An Experimental approach on Association Rule Mining in Supermarket Dataset using Apriori Algorithm

P. Susmitha

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

Abstract— *One of the most popular algorithms is Apriori that is used to extract frequent itemsets from large database and getting the association rule for discovering the knowledge. In this study, we apply the Apriori algorithm to mine association rules in a real-world supermarket transaction dataset. The primary aim is to identify meaningful patterns and associations between items purchased by customers. We pre-process the data and employ the Apriori algorithm to extract frequent itemsets and generate association rules based on the minimum support and confidence thresholds. The results reveal interesting and actionable insights for supermarket managers to optimize product placement, promotions, and customer experience. We achieved a set of high-confidence association rules with strong support that can contribute to enhancing supermarket sales and customer satisfaction.*

I. INTRODUCTION

With the advancement of the innovation of data and the requirement for removing helpful data of finance managers from dataset, information mining and its methods is seemed to accomplish the above objective [1][3]. Information mining is the fundamental course of finding covered up and fascinating examples from gigantic measure of information where information is put away in information stockroom, OLAP (on line logical cycle), data sets and different vaults of data [4]. This information might reach to more than terabytes. Information mining is likewise called (KDD) information revelation in data sets, and it incorporates a joining of procedures from many teaches, for example, measurements, brain organizations, data set innovation, AI and data recovery, and so on [5]. Information mining is a course of recovering profound and significant data for information clients from a lot of sporadic information. Information mining is one of the significant uses of data set. Its capability is to find the secret data of information. Certain individuals compare information mining with information disclosure, yet the data from information mining may not be guaranteed to shape information. In this regard, information revelation can be viewed as the following activity or center part of the exhumed information, gathering, distinguishing and summing up the uncovered data, lastly shaping information.

II. ASSOCIATION RULES

Affiliation Mining is quite possibly of the main datum mining's functionalities and it is the most famous procedure has been concentrated by specialists. Removing affiliation rules is the center of information mining [2]. It is digging for affiliation rules in data set of deals exchanges between things which is significant field of the exploration in dataset. The advantages of these principles are distinguishing obscure connections, creating results which can perform reason for independent direction and expectation [6].

Affiliation rules mirror the relationship and connection between's the two, and are utilized to recover the relationship between's important information things from an enormous number of information. Its motivation is to find incessant thing sets major areas of strength for and rules [5].

The use of affiliation rules — shopping crate investigation. As per the merchandise in the client's shopping container, the standards are found and continuous thing sets is created, that is to say, which products will be bought by the client a few times simultaneously. These affiliation rules can be utilized in traders' showcasing methodologies, like item advancement, item locale division, and so on [7].

The revelation of affiliation rules is partitioned into two stages: recognition the regular itemsets and age of affiliation rules. In the principal stage, each arrangement of things is called itemset, assuming that they happened together more noteworthy than

the base help edge, this itemset is called continuous itemset [4]. Finding continuous itemsets is simple however exorbitant so this stage is a higher priority than second stage. In the subsequent stage, it can produce many standards from one itemset as in structure, if itemset {I1, I2, I3}, its principles are {I1→I2, I3}, {I2→I1, I3}, {I3→I1, I2}, number of those rules is n2-1 where n = number of things. To approve the standard (for example X→Y), where X and Y are things, in view of certainty edge which decide the proportion of the exchanges which contain X and Y to the exchanges A% which contain X, this implies that A% of the exchanges which contain X likewise contain Y. least help and certainty is characterized by the client which addresses limitation of the guidelines. So the help and certainty limits ought to be applied for every one of the principles to prune the standards which it esteems not as much as edges values.

Normal assessment models for regular thing sets:

1) Support: It is one of the two fundamental boundaries of affiliation rules. It is the proportion of the quantity of exchanges containing both x and y in All example of dataset D to all exchanges. Assuming we have two information x and y that should be investigated for relationship, then the comparing support degree is:

$$\text{Support } (X \Rightarrow Y) = P(X \cup Y) = \frac{\text{Count}(X \cup Y)}{|D|} \quad (1)$$

For instance, a help rating of 67% intends that "there is a 67% likelihood that a person in the populace will contain both X and Y".

