Association Rule Mining for Customer Purchase Behaviour Analysis in Restaurant Sales

Narayan Ravikumar
Department of Computer Science
University of Exeter
Exeter, UK
nr433@exeter.ac.uk

Ronaldo Menezes
Department of Computer Science
University of Exeter
Exeter, UK
R.Menezes@exeter.ac.uk

***Abstract*—This study explores customer purchasing behaviour and product associations in the restaurant industry using data mining techniques. Leveraging transaction data from ContaHUB, a multi-tenant ERP system for restaurants in Brazil, exploratory data analysis was conducted extensively and applied association rule mining algorithms. The research compares the performance and effectiveness of Apriori and FP-Growth algorithms in identifying significant product associations. The findings reveal that whilst both algorithms produce similar quality results, FP-Growth demonstrates superior computational efficiency, particularly for large-scale datasets. A comprehensive hyperparameter tuning analysis was conducted, highlighting the critical role of support and confidence thresholds in balancing the discovery of rare associations with computational practicality. The study uncovers several strong product associations, especially in the beverage category, providing actionable insights for stock management, restaurant layouts, and targeted marketing strategies. Additionally, the study presents a case study demonstrating the applicability of the developed analytical pipeline to individual restaurant data, showcasing its potential for personalised inventory management and sales strategies. This research contributes to the field of restaurant analytics by providing a robust framework for customer purchasing behaviour analysis and product association mining, with significant implications for enhancing operational efficiency and customer satisfaction in the restaurant industry.**

***Index Terms*—market basket analysis, association rule mining, Apriori algorithm, FP-Growth algorithm, customer purchasing behaviour, restaurant analytics, data mining, product associations, hyperparameter tuning, inventory management**

I. Introduction

In today's competitive global business environment, it is critical to have a thorough understanding of customer buying behaviour in order to drive business success. This is more relevant in sectors like retail and the restaurant business, where consumers have numerous options every day and their preferences are constantly changing based on the options they have. These market forces have forced businesses to increasingly turn to data-driven strategies to gain insights into consumer preferences and their preferred choices, purchasing patterns, and decision-making processes [1]. Globally, businesses leverage these insights to drive effective marketing and sales strategies, inventory management, and overall business optimisation.

Interestingly, a wide range of influencing factors, including psychological, social, and economic variables [2], drive customer buying behaviour. These factors show up as sales patterns for these businesses, which can provide meaningful data for businesses to analyse and interpret. These sales patterns can uncover trends in offer popularity, seasonal variations in demand, and how specific marketing campaigns can influence consumer choices. A thorough examination of these patterns can help businesses adapt their approaches to meet customer needs, which in turn leads to enhanced customer satisfaction and loyalty.

In this context, one of the most effective tools which has emerged for analysing customer buying behaviour is Market Basket Analysis [3]. This tool, underpinned by data mining and machine learning techniques, helps to discover associations between products that customers tend to buy together. The paramount assumption of Market Basket Analysis is that by discovering these associations, businesses can reveal hidden patterns in consumer behaviour that may not be immediately visible through traditional methods.

Foundational to Market Basket Analysis is the association rule mining, a set of algorithms designed to discover frequent item sets within large datasets and generate rules that define the relationships between these items. These rules are generally in the form of "if-then" statements, such as "if a customer buys product X, they are likely to also buy product Y" [4]. The strength of these associations is defined using metrics like support, confidence, and lift, which explain the different aspects of significance, authenticity, and accuracy of these identified patterns [5].

Though relevant to all industries, Market Basket Analysis finds special significance in the restaurant industry. The world of restaurant service is often dynamic and changing, so understanding customer buying behaviour and patterns is crucial to business success. The restaurant industry can often leverage these techniques to refine their menus, offer popular food combinations, and craft compelling promotional strategies [6]. For example, by identifying and offering dishes that are frequently ordered together, a restaurant can market a special combo to drive additional revenue and margins.

In addition, the insights and the understanding gained from market basket analysis can be leveraged to drive operational efficiency in a highly competitive industry like restaurants. For example, restaurants can identify which ingredients are often

used together in their best-selling dishes, which in turn can be used to optimise their inventory management and reduce waste [7]. These insights can also be helpful to train staff to make the right suggestions based on their ordering patterns, leading to increased average order sizes and customer satisfaction.

The importance of research in this field is extremely crucial. As businesses continue to gather vast amounts of transactional data on a day-to-day basis, the ability to distil meaningful insights from this also becomes highly relevant. Deeper research into customer buying behaviour, sales patterns, and market basket analysis provides businesses with the ammunition and methods to operate in a data-intense environment effectively [8]. It also fills the gap between unprocessed data and actionable strategies, helping businesses to make informed decisions, leading to increased sales and margins.

Ongoing and continued research in this area is important to get aligned with the ever-changing shifts in consumer behaviour, technological advancements, and social benchmarks. New patterns and associations may be discovered, forcing businesses to revise their strategies and approaches towards success. Meaningful research in this field also helps in developing predictive models which can anticipate future changes in consumer buying behaviours [9].

To conclude, the continued study of customer buying behaviour, sales patterns, and market basket analysis, underpinned by methods of association rule mining, represents a crucial area of research which can provide huge benefits for restaurant businesses. The insights and understanding of the complex network of consumer choices and preferences can provide this industry with the knowledge needed to optimise their operations and efficiency, improve customer satisfaction, and eventually help them shine in a competitive marketplace. As businesses continue to generate and accumulate more data than ever before, the relevance of such research in translating this sea of information into practical, actionable business strategies can never be overstated [10].

II. Literature Review

There has been a significant evolution in the field of data mining since its inception in the late 20th century, as association rule mining (ARM) has emerged as a decisive tool to discover hidden patterns in large datasets. ARM was pioneered by Agrawal et al. in 1993, and over the years, it has provided vital support in analysing consumer buying behaviour and creating actionable business strategies [11].

