Detecting Customer Purchasing Patterns using Association Rule Mining

MSc Research in Computing
Data Analytics

Rhutujit Uday Paradkar
Student ID: 20187416

School of Computing
National College of Ireland

Supervisor: Dr. Paul Stynes, Dr. Anu Sahni

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Rhutujit Uday Paradkar
Student ID:	20187416
Programme:	Data Analytics
Year:	2022
Module:	MSc Research in Computing
Supervisor:	Dr. Paul Stynes, Dr. Anu Sahni
Submission Due Date:	19/09/2022
Project Title:	Detecting Customer Purchasing Patterns using Association Rule Mining
Word Count:	6808
Page Count:	21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	19th September 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Detecting Customer Purchasing Patterns using Association Rule Mining

Rhutujit Uday Paradkar
20187416

Abstract

The market basket analysis plays a very important role in driving the business of any organization especially in retail domain. Performing market basket analysis helps the business owners to get a better idea about the different purchasing patterns of the customers. The recommendation systems make use of association rules on a very large scale to find association between the products so that they can be often bought together. The research in this paper mainly focuses on the market basket analysis in the retail sector. The datasets that are selected for the experiments are from a retail shop, a bakery and a online retail shop in UK.The association rule mining algorithms that are used in this paper are apriori, FP-growth and theEclat algorithms. The aim of this experiment is to apply the algorithms to each dataset selected and then use clustering algorithm like the k-means to reduce the size of the dataset by clustering together different association rules. The time and performance of each of the algorithm on the dataset is recorded for comparison. In this experiment the market basket analysis was performed on the datasets along with the K means clustering to reduce the size of the datasets and also principal component analysis was done to reduce the dimensionality of the dataset.

1 Introduction

The retail industry is huge when compared to other type of industries and also it is one of the industries which has a huge turnover estimated in billions. The retail industry itself employs a large number of people in the world. It is therefore crucial for the retail organisations to make use of the modern technologies and increase the revenue. The organisations, therefore make sure that they make personalised recommendations to the customers based on the customer's purchase history. When the companies are successful in making the personalised recommendations to the customers, it automatically increases the customer loyalty and the acquisition. Ferrera et al. (2020). By employing the personalised recommendation strategy, the cost to acquire the customers can be reduced by 50% and the revenue can be increased by upto 20%. These personalised recommendations are done based on the hidden buying patterns of the customers. Finding the hidden purchasing patterns helps the organizations to better plan the product placements in the large supermarkets. This helps these companies to bring down the costs by selling the right products to the customer and to free them from the problem of the information overload. This ensures that the service to the customers is consistent. This will lead to an increase in overall revenue of the retail companies. That is why, it is crucial for these companies to constantly innovate and to keep up in the competition by taking inspiration from different machine learning and data science technologies. Widodo et al. (2021)

The data in the field of retail industry in increasing day by day as a result of introduction of the internet. This data is increasing exponentially and therefore the customers are constantly struggling with the problem of the information overload. This might result in customers buying up the things that they don't need or they would be discouraged from buying anything. In both the cases, the company might end up losing money and that is why it is essential that the customers are made available with the information or the products that they really need. This will encourage the customers to return to the company to buy the products again and again. Building a recommendation system for this purpose can help the companies to overcome this problem Huang (2022). Building recommendation systems can also increase the revenue of the companies by increasing the revenue in the cross selling of the products. Cross selling of the products is a term where the customer is suggested with an item which is suitable for the product that has been just bought by the customer. Some examples of such cross selling in the retail can be , selling laptop bags and wireless mouse with the laptops or selling different car accessories with cars or selling cigarette lighters with the cigarette boxes. In the world of ecommerce, the cross selling happens mostly when the customer is checking out the cart Rawat et al. (2021).

To discover the customer purchasing patterns, the association rule mining algorithms are used on a large scale. These association rules help the companies with the mining the customer purchasing patterns by finding out the relations of the products in the transactions. This ensures that the company has the proper knowledge of the likes of their customers and therefore they can plan better product placement and product bundling strategies Diwandari and Zaky (2021). Association rule mining mainly searches for those occurrences in the data that often appear together. There are different types of association rule algorithms which are used for the frequent pattern discovery in the dataset.

2 Literature Review

This section covers the research in different papers about the challenges faced by different retail companies with the massive data and how the machine learning along with association rule mining helps to increase the revenue of these companies. This section also gives an overview of different aspects of the association rule mining and its incorporation in the retail industry has changed the face of the online stores and the offline retail stores. The challenges with the large datasets are also mentioned and also the techniques to reduce the dataset size and the dimensionality of the datasets is also mentioned.

2.1 Applications of the recommender systems

The recommendations systems act as a filter between the choices of the user and the information available on the internet. As the retail industry and the e-commerce industry is growing exponentially a proper recommendation system needs to be in place that will recommend the relevant items to the specific user groups. As the population that has access to modern technology has increased in the last two decades, so has the information available to them Huang (2022). In today's world, due to availability of the numerous choices of products on the online as well as offline platforms, the customers often get confused with these choices. The customers spend a lot of time while buying , analyzing or comparing different products. As a result ,the customer lose interest and they are unable to make quality decisions. This leads to loss of customers as well as the retail companies Yang and Qiuxiang (2022). That is why it is necessary that the customers are recommended correct products by considering the customer choices and buying patterns of the customers. This can be done by using the historical transactional data of the companies.

The use of the ecommerce websites has become an inseparable part of the people's life. The people spend more time in browsing through the different sites ans stores in an attempt to find something of their choice. The use of recommendation systems can help the people to make intelligent choices while buying the products Li et al. (2021).When the recommendation systems recommend the right products to the customer based on their personal preferences, the customers are more likely to be attracted to that store or ecommerce site as it gives them quality recommendations tailored to their needs and solves the problem of information overload Huang (2022). That is why it is necessary that the retail companies work on the algorithms that can detect the customer purchasing patterns from the data and make useful and appropriate recommendations in order to increase the revenue.