2) Confidence: It is the proportion of the quantity of exchanges including x and y to the quantity of exchanges including y, in particular restrictive likelihood.

$$\text{Confidence } (X \Rightarrow Y) = P(X \mid Y) = \frac{\text{Support}(X \cup Y)}{Support\ (X)} \quad (2)$$

Expecting to be that "85% of the terms containing X contain Y", the certainty is 85%.

Just custom least help, or a blend of custom help and certainty, can decide the continuous thing sets in the data set.

Its center is to recover all regular thing sets, and find all thing sets that are more prominent than or equivalent to the help by setting the base help count and emphasizing constantly.

III. APRIORI ALGORITHM

Apriori calculation is not difficult to execute and exceptionally basic, is utilized to mine all continuous itemsets in data set. The calculation makes many hunts in data set to find successive itemsets where kitemsets are utilized to create k+1-itemsets. Every k-itemset should be more prominent than or equivalent to least help limit to be recurrence. In any case, it is called applicant itemsets [4]. In the first, the calculation filter data set to find recurrence of 1-itemsets that contains just a single thing by including every thing in data set. The recurrence of 1-itemsets is utilized to find the itemsets in 2-itemsets which thus is utilized to find 3-itemsets, etc until there are no more k-itemsets. If an itemset isn't regular, any enormous subset from it is additionally non-continuous; this condition prune from search space in data set.

IV. EXPERIMENTAL RESULTS

We employed a real-world supermarket transaction dataset obtained from a UCI dataset, supermarket chain containing 4627 purchase records and 217 attributes [8]. The dataset included transactional information such as customer ID and the items purchased. Before applying the Apriori algorithm, we conducted data preprocessing steps, including removing duplicate records, handling missing values, and encoding categorical variables. We set the minimum support to 0.05 and the minimum confidence to 0.5, based on empirical observations and domain knowledge. The experiment was performed on a machine with a 2.5GHz processor and 8GB of RAM. We utilized Weka software for Apriori algorithm implementation. The supermarket dataset information is summarized in the figure-1.