The Apriori algorithm was introduced by Agrawal and Srikant in 1994, marking a significant landmark in ARM [12]. This algorithm was designed to identify frequent itemsets by leveraging the anti-monotonicity property, which means that any subset of a frequent itemset must also be frequent. While this was a transformational finding in ARM, its performance on large datasets was found to be limited due to multiple database scans and the generation of a large number of candidate itemsets.

In order to work around these performance limitations, Han et al. suggested the FP-Growth (Frequent Pattern Growth) algorithm in 2000 [13]. FP-Growth deploys a divide-and-conquer strategy and a much more compact data structure called the FP-Tree. This method phases out the need for generating candidates and reduces the number of database scans to two, disregarding the dataset size. Therefore, FP-Growth is often seen outperforming Apriori, especially when dealing with dense, large-scale datasets.

It is often seen that parameter tuning, particularly the minimum support (min_support) and minimum confidence (min_confidence) thresholds, can significantly influence the efficiency of ARM algorithms. The study conducted by Tan et al. on various interestingness measures for association rules provided further insights and understanding into the selection of appropriate thresholds [14]. This study displayed the trade-offs between identifying rare but potentially valuable associations and maintaining computing efficiency.

As datasets accumulate with increasing complexity, studies have focused on adapting ARM algorithms for big data environments. Li et al. suggested a MapReduce-based parallel FP-Growth algorithm, proving its scalability on very large datasets [15]. Furthermore, Moens et al. proposed a distributed version of Apriori using the Spark framework, demonstrating enhanced performance on large-scale transactional data [16].

There have been a number of studies on the application of ARM in understanding consumer buying behaviours and patterns. Chen et al. deployed association rules to analyse cross-category correlations in buying behaviour, predicting insights into product affinities and potential cross-selling opportunities [17]. This work highlighted the relevance of considering both positive and negative associations in developing business strategies.

Tuning parameters like min_support and min_confidence are important in identifying meaningful rules from customer transaction data. Fournier-Viger et al. suggested a flexible and adaptive algorithm that automatically adjusts these parameters based on dataset characteristics, addressing the concerns of parameter selection in diverse retail environments [18].

There is also a strong connection between the evolution of ARM algorithms and technological advancements in computing power and data storage capabilities. In the past, while implementations were constrained by memory and processing speed, modern systems allow for much more sophisticated approaches. For example, the Eclat algorithm introduced by Zaki uses a depth-first search strategy and set intersection operations to efficiently mine frequent itemsets [19]. This vertical data format approach has proven to yield much better results, especially for sparse datasets.

The relevance of ARM and its associated algorithms cuts across multiple industries such as retail, restaurants, e-commerce, and healthcare. In healthcare, Stilou et al. used ARM techniques to examine electronic health records, revealing hidden patterns in patient diagnoses and treatments [20]. Thus, ARM has proven to be quite adaptive and versatile in extracting valuable business insights from complex, multi-dimensional data.

There has been recent work done on including temporal

aspects into ARM. Ale and Rossi suggested a temporal association rule mining framework that considers the lifespan of items and transactions, helping to discover time-sensitive patterns in consumer behaviour [21]. This approach adds a new element to traditional ARM, suggesting methods to capture seasonal trends and changing customer preferences.

The importance and relevance of efficient and scalable ARM algorithms have grown with the volume and variety of data being collected. New and modern approaches are being deployed, such as quantum computing-based algorithms for ARM, as demonstrated by Aimeur et al. [22]. These techniques have proven to push the boundaries of what is achievable in mining association rules from very large datasets.

III. Aims & Objectives

A. Research questions

- How can transaction data be leveraged to identify significant product associations and purchasing patterns in a retail or restaurant setting?
- What are the relative strengths and limitations of different association rule mining algorithms (such as Apriori and FP-Growth) when applied to large-scale transaction datasets?
- How do various hyperparameters affect the quality and practicality of generated association rules in a retail context?
- How can a scalable analytical pipeline be developed that can be customised for different business units or branches within a retail or restaurant chain?

B. Objectives

This research paper intends to generate a structured data preparation and cleaning pipeline to process large-scale data, mostly transactional data so that it will be ready as high quality source for subsequent analysis. The study also conducts an exploratory data analysis to discover key trends in sales and revenue patterns, product popularity and transaction characteristics, providing a base for deeper analytical studies.

The research study also will do a meaningful comparison between multiple association rule mining algorithms mostly leveraging Apriori and FPGrowth so that product associations within customer buying patterns can be identified. A framework for hyper parameter tuning will also be generated, enabling the optimisation of support, confidence and lift thresholds in association rule generation.

The intend is also to design and develop a suite of visualisation tools so that the association rules and the sales patterns can be communicated in a simple way to all stakeholders. This will help the data driven decision making process much more effective. A flexible analytical pipeline will also be created so that data from different business units and during specific time periods can be used to create targeted insights and recommendations.

The practical impacts and benefits of identified association rules on inventory management and sales strategies will be assessed, so that actionable business recommendations can be optimised for inventory management and product placement. The computing capability and scalability of the association rule mining approaches will be also be assessed so that real time decisions can be taken in a live industry production environment.

IV. Methodology

A. Data Acquisition & Preparation

During the Data Acquisition & Preparation phase, the research leverages multiple tables from ContaHUB, a multi-tenant ERP system focused on restaurant industry in Brazil. These tables include Catalogo (product catalogue information), Compra (purchase records), CompraItem (detailed purchase item data), Produto (product details as handled by restaurants), Fornecedor (supplier information), GrupoProduto (product group classifications), Usuario (user data), Venda (sales records), and VendaItem (detailed sales item data).

