بسم الله الرحمن الرحيم

University of Khartoum

Faculty of mathematical science and informatics

Department of statistic

Project Title:

Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods - A Review and Illustration.

A Research submitted in partial fulfillment of the requirement of the B.Sc. degree in the faculty of mathematical science and informatics, Department of Statistics

Participants:

1.Doaa Mahmoud Babiker Mohamed 16-112

2. Shahinda Abdelmonaim Sayed Mohammed 16-228

3. Muzan Ahmed Altigani Mohammed Swar 16-235

Supervisor:

T. Raad Ahmed Shaaban

January 2025

Abstract:
In data mining, frequent pattern mining plays a crucial role in uncovering associations and correlations within large datasets. This research provides a comprehensive review and illustration of fundamental concepts and methods for mining frequent patterns, associations, and correlations, focusing on algorithms like Apriori and FP-Growth. The study explores definitions, applications, and significance of techniques in extracting actionable insights for various domains such as retail and healthcare. By reviewing efficient algorithms and illustrating their application in market basket analysis, the research highlights the advantages and limitations of these methods, including their computational efficiency, scalability, and memory usage. Additionally, the study examines the metrics used to evaluate the strength and reliability of discovered patterns, such as support, confidence, and lift. The findings emphasize the potential of frequent pattern mining in improving business decision-making, enhancing inventory management, and providing personalized recommendations while addressing the challenges of interpretability, scalability, and data preprocessing. The research also includes a practical implementation of market basket analysis using a dataset of 9,835 grocery transactions, demonstrating how the Apriori and FP-Growth algorithms can identify frequent item sets and generate association rules. The results reveal that while both algorithms yield the same results, FP-growth is significantly faster and more efficient for large datasets. This research is a foundational reference for advancing theoretical and applied understanding of frequent pattern mining techniques, offering valuable insights for academic and industry applications.

الملخص

في مجال تعدين البيانات، يلعب استخراج الأنماط المتكررة دورا رئيسيا في الكشف عن العلاقات والارتباطات ضمن مجموعات البيانات الكبيرة تقدم هذا البحث مراجعة شاملة وتوضيحا للمفاهيم والأساليب الأساسية المستخدمة في استخراج الأنماط المتكررة والعلاقات والارتباطات، مع التركيز على خوارزميات مثل Apriori وFP- Growth. تهدف الدراسة إلى استكشاف تعريفات هذه التقنيات وتطبيقاتها وأهميتها في استخراج رؤى قابلة للتنفيذ في مجالات متنوعة مثل البيع بالتجزئة والرعاية الصحية. من خلال مراجعة الخوارزميات الفعالة وتوضيح تطبيقها في تحليل سلة السوق، تسلط البحث الضوء على مزايا وقيود هذه الأساليب، بما في ذلك كفاءتها الحسابية وقابليتها للتوسع واستخدام الذاكرة. بالإضافة إلى ذلك، تدرس البحث المقاييس المستخدمة لتقييم قوة وموثوقية الأنماط المكتشفة، مثل الدعم والثقة والرفع. تؤكد النتائج على إمكانات استخراج الأنماط المتكررة في تحسين عمليات صنع القرار التجارية، وتعزيز إدارة المخزون، وتقديم توصيات مخصصة، مع معالجة تحديات قابلية التفسير وقابلية التوسع ومعالجة البيانات الأولية. يتضمن البحث أيضا تطبيقا عمليا لتحليل سلة السوق باستخدام مجموعة بيانات تحتوي على 9,835 معاملة شراء، مما يوضح كيف يمكن لخوارزميات Apriori وFP- Growth تحديد مجموعات العناصر المتكررة وإنشاء قواعد ارتباطية. تكشف النتائج أن كلا الخوارزميتان تعطيان نفس النتائج، ولكن FP-Growth أسرع وأكثر كفاءة في التعامل مع مجموعات البيانات الكبيرة. يعد هذا البحث مرجعا أساسيا لتطوير الفهم النظري والتطبيقي لتقنيات استخراج الأنماط المتكررة، ويقدم رؤى قيمة لكل من التطبيقات الأكاديمية والصناعية.

Dedication

To our families, the backbone of our lives and the infinite support after Allah.
To our friends, the sweetness of the long journey, the reason we kept on going, trying and evolving. To the ones who crossed this campus before and the ones will come after us. To our country and the angelic souls, we lost for it. To their mothers and fathers who had to go through it and still be standing up so bravely. You will not be forgotten. To the base of this university, the minds that continued to inspire us along the way, to our all teachers during our education journey, you are great, for showing up when it's hard, for being patient with our strayed thoughts, for being the main reason why all of this is becoming real. Thank you. You all are our constants; we are thrilled to represent a small part of this journey. And whenever we go or what we will do we will always be grateful for our constants. We will continue to do our best, amaze and be amazed and no doubt making you forever proud of us which is one of many goals we will achieve sooner or later.

Table of contents:

List OF FIGURES:

List OF TABLES:

Acknowledgment

We express our sincere gratitude to our esteemed supervisor, T. Raad Ahmed Shaaban, for his invaluable guidance, unwavering support, and insightful feedback throughout this research endeavor. We also extend our heartfelt thanks to our families and all those who have offered their encouragement, support, and prayers, making this research possible.

Chapter One: Introduction

Chapter 1: Introduction:

This chapter provides a general overview of the concepts of data mining, frequent analysis, and mining frequent patterns. As well as association rule and correlation, Additionally, will cover the research problem statements, aims, objectives, and the significance of the study, and lastly a simple run over the upcoming chapters combined in the research structure.

1.1 Basic concepts: [1]

1. Data Mining:

It is the process of discovering patterns, relationships, and useful information from large data sets using statistical, mathematical, and computational techniques. It involves extracting hidden knowledge and transforming this knowledge into useful information that supports decision-making and strategic planning to discover previously unknown patterns, predict trends, and understand the structure of data to generate actionable insights.

2. Frequent pattern analysis:

Data mining is a technique used to identify recurring patterns in a dataset, such as itemsets, sequences, or structures. These patterns help uncover relationships and associations within the data, revealing insights that may not be immediately apparent.

3. Mining frequent patterns:

Is a core task in data mining that focuses on finding regularities, associations, or patterns that occur frequently within large datasets. These patterns can reveal important and actionable insights about the data. The process involves identifying item sets, sequences, or substructures that appear together more often than a specified threshold. The goal is to identify significant and interesting relationships in data and a pattern that occurs with a high frequency or support, which can provide valuable insights and knowledge for various applications such as market basket analysis, recommendation systems, and anomaly detection.

4. Association rule:

Associations refer to the relationships between different items or events within a dataset that frequently occur together. For example, in market basket analysis, an association rule might indicate that customers who buy bread also frequently buy butter. By identifying patterns such as frequent itemsets and generating rules of form $A \Rightarrow B$, the strength of these associations is typically measured using key metrics like support, confidence, and lift.

5. Correlation:

Correlation measures the statistical relationship between two or more variables. It determines whether the presence of one item influences the presence of another. Correlations can be positive (both variables increase or decrease together), negative (one variable increases while the other decreases), or zero (no discernible relationship).

1.2 Problem statement

Data is poured into computer networks, the World Wide Web, and various data storage devices daily, some patterns found in a user search query can disclose valuable knowledge that can't be obtained by reading the person's data items alone. So, by reviewing efficient algorithms for discovering frequent patterns, associations, and correlations in large datasets, such as the Apriori Algorithm: which identifies frequent itemset using a level-wise search, or FP-Growth (Frequent Pattern Growth): which uses a compact data structure (FP-tree) to represent the dataset. By illustrating it in a market basket analysis example we will be able to extract meaningful information from these large datasets which are crucial for making informed decisions, improving business processes, and gaining a competitive advantage.

1.3 Research Aim

This research aims to clarify the underlying concepts, explore the implementation of these algorithms, and provide practical examples to illustrate their real-world applicability.

1.4 Research Objectives

1- Discover Efficient Algorithms: Find efficient algorithms and techniques for mining frequent patterns, association rules, and correlations from large datasets that can handle the exponentially growing volume of data and process it promptly.

2- Application of Mining: Apply the mining of frequent patterns, association rules, and correlations to a real-world example.

3- Evaluation and Present Results: Evaluate the effectiveness and reliability of the mining techniques and how they emphasize the potential for improving business processes and outcomes through the application of these techniques.

1.5 Research questions:

1. What are mining Frequent-patterns, Associations, and Correlations?
2. What are the key differences between mining Patterns, association rules, and correlations, and when is each technique more appropriate?
3. what are the different Algorithms and methods being used in mining frequent patterns, Associations, and Correlations?
4. What are the most effective applications of frequent pattern mining in (retail)?
5. How can frequent pattern mining be used to improve business decision-making processes?
6. What are the limitations and challenges associated with interpreting and utilizing frequent patterns in real-world applications?

1.6 The scoop of the research:

This study aims to provide a comprehensive understanding of frequent pattern mining, associations, and correlations by addressing key questions such as the definition and significance of these concepts in data mining. It will explore the distinctions between frequent pattern mining, association rules, and correlations, elucidating when each technique is most suitable for analysis. Additionally, the research will deep dive into the various algorithms and methodologies employed in mining frequent patterns, associations, and correlations, evaluating their strengths and limitations. Furthermore, the study will focus on the practical applications of frequent-making processes and identify challenges in interpreting and utilizing frequent patterns effectively in real-world scenarios.

1.7 Research methodology:

1. Literature Review:

Conduct a thorough review of existing literature on frequent pattern mining, association rules, and correlations to establish a foundational understanding of the concepts and their applications.

2. Comparative Analysis:

Analyze and compare the key differences between frequent pattern mining, association rules, and correlations through theoretical frameworks and case studies.

3. Algorithm Evaluation:

Evaluate the algorithms and methods commonly used in frequent pattern mining, such as Apriori, FP-Growth, and ECLAT, to understand their functionalities, advantages, and limitations.

4. Case Studies and Interviews:

Conduct case studies and interviews with professionals in the retail industry to identify the most effective applications of frequent pattern mining in retail settings and gather insights into its impact on business decision-making processes.

5. Data Collection and Analysis:

Collect retail transaction data or use publicly available datasets to analyze the application of frequent pattern mining techniques in real-world scenarios.

Utilize statistical analysis methods to interpret the data and draw meaningful conclusions about the effectiveness and limitations of frequent pattern mining in improving business processes.

6. Challenges and Limitations Identification:

Identify the limitations and challenges associated with interpreting and utilizing frequent patterns in real-world applications through qualitative analysis of case studies and interviews.

7. Recommendations and Future Directions:

Provide recommendations for improving the application of frequent pattern mining techniques in retail and other industries, considering the identified challenges and limitations.

Propose future research directions to address the gaps and challenges identified in the study, aiming to advance the field of frequent pattern mining and its applications

1.8 Significant of the study

The study will be beneficial to the following: -

1. Manufacturing

- Efficiency Improvement: Optimizes workflows and reduces downtime through pattern identification.
- Quality Control: Pinpoints root causes of defects by analyzing variable associations.

2. Retail

- Customer Insights: Reveals buying behaviors for personalized marketing.
- Inventory Management: Improves stock management by correlating sales trends with inventory levels.

3. Researchers

- New Knowledge Discovery: Uncovers hidden patterns, advancing theoretical and applied research.
- Methodological Innovation: Enhances research methodologies with refined data mining techniques.