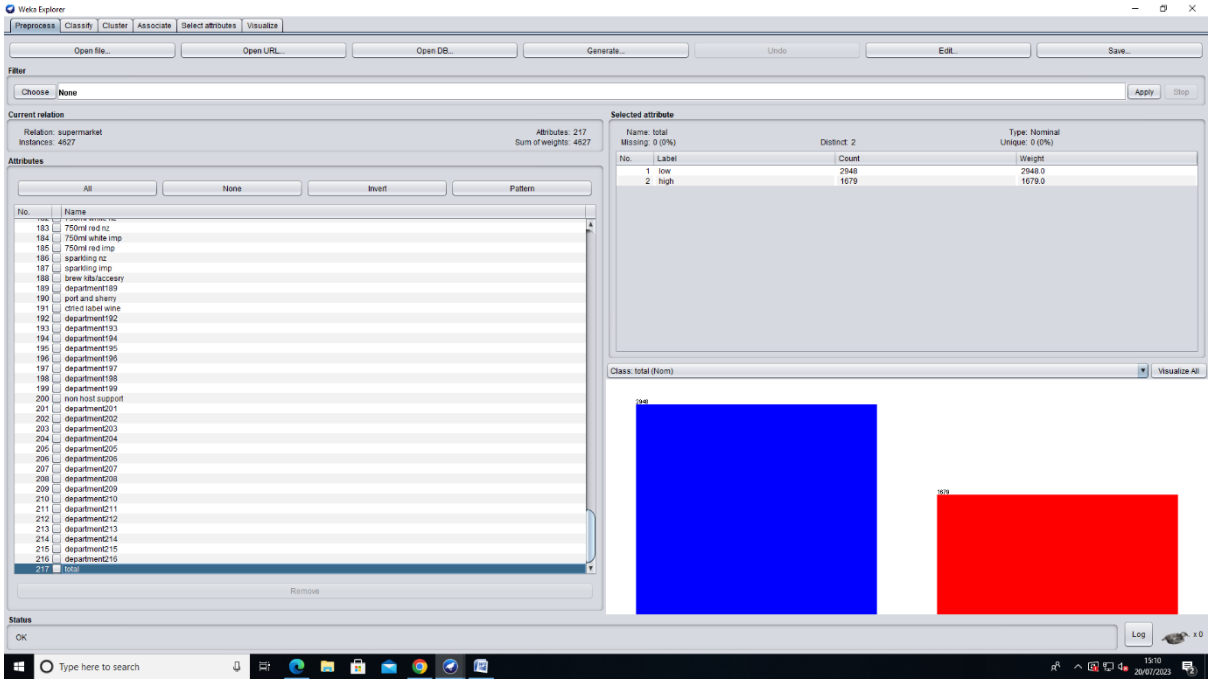

Figure-1: Statistical Summary of the dataset

V. RESULTS AND DISCUSSION

The Apriori algorithm was applied to mine association rules in a supermarket transaction dataset with a minimum support of 0.3 (1388 instances) and a minimum confidence of 0.8. A total of 14 cycles were performed during the execution of the algorithm.

The algorithm generated sets of large itemsets of varying sizes:

Size of set of large itemsets L(1): 25

Size of set of large itemsets L(2): 69

Size of set of large itemsets L(3): 20

The results of the analysis revealed several interesting association rules that could be of great significance to the supermarket's management. Among the rules, the top 10 rules with the highest confidence values are presented below:

1) If a customer purchases biscuits and vegetables, there is an 84% chance they will also buy bread and cake in the same transaction. This association has a lift value of 1.17, indicating a positive correlation between these items.

2) When the total purchase value is high (total=high), there is an 84% probability that the customer will purchase bread and cake as well. This association also has a lift value of 1.17.

3) Customers who buy both biscuits and milk-cream have an 84% likelihood of purchasing bread and cake in the same transaction, with a lift value of 1.17.

4) If a customer buys biscuits and fruit together, there is an 84% chance they will also purchase bread and cake. The lift value for this association is 1.17.

5) When customers purchase biscuits and frozen foods together, there is an 83% probability of them also buying bread and cake. The lift value for this association is 1.16.

6) Customers who buy frozen foods and fruit together have an 83% chance of purchasing bread and cake in the same transaction, with a lift value of 1.16.

7) If a customer purchases frozen foods and milk-cream together, there is an 83% chance they will also buy bread and cake. The lift value for this association is 1.15.

8) When customers buy baking needs and milk-cream together, there is an 83% probability of them also purchasing bread and cake. The lift value for this association is 1.15.

9) Customers who buy milk-cream and fruit together have an 83% likelihood of purchasing bread and cake in the same transaction, with a lift value of 1.15.

10) If a customer purchases baking needs and biscuits together, there is an 83% chance they will also buy bread and cake. The lift value for this association is 1.15.

These high-confidence association rules indicate strong relationships between certain products, suggesting potential opportunities for targeted promotions and improved product placement strategies to enhance customer satisfaction and maximize sales.