The database has a multi-tenant structure where most tables have multi-column primary keys, usually starting with 'Emp' (Company) followed by a tablespecific prefix. As an example, the 'Compra' table has a primary key of 'Emp' + 'Cmp'. Standard Data types were used, with particular focus to date fields being converted to a format of datetime. A custom made utility was implemented to handle values which were absent and that ensure a level of data consistency across all datasets.

Key datasets were merged during the process so that a single unified a unified dataset for sales could be created. This involved joining 'Venda', 'VendaItem', and 'Produto' tables through their respective primary and foreign keys.

B. Exploratory Data Analysis (EDA)

An aggregation of daily sales data was performed so that some potential temporal patterns could be identified. The visualisation techniques used included line plots to display sales trends over a timeline.

Product frequencies were calculated and analysed and thereby a threshold was established to differentiate between products of high-frequency sale and low-frequency sale. Bar plots were generated so that the distribution of product frequencies could be visualised for easier display.

Metrics like average basket size and items per transaction were calculated using the data and histograms were created with kernel density estimation which could be utilised to visually display the distribution of these characteristics.

Then the relationship between different products was analysed and a heat map of this was visualised. A histogram of the top ten product pairs was also created to display this visually for easier understanding.

C. Association Rule Mining

Data transformation was conducted to convert sales data into a transaction format, where each transaction (Venda) was represented as a list of products. There was a mapping exercise done to map Product IDs to product names for a better level of interpretation.

Apriori Algorithm (leveraging a bottom-up approach, generating candidate itemsets and testing them against the data) [23] and FP-Growth Algorithm (deploying a divide-and conquer strategy, using a compact data structure - FP-tree) [24] were both developed, implemented for the datasets and then compared.

Rules were developed from frequent itemsets, with the below given metrics computed: Support (frequency of itemset occurrence), Confidence (conditional probability of consequent given antecedent), and Lift (ratio of observed support to expected support if items were not dependent).

A grid search approach was used to discover the most optimal level of minimum support and confidence thresholds. Overall performance was evaluated based on the number and quality of rules generated.

D. Performance Analysis and Visualisation

The times for execution of both the algorithms, Apriori and FP-Growth were measured and then a comparative study done. Bar plots were used to visually compare the differences in performance.

Kernel density estimations were plotted to compare the lift distributions. The metrics of support vs. confidence were also plotted on a scatter chart to display the rule characteristics. The high impact top rules were extracted based on lift values. These rules were analysed and interpreted from a restaurant industry operations and stock management perspective.

E. Case Study: Restaurant-Specific Analysis

To make the data highly relevant for the specific restaurant names, transactions were filtered by restaurant. The parameters which were optimised through the global analysis were applied to emp-specific data sets and insights drawn. The recommendations made as a result was based on these insights related to individual sales patterns and potential stock optimisation approaches.

V. Results

A. Exploratory Data Analysis

1) Time Series Analysis of Daily Sales: The daily sales graph generated from 2014 to 2024, as shown in figure 1 uncovered some interesting patterns. During 2014-2015, sales showed high variability with notable peaks. This was followed by a elongated period of consistent, low sales from 2016 to early 2021. From 2021 onwards, an positive trend was shown again, although with some fluctuations. The most significant sales peak happened in early 2015.

2) Product Frequency Analysis: While the data of the top 10 selling products was examined, the best seller was found to be 'ACETO BALSAMICO DI MODENA I.G.P PLATINUM', closely followed by 'Babaganuche frasco bergamota' and a downward drop in sales was found after these top performing products, as shown in figure 2. The list includes a broad range of items, including alcoholic beverages, soft drinks, and water products, clearly showing varied consumer preferences.

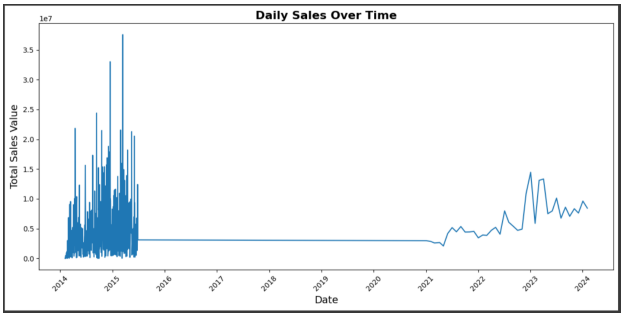

Fig. 1. Daily Sales Over Time

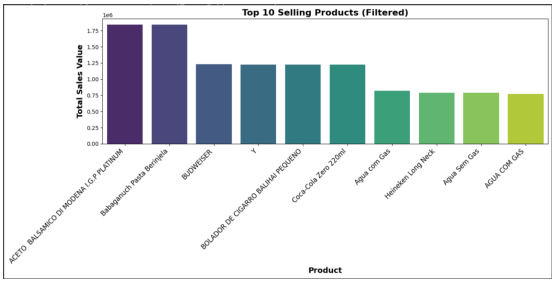

Fig. 2. Top 10 Selling Products (Filtered)

3) Sales by Day of Week: While analysis of weekly sales patterns was done (Figure 3), it showed that Friday was generating the highest sales, closely trailed by Saturday. Sunday consistently showed the lowest sales volume. A general increasing and positive trend is displayed from Monday to Friday, with a slight dip on Saturday before further dipping sales on Sunday. Weekdays (Monday to Thursday) showed relatively stable and positive sales figures.

Fig. 3. Sales by Day of Week

4) Transaction Characteristics: From a transaction characteristics perspective, the pattern of items per transaction showed a skewed pattern, displaying a trend of most transactions involving a small number of items, as shown in figure 4. Most of the transactions had fewer items, though there are very few instances of extremely large transactions.