2.2 Market Basket Analysis

The data mining technique which studies the data and gives the retailers an idea about the customer purchasing patterns is called the market basket analysis. The use of market basket analysis has increased by the retailers in the recent years. The online as well as offline retailers make use of market basket analysis to make better business decisions. The main reason to applying market basket analysis to the data is because of the amount of data and the time required to process it. The retailers want high returns on the investment due to the increasing costs of the sales and marketing Lim (2021). Market basket analysis can be mainly done by using association rule mining where the hidden relationships between the products can be discovered and therefore the product bundling strategies of the retail companies can be improved.

2.2.1 Market Basket analysis for Product placement

Market basket analysis can be used by the companies to make better business decisions based on the output of the market basket analysis. The market basket analysis can be used by the retailers to improve their product placement strategy. The product placement strategies for the retailers can be pitched by the supermarket owners so that they can give the retail companies a premium place at the store so that they can make more sales. Placing the products in a proper place where the most customers are likely to buy the products can help the companies to increase the sales of the company. The companies pay a premium price to showcase their products in the premium places Halim et al. (2019).

Market basket analysis is used to scan through thousands of records for customer purchase patterns. To improve sales, this can be used to create marketing campaigns and bidder campaigns. to find collections of things that are frequently seen in transaction sets. Industry norms are known as association rules or ARs. ARs function as local-type models that search for rule-based patterns for a constrained collection of binary variables by examining specific subsets of a data set (for instance, a subset of variables or observations). They merely provide direction when two item sets co-occur. To put it another way, while assessing how a dependency between two items develops, the rules do not take changes in the relationship between one of the things and another that is not in that relationship into account. There are, however, more modern approaches that look for global models. It is now understood that the combined purchasing behaviours of thousands of customers make up a complex system of market offer components Valle and Ruz (2021).

Common recommenders and pattern-based recommenders do not take into account the sequential information of whatever item is purchased next says Guidotti et al. (2019) They further assert that sequential recommenders ignore aspects like reciprocal influence and assume the independence of the items in a single basket. For each approach to forecast something for a single client, transactional data from a large number of consumers is also necessary. They do not adhere to the World Economic Forum's user-centric approach to data security, which encourages the management of personal data for each and every customer of a data-based company. They claims that according to a study that they studied create a special classifier for each customer and base their forecasts solely on her personal data. By characterizing basket prediction as a classification problem, this predictor respects the user-centric viewpoint. Sadly, this approach presupposes that items that are purchased in large quantities are independent. Additionally, a suggested method for tailored basket prediction only considers co-occurrence and requires a piece of the basket after it to generate the recommendation. Last but not least, the intrinsic difficulty of interpreting prediction models by humans is a drawback of hybrid systems based on neural networks. The interpretability of a predictive model, or the capacity to comprehend the principles underpinning the forecasts, would be of significant use to a retail chain management looking to improve the service and marketing techniques offered. Additionally, because it enables them to better understand their own purchasing habits, customers enjoy interpretability.

2.3 Principal Component Analysis in the retail datasets

The use of the principal component analysis in the field of retail can help the retail companies to use their data effectively and reduce the size of the data. The feature extraction of the data can be more specific and the relationship between the products can be more clear. This will help the stores to make a proper marketing strategy and to make the store image accordingly. Applying the PCA on the dataset can help the companies to find the hidden patterns in the customer purchasing Mittal et al. (2021). In terms of store imaging, the principal component analysis can help the retail companies to pitch the brands offering them premium space in the stores at a premium price. This can be done by understanding the sales for the each brand and then pitch the highest selling product company to pay a premium price for a good product placement so that their sales are higher Fan and Zhang (2022). That is why understanding the customer choice using the association rule mining can help the companies to increase their business by implementing proper advanced technology.

The principal component analysis can be used to even forecast the sales of the stores by studying the important historical features of the data so that the forecast can be more accurate Zhang et al. (2022). The PCA can be very helpful in the feature extraction as it is hepful in reducing the dimensionality of the dataset, thus reducing the overhead for processing. This will help the companies to make efficient use of their resources Mateos-Mínguez et al. (2022).

3 Methodology

The methodology section discusses about the path that is being followed throughout the lifecycle of the project. The methodology followed in this project is CRISP-DM which stands for Cross Industry Standard process for data mining. This methodology follows multiple steps which are as follows:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling

3.1 Business Understanding

Association rule mining plays an important role in the success of retail business. The association rule mining is mainly used to perform market basket analysis and to find the hidden patterns in the customer buying patterns. These patterns are studied and the future recommendations are made to the customers based on the rules found in association rule mining. Association rule mining algorithms can be be used to enhance the existing online and offline retail systems by studying the online data of the customers from different stores and the sites from where the customers buy things. Using association rule mining can help the retail companies to overcome the challenge of next basket recommendations by recommending the customers based on their choice and the goods or services that they have bought or used in the past Agarwal et al. (2022).

Association rules are not only used to maximize the business in retail but they are also in the fields of the medicine, the travelling industries where they are recommended with the next places to visit and the places where they can stay. The most popular algorithm in the association rule mining is the apriori algorithm. The association rules are mostly used by the retail companies on the real time data available through the social media and the browsing history and also the historical data from the supermarkets Exenberger and Bucko (2020). Today's retail companies use the customer purchasing pattern in more smarter way by displaying the ads of their favourite products directly on their mobile devices Arivazhagan et al. (2022).The other advantages to the retail companies in terms of business are as follows:

- Recommendation systems help to retain the customers.

- Helps them to run the business smartly with low cost.
- Increases the number of the return customers.

3.2 Data Understanding

The datasets used in this products are all from the retail domain. There are 3 datasets used for each research and one datasets may have one or more sub-files which can be used as reference files or the rows from the sub-files can be used as the primary keys or foreign keys for joining the files together. The further subsections describe each of the datasets in detail.

3.2.1 Market Basket Dataset

This dataset is taken from Kaggle and is used in the research for the paper.[1] The dataset has 7500 records which consists of the different items sold together. Each row in the dataset represents the list of items that are sold in a single transaction. The data is from a convenience store. The file is in the form of comma separated variable (csv). The dataset consists of only the items that are bought in the transaction in a single row. Every row is a different transaction in this dataset.

