We declare that we have completed this assignment completely and entirely on our own, without any consultation with others. We have read the UAB Academic Honor Code and understand that any breach of the Honor Code may result in severe penalties.

We also declare that the following percentage distribution *faithfully* represents individual group members’ contributions to the completion of the assignment

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Overall Contribution (%)** | **Major Work Items Completed by Me** | **Signature or Initials** | **Date** |
| Venkata Syam Naga Pavan | 20% | Collecting data and pre-processing by including data cleaning and transforming the binary matrices and also implemented the source code part of the project | Sp | 12/5/24 |
| Maruthi &Vijay | 40% | Implementation of Apriori algorithms for associations and also implemented the logistic regressions by analyzing of its limitation and confusion matrix generation | Maruu  &  Vijay | 12/5/24 |
| Sujith &Nishanth | 40% | After implementation decision tree classifier implementation and evaluations by generating the confusion matrices with the logistic regression by drafting discussions, conclusions and insights for the projects by suggesting the future work and practical applications of the project | Sujith  &  Nishanth | 12/5/24 |
|  |  |  |  |  |
|  |  |  |  |  |

**Context-Aware Market Basket Analysis (C-MBA)**

Sajja Pavan Maruthi Samudrala Nishanth Boll

[vsajja@uab.edu](mailto:vsajja@uab.edu) [vsamudra@uab.edu](mailto:vsamudra@uab.edu) [nbolline@uab.edu](mailto:nbolline@uab.edu)

Vijay Deepak Sujith Madhusudan

[vkarnata@uab.edu](mailto:vkarnata@uab.edu) [madhusus@uab.edu](mailto:madhusus@uab.edu)

**Abstract**

This paper explores the concept of market basket analysis, using the Apriori algorithm for association rule mining in combination with machine learning classification models such as Logistic Regression and Decision Tree Classifier. The primary goal is to discover interesting item associations in transactions and use machine learning techniques to rank the discovered rules according to their confidence values. The approach entails pre-sorting data to obtain fre­quent item sets, deriving usable association rules, and using patterns to make these insights clearer.

# **1. Introduction**

Customers’ buying behaviour plays a crucial role when it comes to the competition in the retail sector as well as the use of various marketing techniques and storage of inventory. Market Basket Analysis or MBA is a highly useful method that describes associations between products that are purchased together by consumers and may give detailed information about a consumer’s behaviour. MBA makes transactions transparent to help retailers develop promotions, arrange products

optimally, and seek strategies for increasing sales as well as customers’ satisfaction. This work uses the well-known Apriori algorithm that is used to identify association rules between items.

# **2. Data Collection and Preparation**

## **2.1 Data Collection**

The dataset consists of transactional records representing items purchased in individual transactions. Each row corresponds to a unique transaction, detailing the purchased items. The dataset was meticulously cleaned to ensure that only meaningful and valid records were retained for analysis. Transactions with null values, missing information, or anomalies were excluded.

Below is a brief description of the columns based on the data provided:

* **Sno**: Serial number, likely an index or unique identifier for each row.
* **Invoice No**: Invoice number for the transaction.
* **StockCode**: A unique code representing each product.
* **Description**: Text description of the product.
* **Quantity**: Number of items purchased in this transaction.
* **InvoiceDate**: Date (and possibly time) of the transaction.
* **UnitPrice**: Price per unit of the product.
* **CustomerID**: Identifier for the customer who made the purchase.
* **Country**: Country where the transaction took place.

## **2.2 Data Preprocessing**

Data preprocessing involved multiple steps to prepare the transactional data for mining and predictive modelling:

* **Data Cleaning:** The dataset was cleared of null values, missing entries and duplicate rows as a means of improving its quality. On top of that, the baskets were arranged whereby a basket was a single transaction (Singh et al. 2020).
* **Data Transformation:** Data labels are continuing, so in order to make transactions more understandable, they were converted to a binary matrix into by using the one hot technique. This approach transformed each transaction a vector of 0 and 1 where each item in the basket was reflected by a 1 if present in the transaction basket and 0 if not. This transformation was useful in making the data more appropriate for analysis with the Apriori algorithm.

## **2.3AprioriAlgorithm Implementation**

In this study, the Apriori algorithm is used for the identification of a frequent itemset as well as for the generation of association rules. The following parameters were set:

* Minimum support: 0.01 to take into account only those itemset which are occurred in 1% transactions or more.
* Minimum confidence: 0.2, choosing rules that have at least 20% predictive strength.
* Minimum lift: 1.0, this made rules highly predictive rather than being based on chance.