5) Basket Size Distribution: The filtered basket size distribution actually is similar to the transaction characteristics, displaying a right-skewed pattern. Majority of baskets contain lesser than 1,000 products, with some outlier baskets reaching

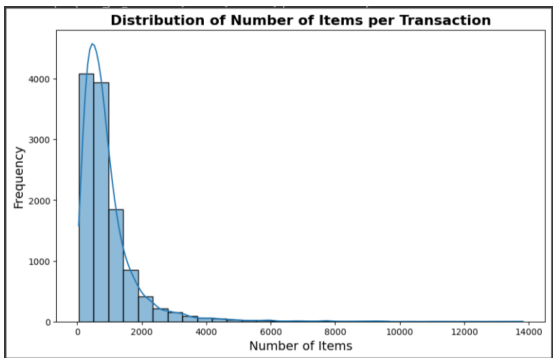

Fig. 4. Distribution of Number of Items per Transaction

up to 6,804 products. The minimum basket size of 47 products showed that a filtering threshold was deployed to highlight analysis on more significant purchases.

6) Product Co-occurrences: A heat map was created to display co-occurrences of the top 20 products and that clearly showed strong associations between certain product pairs. A few products, such as 'Quiabo Fritin' & 'Cachac¸a 51 Insumo', showed high co-occurrence with multiple other items. It was seen that some products were appearing together in transactions as distinct clusters, giving us further insights into complementary product relationships.

7) Top Product Pairs: 'Espumante Brut Fp Prosecco & Revisior' were the two products which were showing as a leading combination of product pairs. It was also seen that there is a gradual decrease in frequency from the top pair to the 10th most frequent pair, as shown in figure 5. A lot of pairs was seen to involve 'Espumante Brut Fp Prosecco', hitting that it's often bought alongside other items, indicating its role as a complement to various products.

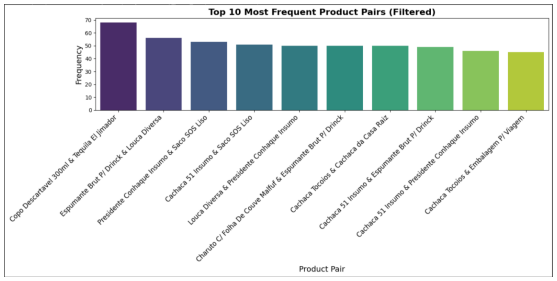

Fig. 5. Top 10 Most Frequent Product Pairs (Filtered)

B. Association Rule Mining

The research analysis did a comparison on two important association rule mining algorithms: Apriori and FP-Growth. These algorithms were then deployed to the retail dataset with the following initial parameters:

- Minimum support: 0.0015
- Minimum confidence: 0.15
- Minimum lift: 2.5

TABLE I
ALGORITHM PERFORMANCE AND RULE GENERATION COMPARISON

Metric	Apriori	FP-Growth
Execution Time (s)	1.46	0.51
Original Rules	34	34
Filtered Rules	26	26

Table I summarises the key performance metrics and rule generation results.

It was clearly demonstrated that the FP-Growth algorithm showed higher level of efficiency, executing at almost three times faster than Apriori. In spite of this performance difference, both algorithms generated an identical number of rules. A filtering process was then used to remove rules with lift values below 2.5, bringing down the number of rules from 34 to 26 for both algorithms.

Figure 6 and Figure 7 illustrate the lift distribution for both algorithms.

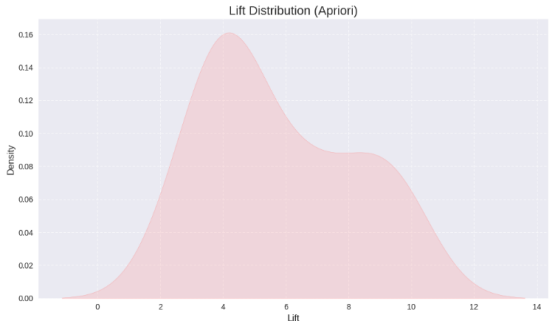

Fig. 6. Lift Distribution for Apriori Algorithms

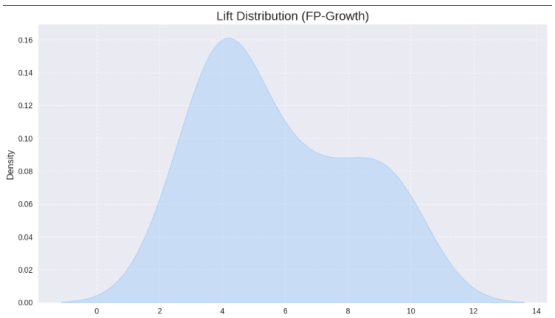

Fig. 7. Lift Distribution for FP-Growth Algorithms

The lift distributions are nearly identical for both algorithms, with the following key statistics:

- Average lift: 5.79
- Highest lift: 9.88
- Lowest lift: 2.60 (due to the minimum lift threshold)
- Standard deviation: 2.38

The above distribution which is largely right-skewed shows that whole most rules have lower lift rules with significantly high lift, again proving strong product associations. The standard deviation of 2.38 also proved significant variability in the strength of associations across the rules.

Figure 8 and Figure 9 present the Support vs Confidence scatter plots for both algorithms.