3.2.2 Instacart Dataset

The dataset consists of seven files which are relational which means that they can be joined together based on the columns. The columns from these can be used as foreign keys in the tables to perform joins between the tables. The dataset is taken from kaggle which is a open source dataset.[2]. The files that are in the dataset consists of the information about the different sections of a supermarket.

1. Information about the aisles and which products are located on which aisle along with the user id
2. The name of the departments
3. The dataset with the details of the product and the details if the products are redordered or not.

3.2.3 Glantus dataset

The dataset is a real world dataset. It is taken from the previous research of a student from the National College of Ireland. As the dataset is real world data and consists the data from a Uk store, proper permission was taken from the college to use the dataset. There are multiple files in this dataset and each file is of approximately 1 gb in size. There is a master file with 'tcg_products' which consists of a column called EAN which is assigned to the unique products This column is further used to join the main transactions file and the products file for further analysis. The features of the dataset are as follows:

- The dataset is in the csv format

[1] https://www.kaggle.com/datasets/devchauhan1/market-basket-optimisationcsv

[2] https://www.kaggle.com/competitions/instacart-market-basket-analysis/data

- The data selected for the analysis is 300000 due to the limitations of machine.
- There are almost 16000 unique produxts in the selected transactions
- Cancelled items in the transactions are shown by negative values

3.3 Data Preparation

This stage is where the data is cleaned and processed to make it ready for the modeling stage.This step involves data preparation steps for multiple datasets. The further subsections demonstrate different data preparation steps that are undertaken and which are specific to the datasets. Each dataset has its unique structure and feature, hence the steps for the data preparation steps differ from each other.

3.3.1 Market Basket Dataset

The preprocessing of the data for the apriori algorithm is done in python.

1. **Importing the packages** - The package called apyori is imported initially to facilitate the generation of the rules for the products. This package is used to run the apriori algorithm.
2. **Creating a list** - As the default file format of the dataset file is .csv it is converted to the list first.

The package that was used for pre processing for the eclat algorithm was done in the R language. The preprocessing that is done on this dataset is:

- **Removing the duplicate values**-The duplicate values from the transactions are removed with the help of the arules package.

3.3.2 Glantus dataset

The steps taken for preprocessing and preparing the data are mentioned below in the detail:

1. **Converting the data to a structured format** - As the raw data set is in the form of semi structured data the 'jsonlite' library is used to create an individual transaction out of the grouped json nodes. Each subnode will have a different row for itself.
2. **Converting the datetime to a categorical variable** - This step is done to make the time variable categorical which will make it easier for us to separate the transactions according to the time of the day.
3. **Removing or ignoring the transactions with the negative values** - Some of the values are negative as they indicate that the product bought before was cancelled in the same transaction. Such transactions are ignored.

3.3.3 The Instacart Dataset

As this dataset contains multiple relational datasets it is necessary that these relations between the different tables should be studied and the tables should be joined accordingly. The data preparation steps are as follows:

1. **Unzipping the files** - When the dataset is downloaded all the files are in the form of zip files and therefore to unzip these all files the 'zipfiles' library is used which unzips all the files.

2. **Combinind two datasets(Merge 1)** - The prior and the order df is combined based on the common column of order_id. The merging of these datasets will enable tracking of the order of the products bought by the specific user.

3. **Combinind two datasets(Merge 2)** - Here , the two datasets, aisles and the products are combined based on the common column of aisle_id.

4. **Combinind two datasets(Merge 3)** - The merged dataset in the step 2 is further joined with the department dataset based on the common column of 'department id'.

3.4 Data Modeling

Data modeling is the step after the data is preprocessed and it is made available for applying different algorithms that are discussed before in this paper. The section below explains the different algorithms that are used in the research along with the principal component analysis. The different data modelling techniques that have been used in the research for different datasets are:

1. Association rule mining(Apriori algorithm, Eclat algorithm)

2. K means clustering algorithm

3. Principal Component analysis

4 Design Specification

The design specification shows an overall architecture of the system and the flow of the different components in the system. The design specification gives us a path to develop a system and sets up a roadmap for the further implementation. The figure 1 shows the overall design of the system. The different components of the architecture are described in the further subsections.

4.1 Acquiring the datasets

This stage of the design specification corresponds to the different datasets that are supposed to be gathered from different sources. The types of the input files for the project may contain csv files, json files,nested json or text files.

4.2 Importing the dataset

This stage involves the importing of the different datasets into the IDE that are intended to be used. In the current project the IDE that are going to be used are

- Jupyter notebook
- R studio

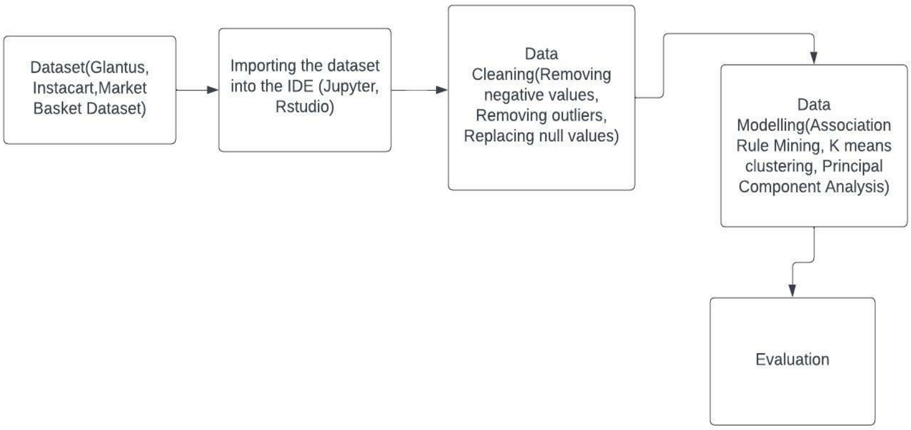

Dataset(Glantus, Instacart,Market Basket Dataset)
Importing the dataset into the IDE (Jupyter, Rstudio)
Data Cleaning(Removing negative values, Removing outliers, Replacing null values)
Data Modelling(Association Rule Mining, K means clustering, Principal Component Analysis)
Evaluation

Figure 1: Design Specification

4.3 Data Cleaning

This step is used to prepare the data before sending it to the modeling. The different data cleaning or the data preparation techniques are as follows:

- Removing the null values
- Removing or ignoring the negative values
- Removing the outliers in the dataset or replacing them
- Categorizing the variables
- Changing the datatypes of the variables for the

4.4 Data Modeling

This step involves applying different modeling algorithms to the dataset like the apriori, Kmeans for the customer segmentation and the dataset reduction. It also involves Principal Component Analysis for the dimensionality reduction.

