

American International University-Bangladesh (AIUB)

Software Project 1

Smart Inventory System using Market Basket Analysis

Submitted By

18-36177-1	Howlader, Snigdho Dip
18-36804-1	Sijan, Asif
18-37371-1	Haque, Md Fahamidul

Department of Computer Science

Faculty of Science and Technology

American International University Bangladesh

August, 2021

Declaration

We declare that this project is our original work. No part of this has been submitted elsewhere partially or fully for the award of any other degree. Any material reproduced in this project has been properly acknowledged.

We declare that this project does not contain any content that discloses the secret of any organization or related parties. American International University-Bangladesh (AIUB) will not be held liable for any such activity, as the project is presented as our original work.

Howlader, Snigdho Dip
18-36177-1
Department of Computer Science

Sijan, Asif
18-36804-1
Department of Computer Science

Haque, Md Fahamidul
18-37371-1
Department of Computer Science

Approval

The thesis titled “Smart Inventory Management using Market Basket Analysis” has been submitted to the following respected members of the board of examiners of the department of computer science in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science on (date of defense) and has been accepted as satisfactory.

Md. Anwarul Kabir
Sr. Assistant Professor and Supervisor
Department of Computer Science
American International University-Bangladesh

Aymed Hossain
Part-Time Lecturer and External
Department of Computer Science
American International University-Bangladesh

Dr. Mahbub Chowdhury Mishu
Head (Undergraduate Program)
Department of Computer Science
American International University-Bangladesh

Professor Dr. Tafazzal Hossain
Dean
Faculty of Science and Technology
American International University-Bangladesh

Dr. Carmen Z. Lamagna
Vice Chancellor
American International University-Bangladesh

Acknowledgement

Firstly, we would like to thank the university and all the individuals who have helped us complete this project successfully.

We would like to show our gratitude to our supervisor, Md. Anwarul Kabir, Sr. Assistant Professor, Department of Computer Science, American International University-Bangladesh for his support and guidance from the start till the end of the thesis. We also want to thank our external supervisor Aymed Hossain, Department of Computer Science, American International University-Bangladesh for his help and presence during our presentation and helping us understand the many intricacies for the project's development.

We would also like to thank Dr. Carmen Z. Lamagna, Vice Chancellor, Professor Dr. Tafazzal Hossain, Professor & Dean, Computer Science and Dr. Md. Mahbub Chowdhury Mishu, Assistant Professor and Head of Undergraduate, Department of Computer Science, American International University, for their generosity and encouragement.

Table of Contents

Thesis/Project Summary 6

Keywords 6

CHAPTER 1: INTRODUCTION 7

1.1 Topic Name 7

1.2 Thesis Background/Problem Analysis 7

1.3 Research Motivation and Objective 7

1.4 Literature Review of Existing studies 7

CHAPTER 2: RESEARCH METHODOLOGY 8

2.1 Conceptual Framework 8

2.2 Data Collection Method 8

2.3 Ethical Issue 8

CHAPTER 3: PROJECT/THESIS PLAN 9

3.1 Project/Thesis Effort Estimation 9

3.2 Project/Thesis Planning 9

3.3 Uncertainties and Risk Analysis 9

3.4 Project/Thesis Execution 10

CHAPTER 4: RESULTS AND ANALYSIS 10

4.1 Solution Description 10

4.2 Solution Verification 11

CONCLUSION 11

REFERENCE 11

APPENDIX 12

Project Summary

Data can be utilized in many ways using technology from today. One of these ways is prediction of sales. The project will deal with this topic in particular. Using data collected over time, regarding purchases from a superstore, the goal is to make predictions. When a customer buys item A, there is a probability that can represent the likeliness of them to buy item B. This probability is very helpful for the superstore in how they should manage their inventory. If a customer buys item A, B and C together, that gives the managers to stock those particular items more.

Using this logic, the idea of our project is to implement a system module that will be able to read through a transactional data set, in order to generate results that will be able to make suggestions for future users. It will make it more convenient for a manager of stock or inventory this way. They can log in to the system, and view the stock accordingly. Selecting a particular item from the stock list will show the items that are likely to be bought by the customer. Also, when placing orders for new stock, suggestions will be made by the system for what other items should be ordered. This will be done once again by the algorithm that has learned what the most likely item sets that are bought by customers are.

Overall, our project will be able to help stock managers make the best out of storing inventory. Wholesale warehouses and food vendors will be able to keep their inventory in check according to what is most likely to be bought by the customers. This not only makes life easier for the stock manager, but also decreases possibilities of items in the inventory going bad from not being sold for too long. Moreover, it is also likely to increase revenue for the users by increasing the chances of item sets being bought together. It is more likely for a user to buy a pair of items when the system is suggesting to them that customers have bought them together in the past.

Keywords

1. Market Basket Analysis
2. Association Rule Mining
3. Data Mining
4. Apriori Algorithm
5. Smart Systems

CHAPTER 1: INTRODUCTION

1.1 Topic Name

The title of the project is Smart Inventory System using Market Basket Analysis. Market Basket Analysis is a data mining technique used by retailers that identifies the strength of association between pairs of products that are purchased together.

1.2 Project Background/Problem Analysis

In Smart Inventory Management System using Market Basket Analysis, an inventory manager can easily get the sales ordered by customer behavior. This also optimizes product placement for them and makes inventory management more convenient.

Finding relations between products sold will let an inventory manager know what products to focus on. Else, they may be facing problems with unsold inventory, or not making the best sales. It should also be noted that without the implementation of an association rule, it might be difficult to understand what kind of products a customer likes to buy together, since not all patterns are apparent.

