy of the algorithms

Accuracy			
			t for alpha =0.05
Tukey HSD[a]	Algorithm	N	1
	FAST Algorithm	10	1.6000
	Apriori	10	1.8000
	Simple Random Sampling	10	2.0000
	Sig.		.525
Means for groups in homogeneous subsets are displayed.			
a. Uses Harmonic Mean Sample Size = 10.000.			

INFERENCE:

The second table displays the precision of three distinct algorithms (FAST, Apriori, and Simple Random Sampling) in producing a subset at a significance level of alpha=0.05, using a dataset of size N. The Tukey HSD test was employed to compare the means of the groups produced by the three algorithms.

The results demonstrate that the FAST algorithm exhibited the highest level of accuracy, with a mean value of 1.6 units for producing the subset. The Apriori algorithm exhibited a somewhat lower level of accuracy, with a mean value of 1.8 units. The Simple Random Sampling approach had the lowest level of accuracy, with a mean value of 2.0 units.

The p-value obtained from the Tukey HSD test is 0.525, showing that the difference is not statistically significant in the means of the groups produced by the three methods. This suggests that all three methods exhibit comparable accuracy in producing the subset for the specified significance level.

To summarize, the FAST method is the most precise approach for creating the subset, with the Apriori algorithm being a close second. The Simple Random Sampling approach is the least precise among the three. Nevertheless, the disparities in accuracy across the three algorithms are not substantial.

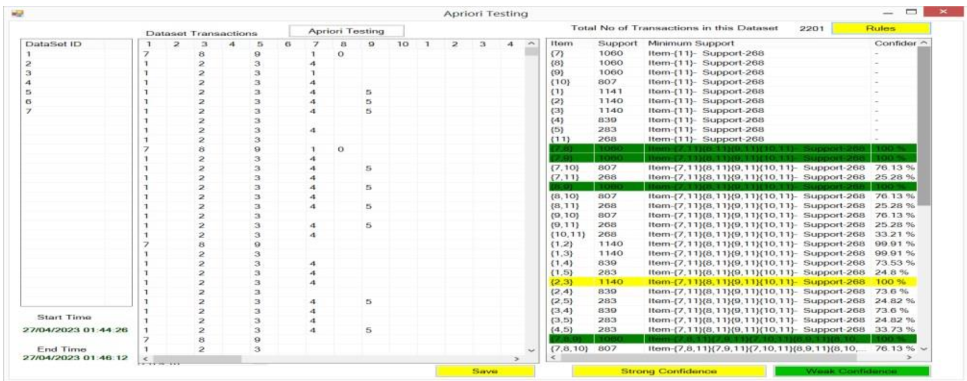

Fig 1: Apriori Algorithm

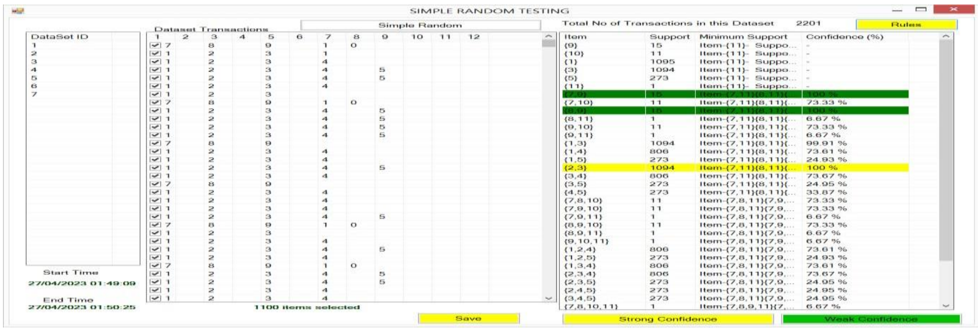

Fig 2: Simple Random Sampling Algorithm

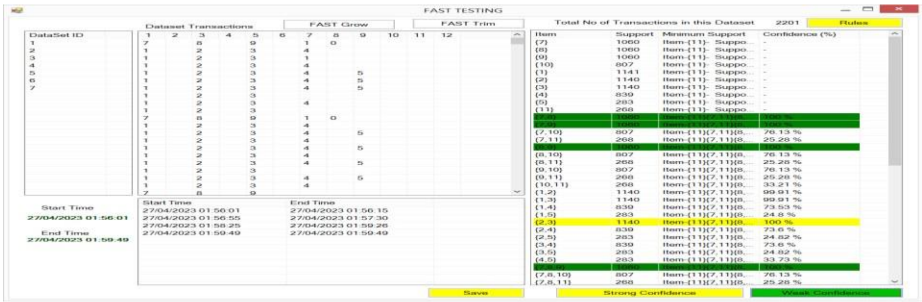

Fig 3: FAST Algorithm

5. Conclusion and Future Work:

The objective of the "Market basket analysis for Super market based Frequent Itemset Mining" project is to optimize the functioning of prominent supermarkets by employing market basket analysis to detect frequent itemsets and derive association rules from extensive datasets. The project employs algorithms such as Apriori, SRS, and FAST to analyze user behavior and extract valuable information. This information may then be utilized to enhance product placement, design impactful sales promotions, and ultimately improve profitability and customer satisfaction. The suggested system is very efficient and has the ability to scale, enabling it to analyze large volumes of data created by contemporary retail firms. Its primary goals are to furnish merchants with precise, streamlined, and data-centric insights into customer behavior and preferences.