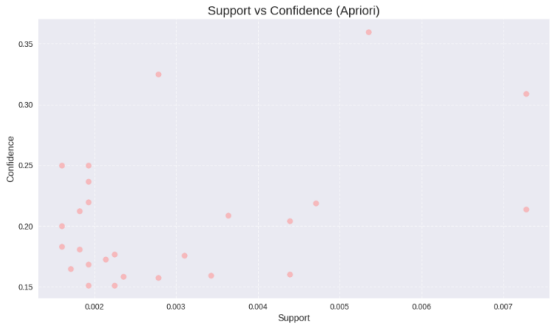

Fig. 8. Support vs Confidence for Apriori Algorithms

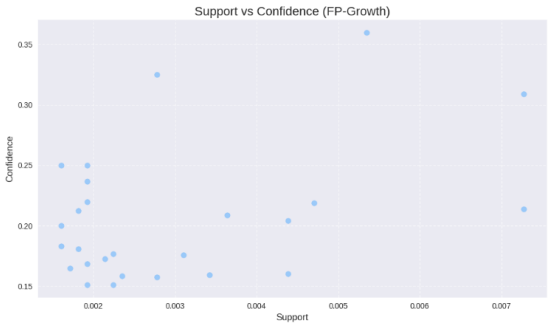

Fig. 9. Support vs Confidence for FP-Growth Algorithms

The plots shows a clustering of rules in the lower left corner, showing generally low support and confidence values. Key observations include:

- Support range: 0.0018 to 0.0073
- Confidence range: 0.15 to 0.36
- Average confidence: 0.21
- Highest confidence: 0.36

The low support values show that these associations occur relatively infrequently, although the associations are strong, in the Fig. 9. Support vs Confidence for FP-Growth Algorithms dataset. The confidence values prove that when the antecedent occurs, the consequent follows between 15% and 36% of the instances.

Table II presents the top five association rules identified by both algorithms.

These associations with high lift associations provide valuable insights for product placement, cross selling approaches and inventory optimisation strategies. Significantly enough, the rule linking Tequila El Jimador and Copo Descartavel 300ml shows in both directions, displaying a strong bidirectional association.

The algorithms showed difference in their execution times, but both the algorithms displayed high consistency in rule generation, with 25 out of 26 rules being common to both algorithms. This overlap of 96.2% shows that both Apriori and FP-Growth are useful in identifying powerful association rules within the dataset, using the same parameters.

TABLE II
TOP 5 ASSOCIATION RULES

Antecedent	Consequent	Support	Confidence	Lift
Cerveja da Casa Puro Malte	Campari Insumo	0.0019	0.2368	9.88
Cachaca Pitu de Garrafa	Pilha	0.0018	0.2125	9.88
Cachaca da Casa Raiz	Cachaca Tocoios	0.0053	0.3597	9.34
Tequila El Jimador	Copo Descartavel 300ml	0.0073	0.3091	9.09
Copo Descartavel 300ml	Tequila El Jimador	0.0073	0.2138	9.09

It was seen that FP-Growth had a superior computational efficiency, but both the algorithms gave similar results in terms of quality and quantity when applied with the same parameters (min support = 0.0015, min confidence = 0.15, min lift = 2.5). Frankly it is the performance requirements rather than the quality of insights generated, which will decide the algorithm to be picked between Apriori and FPGrowth.

The identified association rules, particularly those with high lift values, were seen to offer valuable insights for product association strategies in the restaurant business. However, the low support values indicate that these associations are relatively rare, although their associations may be strong. Restaurant business will need to consider this when deploying strategies based on these rules, so that they can focus on high margin units or while using these for small scale marketing campaigns and promotions.

C. Hyperparameter Tuning Analysis

The impact of various hyperparameter combinations on the performance and effectiveness of the Apriori and FP-Growth algorithms is further analysed. As part of that, five different combinations of minimum support and minimum confidence thresholds were tested to understand their effects on generating rules, execution times and the quality of rules.

The following hyperparameter combinations were evaluated:

TABLE III
HYPERPARAMETER SETS TESTED

Set	Minimum Support	Minimum Confidence
1	0.001	0.10
2	0.002	0.15
3	0.0015	0.20
4	0.0005	0.25
5	0.003	0.05

The number of rules generated varied significantly across different hyperparameter sets:

Set 4 (0.0005, 0.25) is seen to produce the highest number of rules (423), showing that very low support and high confidence thresholds capture rare but potentially strong associations.

Further, in more of a converse trend, Set 2 (0.002, 0.15) and Set 3 (0.0015, 0.20) produced the fewest rules (19 each), suggesting a more balanced approach. Both algorithms consistently generated the same number of rules for each parameter

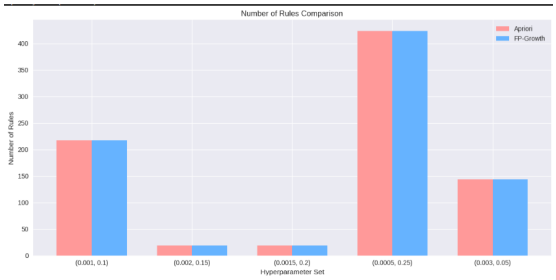

Fig. 10. Number of Rules Generated for Different Hyperparameter Sets

set, showing their equivalence in rule discovery, as shown in figure 10.

The execution times for both algorithms varied across the hyperparameter sets:

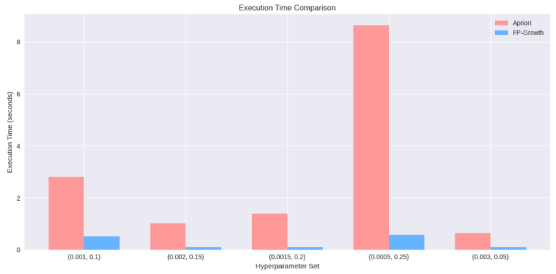

Fig. 11. Execution Time Comparison for Different Hyperparameter Sets

It was seen that execution speed of FP-Growth was far better in terms of executon speed across all parameters, as shown in figure 11. While FP-Growth maintained relatively stable performance, the execution time for Apriori increased significantly for Set 4 (0.0005, 0.25), consuming over 8 seconds.The fastest execution time was seen in Set 5(0.003, 0.05), more likely due to the higher support threshold reducing the number of candidate itemsets.