This project will try to mitigate these issues by letting inventory managers view the relevant items according to their sales. It will also show the items that each specific item is most likely to be bought with.

1.3 Research Motivation and Objective

The goal of this study is to make management of inventory easier for all parties involved. The main motivation is to make sure that products never go to waste or are thrown away. Such items will take up important space in a store or warehouse that can be utilized for ones more likely to be sold or put into good use. The concerns of this study mainly deals with a few particular questions:

- What kind of different items do customers buy simultaneously?
- How related are certain items in a transactional scenario?
- What factors might affect the higher or lower sales of some products?

The project uses the transactional data that comes from everyday purchases between a customer and a store, and makes a predictional output based on that. The objectives are to make life easier for owners and managers of inventory warehouses. They will be able to minimize the wastage of their stock, which can contain raw materials as well.

The main concepts that have been explored to make this project come to life was association rule mining, which was able to give outputs of a predictive nature.

1.4 Literature Review

Market Basket Analysis, or MBA for short, is a data mining technique that is used in various fields. It is also known as association rule learning or affinity analysis. The main goal of MBA in marketing is to ensure that the behavior of the customer is understood in regards to the products they buy. [1] There are many different techniques that exist to implement this,

each differing from one another in terms of complexity. Association rule mining is one of these, and it is used to discover desirable relationships in huge datasets. [2]

A large number of algorithms exist for association rule mining. All existing algorithms work for static data. They try to find association rules on different types of metrics such as support, confidence and lift. An improved algorithm has been proposed in 2016 which performs periodic mining using the KDD process. [2]

The three main elements that are concerned when discussing association rule learning are:

- **Support**

  Tells the significance of an itemset in a database. It is a measure of the proportion of transactions in which an itemset is present, by the total number of transactions. An easy way to describe it is the percentage of times it appears across a list of transactions. If it appears four times across ten transactions, its support value is 40%. [4]

$$S(A) = \frac{Amount\ of\ transactions\ A\ has\ appeared\ in}{Total\ number\ of\ transactions}$$

- **Confidence**

  The likeliness of an item being bought with regards to another item. The confidence of item X against Y, is the support value for X,Y by the support value for X. [4]

$$S\{X \text{ -> } Y\} = \frac{Support\ of\ X, Y}{Support\ of\ X}$$

MBA is based upon the theory that if you buy a certain group of items, you are more likely to buy another group of items.

MBA is an advanced analytics and data mining technique that helps us know consumer behavior. To reiterate, this refers to the behavior by which a consumer decides to purchase a product. When attempting implementation of Market basket Analysis in the context of a super shop, the product to customer purchasing patterns first we need a data set of transactions. Using those transaction data we can know which lists all items bought by a customer in a single purchase. [2]

Then we check what product is purchased with what product and this is called Association Rule Mining. Association rules are “If and then” it helps us to show the probability of relationships between data items. [4]

Market Basket Analysis can be applied in various ways, such as:

1. Developing bundled offers based on products sold in sets
2. Forming associations of products in stores or shops
3. Inventory management depending on what products are sold together

These three uses are possibly the most common ways that MBA is used today. There are deeper and more improved ways that they are implemented today as well. [6] Those topics will not be studied in this project, however.

The following section will show the steps in association rule calculation and mining:

1) A scan is performed across the whole database to get a 1-hot-encoding. This is then used to measure the support and it is compared with the minimum support required. This value is then included in a large itemset.
2) The itemset or itemsets that are not included in large itemsets gets discarded during the next iteration.
3) On the second iteration, the comparison is made to form a second itemset candidate. This will cause the system to use the results of the large itemset in the iteration.
4) Now the previous itemset is added with the new itemset accordingly with their relevance based on support.
5) This process continues until the whole itemset is null, or there are no longer candidate itemsets to be formed. [5] .

Explained in a broader format, apriori algorithm functions based on candidate itemsets:

1) A candidate itemset needs to be formed. Candidate K-itemset will be formed, which will be done using combinations of item k-1 from the previous iteration.
2) The support value will be measured for each candidate K-itemset. These support values will be obtained from scanning the rest of the data. This is an important feature of the Apriori algorithm. It requires scanning the entire database for as much as the longest itemset.
3) A high frequency pattern is made. The highest frequencies containing the K-item or itemset is judged from the candidate K-itemset. If the support is greater than the least support, then it is discarded.
4) If there are no new frequency patterns formed, the whole process ends up stopping. Otherwise, it will go to part 1 and add to the K value. [5]

CHAPTER 2: RESEARCH METHODOLOGY

2.1 Conceptual Framework

For the following project, The Dynamic Systems Development Method (DSDM) will be followed. The goal of this software is to help customers read their consumer's behavior and show them the appropriate relations between products purchased from the customer's organization. The plan is to execute this primary feature in a very small time frame which fits perfectly with DSDM's fundamental technique 80:20 rule.

And applying another DSDM technique, MoSCoW rules - the priorities can be set on several other features. Here:

Must have:

1. Generate relations among products using information from customer's invoices.
2. Finding frequently bought together items.

Should have:

1. Show products that need restocking according to total revenue and stocks.

Could have:

1. Stock suggestions according to price & popularity.

Want to have but won't have:

1. A properly functioning backend of the website using relevant framework
2. A means for the user to run the mining process according to their wishes

We will add points on this list in later phases of our project.

Since we'll be developing our main features within the first 20% time of our project- rest of the time will be spent on improving our software's user experience according to the customer, and optimizing our software since we'll be working with large amounts of data.