1.9 Research structure

- **Chapter one**: Introduction – Provides an overview of the study, including background, problem statement, objectives, significance, and structure.
- **Chapter two**: Literature Review – Reviews existing literature related to data mining techniques, identifying key theories, models, and research findings.
- **Chapter three**: Research Methodology – Describes the research design, data collection methods, and data analysis techniques used in this study.
  Results – Presents the findings of the study, including data analysis and interpretation.
- **Chapter four**: implementation and result – with an illustration of a real-world example and describing the way of implementing algorithms and the final result.
- **Chapter five**: conclusion and recommendation – provides an overall conclusion of the research and a recommendation for future research.

Chapter Two: Literature Review

Chapter 2: literature review:

This chapter provides the research's main definitions and illustrates the algorithms that are commonly used in mining frequent pattern association rules and correlation walks through related papers and discusses its findings.

2.1 Frequent Pattern Mining:

Frequent patterns are defined as combinations of items, sequences, or structures within datasets. Identifying these patterns is essential in applications like market basket analysis, where insights into commonly purchased products (e.g., milk and bread) can inform inventory and marketing strategies. [2]

2.2 Association Rule Mining:

Association rule mining is a fundamental technique in data mining that aims to uncover hidden patterns and relationships within datasets. It involves two key steps: identifying frequent items that meet a minimum support threshold and generating rules from these items based on a minimum confidence threshold. [3]

One of the most prominent algorithms for this process is the Apriori algorithm, which iteratively identifies frequent items and uses them to generate association rules. This foundational method has inspired numerous extensions, such as sequential pattern mining and episode mining, to address more complex data scenarios.

Association rules vary in complexity and application.

2.2.1 Different Types of Association Rules:

- Single-dimensional rules focus on a single attribute, such as predicting that customers who buy bread are also likely to buy butter.
- Multidimensional rules incorporate multiple attributes, like linking age and income to purchasing behavior.
- Boolean rules identify associations based on the presence or absence of items, while quantitative rules handle continuous variables by grouping them into intervals.
- Additionally, correlation rules go beyond simple associations to highlight statistically significant relationships, refining results to produce more actionable insights.

2.2.2 association rule mining challenges such as:

Despite its utility, computational complexity, scalability, and result interpretability. Advances in algorithmic efficiency, including pruning strategies and enhanced data structures, have addressed many of these issues, making it possible to extract meaningful patterns from large, complex datasets. These methodologies are indispensable for understanding data structures and behaviors,

with applications spanning from market basket analysis to bioinformatics and web analytics. [4]

The review delves into the evolution of association rule mining, focusing on its foundational principles, diverse rule types, and advancements in overcoming key challenges.

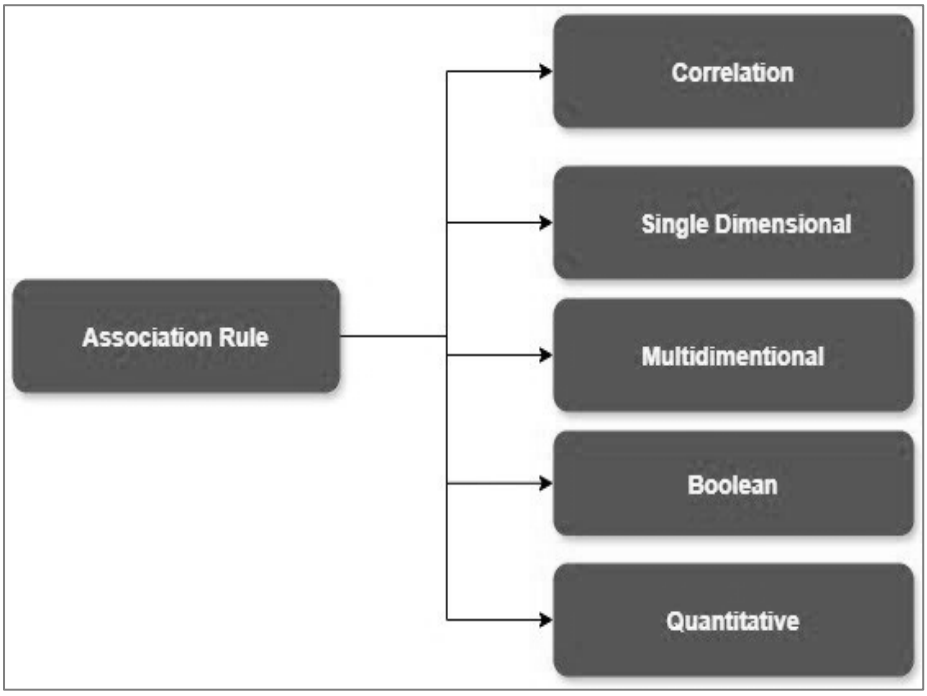


Figure 1: Shows the types of association rules

2.3 Correlation analysis:

Correlation analysis is a crucial extension of association rule mining, enhancing the ability to uncover insightful patterns in large and complex datasets; offering practical value in domains such as market analysis, recommendation systems, and decision support systems.
The inclusion of correlation measures like lift and Chi-Square in association rule mining addresses critical shortcomings of the traditional support-confidence framework. These measures allow for a more refined evaluation of rules, distinguishing between coincidental associations and meaningful relationships. [5]

Correlation analysis in data mining is essential for identifying meaningful relationships between variables, especially in datasets with diverse data types such as real-valued, nominal, and ordinal data. While traditional statistical correlation measures work well for real-valued data, nominal and ordinal types require specialized techniques to ensure accuracy and relevance.

-For mixed data types, innovative methods like **A-correlation** are introduced to assign real values to nominal data in a way that maximizes correlation, maintaining a natural

interpretation of the coefficient. Additionally, **Cramer's V-statistic** is employed for nominal/nominal and nominal/ordinal data to uncover significant relationships.

-When dealing with purely nominal data, traditional correlation measures like Pearson's are unsuitable due to the lack of inherent order. Instead, **Cramer's V-statistic** and matching-based methods are applied, while rank-based measures, such as Spearman's correlation, are used for ordinal data to account for their natural ordering.

-For real-valued and ordinal data, rank-based techniques remain effective, but the paper also highlights the role of domain expertise in assigning real values to ordinal categories. This approach improves accuracy, particularly when the ordinal data reflects an underlying continuum, such as satisfaction levels or academic grades. [6]

By adapting correlation methods to suit different data types, these techniques enhance the versatility and effectiveness of correlation analysis in data mining, enabling the extraction of valuable and reliable insights from complex datasets.

2.4 Apriori Algorithm:

The Apriori algorithm, introduced by R. Agrawal and R. Srikant in 1994, is a foundational method in frequent pattern mining. It operates iteratively, leveraging prior knowledge of frequent itemset properties. It is widely used in many applications such as market basket analysis, bioinformatics and web mining.[7]

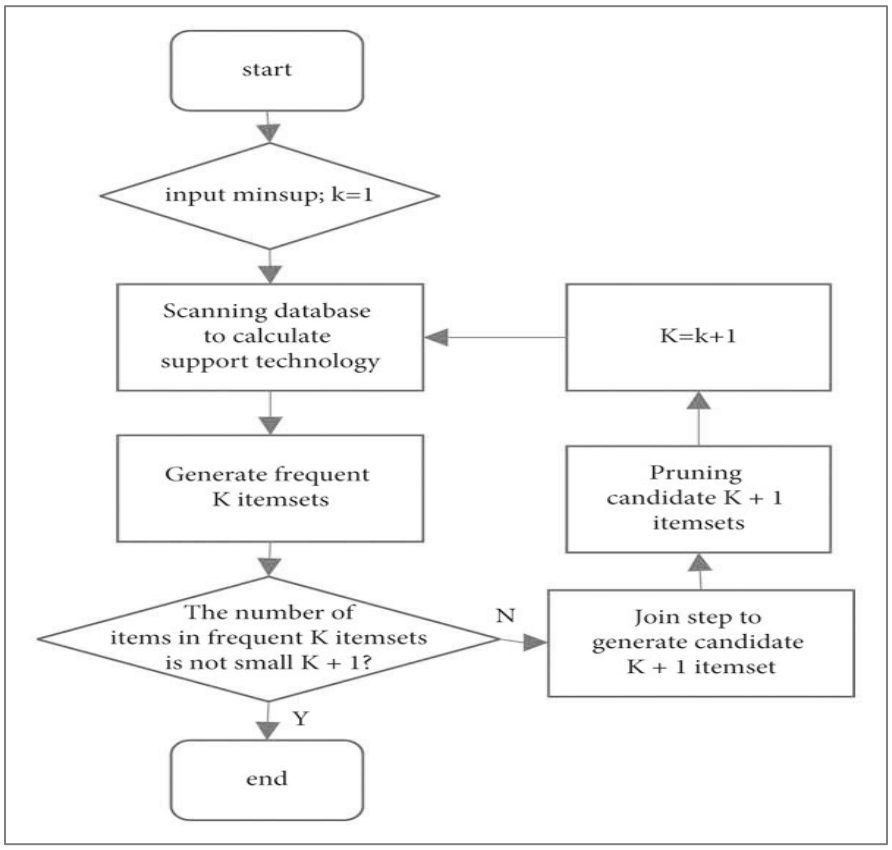


Figure 2: Shows the steps of the Apriori algorithm in a diagram

Given the dataset:

TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

Table 1: shows the given itemsets

The main point is to generate frequent itemsets to determine the support count for each candidate itemset which is considered to be complex process.

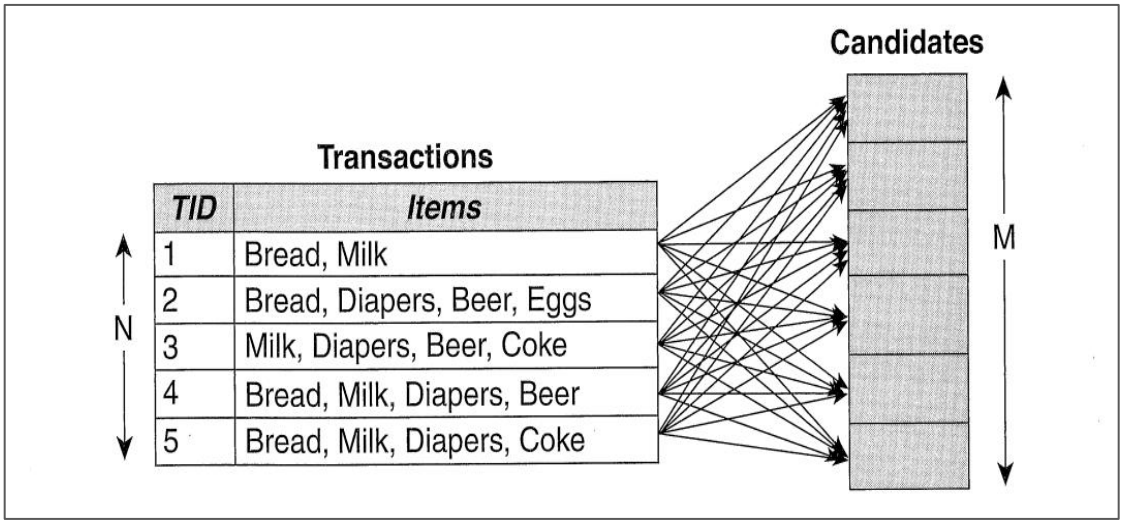


Figure 3: shows the candidates items

The Apriori algorithm reduce the large number of candidates itemsets explored during frequent item generation, The algorithm begins by scanning the database to identify frequent items that meet a minimum support threshold. These items are then used to generate larger candidate itemset in subsequent iterations. This process continues until no new frequent itemset can be identified.