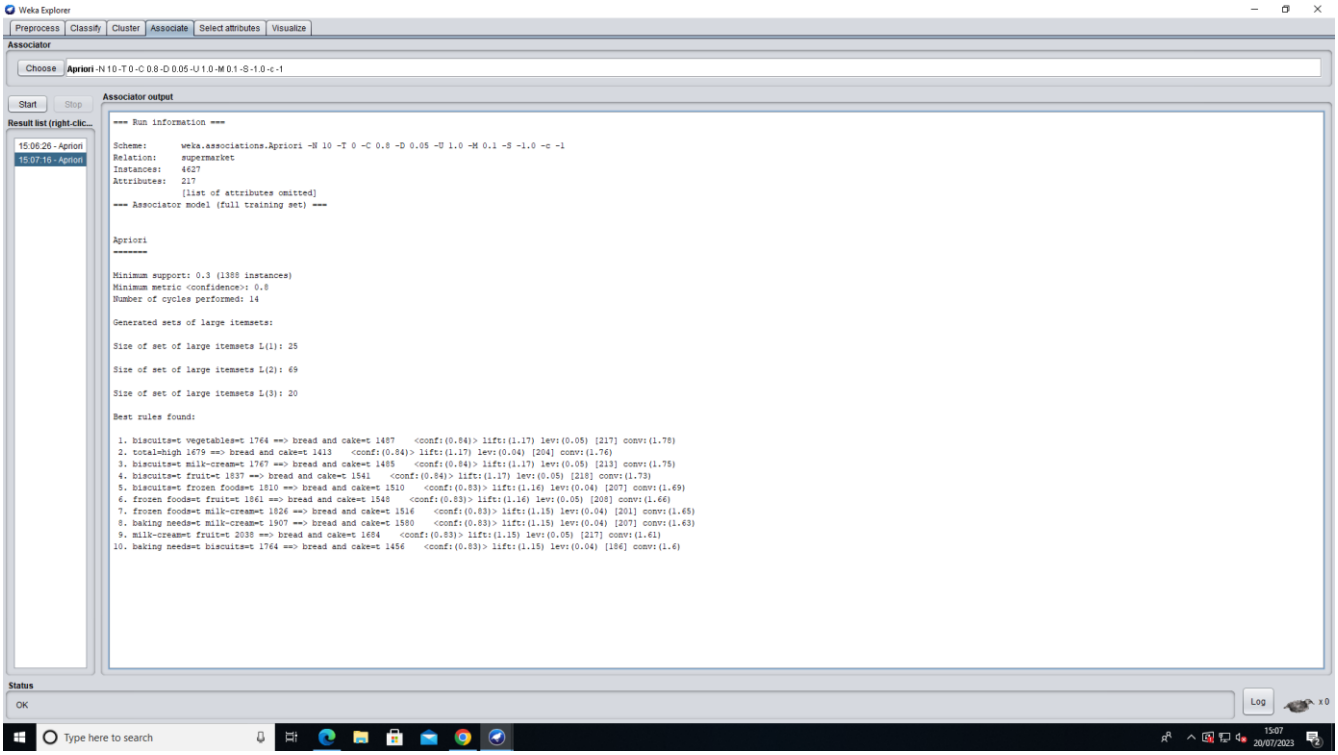

Figure-2: Screen shot of Experimental Results

VI. CONCLUSION

In conclusion, our study demonstrates the effectiveness of the Apriori algorithm in extracting association rules from a supermarket transaction dataset. The discovered rules shed light on purchasing behavior and uncover valuable product associations, which can guide supermarket managers in decision-making processes. By leveraging these insights, supermarkets can optimize product placement and promotions, leading to increased sales and customer satisfaction. In the future, it would be beneficial to explore more advanced association rule mining techniques and consider other relevant factors like customer demographics and seasonal variations to further enhance the analysis's accuracy and utility.

REFERENCES

[1] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[2] F. H. AL-Zawaidah, Y. H. Jbara, and A. L. Marwan, “An Improved Algorithm for Mining Association Rules in Large Databases,” Vol. 1, No. 7, 311-316, 2011

[3] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.

[4] J.Han and M.Kamber,”Data Mining concepts and Techniques”, the Morgan Kaufmann series in Data Management Systems, 2nd ed.San Mateo, CA; Morgan Kaufmann, 2006.

[5] N.Michael, “Artificial Intelligence – A Guide to Intelligent Systems”, 2nd Edition, Addison Wesley 2005

[6] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” in ACM SIGMOD Record, vol. 22, pp. 207–216, 1993

[7] R. Srikant, “Fast algorithms for mining association rules and sequential patterns,” UNIVERSITY OF WISCONSIN, 1996.

[8] UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/ datasets.