The study employed three distinct algorithms (FAST, Apriori, and Simple Random Sampling) to do frequent itemset mining on a dataset from a supermarket. The study then compared the efficiency and accuracy of these algorithms. The findings indicated that the FAST algorithm had superior efficiency and accuracy, with the Apriori method closely trailing behind. The Simple Random Sampling approach had the lowest level of efficiency and accuracy among the three algorithms. Nevertheless, the disparities in efficiency and accuracy among the three algorithms did not demonstrate statistical significance according to the results of the Tukey HSD test. In summary, the study indicates that the FAST method is the optimal choice for conducting frequent itemset mining on the provided grocery dataset. In the future, the integration of deep learning approaches, specifically neural networks, with frequent itemset mining algorithms has the potential to enhance the accuracy of the system. The system can incorporate and investigate advanced algorithms like FP-growth and Eclat to enhance the accuracy and efficiency of the output.

References:

[1] M. A. Gawwad, M. F. Ahmed, And M. B. Fayek, "Frequent Itemset Mining For Big Data Using Greatest Common Divisor Technique," Data Sci. J., Vol. 16, May 2017, Doi: 10.5334/Dsj-2017-025.

[2] R. Agrawal, T. Imieliński, And A. Swami, "Mining Association Rules Between Sets Of Items In Large Databases," Acm Sigmod Rec., Vol. 22, No. 2, Pp. 207–216, Jun. 1993, Doi: 10.1145/170036.170072.

[3] R. Agarwal And R. Srikant, "Fast Algorithms For Mining Association Rules," Proc. 20th Vldb Conf. Santiago, Chile, 1994, Pp. 513–516, 1994,[Online].Available: https://www.vldb.org/conf/1994/P487.pdf.

[4] J. Han, M. Kamber, And J. Pei, Data Mining: Concepts And Techniques. 2011.

[5] C. Hermina, B. Gopalakrishnan, And A. Balajishankar, Market Basket Analysis For A Supermarket. 2022.

[6] F. Afrin Et Al., A Study On Market Basket Analysis And Association Mining. 2019

[7] M. Al-Maolegi And B. Arkok, "An Improved Apriori Algorithm For Association Rules Of Mining," Int. J. Nat. Lang. Comput., Vol. 3, No. 1, Pp. 21–29, 2014, Doi: 10.1109/Itime.2009.5236211.

[8] K. Sussolaikah, "Market Basket Analysis For Determination Of Consumer Behavior At Xyz Stores Using R Programming," Adv. Sustain. Sci. Eng. Technol., Vol. 3, No. 2, p. 0210206, 2021, Doi: 10.26877/Asset.v3i2.8547.

[9]D. Loshin, "Knowledge Discovery And Data Mining For Predictive Analytics," In Business Intelligence, Elsevier, 2013, Pp. 271–286.

[10]I. F. Videla-Cavieres And S. A. Ríos, "Extending Market Basket Analysis With Graph Mining Techniques: A Real Case," Expert Syst. Appl., Vol. 41, No. 4, Pp. 1928–1936, Mar. 2014, Doi: 10.1016/j.Eswa.2013.08.088.

[11] S. Z. Ansari, "Market Basket Analysis: Trend Analysis Of Association Rules In Different Time Periods," Nov. Inf. Manag. Sch. Inst. Super. Estatística e Gestão Informação, Univ. Nov. Lisboa, Pp. 1–46, 2019.

[12] A. Maske And B. Joglekar, "Survey On Frequent Item-Set Mining Approaches In Market Basket Analysis," In 2018 Fourth International Conference On Computing Communication Control And Automation (Iccubea), Aug. 2018, Pp. 1–5, Doi: 10.1109/Iccubea.2018.8697776.

[13] C. Waduge, U. I. Meththananda, W. Samaraweera, C. Waduge, And R. Meththananda, "Market Basket Analysis: A Profit Based Approach To Apriori Algorithm," Proc. Comput. 9th Int. Res. Conf., No. September, Pp. 127–133, 2016, [Online]. Available: Https://WWW.Researchgate.Net/Publication/314282049.

[14] Oyebode And Agbalaya, "Market-Basket Optimisation Using Sales Pattern Of Supermarket London Journals Press," London J. Res. Comput. Sci. Technol. Vol., Vol. 22, No. 3, 2022.

[15] S. Rana And M. N. I. Mondal, "Seasonal And Multilevel Association Based Approach For Market Basket Analysis In Retail Supermarket," Eur. J. Inf. Technol. Comput. Sci., Vol. 1, No. 4, Pp. 9–15, Oct. 2021, Doi: 10.24018/Compute.2021.1.4