The algorithm uses a pruning technique that significantly reduce the number of combinations that need to be evaluated. then the algorithm generates association rules that shows how items related to one another and the strength of these relationships.[8]

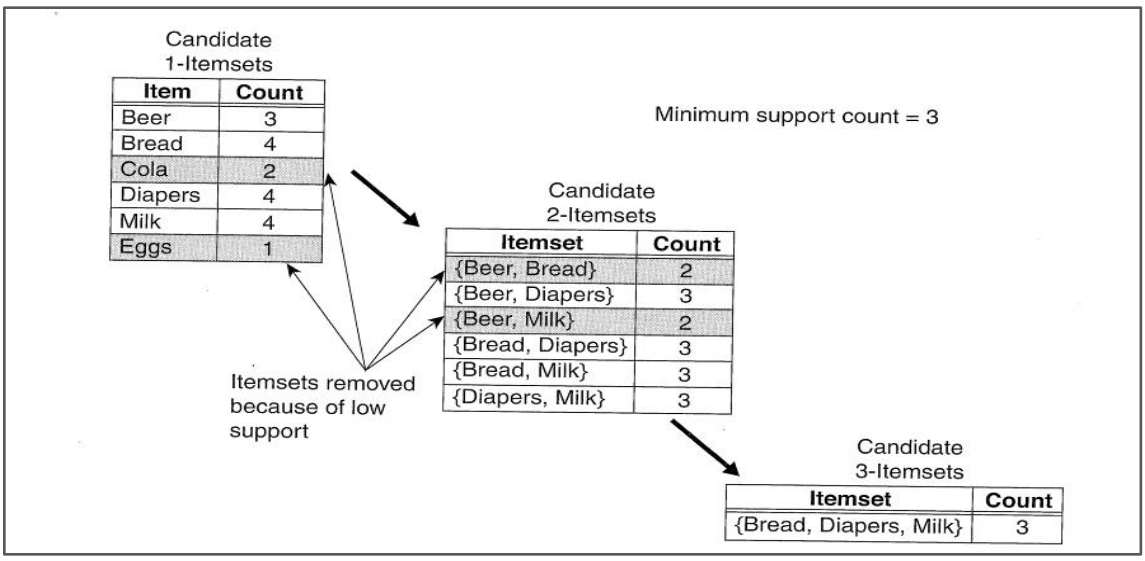


Figure 4: shows the process of support count and comparison

2.5 FP -Growth Algorithm:

The FP-Growth algorithm, introduced by Han et al. in 2000, addresses the limitations of the Apriori algorithm by eliminating the need for costly candidate generation. It uses a compact data structure called the Frequent Pattern Tree (FP-tree) to represent the database in a compressed form. It is more used in large datasets as it reduces both memory usage and computational overhead.[9]

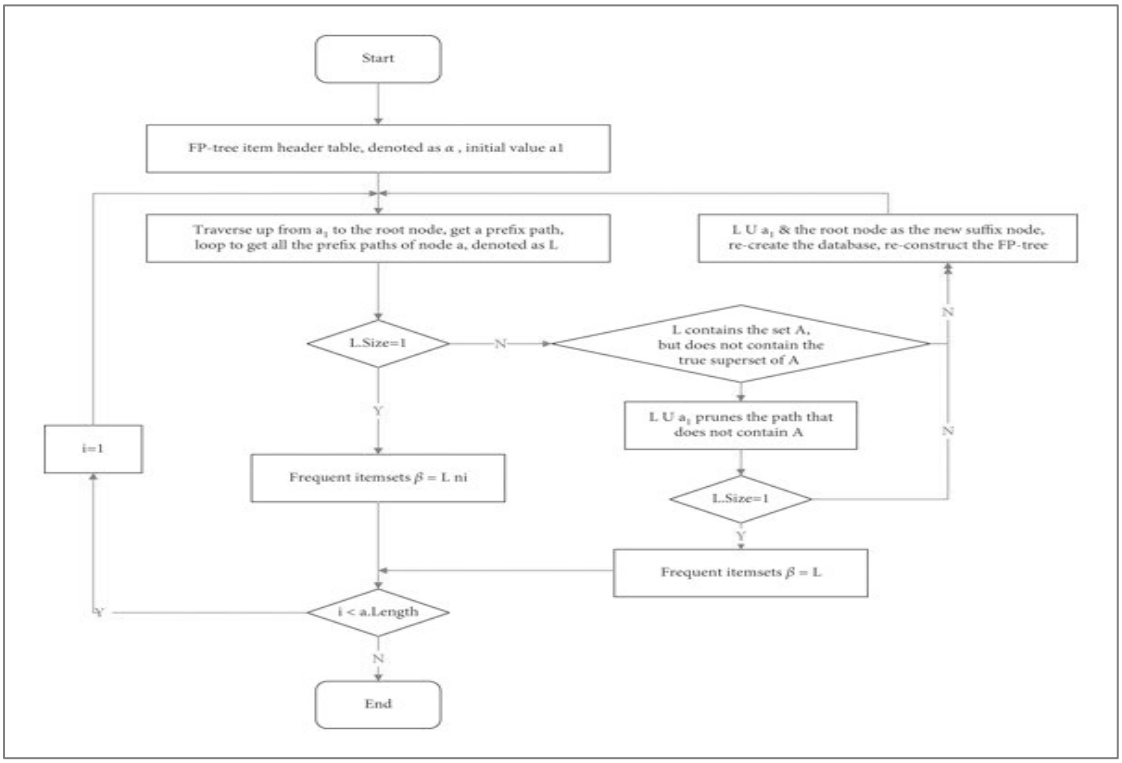


Figure 5: Show the steps of the FP-Growth algorithm in a diagram

The FP tree is built by starting with a root node and adding nodes for each transaction, each node represents an item and its frequency count. Nodes are linked to represent shared itemset, allowing efficient traversal and pattern discovery. This structure enables the algorithm to mine the complete set of frequent itemset directly from the tree, significantly improving computational efficiency.

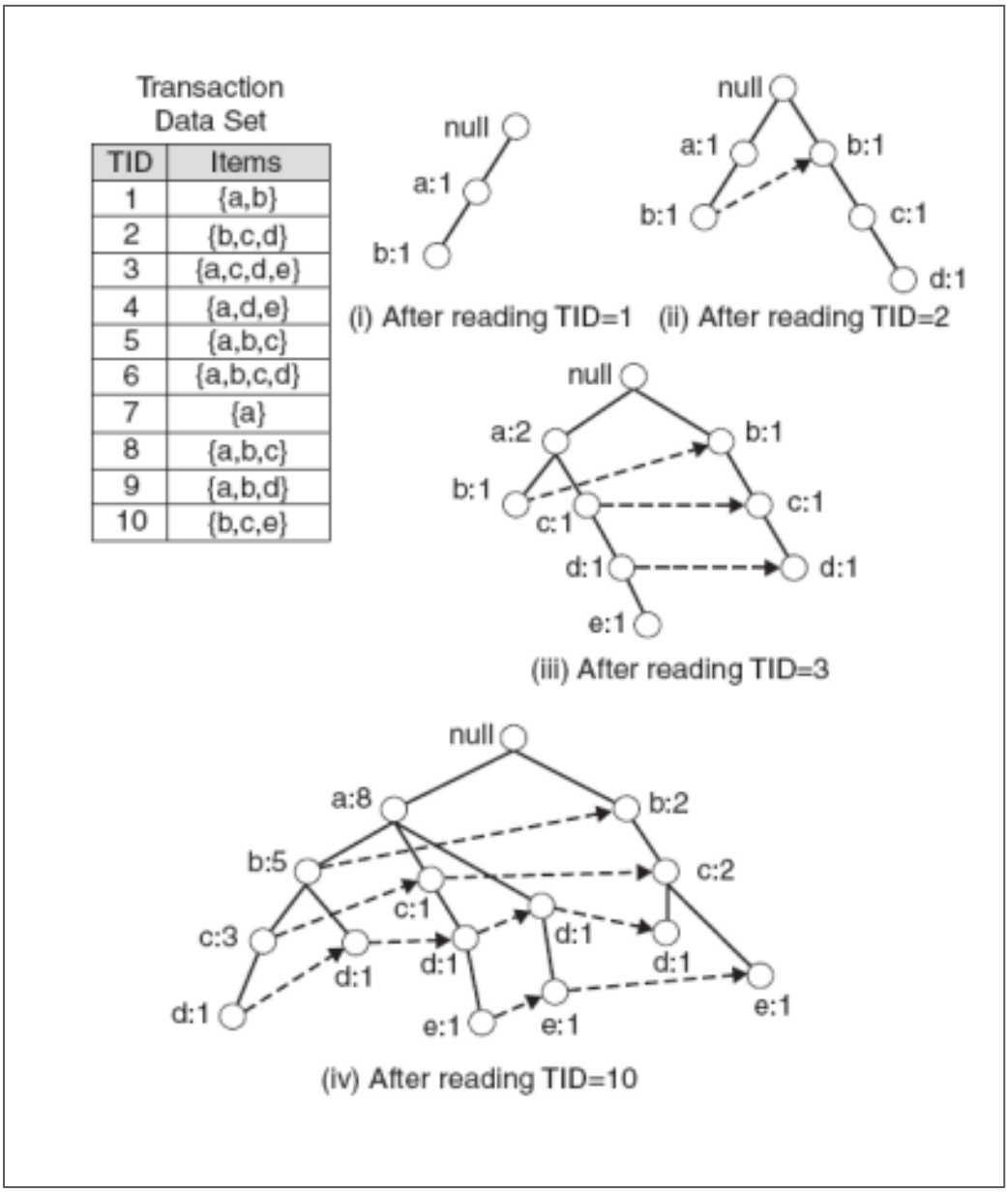


Figure 6: shows the creation of FP-tree

Once the FP-tree is constructed, frequent itemsets can be generated reclusively starting from the bottom of the tree working upwards to find all combinations of itemsets that reach the minimum support threshold.[10]

2.6 Challenges and Limitations

Frequent pattern mining, association rule mining, and correlation analysis are critical methodologies in data mining, yet they face several overarching challenges:

1. Evolving Data Dynamic: [11]

With the continuous growth of global data—from 33 ZB in 2018 to a projected 175 ZB by 2025—adapting mining techniques to handle dynamic and streaming datasets remains a major challenge. Traditional algorithms often struggle with real-time data processing, necessitating advancements in incremental and adaptive learning techniques.

2. Integration into Real-World Applications:

Despite their theoretical robustness, mining algorithms often fall short when applied in real-world scenarios. Issues such as domain-specific customization, integration with existing systems, and stakeholder usability hinder their deployment. Addressing these concerns involves tailoring solutions to industry-specific needs and improving algorithm interpretability.

3. Ethical and Privacy Concerns:

Data mining raises concerns about data security, privacy breaches, and ethical usage, especially when handling sensitive information such as medical or financial records. Solutions involve incorporating privacy-preserving data mining techniques, adhering to regulations, and fostering transparency in data use.

4. Balancing Performance and Interpretability:

Advanced techniques often achieve high accuracy but lack interpretability, making it challenging for stakeholders to trust and apply results. Hybrid models and visualization tools can bridge this gap by presenting findings in an accessible and actionable manner.

2.7 Related Papers:

2.7.1 Paper 1: Application of Data Mining Techniques in Healthcare: Identifying Inter-Disease Relationships through Association Rule Mining [12]:

In this paper, the researcher discusses the application of data mining techniques in healthcare departments to explore the relationship between diseases using the patient hospital visits data with the usage of the international classification of diseases (ICD-10) that provides a system to store, retrieve, and analyze healthcare information.

The paper focuses on using the FP-growth algorithm and association rule mining to seek hidden disease patterns within the data set.

It starts first with collecting the data (patient visits) from X hospital and then preparing the data for applying the association rule mining that uses the FP-

growth algorithm by determining the support threshold and confidence threshold to generate frequent item sets based on the ICD-10 diagnosis codes.