4.5 Evaluation

It involves different evaluation techniques like the execution time, the speed to validate the claims made during the research.

5 Implementation

The implementation shows the different steps that have been taken to complete the project.

5.1 Apriori algorithm

The association rule mining algorithm is one of the oldest algorithm in the association rule mining. This algorithm is suitable for the datasets where large number of transactions take place on a daily basis. This algorithm works on the BFS(Breadth first search) approach Tang et al. (2020). The apriori algorithm was first introduced in the year of 1994. There are certain steps that are followed in the apriori algorithm The apriori algorithm has the initial knowledge of the dataset as it scans the database initially Kuruppu and Galappaththi (2021). This approach used by the apriori algorithm makes it more reliable than the other association rule algorithms due to its approach towards the dataset.

If the use of 'apyori' package is intended then the data should be standardized first and the data should be in the form of the list.

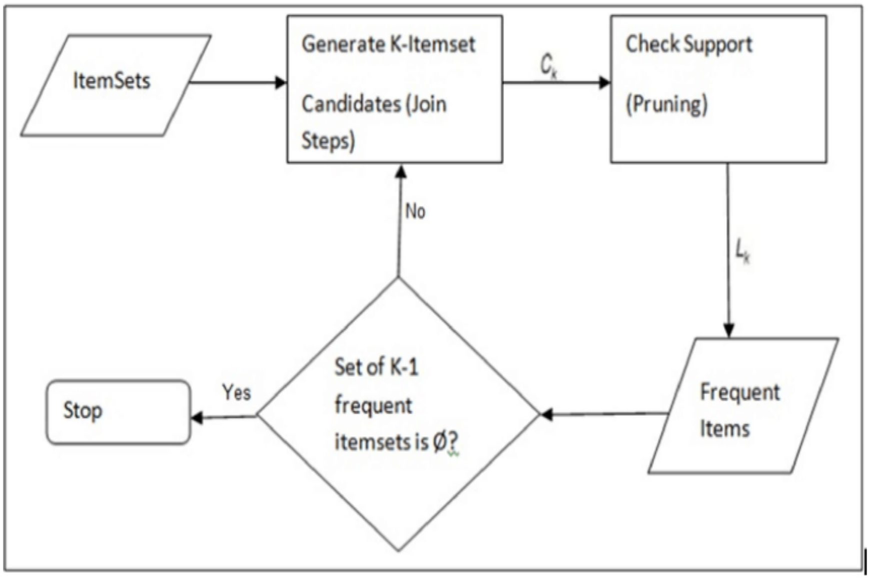


Figure 2: Processes in Apriori Kuruppu and Galappaththi (2021)

The fig.2 shows the flowchart about the processes that take place in the apriori algorithm. The main drawback of the apriori algorithm is that it gets slower when as the size of the dataset increases. Therefore, it is necessary that we follow some othe methodology like clustering with kmeans or principal component analysis to reduce the dimensionality of the dataset.

For the market basket dataset, there are only products from the transaction and therefore there is no need of the standardization. The products in the dataset are used as columns and the transaction numbers are selected as rows. After these transformations are performed then the data is sent to the function that generates the association rules.

5.2 K means clustering algorithm

The Kmeans clustering is used in this research for two datasets. The purpose of using the K-means algorithm in both the datasets is different but the process to find the optimum number of clusters is the same. The process used is the Elbow method which is found out by the plot called silhuotte plot.

5.2.1 K-means clustering for the customer segmentation

The k means algorithm is mostly used by the retail companies to find the customers that are valuable by tracking their choices so that they can retain those valuable customers Parikh and Abdelfattah (2020). It is alo mentioned in the future work of the state-of-the-art that using K-means is more suitable when trying to increase the output of the business Mohapatra et al. (2021).

The clustering using K-means was used in the instacart dataset to segment the customers according to their choices and their spending values. The crosstab library from the pandas was used to combine the values from the aisles and the user_id tables. The sklearn.cluster library of python is used to make the k means cluster after finding the optimal number of clusters from the elbow method. After the K means clustering is completed there are 5 clusters generated where the cluster with highest frequency of the user id's was selected. This cluster is the cluster with the users that spend high in the store.

5.2.2 K-means Clustering for reducing the size of the dataset

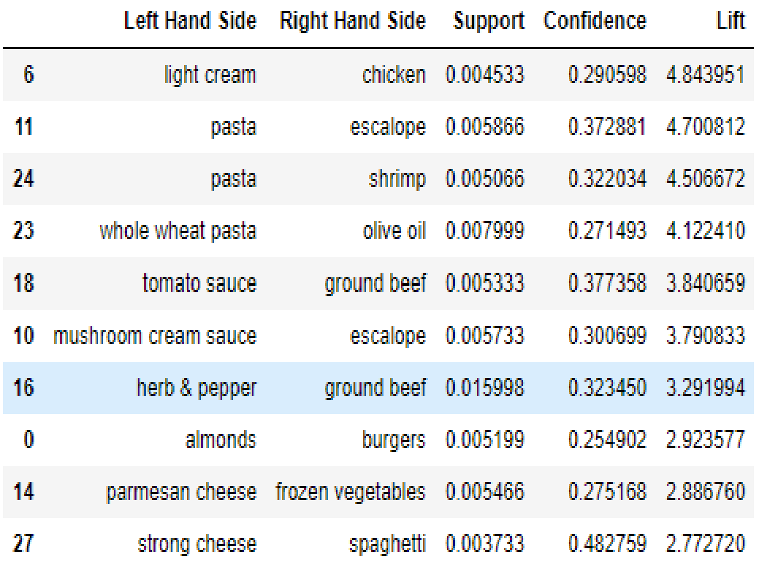
The dataset is similr to the data used in the The glantus dataset is huge when compared to other datasets that are used in the dataset. In this dataset the K means is performed by first finding the average price of the products and the frequency of the products around that price is calculated. The silhuotte plot is generated to find the optimal number of the clusters which is found to be 4. The dataset used in this research is taken from the research done by the authors of the paper Kanhere et al. (2021). The clustering output is then mapped to the individual transactions again and then the association rules are calculated again. The evaluation and the discussion of the results is done in the further sections of the paper.