The average lift values for rules generated by each algorithm varied across the hyperparameter sets:

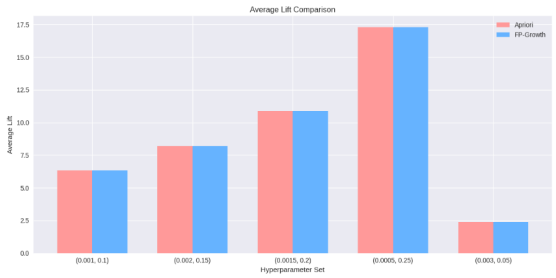

Fig. 12. Average Lift Comparison for Different Hyperparameter Sets

The highest average lift (17.28) which was part of the rules produced by Set 4 (0.0005, 0.25), showed strong associations but potentially at the cost of lower support, as shown in figure 12. Set 5 (0.003, 0.05) showed the lowest average lift (2.40), which indicated that whiled these rules are more frequently occurring, they may represent not so strong associations. But it was seen that both algorithms consistently produced rules with identical average lift values for each parameter set.

The efficiency of both algorithms in terms of rules generated per second showed interesting patterns:

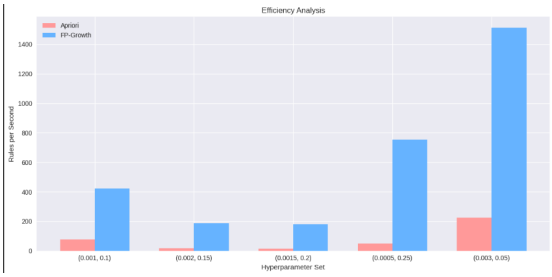

Fig. 13. Efficiency Analysis: Rules Generated per Second

It was seen that FP-Growth showed a much higher level of efficiency across all parameter sets (Figure 13), producing more rules per second than Apriori. The efficiency gap between FP-Growth and Apriori was most significant for Set 5 (0.003, 0.05), where FP-Growth produced rules at a much higher rate. The efficiency of Apriori was much more stable across parameter sets, whilst FP-Growth showed more variability, peaking in efficiency for Set 5.

Based on the analysis, the following conclusions can be drawn:

- Set 4 (0.0005, 0.25) is seen to be most optimal for finding rare but strong associations. But this comes at expense of much higher level of execution times, especially for Apriori.
- Set 3 (0.0015, 0.20) offers a good compromise generating a manageable number of rules with higher average lift. This brings a level of balance between rule quantity, quality, and execution time.
- Set 5 (0.003, 0.05) provides the fastest performance, but with lower average lift values.
- It was seen that FP-Growth showed high level of performance in terms of speed and efficiency, compared to Apriori. That makes it a perfect fit and choice for larger datasets, specially when quick results are needed.

In summary, the choice of hyperparameters significantly impacts the number of rules generated, quality of the rules generated, and the computational efficiency of the algorithms. The optimal choice and fit depends on the specific needs and demands of the analysis, which brings in a balance across discovering rare associations, maintaining rule quality, and ensuring computational feasibility.

D. Case Study: Emp-Specific Analysis

The sales data of Emp 1 was analysed with an intent of proving the practical application of the association rule mining approach. This case study provides valuable insights into individual sales and revenue patterns and product associations while demonstrating how the methodology can be applicable in real-world scenarios.

From 4th January 2021 to 2nd February 2024, over a three year period, it was seen that Emp 1 showed a high level of sales activity. With 2,556 transactions and an average of 229.20 items per transaction, this emp's performance provides a rich dataset for analysis. The diversity and range of their sales is demonstrated in the 30 unique products sold during this time.

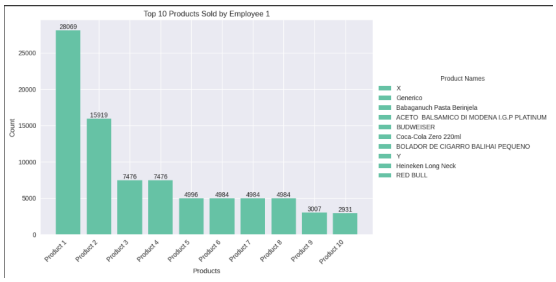

Fig. 14. Top 10 Products Sold by Emp 1

When the top selling products and their sales patterns were examined, as shown in Figure 14 , Product X was seen as the clear leader with 28,069 units sold, followed by Generico at 15,919 units. ACETO BALSAMICO DI MODENA I.G.P PLATINUM and Babaganuch Pasta Berinjela are at third place, each with 7,476 units sold. This balance of high volume products with more moderate sellers give a more balanced trend of customer preferences and a great opportunity for finding product associations.

A filtering process was applied to remove high-frequency items, which is often a common practice in association rule mining to discover less obvious but potentially valuable relationships. This filtering reduces the average basket size from 229.20 to 211.99 items, indicating that there is a balanced mix of both common and less common products in most of the transactions.

Once the filtering process was done, the focus was shifted to products like BUDWEISER (4,996 units), Coca-Cola Zero 220ml (4,984 units), and BOLADOR DE CIGARRO BALIHAI PEQUENO (4,984 units). This provided a much clearer view of the more popular items that aren't dominated by other top products, which give opportunities for other meaningful association rule discovery.

The sales pattern showed by Emp 1 shows a very interesting case for association rule mining. The high average basket size tells us that numerous opportunities exist for discovering product associations. Moreover, the diverse product types ranging from food items like Babaganuch Pasta Berinjela to beverages such as BUDWEISER and Coca-Cola Zero, and even smoking accessories like BOLADOR DE CIGARRO BALIHAI PEQUENO, present real potential for discovering cross-category product associations.