These thresholds were chosen to see if rules can meaningfully and efficiently be generated within certain limitations, while taking into account computational complexity. The list of generated association rules was valid for further analysis, and acted as a core of the prediction models (Edastama et al. 2021).

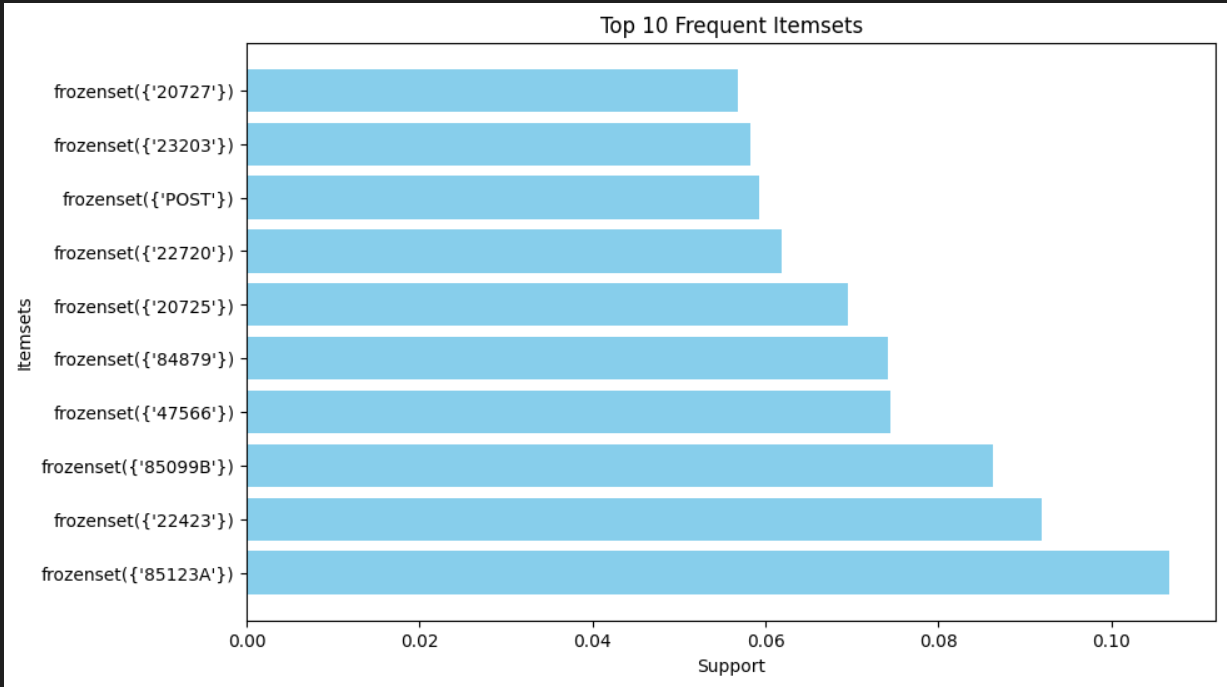


Figure 1: Frequent itemset

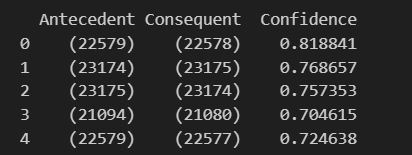


Figure 2: Rules for confidence

# **3. Prediction Models and Results**

To enhance the utility of association rule mining, predictive modelling was employed using two classification algorithms: A binary classification analysis will include logistic regression and the decision tree classifier techniques. These models sought to estimate the probability of the association rules to be of high confidence, where confidence refers to rules with a score of greater than 0.8. The data set contained association rules paired with their confidence and support levels, and a target variable which was either high or low confidence.