As the following project falls under Adaptive environment, linear models like waterfall won't suit it. Since the main feature can't be divided into small parts, incremental models won't be a wise choice either. That leads to the iterative model's subset agile model.

To assure the customers' delight, The most cutting-edge technologies currently available were employed. And to ensure that, the go to model was DSDM because it will be using timeboxes to improve the UI of the product. And, because of the major feature being so particular, there will be plenty of opportunity to test it and enhance its efficiency in subsequent revisions. Overall, the DSDM technique will be an excellent match for the project.

2.2 Data Collection Method

The data used for mining in this project was collected from an open source. It was accessed from the Google certified subsidiary, Kaggle. The data set used was a transactional data set of items that customers have bought over time, and contains a large amount of such data. For this particular study, the data used contained more than 9,800 rows of data.

Each of the rows in this data set contain a particular transaction, which has any number of products ranging from 1 to 37. These data would be later used to convert into a one hot encoded data set to use for implementation of the Apriori algorithm.

2.3 Ethical Issue

The work carried out in this research was voluntary, and did not consist of any questionable motivations. The tools involved were all open source. The framework used for creating the user interface, the data mining tools and the data set that was used were all open-source.

The intentions for creating this project and doing the necessary research for it was done for learning purposes only. This was done for completing a course in the core curriculum of those involved in this paper.

None of the resulting effects of the project can be harmful, and only contains the best of intentions for the community.There was no bias in collection of the data, as the dataset that was used was a ready-made open source dataset. The research data that was used was realistic as it just involved a large dataset of transactions from a super shop or supermarket.

CHAPTER 3: PROJECT/THESIS PLAN

3.1 Project Effort Estimation

For Project, Estimate the project development effort (time, budget, resources, etc.) using the standard project estimation techniques such as COCOMO.

Using basic COCOMO model,

Estimated size of project = 4 KLOC (4000 lines of code)
Effort, $E = a(KLOC)^b$
$=2.4(4)1.05 = 10.3$ person-months
Time, T = c(Effort)d
$=2.5*(10.3)0.35 = 5.7$ months
Person required = Effort/time = 10.3 / 5.7 = 2 persons

3.2 Project/Thesis Planning

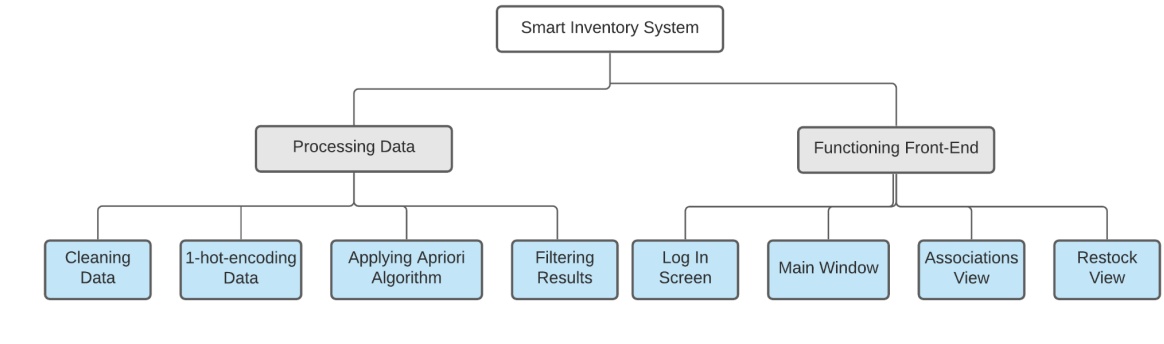


Fig 1: Work Breakdown Structure

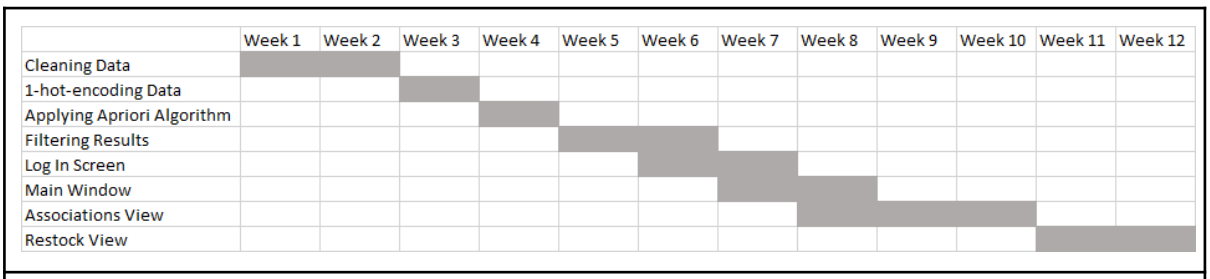


Fig 2: Gantt Chart

The project is developed by a backend programmer to process all the data. It will require some significant time as the data processing is a long procedure. There is also a risk of errors and bugs encountering, which will leave some of the allotted time wasted.

The other programmer involved developed the front end interfaces using web technologies. It involves robust user interfaces. It consists of a log in screen, and some specialized tables

structures in order to display the data.