No.	ICD-10 Code	ICD Description	Frequencies
1	Z09.8	Follow-up exam after other treatment for other conditions	9383
2	I10	Essential (primary) hypertension	4102
3	E11	Non-insulin-dependent diabetes mellitus	3761
4	I63.9	Cerebral infarction, unspecified	2675
5	R50.9	Fever, unspecified	2474
6	K30	Dyspepsia	1687
7	H26	Another cataract	1384
8	A91	Dengue hemorrhagic fever	1318
9	A01.0	Typhoid fever	1284

Table 2: Shows the Frequencies of 10 Most Common Diseases

By using the binary format to highlight the patterns and the relationships within the diseases that frequently co-occur with the patient's hospital visits.
The results show the strength of the relationships among the different ICD-10 coded diseases that can help in healthcare management and planning and encourage data-driven methods in healthcare.

2.7.2 Paper 2: Survey on frequent item set mining approaches in market basket analysis [13]

Market Basket Analysis (MBA) is widely used to generate frequent item sets through different association rules. By arranging items adjacently, it becomes easier for customers to make purchases, enhancing their shopping experience. MBA aims to discover purchasing patterns by analyzing transactional data to identify associations between items frequently bought together. This approach is instrumental for retailers, as it informs strategies like shelf arrangement to improve sales and customer satisfaction. Various studies have focused on the application of association rule mining to uncover hidden relationships within large datasets. This method has proven essential for understanding consumer buying behavior and optimizing store layouts to drive sales. Key metrics in association rule mining include support, which measures the frequency of an item set across all transactions, and confidence, which evaluates the likelihood of a customer buying a specific item given that another item is already purchased. These metrics provide valuable insights that help businesses refine their marketing and sales strategies, contributing to decision-making support and enhancing customer relationship management.

2.7.3 Paper 3: Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms [14]:

In this paper, the researcher discusses the concept of web usage mining using data mining techniques to discover usage patterns from web data. By applying these techniques

to the data helps in understanding the web user browsing behavior at a website and discovering usage patterns from web data to better serve the wed user needs.
The researcher focuses on the web server data mining mainly on the server log files that include IP address, page references, and access time, and mainly dive in using the access time of the users to access the user logs.
The main step to operating any algorithm is building a processing model that converts the log files that are normally in ASCII format into a database-like format. Firstly, using the Apriori algorithm to generate candidate item sets shows the efficiency of the algorithm in finding all the frequent item sets but it also shows the main drawback of the Apriori algorithm which is the high cost of handling a huge number of candidates sets the repeatedly scan of the database.
For the researcher to overcome this limitation, he discusses the usage of the FP-growth algorithm which shows a huge improvement that shows in using a compact data structure and limiting the repetitive database scanning. And finally, by addressing the implantation of these algorithms it opens the path for further in web mining techniques.

2.7.4 Paper4: Data Mining market basket analysis using hybrid dimension association rule, case study in minimarket x [15]

This paper introduces Market Basket Analysis (MBA) in the context of a minimarket to analyze customer shopping habits and identify associations and correlations among items in their shopping baskets. The authors explore MBA using a hyper-dimensional approach and employ the Apriori algorithm to identify frequently purchased items that are commonly bought together by customers. The primary goal is to determine frequent itemset, providing insights into customer purchasing behavior. The application is designed to perform multi-dimensional data mining, enhancing the analysis beyond traditional single-dimensional methods. Market Basket Analysis is a process that examines buyer habits to uncover relationships between different items in a shopping cart. Multi-dimensional association rules are used, allowing the extraction of information across multiple attributes or dimensions, offering a more comprehensive view compared to single-dimensional rules. Specifically, multi-dimensional rules, such as two-dimensional association rules, explore relationships between items across different aspects, while inter-dimensional association rules focus on connections without repeating predicates. Basket Association rules that method that allows extracting information in terms of some attributes or dimensions, compared to a single dimension. For example, Age (x, "13-18") work ^ (x, "student") buy (x,"Magazine") [support = 3%, confidence = 71%]. The relationship shows that as the sample of buyers, 3% are aged 13 to 18 years, work as a student, and buy a magazine. There is a 71% chance that the buyer at this age, working as a student and this group will buy a magazine. o Hybrid-dimension association rules (repeated predicates) Age (x, "13-18") buy ^ (x, "magazine") buy (x, "Newspaper") [support = 2%, confidence = 60%] These relationships show that the visitors of the store, 2% are aged 13 to 18 years, there is a 60% chance of a buyer at this age if they buy a magazine also buys a newspaper.

Study	Focus	Contribution	Strength	Weakness
Data Mining Market Basket Analysis Using Hybrid-Dimension Association Rules	Apriori algorithm, hybrid-dimension approach	Demonstrates practical application in retail; highlights support and confidence	- Introduces multi-dimensional association rules. - Provides a case study in a minimarket	- Limited to a specific case study. - Does not address scalability for larger datasets.
Survey on Frequent Item-Set Mining Approaches in Market Basket Analysis	Apriori algorithm	Theoretical overview of frequent itemset mining; highlights scalability issues	-Comprehensive review of MBA techniques. - Highlights key metrics like support and confidence. - Emphasizes practical retail applications.	-Lacks in-depth analysis of algorithm performance. - Does not address challenges like scalability or memory usage.
Implementation of Web Usage Mining Using APRIORI and FP Growth Algorithms	Apriori and FP-Growth algorithms	Compares efficiency of both algorithms in web usage mining	-Demonstrates the application of both algorithms. - Highlights the efficiency of FP-Growth over Apriori. - Focuses on web server log data	- Limited to web usage data. - Does not explore other data types or domains.
Application of Data Mining Techniques in Healthcare	FP-Growth algorithm, ICD-10 disease classification	Identifies co-occurring diseases; aids healthcare management	-Focuses on healthcare, a critical domain. - Uses ICD-10 codes for disease classification. - Demonstrates practical application of FP-Growth.	-limited to binary data format. - Does not explore other algorithms or hybrid approaches

Table 3: Shows the critical analysis of related papers, summarizing their focus, contributions, strengths, and weaknesses.

Chapter Three: Research Methodology

Chapter 3: Research Methodology

This Chapter provides a quick review of the methodology used in the study. Starting with the research design, discusses the data collection and preprocessing steps, ensuring the dataset is suitable for analysis. The algorithms and techniques utilized for frequent pattern mining, specifically the Apriori and FP-Growth algorithms. Finally, the chapter outlines the testing and evaluation procedures, including the metrics used to assess the performance and relevance of the extracted patterns.

3.1 Research Design

This study aims to analyze and compare the Apriori and FP-Growth algorithms for mining frequent patterns, associations, and correlations in large datasets. The primary purpose of this design is to synthesize existing knowledge and research findings related to the two algorithms, evaluating their objectives, advantages, limitations, memory usage, and metrics used for algorithm evaluation. By conducting a thorough review of the literature, the study aims to provide a comprehensive understanding of the effectiveness and efficiency of these algorithms in different data mining applications, particularly in the context of market basket analysis.

A theoretical review is the most appropriate design for this study because:

- We can explore the nature of the research, where it focuses on reviewing previously published results to gain insights into the algorithms' performance and limitations across different contexts.
- A theoretical review enables the identification of similarities and differences between the Apriori and FP-Growth algorithms, based on various factors like computational efficiency, scalability, and application domains.
- The study uses a lot of published research studies, and research papers that have already conducted experiments on these algorithms, which allows for a thorough examination without the need for additional experimental work.

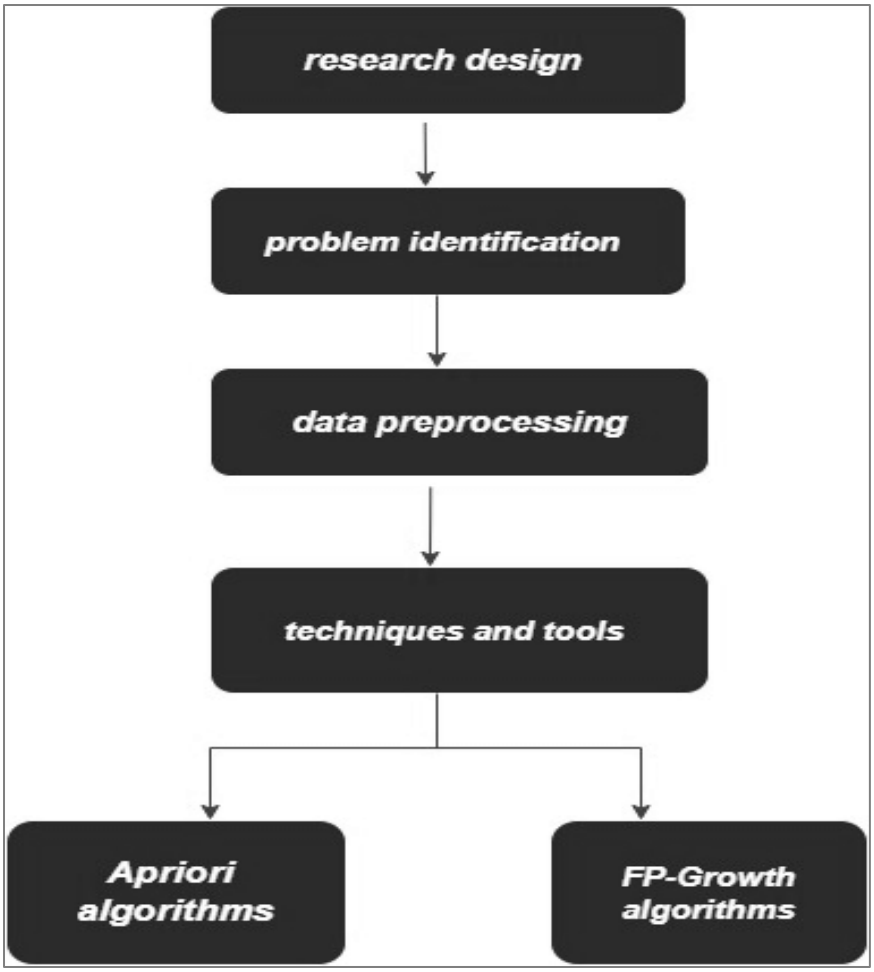


Figure 7: Show the different levels of research design from problem identification to techniques and tools

3.2 Problem identification:

We analyze a large dataset to illustrate how mining frequent patterns can provide actionable insights in specific domains such as retail (e.g., market basket analysis) and healthcare (e.g., inter-disease relationships).

3.3 Data collection:

The data can be collected by several methods:

1. Automated data extraction: where data is collected using web scraping (APIs) or collecting transaction logs.
2. Surveys and questionnaires: where data is collected by interviewing the target group in a specified domain.
3. Publicly available data: where data is provided in online public repositories such as **Kaggle** which is a website which is a subsidiary of Google that provides a community for data scientists and machine learning engineers that allow users to find and publish datasets.

3.4 Data preprocessing:[16]

The huge real-world datasets face many problems such as incomplete data (Missing Values), Noise, and Inconsistency, the preparation of these data guarantees more efficient and intelligent data analysis.