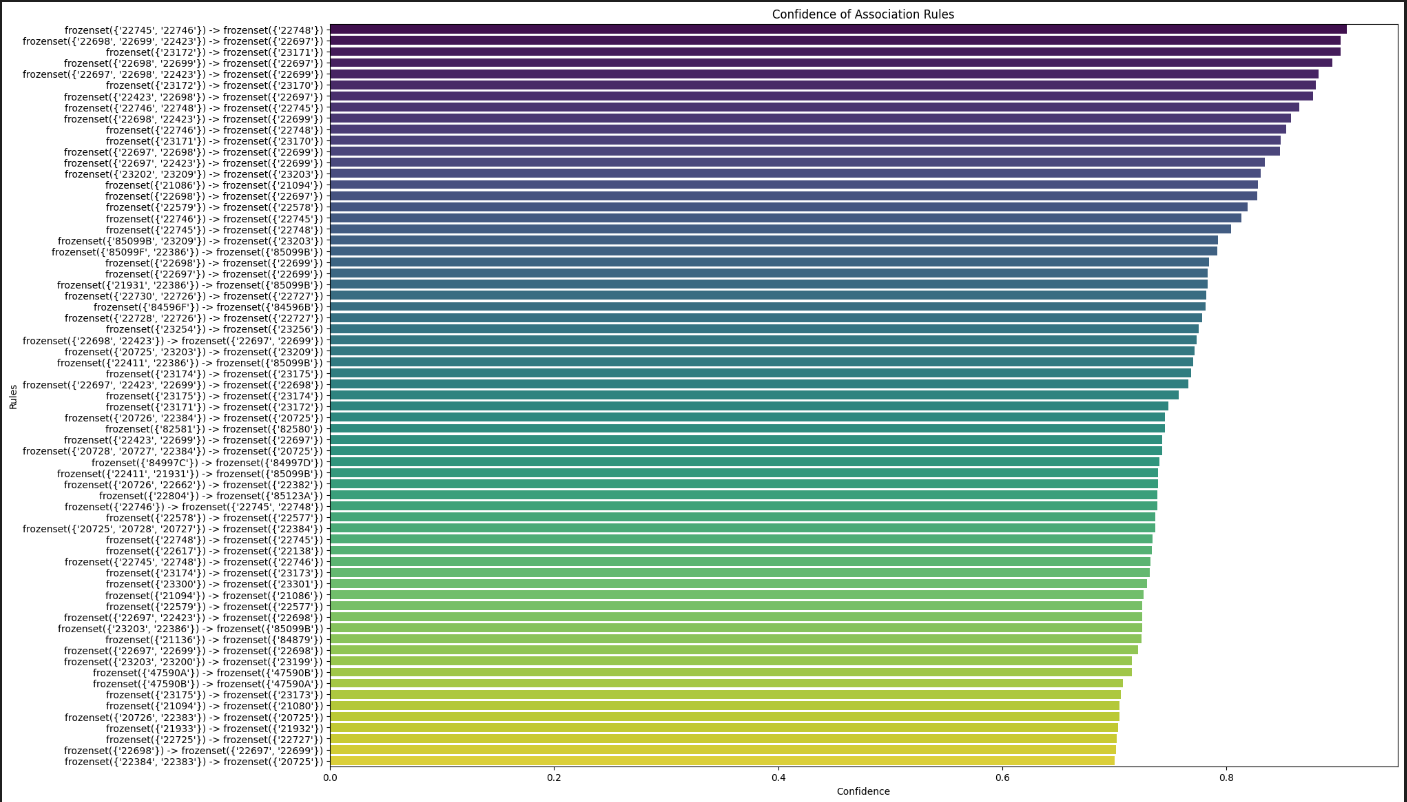


Figure 3: Confidence of association rules

## **Logistic Regression Model**

The Logistic Regression model achieved an overall accuracy of **78.57%**, suggesting a fair ability to differentiate between high- and low-confidence rules. The confusion matrix revealed the following:

* **True Positives (TP):** 11
* **False Positives (FP):** 0
* **False Negatives (FN):** 3
* **True Negatives (TN):** 0

While the model performed well in predicting high-confidence rules, it struggled to correctly identify low-confidence ones, as indicated by the absence of true negatives and the presence of false negatives. This bias towards the positive class implies that the Logistic Regression model may not effectively capture the more complex relationships in the dataset (Zhang et al. 2021).

## **Decision Tree Classifier Model**

The Decision Tree Classifier demonstrated superior performance, achieving a **perfect accuracy of 100%**. Its confusion matrix was as follows:

* **True Positives (TP):** 11
* **False Positives (FP):** 0
* **False Negatives (FN):** 0
* **True Negatives (TN):** 3

Unlike Logistic Regression, the Decision Tree correctly classified all instances, making no misclassifications. Decision Trees excel at identifying patterns and thresholds in datasets, making them particularly suitable for classification tasks involving rule-based data (Charbuty et al. 2021).

## **Visualization Analysis**

The performance of both models was visualized using confusion matrices:

1. **Logistic Regression Confusion Matrix:** Highlighted accurate predictions for high-confidence rules but notable misclassification of low-confidence ones (false negatives).
2. **Decision Tree Confusion Matrix:** Demonstrated flawless performance with no misclassifications.