3.3 Uncertainties and Risk Analysis

S/N	Risk Description	Probability	Impact	Mitigation Plan
1	Unrealistic time estimate	20%	Significant	Take multiple estimation
2	Inefficient algorithm complexities	30%	Significant	Hire an expert
3	Data processing errors	30%	Moderate	Use a cleaner dataset
4	Poor productivity	40%	Low	Short iterations

3.4 Project/Thesis Execution

Project Deliverable	
1. Preliminary Project/Thesis Plan	2021.03.15
2. Requirements Specification	2021.03.21
3. Analysis [Object model, User interface]	2021.04.15
4. Source Code	2021.05.13
5. Test Plan	2021.07.07
6. Final Product / Demo	2021.07.29

CHAPTER 4: DESIGN AND IMPLEMENTATION

4.1 Diagrams

The following are the diagrams that represent the system that has been developed during the phase of this project.

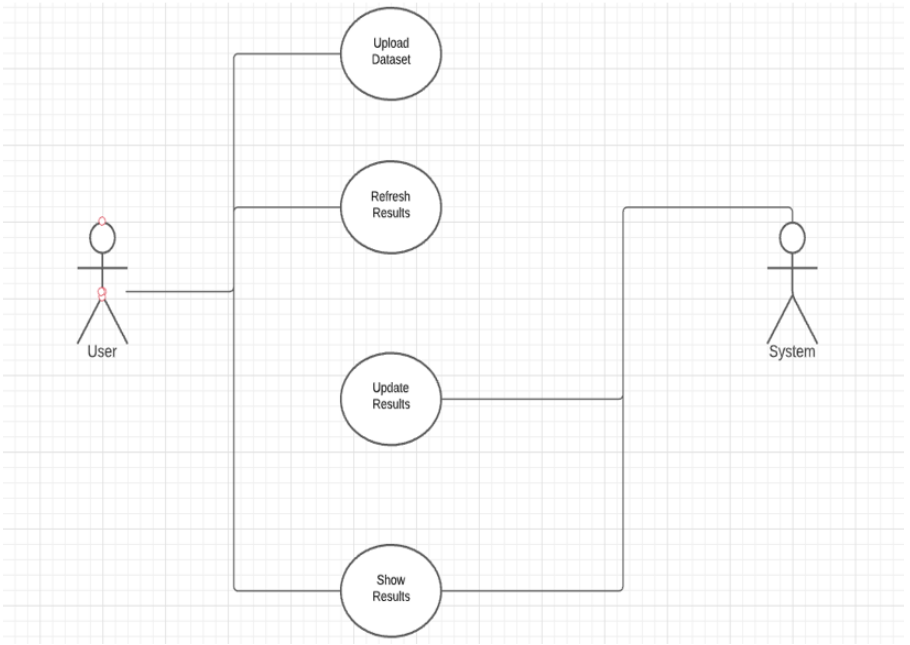


Fig 3: Use Case Diagram 1

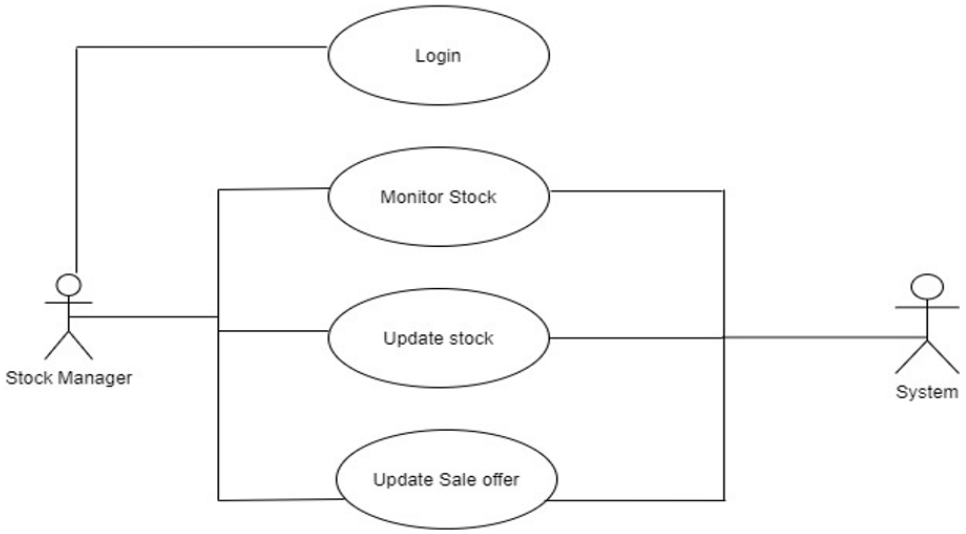


Fig 4: Use Case Diagram 2

4.1.2 Class Diagrams

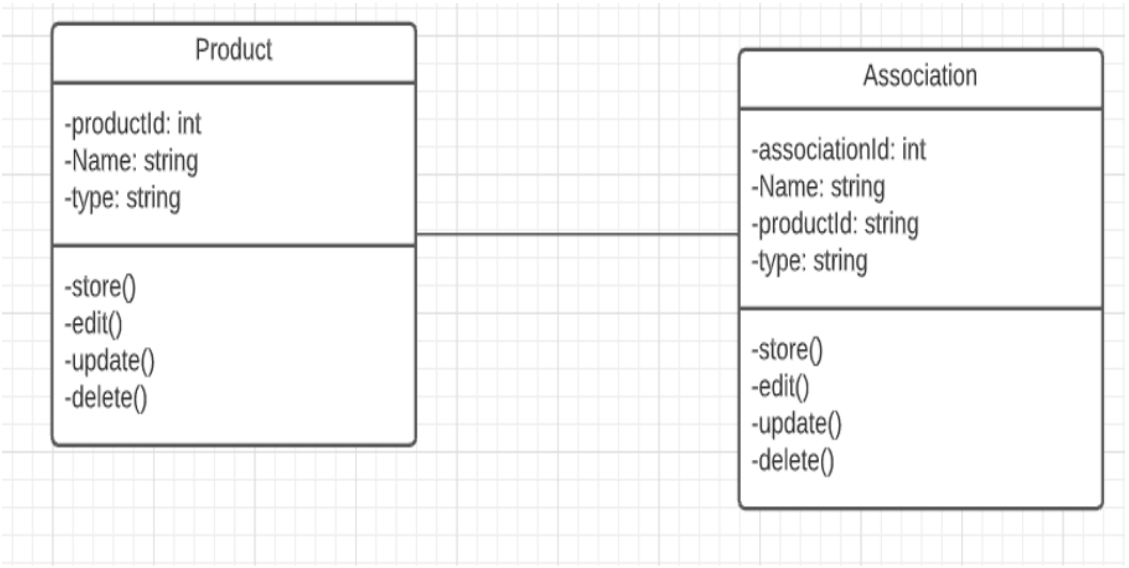


Fig 5: Class Diagram 1

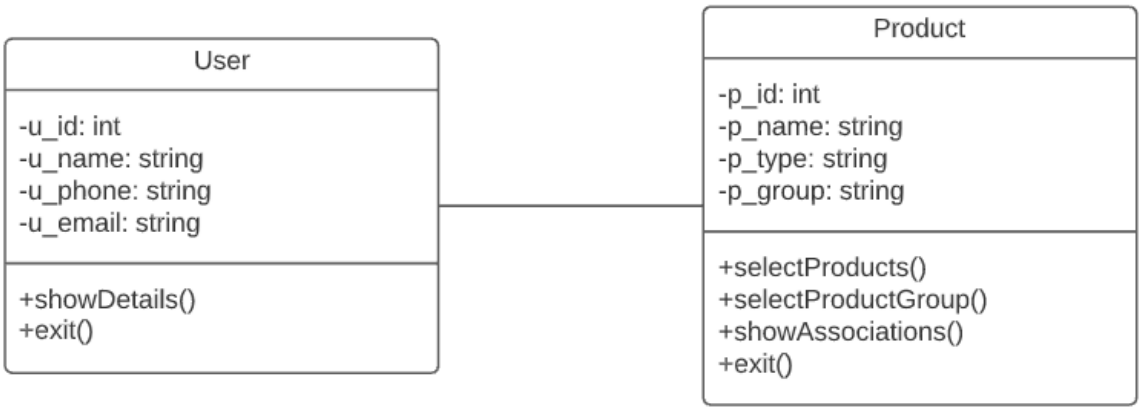


Fig 6: Class Diagram 2

4.1.3 Sequence Diagrams