5.3 Evaluation

This steps puts together the results and thus compares the performance of the different algorithms in association rule mining.

5.3.1 Experiment 1

The first experiment is to replicate the research conducted in the base paper Mohapatra et al. (2021). This paper mainly focuses on implementing the apriori and the eclat algorithms to the market basket optimisation dataset. This dataset consists of 7500 records and each row in the dataset represents the list of items in a transaction. The table in below table below shows the results of the apriori algorithm that is applied on the market basket dataset from the base paper. The minimum length of the products was set to 2 which means that there will be one antecedent and one consequent. The transactions are organized by decreasing order of the lift. The lift of the of the the first transaction is highest with 4.84%.



	Left Hand Side	Right Hand Side	Support	Confidence	Lift
6	light cream	chicken	0.004533	0.290598	4.843951
11	pasta	escalope	0.005866	0.372881	4.700812
24	pasta	shrimp	0.005066	0.322034	4.506672
23	whole wheat pasta	olive oil	0.007999	0.271493	4.122410
18	tomato sauce	ground beef	0.005333	0.377358	3.840659
10	mushroom cream sauce	escalope	0.005733	0.300699	3.790833
16	herb & pepper	ground beef	0.015998	0.323450	3.291994
0	almonds	burgers	0.005199	0.254902	2.923577
14	parmesan cheese	frozen vegetables	0.005466	0.275168	2.886760
27	strong cheese	spaghetti	0.003733	0.482759	2.772720

Figure 3: Apriori algorithm

To explain the table further we the transaction with pasta and escalope is considered. The support value for the second transaction in the table is 0.005866. The number of transactions that contain pasta is divided by the total number of transactions in the dataset. The confidence level of 0.372881 of the transaction indicates that the out of all the transactions that contain pasta 37% of that transaction contain escalope. In the end, the lift of 4.7 explains that out of all the transaction the customers are 4.7 times more likely to buy escalope with pasta than the escalope alone. The parameters that were set for this experiment are as follows:

Parameter	Value
min_support	0.003
min_confidence	0.25
min_lift	2
min_length	2
max_length	2

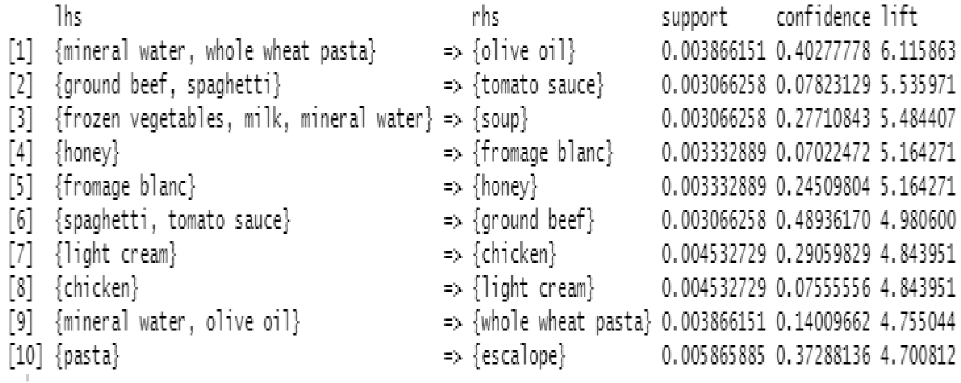
Table 1: Parameters for Apriori Algorithm

This information helps the retailers with making the proper bundling strategies and make the cross-selling of the products more profitable.

The table below shows the output of the eclat algorithm. The parameters that were used for creating the table below are as follows:

Parameter	Value
min_support	0.003
min_confidence	0.25
min_lift	2
min_length	2
max_length	2

Table 2: Parameters for Eclat Algorithm



```
     lhs                                     rhs                 support     confidence lift
[1]  {mineral water, whole wheat pasta}      => {olive oil}        0.003866151 0.40277778 6.115863
[2]  {ground beef, spaghetti}                => {tomato sauce}     0.003066258 0.07823129 5.535971
[3]  {frozen vegetables, milk, mineral water} => {soup}            0.003066258 0.27710843 5.484407
[4]  {honey}                                 => {fromage blanc}    0.003332889 0.07022472 5.164271
[5]  {fromage blanc}                         => {honey}            0.003332889 0.24509804 5.164271
[6]  {spaghetti, tomato sauce}               => {ground beef}      0.003066258 0.48936170 4.980600
[7]  {light cream}                           => {chicken}          0.004532729 0.29059829 4.843951
[8]  {chicken}                               => {light cream}      0.004532729 0.07555556 4.843951
[9]  {mineral water, olive oil}              => {whole wheat pasta} 0.003866151 0.14009662 4.755044
[10] {pasta}                                 => {escalope}         0.005865885 0.37288136 4.700812
```

Figure 4: Eclat Algorithm

The results that are shown in the table are organized in descending order of the lift. It shows only the top 10 transactions. The confidence of 0.4027 in the first transaction indicates that 40.27% of the transactions that contain mineral water and whole wheat pasta in them also contain olive oil. Also the lift of 6.11 explains that the people who buy olive oil are 6.1 times more likely to buy it with mineral water and whole wheat pasta than the olive oil alone.

5.3.2 Experiment 2 - Association rules on the glantus dataset

The second experiment is performed on the real world dataset of the company called Glantus. The dataset is used to get more clear understanding of the working of the apriori and the eclat algorithms. Using the dataset of this size also facilitates the comparison of the two algorithms in terms of the size and the speed of the algorithms. This dataset by glantus contains almost 1 million records. But due to system limitations only the top 300,000 records are taken for the purpose of applying the association rule algorithms.