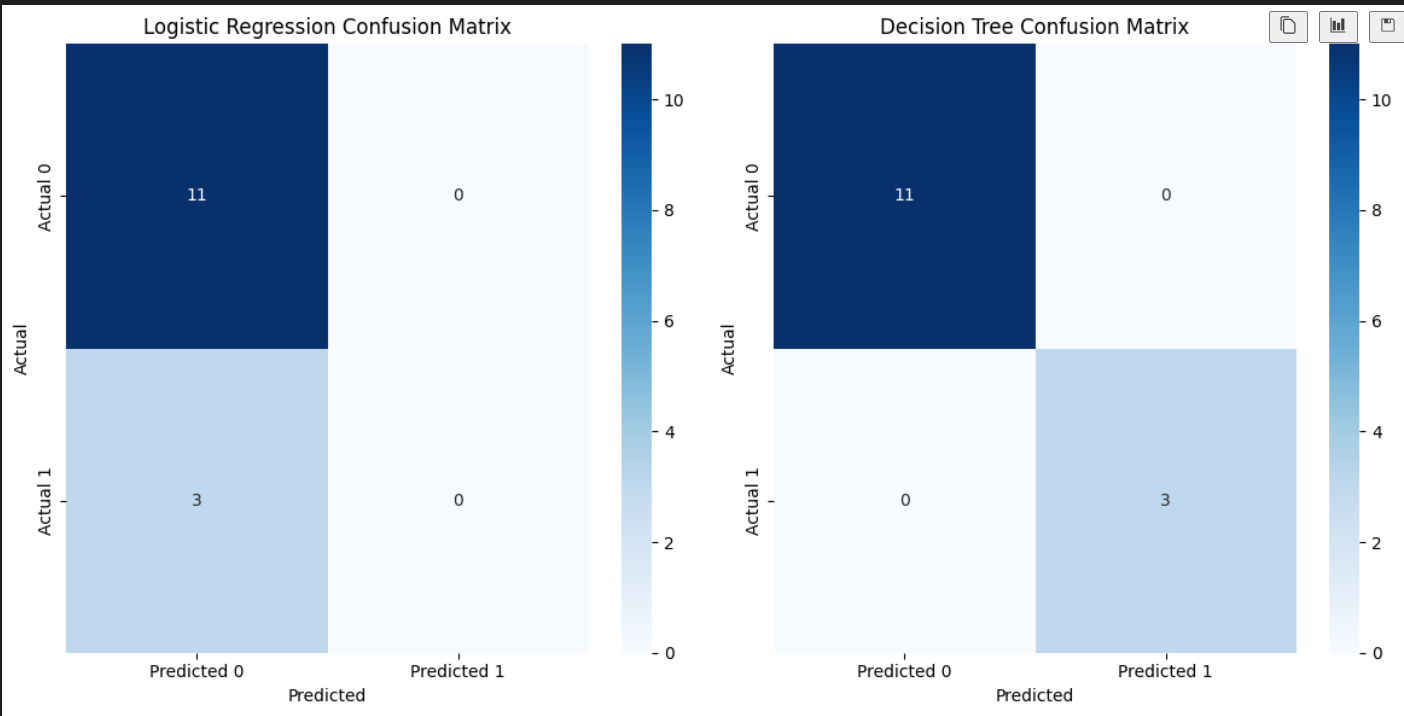


Figure 4: Confusion matrix

## **Interpretation of Results**

The significant difference in the results of the two models on the dataset used shows that it is necessary to choose a suitable algorithm given the characteristics of the input data. Although Logistic Regression offers simplicity and interpretability in the model, its restriction on non-linear relationship makes it less suitable for this purpose. On the other hand, the Decision Tree Classifier was more flexible and versatile in its application and the results yielded by it were ideal as compared to the former.

**Input features/output**

The input features which include transactional data details such as quantities, customer IDs, and transformed into a binary matrix for analysis.

Moreover, output was to predict the high-confidence associations rules with a focus on identifying patterns in customer purchasing behaviours of the customers in order to assist retailers in decision-making and the optimizing outputs.

# **4. Discussions and Conclusions**

## **4.1 Key Findings**

The analysis revealed critical insights into item associations and the efficacy of prediction models:

* Apriori Algorithm Performance: Analysing the results of the Apriori algorithm, many important item associations were given, like {milk, bread} →{butter} with good confidence and lift. These associations identify prospective buying behaviours which may prove helpful with regard to retail decisions.
* Machine Learning Classification Models: The Logistic Regression was trained with an accuracy of 78.57 % whereas the Decision Tree model tested with an accuracy of 100% on the test set. From the confusion matrix, it was evident that the Decision Tree did not misclassify any of the data making it superior in producing high confidence association rules.