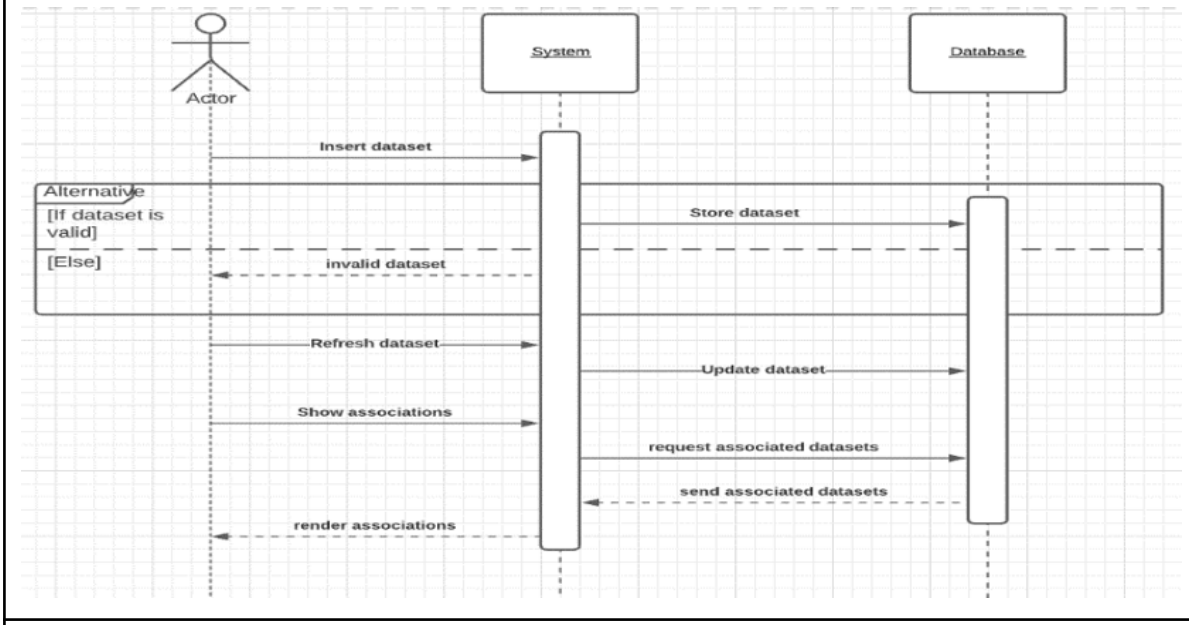


Fig 7: Sequence Diagram 1

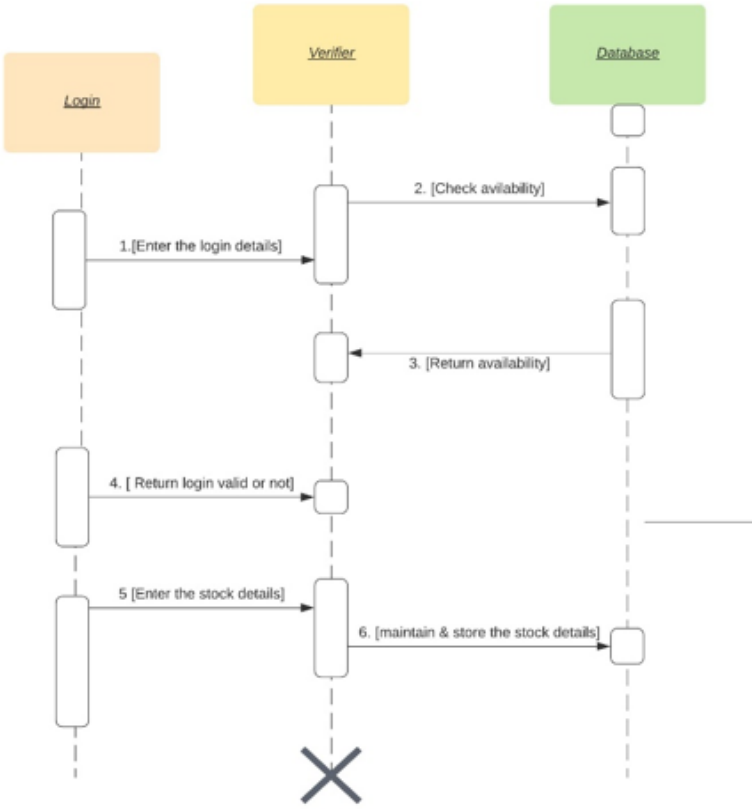


Fig 8: Sequence Diagram 2

4.1.3 Activity Diagram

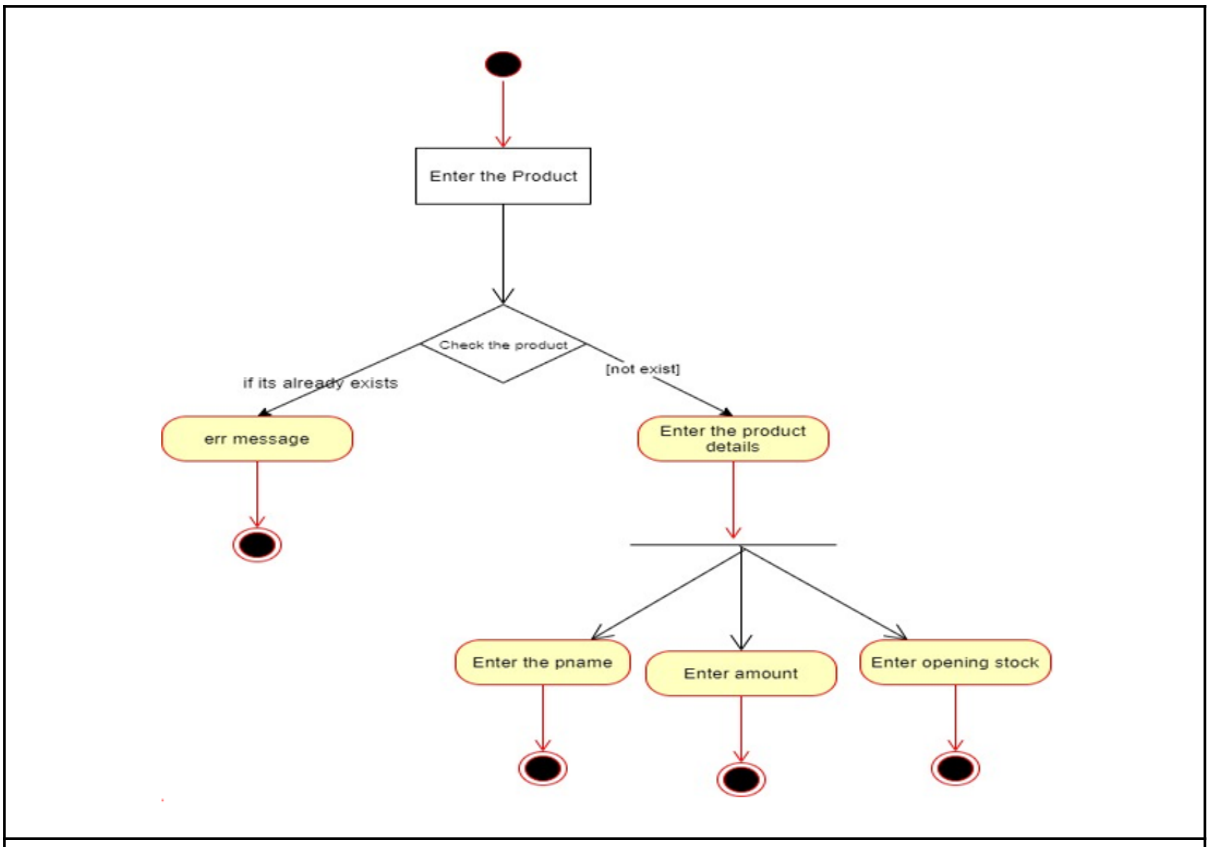


fig 9: activity diagram 1

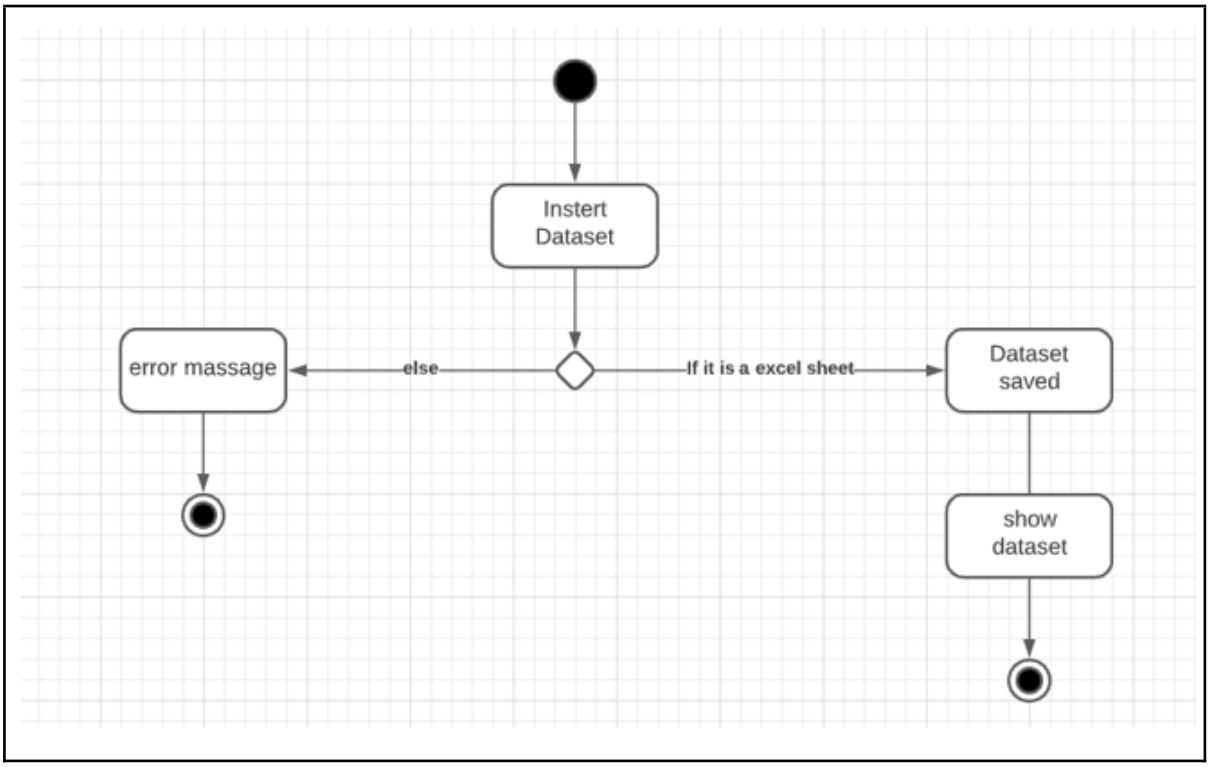


fig 10: activity diagram 2

4.2 Backend Calculations

The backend of this project consists of data processing and filtering in many steps. It is all done in python and two of its libraries. The libraries used are pandas and mlxtend. The code is written with the intention of making it reusable. This is why functions have been created where files can be passed as parameters and the code can generate results accordingly. As the processing takes a huge time, this is not doable by the user and can be handled by the backend only.

There are three files that have done all the processing. One creates a list of the unique items present in the dataset. Then, a 1-hot-encoded file can be generated from it.

```
def one_hot_encode(groceries_DataFrame):
    groceries_DataFrameUnique = pd.read_csv("files/outputs/unique-stock.csv")
    groceries_list = []