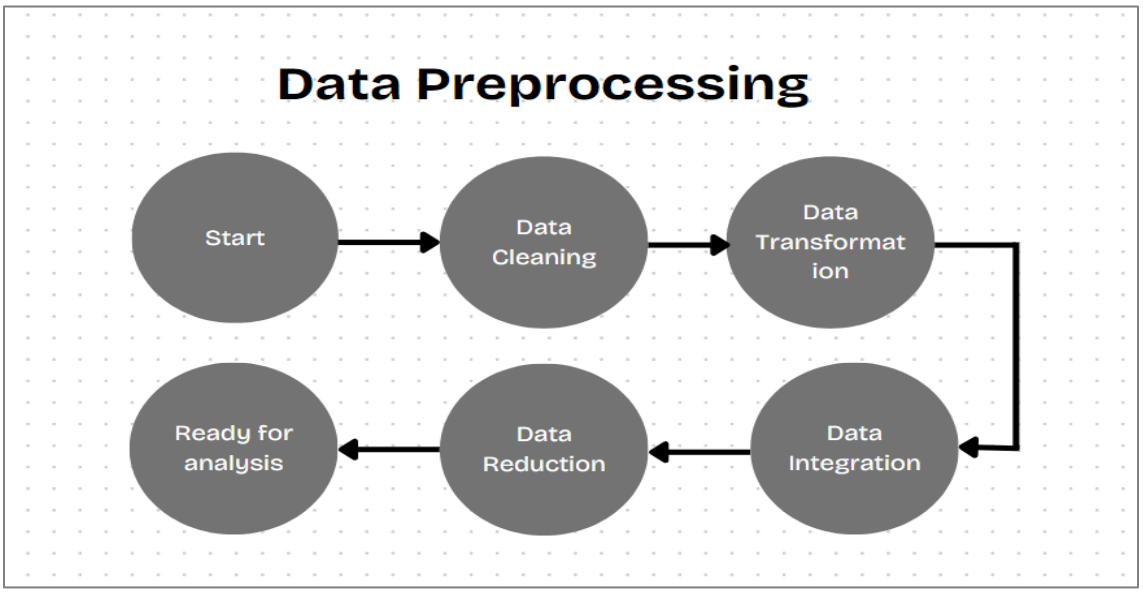


Figure 8: Shows the four stages of data preprocessing

1. Data cleaning:

The data cleaning attempts to fill in the missing values in the datasets and remove noise and outliers to achieve the consistency of the data.

1. Missing values can be handled by:
   - ignoring the whole tuple that contains a missing value.
   - fill in the missing values manually
   - use global constant to fill in the missing value
   - use the attribute mean to fill in the missing value
   - use the attribute mean for all samples belonging to the same class in the given tuple
   - use the most probable value to fill in the missing value.
2. Noisy data can be handled by:
   - binning methods: to smooth a sorted data value by consulting the values around its "Neighborhood".
   - clustering: Identifying outliers can be done by clustering similar values into groups or clusters.
   - regression: smoothing the data by fitting it into a function.
4. Inconsistence data: some data inconsistency can be corrected manually using external references, or using Knowledge engineering tools such as "WEKA" to find values that do not go with the functional constraints, finally for the issue of data inconsistencies due to databases integration can be handled by redundancies.

2. Data integration:

The process of data integration may lead to inconsistency in data and cause redundancies in the resulting datasets. These redundancies can be detected by correlation analysis, Numerical data are handled by correlation coefficient (Pearson's coefficient) while categorical data is handled by the chi-square test.

3. Data transformation:

The process of transforming data into forms appropriate for mining, and it's done by:

- smoothing: remove the noise from the data.
- Aggregation: Appling aggregation operations on data.
- Generalization of the data: where low-level data are replaced by higher-level concepts.
- Normalization: to make data fall within a small specific range (-1,1) (1,0).

4. Data reduction:

This is done to obtain a smaller representation of the data set that produces the same analytical results. It can be done by:

- data cube aggregation: applying aggregation operations on the data.
- dimension reduction: remove redundant and irrelevant data.
- data compression: use encoding to reduce the data size.
- numerosity reduction: replacing data with smaller representation data.
- Discretization and concept hierarchy generation: replace data with higher conceptual levels.

3.5 Methods and Techniques

3.5.1 Algorithms and Tools:

- **Apriori Algorithm [1]:**

Steps and Logic:

- Initialization: Start by identifying all the individual items in the dataset that meet the minimum support threshold.
- Candidate Generation: Generate candidate itemset of length k from frequent itemset of length k-1.
- Support Counting: Count the occurrences of each candidate itemset in the dataset.
- Prune: Eliminate candidate items that do not meet the minimum support threshold.
- Repeat: Repeat steps 2-4 until no more candidate itemset can be generated.

Algorithm: Apriori. Find frequent itemsets using an iterative level-wise approach based on candidate generation.

Input:

- D, a database of transactions;
- *min_sup*, the minimum support count threshold.

Output: L, frequent itemsets in D.

Method:

```
(1)   L1 = find_frequent_1-itemsets(D);
(2)   for (k = 2; Lk−1 ≠ ϕ; k++) {
(3)       Ck = apriori_gen(Lk−1);
(4)       for each transaction t ∈ D { // scan D for counts
(5)           Ct = subset(Ck, t); // get the subsets of t that are candidates
(6)           for each candidate c ∈ Ct
(7)               c.count++;
(8)       }
(9)       Lk = {c ∈ Ck|c.count ≥ min_sup}
(10)  }
(11)  return L = ∪k Lk;
```

Figure 9: Shows the step of pseudocode of the Apriori Algorithm

- **FP-Growth Algorithm**

1. Construction of the FP-tree:
   - Scan Dataset: Perform an initial scan of the dataset to identify frequent items and their support counts.
   - Sort and Prune: Sort items in each transaction in descending order of their frequency and remove infrequent items.
   - Build FP-tree: Insert transactions into the FP-tree, creating branches for each transaction path. Update counts for common prefixes.
2. Traversal of the FP-tree:
   - Conditional Pattern Base: For each frequent item, extract the paths in the FP-tree that contain the item.
   - Conditional FP-tree: Construct a conditional FP-tree from the conditional pattern base.
   - Recursive Mining: Recursively mines the conditional FP-tree to find frequent patterns.

Algorithm: **FP_growth.** Mine frequent itemsets using an FP-tree by pattern fragment growth.

Input:

- D, a transaction database;
- *min_sup*, the minimum support count threshold.

Output: The complete set of frequent patterns.

Method:

1. The FP-tree is constructed in the following steps:
   - **(a)** Scan the transaction database D once. Collect F, the set of frequent items, and their support counts. Sort F in support count descending order as L, the *list* of frequent items.
   - **(b)** Create the root of an FP-tree, and label it as "null." For each transaction *Trans* in D do the following.
     Select and sort the frequent items in *Trans* according to the order of L. Let the sorted frequent item list in *Trans* be $[p|P]$, where p is the first element and P is the remaining list. Call insert_tree($[p|P]$, T), which is performed as follows. If T has a child N such that $N.item\text{-}name = p.item\text{-}name$, then increment N's count by 1; else create a new node N, and let its count be 1, its parent link be linked to T, and its node-link to the nodes with the same *item-name* via the node-link structure. If P is nonempty, call insert_tree(P, N) recursively.
2. The FP-tree is mined by calling FP_growth(*FP_tree*, *null*), which is implemented as follows.

```
procedure FP_growth(Tree, α)
(1)  if Tree contains a single path P then
(2)      for each combination (denoted as β) of the nodes in the path P
(3)          generate pattern β ∪ α with support_count = minimum support count of nodes in β;
(4)  else for each a_i in the header of Tree {
(5)      generate pattern β = a_i ∪ α with support_count = a_i.support_count;
(6)      construct β's conditional pattern base and then β's conditional FP_tree Tree_β;
(7)      if Tree_β ≠ ∅ then
(8)          call FP_growth(Tree_β, β); }
```

Figure 10: shows the steps of the pseudocode of the FP-Growth Algorithm

- **Comparison between algorithms [17]**

Algorithm	Objectives	Advantages	Limitation	Memory usage
Apriori	priori is a data mining methodology that recognizes frequent patterns and association rules in big transactional databases	-simple and easy to understand algorithm -Multiple scans	-Time-consuming due to expensive Candidate generation -Excessive usage of memory	Higher
FP-Growth	FP-growth proposes a method for Finding frequent patterns without candidate generation	-Faster than Apriori -No candidate generation	-Running time complexity -FP-tree mapping complexity	Lower

Table 4: Shows the comparison between Apriori and FP-growth algorithms

3.5.2 Metrics for Association Rules [1]

Association rule mining is a data mining process used to discover meaningful relationships, patterns, and correlations within large datasets. The effectiveness of these relationships is measured using various evaluation metrics. Below are five commonly used metrics in association rule mining and correlation:

1. **Support**

- Support measures the proportion of transactions that contain both the antecedent and consequent of a rule, showing how often the rule occurs and its relevance. It determines the minimum frequency at which an item must be considered significant. A low support value helps capture rare patterns but may include noise, while a high value focuses on popular patterns, potentially overlooking subtle associations.

- Mathematically, Support is calculated as:

$$Support(A \Rightarrow B) = P(A \cup B)$$

Where:

$P(A \cup B)$: represents the proportion of transactions in the dataset that contain both A and B

2. Confidence:

- The likelihood that an item B is purchased when item A is purchased. Higher confidence values indicate stronger association rules.
- Selection: Balances between precision and overgeneralization; a value too high may overlook useful patterns, while too low may allow spurious rules.
- Mathematically, Confidence is calculated as:

$$Confidence(A \Rightarrow B) = P(B|A) = \frac{support\_count(A \cup B)}{support\_count(A)}$$

3. Conviction:

- Used to measure the degree of dependence of the consequent on the antecedent in association rule mining. It considers the occurrence of the consequent when the antecedent is not present and helps assess how strong the rule is in terms of predictive power.
- The mathematical formula for Conviction is:

$$Conviction(A \Rightarrow B) = \frac{1 - P(A)}{1 - P(A \Longrightarrow B)}$$

4. Lift:

- Measures the strength of association between two items. A lift value greater than 1 indicates a positive correlation, while a value less than 1 indicates a negative correlation.
- Mathematically, lift is calculated as:

$$lift(A, B) = \frac{P(A \cup B)}{P(A)\ .\ P(B)}$$

Where:

$P(A \cup B)$: is the probability of both items A and B occurring together.

$P(A)\ .\ P(B)$: is the product of their probabilities, representing their expected co-occurrence under independence.

5. Chi-Square Test:

- Tests the independence of two items. A high Chi-Square value suggests a strong association between the items.
- we take the squared difference between the observed and expected value for a slot (A and B pair) in the contingency table, divided by the expected value.

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Where:

Observed: The actual frequency of an event in the dataset.

Expected: The frequency expected under the assumption that the two variables are independent.

3.5.3 Software and Tools

- **Python** is a high-level, general-purpose programming language known for its simplicity, versatility, and ease of use. It is widely used in various fields, including data analysis, machine learning, web development, and automation. Python emphasizes readability with its clear and concise syntax, making it an excellent choice for beginners and experienced developers alike.
- A Python library refers to a collection of related modules linked together. It contains a code bundle that developers can use repeatedly in different programs
- There are libraries used in the research:
- **Pandas:** Used for data manipulation and preprocessing. Provides robust data structures like Data Frames, enabling seamless handling of transaction datasets.
- **NumPy:** Essential for numerical computations and matrix operations, which are often required in algorithms like matrix factorization.
- **extend:** Contains specific modules for frequent pattern mining and association rule generation, such as the Apriori and FP-Growth implementations.
- **Matplotlib and Seaborn:** Used for data visualization to present insights like item associations and transaction patterns graphically.

✓ **Why Python?**

- **Open Source:** Freely available with an active community contributing to its development.
- **Rich Ecosystem:** A wide range of libraries and frameworks for different domains, including data science, data mining, machine learning, and visualization.
- **Cross-Platform:** Works on all major operating systems (Windows, macOS, Linux).
- **Easy to Learn:** Python's syntax is straightforward, resembling natural language.