The first step is to apply apriori algorithm on the dataset. The 'arules' package is used for the purpose of applying association rule learning algorithms on the given data. The track of the time is kept by using the Sys.time() variable from the R language. The parameters that were set for running the association rules were as follows:

The apriori algorithm works in such a way that it scans the whole dataset initially which means it works in BFS(Breadth First Search) pattern. The use of apriori algorithm is useful where large datasets are supposed to be used.

Parameter	Value
Support	0.002
Confidence	0.005
min_length	2
max_length	10

Table 3: Parameters for Apriori Algorithm

Also, along with the apriori algorithm, the output of the eclat algorithm is also taken and then the outputs of the eclat and the apriori algorithms are compared. The eclat algorithm gives more accurate results than the apriori algorithm, but in some cases it misses on some rules as it searches for the results in the BFS(Breadth first search) manner. This leads to generation of the rules which are different from the apriori algorithm but in this case as the number of transactions is greater the amount of time taken by the eclat algorithm is slightly greater than the apriori algorithm. After the execution times for the algorithms were found out, the apriori algorithm was found to be 53% faster than the eclat algorithm.

5.4 Experiment 3 - K-means clustering on the glantus dataset

This experiment aims at reducing the processing overhead of the algorithm by reducing the dataset size using the clustering approach for association rule mining. For this purpose, the K-means algorithm is used to put together different items into clusters. The clusters are computed using the price parameter of the product. The optimal value of the K is selected based on the Silhouette plot. This is called the elbow method. In this method the optimal cluster is selected based on the bend of the graph.

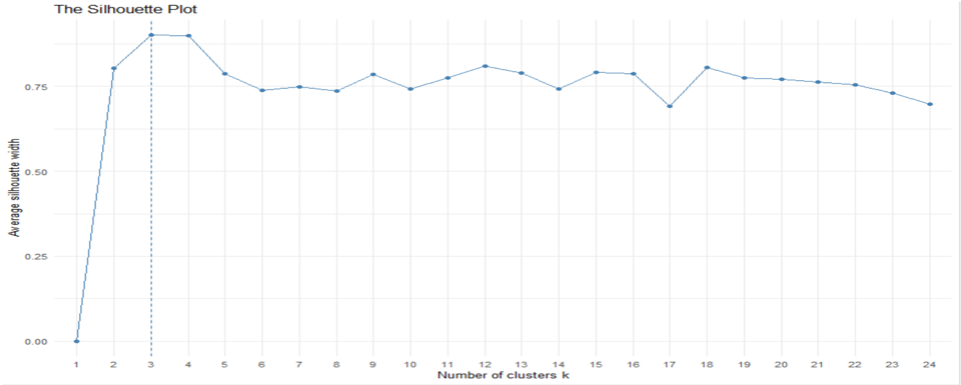
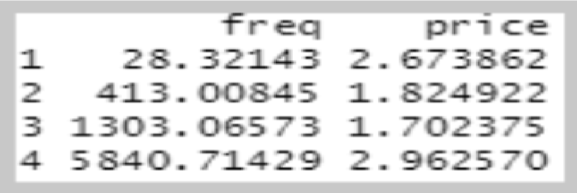


Figure 5: Silhuotte Plot

As it can be seen in the fig 5 the bend of the plot is at the cluster 4. Therefore the optimal number of the clusters is supposed to be 4. The optimal number of clusters is found out by using the function fviz_nbclust and the transactions are sent to this function. The silhuotte plot generated by this method is further used for generating the clusters.

The clusters are then generated by passing the cluster parameter as 4 to the kmeans function in R. The results of the cluster function are shown in the table below:



```
        freq    price
1   28.32143 2.673862
2  413.00845 1.824922
3 1303.06573 1.702375
4 5840.71429 2.962570
```

Figure 6: The Frequency and Price table

For the 1st cluster, the number of the products is 14728 and the frequency is 28.32 and as it can be seen in the figure the price contribution is also not very high. Therefore, it can be said that these products do not add any significance to the business or the sale in the market. Whereas, the products in other clusters with the high frequency and average price add a lot of value to the business. That is why, these products are considered while creating the further association rules. The count of products for cluster 2, 3 and 4 are 1065, 213 and 7 respectively. These products make about 7% of the total data though they are more significant as compared to the other 14,728 products.

The rules that are generated by the data in this reduced dataset are similar to those generated using the whole dataset. Also, as the support value is increased the for the apriori and the eclat algorithms. Therefore, it can confirmed that the association rule mining combined with K-means clustering algorithm can reduce the size of the dataset to a great extent and thus the execution time is also reduced significantly. That is why, the clustering of the products based on the price can be used to further reduce the size of the dataset.

5.5 Experiment 4 - Principal Componenent analysis to reduce dimensionality of the dataset

The dataset used for this purpose is the instacart dataset. The experiment was done to segment the customers according to their spending strategies so that the store can concentrate on the highest spending customers. K-means clustering algorithm was used for clustering the group of the customers.

The reduction of dimensionality is to reduce the features that really impact the outcome of the model to only specific features. PCA is performed on the data to reduce the 134 dimensions to only the features that will explain the most variance in the data. For performing the principal component analysis the elbow method is used to find the optimal number of 'k' clusters.

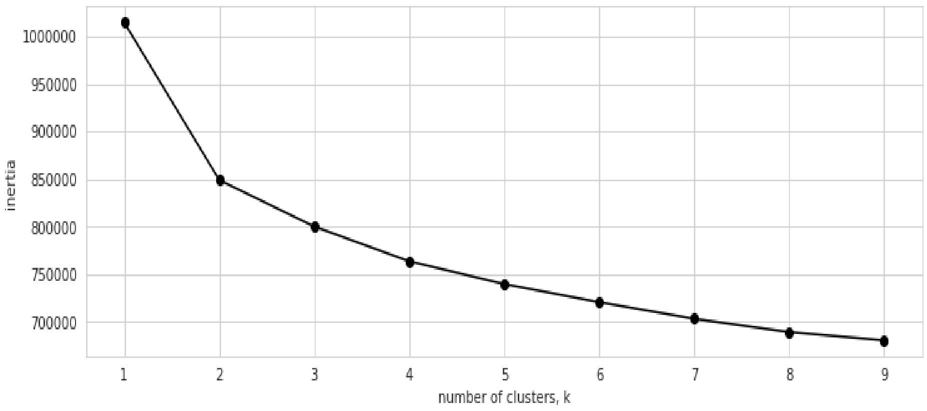


Figure 7: Elbow plot for PCA

As seen in the figure, the elbow graph shows that the optimal value of k is between 5 and 6. Therefore the optimal value of k will be k=5 or k=6. The values beyond these will not explain the variability in the data. The value of n_components is changed constantly and only the first 6 components can be only used to explain more than 50% variability in the dataset.