This analysis shows us that there is great value in applying association rule mining techniques to individual sales data. By way of this, personalised insights can be discovered that could help in informed business decisions like targeted sales strategies and optimisation of inventory management. For example, the relationship between ACETO BALSAMICO DI MODENA I.G.P PLATINUM and other products in the basket could hight opportunities to improve cross selling strategies.

As the analysis delves deeper, the broad range of sales data from Emp 1 offers a great base and foundation for testing and refining the association rule mining algorithms. The balance and a mix between high-volume and moderate-selling products, along with the large basket sizes offers both challenges and opportunities for uncovering meaningful product associations to drive sales and margins.

VI. Discussion

There are several high impact insights which are drawn from this analysis of customer purchase behavior and product associations. By exploring two prominent association rule mining algorithms, Apriori and FP-Growth, this study has uncovered their strengths and limitations when it is deployed in large-scale transaction datasets.

A comparative study between Apriori and FP-Growth algorithms shows that both these product high quality results and similar in nature. But it is also seen that FP-Growth consistently shows superior results to Apriori in terms of execution speed. This efficiency becomes more prominent when it is applied to larger datasets of when lower support thresholds are used. This is largely due to its compact data structure, which allows for must quicker way of processing frequent datasets [23]. But it is worth mentioning that despite the difference in execution and performance, both algorithms do provide a high level of consistency in the way rules are generated.

However, it is noteworthy that despite the difference in execution times, both algorithms demonstrated high consistency in rule generation, with a 96.2% overlap in the rules produced. This suggests that for smaller datasets or when computational resources are not a constraint, either algorithm could be effectively employed [24].

The hyperparameter tuning analysis uncovered the level of impact of support and confidence thresholds on the number and quality of rules generated. When the lower support thresholds (e.g., 0.0005) is combined with higher confidence thresholds (e.g., 0.25), it results in a larger number of rules with higher average lift values. This configuration is particularly helpful when it comes to discovering rare but strong associations. But this is achieved at the expense of increased computational time, especially for the Apriori algorithm. On the other hand, the converse also holds true., higher support thresholds (e.g., 0.003) with lower confidence thresholds (e.g., 0.05) led to faster execution times but produced rules with lower average lift values. This configuration might be more relevant when it comes to identifying broader and diverse trends in customer buying patterns.

The correct choice of hyperparameters depends on the specific business requirements, provided by the right balance between the discovery of rare associations, rule quality, and computational feasibility.

There are several strong product associations which were uncovered through this analysis, particularly in the beverage category. For example, the high lift value (9.88) for the association between 'Cerveja da Casa Puro Malte' and 'Campari Insumo' suggests a strong tendency or probability for these products to be bought together. These kinds of insights are really helpful in optimising inventory management, store lay out design and for creating targeted marketing campaigns [25][26].

However, it is also to be noted that many of the strong associations had low support values, which means that they occur infrequently although their associations are strong. It also suggests to us that these results must be interpreted carefully, and both lift and support metrics have to be considered in tandem when making business decisions based on these product associations.

The analysis done based on the emp-specific date showed the practical applicability of the analytical pipeline to individual level data. The analysis of Emp 1's individual sales data uncovers unique selling patterns and product associations which were not very clear in the aggregate data. This highlights the potential of using association rule mining for customised inventory management and sales strategies at the emp or restaurant level.

This research analysis does provide valuable insights into the purchasing behaviour of consumers, however there are several limitations which need to be addressed in future research. It is to be noted that this analysis was done on a subset of the available data due to computing constraints. More work is needed to explore techniques for scaling the analysis to even larger datasets, potentially leveraging advanced distributed computing frameworks [27].

Secondly, the current study is primarily based on product associations without considering temporal and seasonal patterns [28] or customer demographics. These factors need to be included which can provide more balanced results and insights into the real customer behaviour.

Additionally, it has to be kept in mind that the study doesn't include a real-world implementation of these insights although it does demonstrate the powerful potential of association rule mining for inventory management and sale strategies. Future research could include advanced validation techniques like A/B testing of marketing strategies or restaurant lay outs based on the associations so that real world impact of these insights and recommendations could be adequately tested [29].

VII. Conclusion

This research study has developed and deployed a robust analytical framework for customer purchase behaviour and product association mining in the restaurant industry. The structured and methodical analysis of Apriori and FP-Growth algorithms demonstrates that whilst both algorithms produce similar and high quality results, FP-Growth provides better computational efficiency, making it more relevant for large-scale datasets. Further, the analysis of the hyper parameter tuning exercise shows that support and confidence thresholds play a significant role in providing the right balance between discovery of rare associations and computational practicality. Therefore, it is important to carefully pick these parameters based on specific business needs and availability of computing resources. A set of strong product associations were discovered, particularly in the beverage category. This provides actionable and meaningful insights to optimise inventory management, restaurant lay outs and to craft targeted marketing strategies. However, the low support values seen for many strong product associations highlights the need for careful interpretation and validation of these insights before they are implemented in a real-world business environment.

The case study conducted based on emp-specific data demonstrated the potential of applying association rule mining techniques at lower levels, which can also be directed towards customised inventory management and sales strategies. This clearly illustrates the flexibility and scalability of the developed analytical pipeline and framework. In summary, this research study makes a meaningful contribution to the field of restaurant analytics through a framework for customer purchase behaviour analysis and product association mining. These insights have significant and positive implications for restaurant strategy and operations, which in turn can lead to better customer satisfaction, optimised inventory management and increases sales. There is tremendous potential to extend this research to further areas like integrating temporal data to reveal seasonal trends, including customer demographic information for more balanced segmentation, and real-world deployment of recommended strategies derived from these associations. These approaches could significantly improve the impact and relevance of association rule mining in restaurant business.