Chapter four: implementation and results

Chapter 4: implementation and results

This chapter provides an applied example that will be reviewed in which it illustrates the frequent pattern and association rules implementation, how the algorithms (**Apriori, FP-growth**) are applied, and a comparison between the efficiency of these algorithms in achieving the required results. Finally, address the conclusions based on what has been drawn from the results.

4.1 Market basket analysis (MBA):

Market basket analysis is one of the most popular retail applications of frequent pattern mining, works on analyzing transactions that each shows the products that have been bought by customers and it's referred to as itemset. These itemsets are analyzed to identify any patterns that guarantee a relationship between the items that are bought together. The process of discovering frequent patterns from data sets that are continuously collected provides a large base of transactional datasets which facilitate the process of discovering the correlation between items.

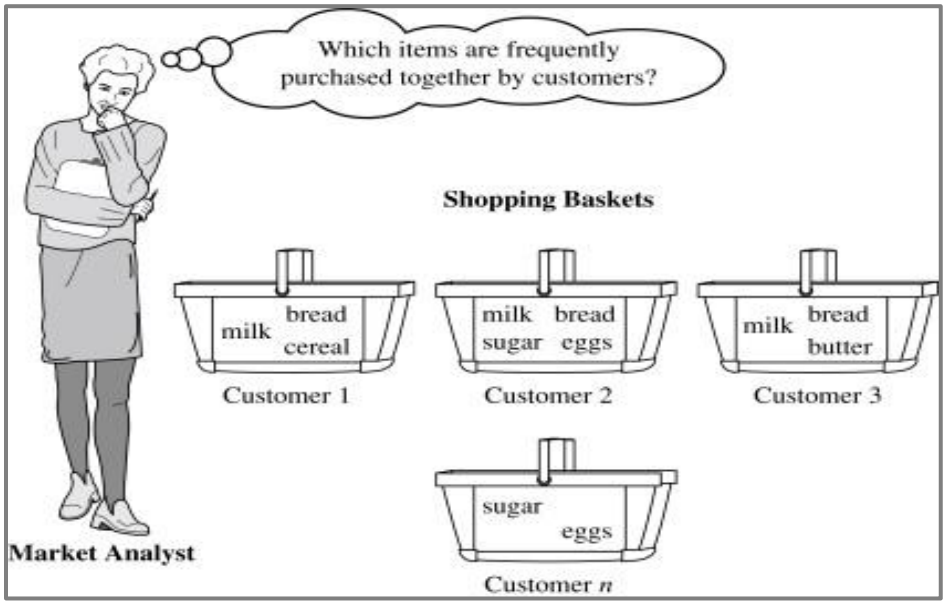


Figure 11: describes the idea of market basket analysis and frequent pattern mining

The real value behind the frequent pattern mining in the retail industry is to help in understanding the customers' purchasing behaviors which companies benefit from such as Online platforms like Flipkart and Amazon use this technique for personalized recommendations, while physical stores can optimize layouts and inventory management. For example, supermarkets often bundle low-selling items like specialty cheeses with popular items like crackers or bread at a discounted price, encouraging customers to try the cheese.

Overall market basket analysis and the storage of data in digital records made it easier to process and analyze large volumes of data that are used in many aspects in real-world.

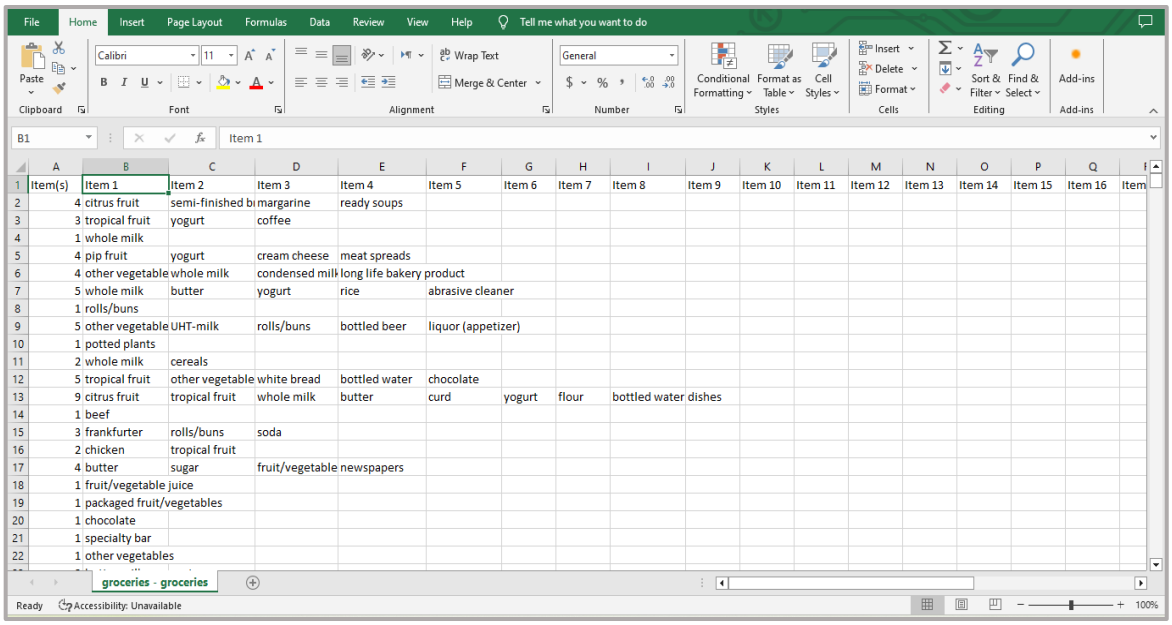
4.2 Case study:

The research" [paper **"Market Basket Analysis Using Association Rule Mining and Apriori/FP-Growth Algorithm by Aman Desai-(19IT031), Hardik Gandhi-(19IT037), Jalpesh Vasa"** [18] provides a full explanation of how to implement Apriori and FP-growth algorithm in a given data set by using Python and representing results to concluded.

The practical example will be illustrated step by step in the sections below:

4.2.1 Data description:

In this example, a dataset of 9835 transactions of customers shopping for groceries was used, the data contains 169 unique items. It was provided by Dr. Aruna Malapati, Assistant Professor in BITS Pilani, Hyderabad Campus as a Course Assignment for CS F415- Data Mining (2019). The grocery data are listed starting on the left side ordering the items within each basket.[19]



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Item(s)	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16
2	4	citrus fruit	semi-finished bi	margarine	ready soups												
3	3	tropical fruit	yogurt	coffee													
4	1	whole milk															
5	4	pip fruit	yogurt	cream cheese	meat spreads												
6	4	other vegetable	whole milk	condensed milk	long life bakery	product											
7	5	whole milk	butter	yogurt	rice	abrasive cleaner											
8	1	rolls/buns															
9	5	other vegetable	UHT-milk	rolls/buns	bottled beer	liquor (appetizer)											
10	1	potted plants															
11	2	whole milk	cereals														
12	5	tropical fruit	other vegetable	white bread	bottled water	chocolate											
13	9	citrus fruit	tropical fruit	whole milk	butter	curd	yogurt	flour	bottled water	dishes							
14	1	beef															
15	3	frankfurter	rolls/buns	soda													
16	2	chicken	tropical fruit														
17	4	butter	sugar	fruit/vegetable	newspapers												
18	1	fruit/vegetable	juice														
19	1	packaged fruit/vegetables															
20	1	chocolate															
21	1	specialty bar															
22	1	other vegetables															

Figure 12: shows an Excel sheet that contains all the 9835 items that were collected

4.2.2 Data preprocessing:

Start by loading the dataset:

```
In [2]:
dataset=pd.read_excel("D:\Aman\Python\groceries.xlsx",header=None)
groc_data = pd.DataFrame(dataset)
groc_data
```

Figure 13: shows the code to load the data

Out[2]:

	0	1	2	3	4	5	6	7	8	9	...	22	23	24
0	citrus fruit	semi-finished bread	margarine	ready soups	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
1	tropical fruit	yogurt	coffee	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
2	whole milk	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
3	pip fruit	yogurt	cream cheese	meat spreads	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
4	other vegetables	whole milk	condensed milk	long life bakery product	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9830	sausage	chicken	beef	hamburger meat	citrus fruit	grapes	root vegetables	whole milk	butter	whipped/sour cream	...	NaN	NaN	NaN
9831	cooking chocolate	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
9832	chicken	citrus fruit	other vegetables	butter	yogurt	frozen dessert	domestic eggs	rolls/buns	rum	cling film/bags	...	NaN	NaN	NaN
9833	semi-finished bread	bottled water	soda	bottled beer	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN
9834	chicken	tropical fruit	other vegetables	vinegar	shopping bags	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN

Table 5: shows the loaded data in the Python compiler

Then the data is encoded to (0,1) for an item in a particular transaction if it was purchased within it puts its entry as 1, and if not put as 0. The encoding transforms the categorical data into numerical data so that the algorithms can process it effectively

.

In [3]:
```
encoding = []

for i in range(0, 9835):
    encoding.append([str(groc_data.values[i,j]) for j in range(0, 32)])

# conveting it into an numpy array
encoding = np.array(encoding)
```

In [4]:
```
te = TransactionEncoder()

groc_data = te.fit(encoding).transform(encoding)
groc_data = pd.DataFrame(groc_data, columns = te.columns_)
groc_data=groc_data.drop(['nan'],axis=1).astype('int')
groc_data
```

Figure 14 : shows the code for decoding the data into 1,0

Out[4]:

	Instant food products	UHT-milk	abrasive cleaner	artif. sweetener	baby cosmetics	baby food	bags	baking powder	bathroom cleaner	beef	...	turkey	vinegar	waffles	whipped c
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
9830	0	0	0	0	0	0	0	0	0	1	...	0	0	0	
9831	0	0	0	0	0	0	0	0	0	0	...	0	0	0	
9832	0	0	0	0	0	0	0	0	0	0	...	0	0	0	
9833	0	0	0	0	0	0	0	0	0	0	...	0	0	0	
9834	0	0	0	0	0	0	0	0	0	0	...	0	1	0	

Table 6: shows the data after decoding

4.2.3 Algorithms implementation:

- **Apriori algorithm:**
  This code snippet implements association rule mining using the Apriori algorithm:

```
frequent_items = ap(groc_data, min_support=0.001, use_colnames=True)
# frequent_items = frequent_items.drop([60, 61], axis=0)

most_pop_items = frequent_items.sort_values('support', ascending=False)
# most_pop_items

most_pop_items = most_pop_items.head(15)
most_pop_items
```

Figure 15: shows code Identifying Frequent Itemsets using Apriori

	support	itemsets
154	0.255516	(whole milk)
96	0.193493	(other vegetables)
115	0.183935	(rolls/buns)
130	0.174377	(soda)
155	0.139502	(yogurt)
10	0.110524	(bottled water)
116	0.108998	(root vegetables)
147	0.104931	(tropical fruit)
125	0.098526	(shopping bags)
122	0.093950	(sausage)
99	0.088968	(pastry)
28	0.082766	(citrus fruit)
9	0.080529	(bottled beer)
89	0.079817	(newspapers)
18	0.077682	(canned beer)

Table 7: show the top 15 Most Frequent Items (Support Values)

The dataset included 9,835 transactions. The Apriori algorithm was applied with a minimum support value of 0.001, iterating until this threshold was reached. It took about 2 minutes and 45 seconds to identify the frequent itemsets on a Windows 10 64-bit system with an Intel i5-9300H processor and 8GB of RAM. It identified notable patterns and relationships, highlighting that "whole milk" topped the list of most purchased items, followed by "other vegetables," "rolls/buns," and "soda."