    for i in groceries_DataFrameUnique.index:
        groceries_list.append(groceries_DataFrameUnique.iloc[i][1])
    groceries_list = [groceries_list]

    encoded_DataFrame = pd.DataFrame(index = groceries_DataFrame.index, columns=groceries_list)
    encoded_DataFrame.fillna(value=0, inplace=True)
    #display(encoded_DataFrame)
    for i in groceries_DataFrame.index:
        for j in range(1, len(groceries_DataFrame.iloc[i])):
            if groceries_DataFrame.iloc[i][j] != 0 and groceries_DataFrame.iloc[i][j] != '0':
                for k in list(encoded_DataFrame.columns):
                    if str(groceries_DataFrame.iloc[i][j]) == str(k)[2:-3]:
                        encoded_DataFrame.loc[i, str(groceries_DataFrame.iloc[i][j])]=1
    return encoded_DataFrame
display(encoded_DataFrame)
```

Fig X: Algorithm for creating a 1-hot-encoded dataset

```
def implement_apriori(encoded_DataFrame, supportval):
    from mlxtend.frequent_patterns import apriori
    from mlxtend.frequent_patterns import association_rules

    frequent_ItemSet = apriori(encoded_DataFrame, min_support=0.005, use_colnames=True)
    frequent_ItemSet['length'] = frequent_ItemSet['itemsets'].apply(lambda x: len(x))
    #display(frequent_ItemSet)

    filtered_frequent_ItemSet = frequent_ItemSet[(frequent_ItemSet['length'] >= 2) & (frequent_It
emSet['support'] >= supportval)]
    list_relations_frozensets = filtered_frequent_ItemSet['itemsets'].tolist()
    list_relations_frozensets = [re.findall(r"'(.*?)'", str(x)) for x in list_relations]
    filtered_frequent_ItemSet['itemsets'] = list_relations_frozensets
    filtered_frequent_ItemSet.to_csv('files/outputs/frequent-data.csv', index = None, header=Tru
e)
    return filtered_frequent_ItemSet
```

Fig X: Algorithm for to implement apriori on the 1-hot-encoded dataset

4.3 User Interface

This project used a javascript library called react for representing GUI. React (or ReactJS) is a free and open source front end library developed by Jordan Walke and maintained by facebook. React's main use cases are building interactive user interfaces, seamless single page app and building reusable UI components. This project also used two popular React framework named Ant.js and Material UI.