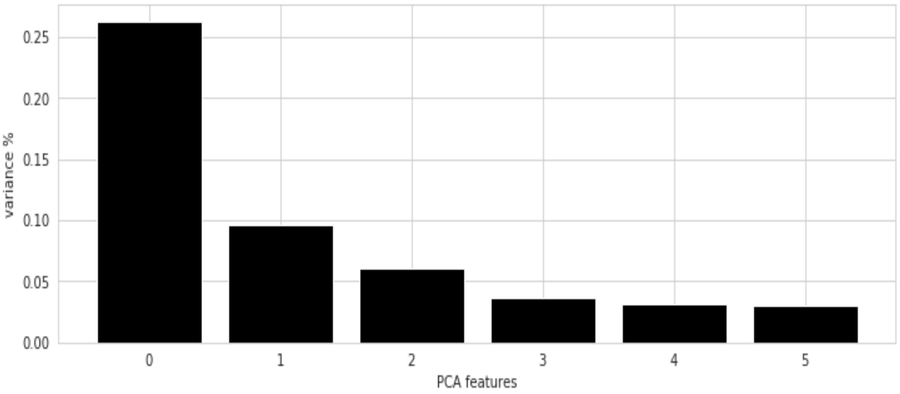
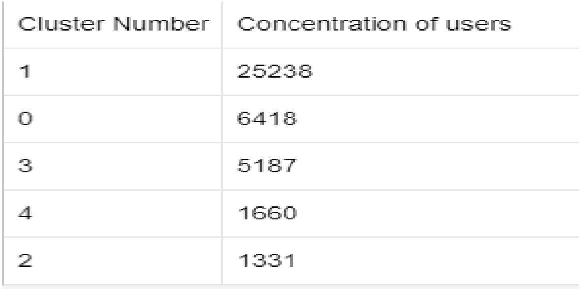


Figure 8: PCA Features

The figure 8 shows the selected pca features that explain the most variance in dataset. Therefore, the number of features which was 134 before has come doen to only 6 features and therefore these features can be used to further perform K-means clustering for the segmentation of the customers.

The value for number of clusters was selected as 5. The scatterplot is plotted to show the cluster to which every user is mapped. Total 39834 unique users were identified from the user_by_aisle dataset. The table below shows the concentration of the users from the 39834 users.



Cluster Number	Concentration of users
1	25238
0	6418
3	5187
4	1660
2	1331

Figure 9: Concentration of the customers

The concentration of the users shows the number of customers that are active and are among the highest spending in the users. The cluster numbers are then assigned to the individual users in the table. From the table above we get an idea about the customer choices and the highest spending customers. That is why, the retail companies can concentrate on these customers and then put the low spending customers on the low priority.

6 Discussion

This section aims to discuss about the evaluation techniques and the improvements and the future work that is carried on the work of the state of the art Mohapatra et al. (2021). There are 4 experiments performed in the research totally.

In the experiment 1 the replication of experiments in the base paper is done. There are two algorithms which are implemented by the authors in the paper:

- Apriori
- Eclat

In this experiment , the market basket dataset with 7500 records is used. The implementation of the apriori algorithm has been done in three steps. The apyori package is used to generate the frequent itemsets. To generate the rules, jupyter notebook is used and initially the dataset is converted into the list. The results that are generated by the apyori package are similar to those in the research paper. The minimum support was varied continuously and the observations are that the increase and decrease in the values of the support affects the execution time. The more the value of the minimum confidence, less is the execution time for the algorithms. The main drawback of the apriori algorithm is that it scans for all the patterns that are possible in the dataset which leads to the generation of rules of which some may be insignificant or may be related to the transactions that don't make any difference to the overall revenue of the company. The eclat algorithm takes the same approach like the eclat algorithm but the difference between them is that the eclat operates on thr transpose of the dataset that is used for the apriori. The 'arules' package was used from the R language. The eclat algorithm is more faster than the apriori when the dataset size is small. In the experiment 2 as well the apriori and the eclat algorithms are applied. The scalability issue with the association rule mining algorithms can be seen in the experiment as well. This dataset is much more bigger than the one used in the experiment 1. The eclat algorithm is found to be slower than the epriori algorithm. The execution times of both the algorithms are calculated and the apriori was found to be faster than the apriori algorithm. As, the dataset size increases, the execution time of the eclat increases and therefore it would be beneficial to switch to apriori algorithm.

In the experiment 3 , the k means clustering was implemented on the glantus dataset. The glantus dataset is huge and therefore the processing overhead is large for the association rules. The clustering of the products based on the price and the frequency of the unique products brought down the execution time. After the clustering when the cluster numbers are assigned to the association rules, it is seen that the execution time of the algorithms is down. The clustering approach is seen to be successful in reducing the size of the dataset by 93% without varying the output of the association rules.

In the experiment 4, the principal component analysis is carried out on the instacart dataset. This was done as preprocessing step before applying kmeans on the dataset. This is the future work of the base paper Mohapatra et al. (2021). The issue with the market basket dataset is that it does not have much features other than the products in the transactions. On the other side the instacart dataset has the dataset with more features and therefore it is more beneficial to apply the principal component analysis and the k means algorithm. Initially the 134 features of the dataset are brought down to 6 features which actually explain more than 50% of the variance of the dataset. When calculated the 51.7% of the variance is explained by only the 6 features of the dataset. After the PCA is applied , the elbow method is used to calculate the optimal number of clusters which was found to be 5. After applying the k means clustering, out of all the clusters the cluster with the highest frequency of the customers was selected. Applying the customer segmentation makes sure that the companies have the proper knowledge of the customers and the customer choices so that they can take wise business decisions Mittal et al. (2021).