- **FP-Growth algorithm:**

The paper implements several measures for association rule mining, including support, confidence, and lift, along with code execution. Now apply the FP-Growth Algorithm with a minimum support value of 0.001. This was analyzed with the same datasets and support threshold, and since FP-Growth scans the data only twice, the entire process was completed in just 2-3 seconds on the same system

```
freq_items = fpgrowth(encod_df , min_support = 0.005 , use_colnames = True)
freq_items
```

Figure 16 : show the FP-growth code of support itemsets

	support	itemsets
0	0.082766	(citrus fruit)
1	0.058566	(margarine)
2	0.017692	(semi-finished bread)
3	0.139502	(yogurt)
4	0.104931	(tropical fruit)
...	...	...
13417	0.001220	(nuts/prunes, whole milk)
13418	0.001017	(nuts/prunes, rolls/buns)
13419	0.001220	(tidbits, rolls/buns)
13420	0.001017	(tidbits, soda)
13421	0.001322	(whole milk, cooking chocolate)

13422 rows × 2 columns

Table 8: shows that itemsets and support

The dataset contains 13,422 rows with two columns: itemsets and their corresponding support values. Single items generally have support values significantly below the min_support threshold (0.001). However, when items are associated with others, their combined support values increase, often approaching or exceeding the min_support threshold.

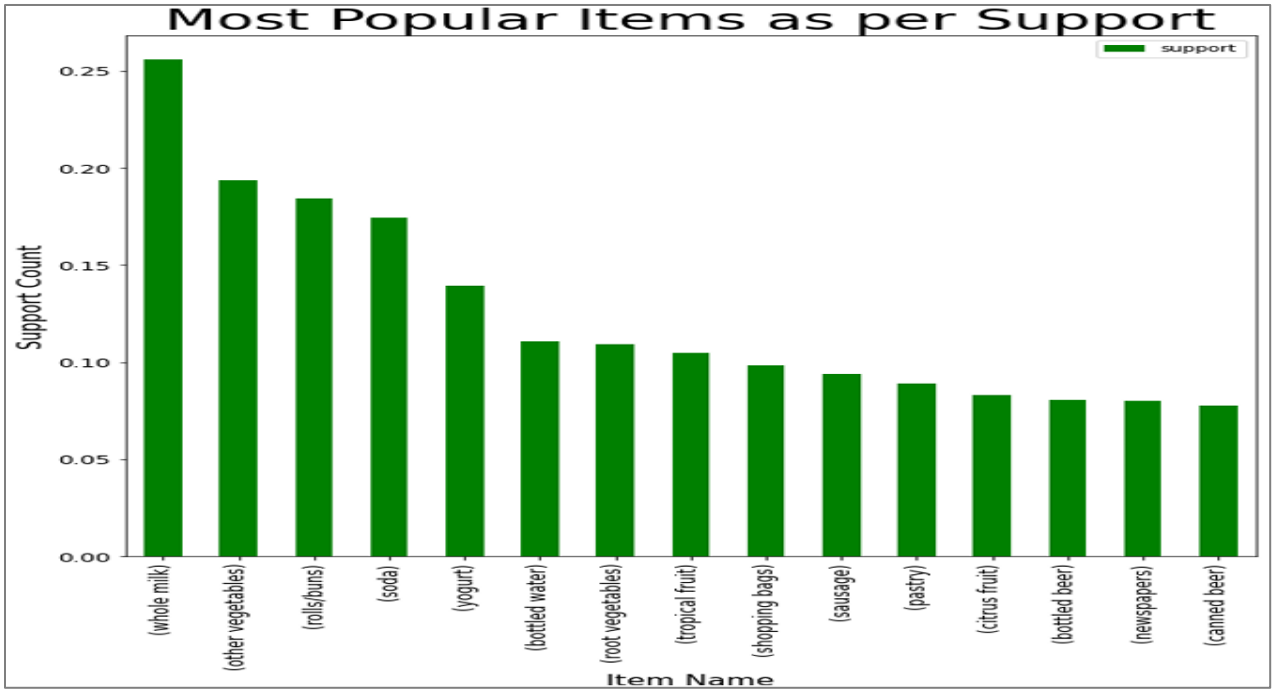


Figure 17: shows the top 15 most frequent itemsets in ascending order as per support.

Shows the popularity of items based on support counts, from ascending order the whole milk (0.255516) leading, followed by other vegetables, rolls/buns, …, and canned beer (0.077682). These insights are useful for optimizing inventory management and marketing strategies.

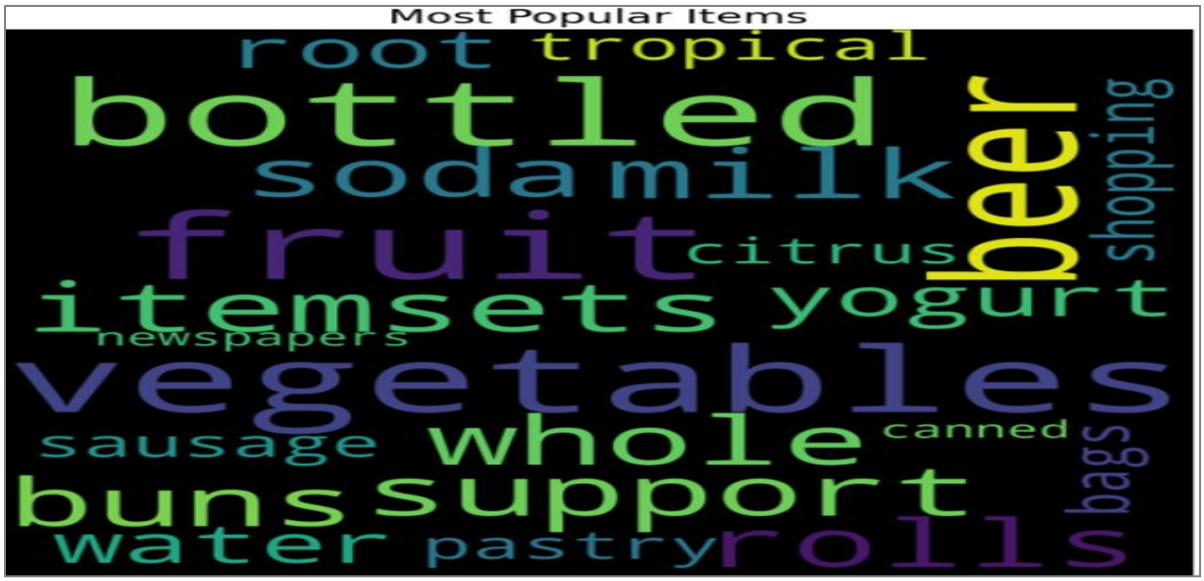


Figure 18 : show the top 15 most popular items in the form of the word cloud.

The largest words, such as "bottled," "beer," "fruit," and "vegetables," indicate they are the most frequently chosen items. Other notable items include "soda," "milk," "yogurt," and "water." This visual depiction provides a clear and engaging way to understand consumer preferences and trends.

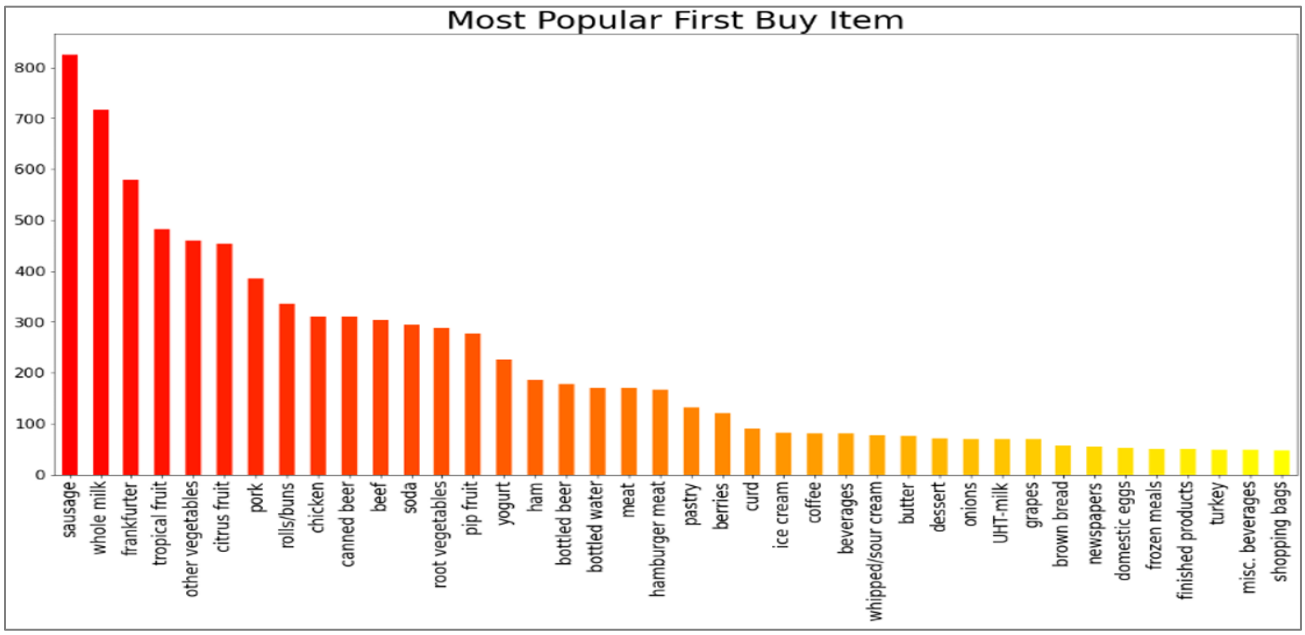


Figure 19 : shows the top 40 first-buy items from the dataset

The bar chart reveals consumer preferences, with sausages leading the purchases at over 800, while items like shopping bags and turkey lag under 100. The red-to-yellow gradient enhances the visual appeal, highlighting popular and less popular items.

	antecedents	consequents	support	confidence	lift	leverage	conviction
887	(root vegetables, tropical fruit, yogurt)	(whole milk)	0.005694	0.700000	2.739554	0.003616	2.481613
842	(root vegetables, pip fruit, other vegetables)	(whole milk)	0.005491	0.675000	2.641713	0.003412	2.290720
375	(whipped/sour cream, butter)	(whole milk)	0.006711	0.660000	2.583008	0.004113	2.189659
726	(whipped/sour cream, pip fruit)	(whole milk)	0.005999	0.648352	2.537421	0.003635	2.117126
377	(yogurt, butter)	(whole milk)	0.009354	0.638889	2.500387	0.005613	2.061648
370	(root vegetables, butter)	(whole milk)	0.008236	0.637795	2.496107	0.004936	2.055423
454	(tropical fruit, curd)	(whole milk)	0.006507	0.633663	2.479936	0.003883	2.032240
828	(root vegetables, whole milk, citrus fruit)	(other vegetables)	0.005796	0.633333	3.273165	0.004025	2.199566
849	(yogurt, pip fruit, other vegetables)	(whole milk)	0.005084	0.625000	2.446031	0.003005	1.985291
475	(domestic eggs, pip fruit)	(whole milk)	0.005389	0.623529	2.440275	0.003181	1.977536

Table 9 : shows the top 10 items that have the highest confidence and support value greater than or equal to 0.005.