## **4.2 Insights for Retailers**

The findings have several practical implications for retail operations:

* Inventory Optimization: The common item pairs help a retailer to restock products and avoid situations where customers demand products that are not available or are stacked with products they rarely buy. For instance, if {milk, bread} associates with {butter}, it will help the retailers to keep these items together to facilitate purchases.
* Personalized Marketing: High lift value association rules can also be used in creating marketing initiatives, for example, joint sales promotions or promotions. These efforts can boost customer satisfaction and trigger further sales, as experience has shown. For example, promoting {bread, jam} as a set can win the loyalty of consumers and boost sales (Poli et al. 2023).

**Experimental setting**

Two models which we have used Logistic Regression and Decision tree classifier are trained to associate the rules in to high and low confidence rules in order to indicates the superior performance.

## **4.3 Limitations**

Despite the valuable insights, the study had limitations that warrant consideration:

* **Dataset Size and Scope:** The analysis was conducted on a relatively small and static dataset. This limits the generalizability and scalability of the findings to larger, dynamic datasets often encountered in real-world retail settings.

# **5. Future Work**

The potential for further enhancing the current approach to market basket analysis (MBA) is vast. Below are several avenues for future work:

## **5.1 Real-time Market Basket Analysis**

Another prospective direction is applying real-time data streaming approaches that can support on-the-fly rule generation and evaluation. In the constantly changing today’s environment, virtual MBA, instant at Apache Kafka or Spark Streaming, can be applied to emerging ones to respond simultaneously to the customer’s behaviour. For example, real-time analysis can be used to recommend related products during the e-shopping processes to make consumers’ experiences more relevant (Zamil et al. 2020).

## **5.2 Advanced Techniques**

To push the boundaries of traditional MBA, advanced methodologies can be explored:

* Deep Learning Models: The neural network, and in particular, the recurrent and convolutional neural networks, can reveal more complex, non-linear dependencies between items. These models could incorporate temporal properties, thus providing more dynamism that corresponds with the changing customer behaviour at different time stamps (Ghous et al. 2023).
* Improved Visualization Tools:
* Organizing special informative displays can be helpful in delivering the data to the stakeholders in an engaging and easy to understand format. Tableau or Power BI could be improved with additional custom visuals, like dynamic network graphs showing association rules and heatmaps for itemset (Alawadh et al. 2022).

## **5.3 Broader Applications**

As much as MBA has its usefulness in retail, it is not limited to this aspect of business. In e-commerce, it can help to enhance recommendations based on the user’s preferences, clicks, or purchase history. Likewise, in medicine, MBA can search for patterns like diseases that often accompany each other or medications that should not be used concurrently. Thus, MBA used in connection with apply of these techniques in the various spheres permitted to eliminate shortcomings observed in a decision-making process and to increase the efficiency of operation.

# **Acknowledgement**

We express our gratitude to the open-source community for providing the dataset used in this study. The transactional dataset was sourced from GitHub, a platform that fosters collaboration and knowledge sharing. The availability of this dataset enabled the exploration of frequent pattern mining and predictive modelling, contributing significantly to the study's outcomes.

# **References**

Singh, S.K. and Dwivedi, D.R.K., 2020. Data mining: dirty data and data cleaning. *Available at SSRN 3610772*.

Nguyen, Q.H., Ly, H.B., Ho, L.S., Al-Ansari, N., Le, H.V., Tran, V.Q., Prakash, I. and Pham, B.T., 2021. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, *2021*(1), p.4832864.

Edastama, P., Bist, A.S. and Prambudi, A., 2021. Implementation of data mining on glasses sales using the apriori algorithm. *International Journal of Cyber and IT Service Management*, *1*(2), pp.159-172.

Charbuty, B. and Abdulazeez, A., 2021. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), pp.20-28.

Poli, N.S. and Sikder, A.S., 2023. Predictive Analysis of Sales Using the Apriori Algorithm: A Comprehensive Study on Sales Forecasting and Business Strategies in the Retail Industry.: Predictive Analysis of Sales Using the Apriori Algorithm. *International Journal of Imminent Science & Technology*, *1*(1), pp.1-16.

Zamil, A.M.A., Al Adwan, A. and Vasista, T.G., 2020. Enhancing customer loyalty with market basket analysis using innovative methods: a python implementation approach. *International Journal of Innovation, Creativity and Change*, *14*(2), pp.1351-1368.

Ghous, H., Malik, M. and Rehman, I., 2023. Deep Learning based Market Basket Analysis using Association Rules. *KIET Journal of Computing and Information Sciences*, *6*(2), pp.14-34.

Alawadh, M.M. and Barnawi, A.M., 2022. A survey on methods and applications of intelligent market basket analysis based on association rule. *Journal on Big Data*, *4*(1).