7 Conclusion and Future Work

In this paper , first the research in the base paper was replicated. The association rules were generated by using the algorithms like the apriori and the eclat. The dataset was acquired from kaggle. The language used for the apriori algorithm for the first experiment was python and the for the eclat the language used was R. The reason for using different languages for both the algorithms was the unavailability of the proper library for the eclat algorithm in the python libraries. The libraries in the R language are more suitable for performing the association rule mining.

In the glantus dataset as well, the apriori and the eclat algorithms faced the challenge of scalability as the dataset is huge when compared with other dataset. The problem of the scalalbility is seemed to solved by applying clustering to the dataset. Application of the clustering on the dataset seemed to have brought down the size of the dataset by 93% than the original dataset. The results of the association rule mining were not seen to be changed when the clustering was applied to the dataset.

On the instacart dataset, the principal component analysis was done which brought down the number of features from 134 to 6 which really explained the variance in the data. After the PCA, the K means was applied by using the selected features by finding the optimal number of clusters using the elbow method. This helps to implement the dimensionality reduction on the dataset and reduce the processing time without changing the results. The use of K means allows the companies to concentrate on the valuable and the loyal customers and thus increase the business. For the future work , the PCA may be applied on the glantus dataset and thencompare the results of the dataset reduction using the clustering approach and the PCA.

References

Agarwal, V., Gupta, R. and Tiwari, A. (2022). Applicability of association rule mining in recommendation system for big data analysis, *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 1810–1813.

Arivazhagan, B., Pandikumar, S., Sethupandian, S. B. and Subramanian, R. S. (2022). Pattern discovery and analysis of customer buying behavior using association rules mining algorithm in e-commerce, *2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, pp. 1–5.

Diwandari, S. and Zaky, U. (2021). Analysis of customer purchase behavior using association rules in e-shop, *2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pp. 144–149.

Exenberger, E. and Bucko, J. (2020). Study of customer behavior in online b2b shopping, *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, pp. 1301–1305.

Fan, D. and Zhang, S. (2022). Conditional-robust-profit-based decision model for flexible contract between electricity retailer and customer, *2022 IEEE 5th International Electrical and Energy Conference (CIEEC)*, pp. 3225–3230.

Ferrera, R., Pittman, J. M., Zapryanov, M., Schaer, O. and Adams, S. (2020). Retailer's dilemma: Personalized product marketing to maximize revenue, *2020 Systems and Information Engineering Design Symposium (SIEDS)*, pp. 1–6.

Guidotti, R., Rossetti, G., Pappalardo, L., Giannotti, F. and Pedreschi, D. (2019). Personalized market basket prediction with temporal annotated recurring sequences, *IEEE Transactions on Knowledge and Data Engineering* **31**(11): 2151–2163.

Halim, S., Octavia, T. and Alianto, C. (2019). Designing facility layout of an amusement arcade using market basket analysis, *Procedia Computer Science* **161**: 623–629. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.
URL: *https://www.sciencedirect.com/science/article/pii/S1877050919318769*

Huang, G. (2022). E-commerce intelligent recommendation system based on deep learning, *2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, pp. 1154–1157.

Kanhere, S., Sahni, A., Stynes, P. and Pathak, P. (2021). Clustering based approach to enhance association rule mining, *2021 28th Conference of Open Innovations Association (FRUCT)*, pp. 142–150.

Kuruppu, S. and Galappaththi, K. (2021). A data mining approach to identify associations between job titles and skills in job vacancies.

Li, H., Wu, Y. J., Zhang, S. and Zou, J. (2021). Temporary rules of retail product sales time series based on the matrix profile, *Journal of Retailing and Consumer Services* **60**: 102431.
URL: *https://www.sciencedirect.com/science/article/pii/S0969698920314387*

Lim, T. (2021). K-means clustering-based market basket analysis: U.k. online e-commerce retailer, *2021 International Conference on Information Technology (ICIT)*, pp. 126–131.

Mateos-Mínguez, P., Arranz-López, A. and Soria-Lara, J. A. (2022). Analysing the spatial impacts of retail accessibility for e-shoppers' groups, *Transportation Research Procedia* **60**: 544–551. New scenarios for safe mobility in urban areasProceedings of the XXV International Conference Living and Walking in Cities (LWC 2021), September 9-10, 2021, Brescia, Italy.
URL: *https://www.sciencedirect.com/science/article/pii/S2352146521009716*

Mittal, R., Mittal, A., Singh, J., Rattan, V. and Malik, V. (2021). Principal component analysis based feature selection driving store choice: A data mining approach, *2021 International Conference on Computing, Communication and Green Engineering (CCGE)*, pp. 1–3.

Mohapatra, D., Tripathy, J., Mohanty, K. K. and Nayak, D. S. K. (2021). Interpretation of optimized hyper parameters in associative rule learning using eclat and apriori, *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 879–882.

Parikh, Y. and Abdelfattah, E. (2020). Clustering algorithms and rfm analysis performed on retail transactions, *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, IEEE, pp. 0506–0511.

Rawat, S., Tyagi, U. and Singhal, S. (2021). Recommender systems in e-commerce and their challenges, *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pp. 1598–1601.

Tang, K.-T., Sun, Y., Lee, P.-H. and Huang, Q. (2020). Apply apriori algorithm in supermarket layout research, *2020 International Conference on Modern Education and Information Management (ICMEIM)*, pp. 521–524.

Valle, M. A. and Ruz, G. A. (2021). Finding hierarchical structures of disordered systems: An application for market basket analysis, *IEEE Access* **9**: 1626–1641.

Widodo, I., Ulfah, H. and Anggraeni, K. (2021). Redesign super market layout analysis based on hidden customer purchase behaviour, *2021 IEEE 8th International Conference on Industrial Engineering and Applications (ICIEA)*, pp. 261–264.

Yang, L. and Qiuxiang, Z. (2022). Research on e-commerce user interest recommendation method based on tf-idf algorithm, *2022 2nd Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS)*, pp. 291–295.

Zhang, H., Song, H. and Song, L. (2022). Research on precise demand forecast of new retail based on principal component analysis model, *International Journal of Social Science and Education Research* **5**(2): 413–416.