VIII. Declaration

Declaration of Originality. I am aware of and understand the University of Exeter's policy on plagiarism and I certify that this assignment is my own work, except where indicated by referencing, and that I have followed the good academic practices.

Declaration of Ethical Concerns. This work does not raise any ethical issues. No human or animal subjects are involved neither has personal data of human subjects been processed. Also, no security or safety critical activities have been carried out

References

[1] AlShamsi, A. Y. (2022). Understanding customer behaviour in restaurants based on data mining prediction technique.

[2] Maheswari, K., Priya, P. P. A. (2017). Predicting customer behavior in online shopping using SVM classifier. In *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)* (pp. 1-5). IEEE. https://doi.org/10.1109/ITCOSP.2017.8303085

[3] Kaur, M., Kang, S. (2016). Market basket analysis: Identify the changing trends of market data using association rule mining. *Procedia Computer Science*, 85, 78–85. https://doi.org/10.1016/j.procs.2016.05.180

[4] Kurnia, Y., Isharianto, Y., Giap, Y. C., Hermawan, A. (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 Physics: Conference Series*, 1175(1), Article 012047. https://doi.org/10.1088/1742-6596/1175/1/012047

[5] Jorge, A. M., Azevedo, P. J. (2005). An experiment with association rules and classification: Post-bagging and conviction. In *International Conference on Discovery Science* (pp. 137-149). Springer.

[6] Ting PingHo, T. P., Pan, S., Chou ShuoShiung, C. S. (2010). Finding ideal menu items assortments: An empirical application of market basket analysis.

[7] Gómez-Talal, I., González-Serrano, L., Rojo-Álvarez, J. L., Talón-Ballestero, P. (2024). Avoiding food waste from restaurant tickets: A big data management tool. *Journal of Hospitality and Tourism Technology*, 15(2), 232-253.

[8] Schroeder, R. (2016). Big data business models: Challenges and opportunities. *Cogent Social Sciences*, 2(1), Article 1166924.

[9] Bourlakis, M., Papagiannidis, S., Fox, H. (2008). E-consumer behaviour: Past, present and future trajectories of an evolving retail revolution. *International Journal of E-Business Research*, 4(3), 64-76.

[10] Solnet, D., Boztug, Y., Dolnicar, S. (2016). An untapped gold mine? Exploring the potential of market basket analysis to grow hotel revenue. *International Journal of Hospitality Management*, 56, 119-125.

[11] Agrawal, R., Imieliński, T., Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (pp. 207-216). ACM.

[12] Agrawal, R., Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)* (Vol. 1215, pp. 487-499).

[13] Han, J., Pei, J., Yin, Y. (2000). Mining frequent patterns without candidate generation. *ACM SIGMOD Record*, 29(2), 1-12.

[14] Tan, P. N., Kumar, V., Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, 29(4), 293-313.

[15] Li, H., Wang, Y., Zhang, D., Zhang, M., Chang, E. Y. (2008). PFP: Parallel FP-growth for query recommendation. In *Proceedings of the 2008 ACM Conference on Recommender Systems* (pp. 107-114). ACM.

[16] Moens, S., Aksehirli, E., Goethals, B. (2013). Frequent itemset mining for big data. In *2013 IEEE International Conference on Big Data* (pp. 111-118). IEEE.

[17] Chen, Y. L., Tang, K., Shen, R. J., Hu, Y. H. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems*, 40(2), 339-354.

[18] Fournier-Viger, P., Lin, J. C. W., Vo, B., Chi, T. T., Zhang, J., Le, H. B. (2017). A survey of itemset mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(4), Article e1207.

[19] Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3), 372-390.

[20] Stilou, S., Bamidis, P. D., Maglaveras, N., Pappas, C. (2001). Mining association rules from clinical databases: An intelligent diagnostic process in healthcare. *Studies in Health Technology and Informatics*, 84(2), 1399-1403.

[21] Ale, J. M., Rossi, G. H. (2000). An approach to discovering temporal association rules. In *Proceedings of the 2000 ACM Symposium on Applied Computing* (pp. 294-300). ACM.

[22] Aïmeur, E., Brassard, G., Gambs, S. (2013). Quantum speed-up for unsupervised learning. *Machine Learning*, 90(2), 261-287.

[23] Shang, X., Sattler, K. U., Geist, I. (2004). SQL based frequent pattern mining with FP-growth. In *International Conference on Applications of Declarative Programming and Knowledge Management* (pp. 32-46). Springer.

[24] Bala, A., Shuaibu, M. Z., KaramiLawal, Z., Zakari, R. I. Y. (2016). Performance analysis of apriori and fp-growth algorithms (association rule mining). *International Journal of Computer Technology & Applications*, 7(2), 279-293.

[25] Glanz, K., Bader, M. D., Iyer, S. (2012). Retail grocery store marketing strategies and obesity: An integrative review. *American Journal of Preventive Medicine*, 42(5), 503-512.

[26] Khasanah, A. U., Baihaqie, M. R. Q. (2024). Analysis of consumer characteristics on retail business with clustering analysis method and association rule for selling improvement strategy recommendations. *OPSI*, 17(1), 249-257.

[27] Masih, S., Pathak, D., Rahatekar, N. (2014). Data mining techniques in parallel and distributed environment: A comprehensive survey. In *Proceedings of the International Conference on Data Mining Techniques in Parallel and Distributed Environment*.

[28] Hariharan, S., Kannan, M., Raguraman, P. (2013). A seasonal approach for analysis of temporal trends in retail marketing using association rule mining. *International Journal of Computer Applications*, 71(13).

[29] Siroker, D., Koomen, P. (2015). *A/B testing: The most powerful way to turn clicks into customers*. John Wiley Sons.