{Yogurt, Butter} as Antecedents (items that are frequently bought together) and {Whole, Milk} as Consequents (items likely to be purchased when the antecedents are bought) having support count as 0.009354 and confidence of roughly 64%. So, we can say that the chances of buying whole milk are 64% if the customer purchases Yogurt and Butter.

	antecedents	consequents	support	confidence	lift	leverage	conviction
210	(whole milk, tropical fruit)	(root vegetables, yogurt)	0.005694	0.134615	5.212371	0.004602	1.125712
213	(root vegetables, yogurt)	(whole milk, tropical fruit)	0.005694	0.220472	5.212371	0.004602	1.228567
214	(whole milk, yogurt)	(root vegetables, tropical fruit)	0.005694	0.101633	4.828814	0.004515	1.089703
209	(root vegetables, tropical fruit)	(whole milk, yogurt)	0.005694	0.270531	4.828814	0.004515	1.294059
145	(whole milk, other vegetables)	(root vegetables, pip fruit)	0.005491	0.073370	4.716272	0.004326	1.062390
142	(root vegetables, pip fruit)	(whole milk, other vegetables)	0.005491	0.352941	4.716272	0.004326	1.429801
9	(ham)	(white bread)	0.005084	0.195312	4.639851	0.003988	1.190407
8	(white bread)	(ham)	0.005084	0.120773	4.639851	0.003988	1.107758
164	(root vegetables, tropical fruit)	(whole milk, other vegetables)	0.007016	0.333333	4.454257	0.005441	1.387748
169	(whole milk, other vegetables)	(root vegetables, tropical fruit)	0.007016	0.093750	4.454257	0.005441	1.080224

Table 10: shows the top 10 items that are bought together having the highest lift.

List of association rules, highlighting relationships between different product combinations. Each rule consists of consequences. Metrics that provide insights into the strength and significance of these associations. For instance, the rule "(whole milk, tropical fruit) => (root vegetables, yogurt)" with a high lift value of 5.212371 suggests a strong association between these product groups.

4.2.5 Results:

This research highlights the value of data mining techniques, particularly Market Basket Analysis, in boosting retail profitability and efficiency. Using algorithms like Apriori and FP-Growth, retailers can identify frequent itemsets and derive association rules to better understand customer purchasing behavior. These insights support strategies such as product placement, targeted promotions, and optimized pricing

The comparison between Apriori and FP-Growth algorithms demonstrates that both yield identical results, but FP-Growth is significantly more efficient for large datasets, completing the analysis in under 5 seconds compared to Apriori's 2 minutes and 45 seconds.

1. **Strengths:**
   - **Comprehensive Coverage:** The research effectively covers key concepts like Market Basket Analysis, association rule mining, and algorithms like Apriori and FP-Growth.
   - **Practical Focus:** It emphasizes real-world applications, such as improving business decisions in retail through optimized product placement and targeted promotions.
   - **Algorithm Comparison:** The research highlights the efficiency of FP-Growth over Apriori for practical use.
   - **Clear Explanations:** Key concepts such as support, confidence, and lift are explained clearly.
   - **Case Study:** A case study illustrates the practical application of the techniques.

2. **Weaknesses:**

- **Lack of Comprehensive Analysis of Challenges**: It does not deeply address challenges such as memory limitations or handling sparse data when working with massive datasets.
- **Over-Simplification of Algorithm Performance**: Although it compares the efficiency of FP-Growth and Apriori, the research does not explore how performance may vary with different data types, such as high-dimensional or large datasets.
- **Absence of Alternative Algorithms**: The focus is primarily on Apriori and FP-Growth, missing an exploration of other algorithms or hybrid approaches, which would offer a more comprehensive comparison.
- **No Discussion on Scalability**: There is minimal discussion on how these techniques scale when applied to exponentially larger datasets, which is crucial as retail databases continue to grow.

3. **Challenges and limitations:**

- **Memory and Scalability:** Both Apriori and FP-Growth face challenges with memory consumption, especially Apriori. Scalability issues arise when dealing with massive datasets, impacting performance.
- **Processing Time and Resource Intensity**: The sheer volume of big data leads to increased processing time and demands significant computational resources.
- **Real-time Processing:** The algorithms are not well-suited for real-time data processing, which is critical in many industrial settings.
- **Integration and Interpretability:** Adapting and interpreting the results for specific industrial needs and stakeholders can be complex.

Chapter five: Conclusion and recommendations

Chapter 5: Conclusion and Recommendations:

This chapter provides a conclusion regarding the concepts that have been covered in this research, additionally with some recommendations related to improvements and future research.

5.1 Conclusion:

This project provides a comprehensive exploration of data mining techniques, with a particular focus on frequent pattern mining, association rule mining, and correlation analysis. It delivers a detailed analysis of the Apriori and FP-Growth algorithms, emphasizing their methodologies, practical applications, and efficiency in market basket analysis. Frequent pattern mining is pivotal in uncovering hidden relationships within large datasets, enabling businesses to gain deeper insights into customer behavior and make data-driven decisions that significantly impact their profitability. By applying these techniques to market basket analysis, the project demonstrates their effectiveness in understanding purchasing behavior, optimizing inventory management, and enhancing overall profitability. The comparative analysis reveals that while Apriori and FP-Growth produce identical results, FP-Growth outperforms speed and efficiency, especially for larger datasets, due to its compact data structure and reduced database scans. The insights derived from frequent pattern mining have broad applications across various domains, including fraud detection, network security, and scientific research. As data volumes grow exponentially, scalable and efficient algorithms for frequent pattern mining will be essential for extracting valuable insights from increasingly large and complex data streams. This project not only highlights the strengths and limitations of existing algorithms but also provides a clear vision for future advancements in the field, emphasizing the need for integration with business objectives, enhanced visualization tools, and the exploration of advanced algorithms to refine and expand the applications of pattern mining techniques.

5.2 Recommendation:

1. Algorithm Selection

Selecting the algorithm is vital for achieving efficient and meaningful results in pattern mining. The Apriori algorithm is best suited for small datasets, educational applications, and scenarios where understanding the step-by-step candidate generation process is critical. On the other hand, FP-Growth is more suitable for larger datasets, real-time analytics, and high-dimensional data, delivering significant improvements in efficiency and scalability.

2. Real-World Integration

Frequent pattern mining techniques must align closely with business objectives to maximize their impact. In retail, these techniques can inform the design of effective promotions, such as product bundling, and optimize inventory management based on co-purchased items. In e-commerce, they can enhance recommendation systems and enable targeted marketing campaigns. Similarly, in healthcare frequent pattern mining can uncover co-occurring symptoms or medication patterns, aiding in diagnosis and treatment strategies.

3. Visualization for Decision Making

Clear and impactful presentation of results is essential for decision-making. Using graphical representations such as network diagrams, bar charts, and heat maps can effectively highlight relationships and patterns. Interactive dashboards, built with tools like Tableau or Power BI, enable stakeholders to explore data dynamically, providing deeper insights and aiding in strategic planning.

4. Future Research and Enhancement

Advancing frequent pattern mining involves incorporating new dimensions and methodologies. Temporal pattern analysis can reveal how associations evolve, such as seasonal trends in purchasing behavior. Combining transactional data with contextual information like demographics and geographic data can provide a richer understanding of customer behaviors. Additionally, exploring advanced algorithms like ECLAT or hybrid approaches can offer new opportunities to refine and expand the application of pattern mining techniques.

5. Bundling Offers

Create product bundles based on frequently co-purchased items. For instance, if (yogurt) and (strawberry) are often bought together, offer them as a discounted bundle to increase sales volume.

6. Customer Segmentation

Involves dividing customers into groups based on their purchasing behaviors to tailor marketing strategies effectively. Through behavioral segmentation, retailers can identify specific buying patterns, such as customers who frequently purchase (organic products), and target them with promotions for new organic items. Additionally, by using the lift metric, businesses can pinpoint high-value customers who buy high-margin items like (water) and (cheese), allowing for focused retention strategies and personalized offers to enhance customer loyalty and profitability.

7. Future research and development

In market basket analysis should focus on hybrid approaches that integrate the strengths of both Apriori and FP-Growth algorithms to enhance efficiency and accuracy in identifying frequent patterns. Additionally, exploring temporal analysis can provide valuable insights into how customer purchasing behavior evolves over time, such as seasonal trends in buying habits (e.g., increased purchases of winter clothing during colder months). These advancements will enable retailers to better anticipate customer needs, optimize inventory, and tailor marketing strategies to dynamic consumer preferences.

Reference: -

1. Han, J., Kamber, M., & Pei, J. (2012), **Data Mining Concepts and Techniques,** 3rd edition, British Library, ISBN 978-0-12-381479-1, ch:6, pp:243-270.
2. Chowdary, A., Chamarti, S., Reddy, A. L., Babu, Y. M., & Radha, K. (2019, April). **Mining frequent patterns, associations, and correlations.**
3. Slimani, T. (2014, February). **Efficient analysis of pattern and association rule mining approaches,** College of Computer Science and Information Technology, Taif University, KSA, and LARODEC Lab.
4. Shajeeah, M., Safana, N., Fathimath Rajeela, K. A., Kadeejath Sajida, & Ansari, Z. (2014, December). **A framework to discover association rules using frequent pattern mining.** Department of Computer Science Engineering, P. A. College of Engineering, Mangalore, India.
5. Tanksali, P. R. (2016, June). **Approaches for mining frequent itemsets and minimal association rules.** Department of Information Engineering, Padre Conceicao College of Engineering, Verna, Goa, India.
6. Rayward-Smith, V. (2007). Statistics to measure correlation for data mining applications. Computational Statistics and Data Analysis, 51(8), 3968 – 3982
7. Yang, X. S., Sherratt, S., Dey, N., & Joshi, A. (2024). **Proceedings of Ninth International Congress on Information and Communication Technology**, volume 5. London, pp. 105-109.
8. Agrawal, R., & Srikant, R. (1994). ***Fast algorithms for mining association rules in large databases*.**
9. Cafaro, M., Epicoco, I., & Pulimeno, M. (2019). **Data mining: Mining frequent patterns, associations rules, and correlations.** University of Salento, Lecce, Italy.
10. Han, J., Pei, J., & Yin, Y. (2000)**.** ***Mining frequent patterns without candidate generation*.**
11. Kaura, M., & Kang, S. (2016). **Market Basket Analysis: Identify the changing trends of market data using association rule mining,** Bhai Gurdas Institute of Engineering and Technology, Sangrur, India.
12. Nsabimana, T. (2015). **Association rule mining using the Apriori algorithm: A simulation study using retail market basket data**, Doctoral dissertation, University of Buea, Cameroon.
13. Haryanto, H., Winarto, H., & Juliane, C. (2024). **Application of data mining techniques in healthcare: Identifying inter-disease relationships through association rule mining**, STMIK LIKMI, Jawa Barat, Indonesia.
14. Maske. A., joglekar.B,(2018), **Survey on frequent item set mining approaches in market basket analysis**, fourth international conference on computing communication control and automation.
15. Kumar, B. S., & Rukmani, K. V. (2010). **Implementation of web usage mining using Apriori and FP Growth Algorithms,** Department of Computer Science, C.S.I. College of Engineering, Ketti, The Nilgiris, pp. 400-404.
16. Setiabudi, H. D., Budhi, S. G., Purnama, J. W., & Noertajahyana, A. (2011). **Data Mining market basket analysis using hybrid dimension association rule, case study in minimarket x.** IEEE, 978 1-4244-9983.

17. Malik, J. S., Goyal, P., & Sharma, A. K. (2010, February). **A Comprehensive Approach Towards Data Preprocessing Techniques & Association Rules**. IES-IPS Academy, Rajendra Nagar Indore, India

18. Han, J., Pei, J., (2000). **Mining Frequent Patterns by Pattern Growth: Methodology and Implication**, School of Computing Science Simon Fraser University, Burnaby, Canada.

19. Desai, A., Gandhi, H., & Vasa, J. (2020). **Market Basket Analysis Using Association Rule Mining and Apriori/FP-Growth Algorithm**. India.

20. Desai, A. (2020). **Market Basket Analysis Using Association Rule Mining and Apriori/FP-Growth Algorithm**
https://github.com/AmanDesai10/Market-Basket-Analysis