For front-end hosting, Firebase was used. Firebase is titled as 'backend as a service', which includes features like database, api and hosting altogether. For the following project, only their hosting functionalities were used to deploy the app.

```
"dependencies": {
  "@material-ui/core": "^4.11.4",
  "@material-ui/data-grid": "^4.0.0-alpha.27",
  "@testing-library/jest-dom": "^5.11.4",
  "@testing-library/react": "^11.1.0",
  "@testing-library/user-event": "^12.1.10",
  "antd": "^4.16.2",
  "firebase": "^8.5.0",
  "react": "^17.0.2",
  "react-dom": "^17.0.2",
  "react-router-dom": "^5.2.0",
  "react-scripts": "4.0.3",
  "web-vitals": "^1.0.1"
},
```

fig 11: All frontend dependencies used in frontend app

Firebase was also used for user-authentication. After logging in, users will find three separate tabs named inventory, top associations and restock. Inventory shows the current quantity and associated products of a product. Top associations shows the most popular items that were bought together previously as a group. And for the restocking sections, items were selected according to two threshold points. One is if a product's revenue was higher than a certain limit. And another one is if the product's quantity is less than a certain limit where the limit is calculated by dividing the item's current quantity with how many times it was bought.

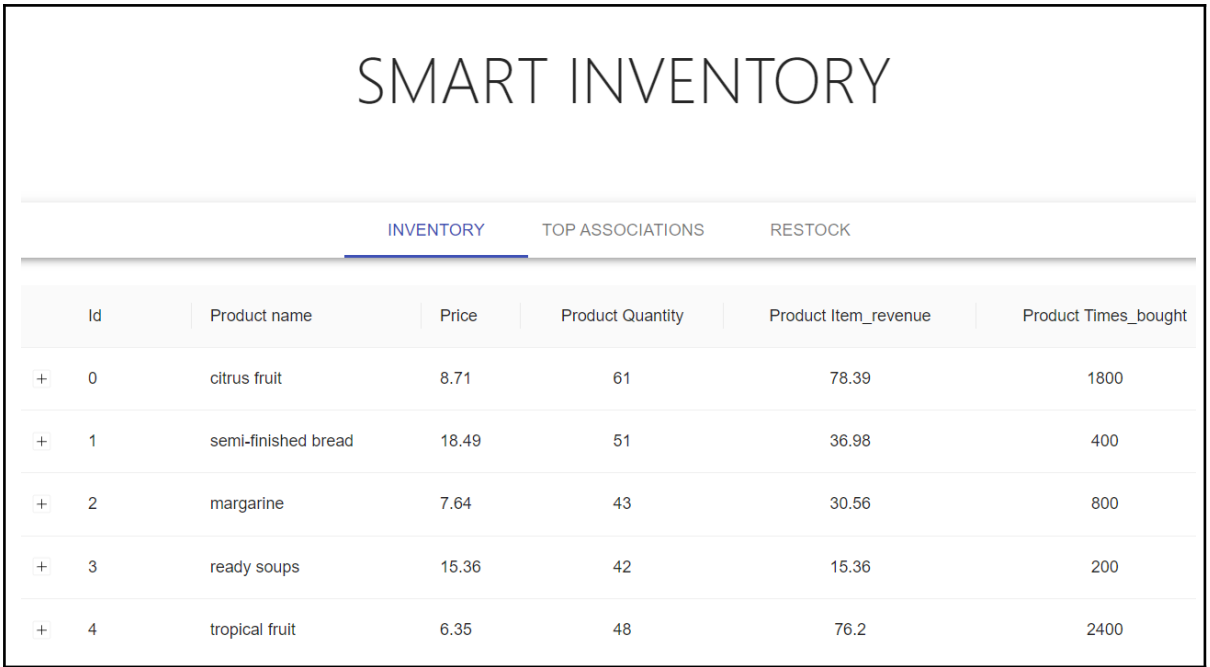


fig 12: Inventory section of the application

CHAPTER 5: RESULTS AND ANALYSIS

5.1 Solution Description.

The solution that was proposed initially is achieved as the developed system mitigates the problems mentioned. The resulting project is able to display the relevant itemsets that are being sold most likely with another item, or a set of other items. The appropriate values are also displayed including the support value for them.

The data that was used consisted of 9850 rows, in a .csv format file. The dataset consisted of columns, each containing a singular item per row. The data was appropriately processed, in multiple steps. Firstly, the data cannot contain any empty cells. This cleaned data needs to be further implemented in additional steps.

The unique items are then extracted from the whole dataset. The unique items present in the dataset are then extracted using different algorithms to make it suitable for apriori.

25	white bre	[]
26	bottled w	[['yogurt', 'bottled water', '3.0%'], ['other vegetables', 'bottled water', '3.0%'], ['bottled water', 'rolls/buns', '4.0%']]
27	chocolate	[]
28	curd	[['curd', 'whole milk', '4.0%']]
29	flour	[]
30	dishes	[]
31	beef	[]
32	frankfurte	[]
33	soda	[['yogurt', 'soda', '3.0%'], ['whole milk', 'soda', '4.0%'], ['rolls/buns', 'soda', '5.0%'], ['pastry', 'soda', '3.0%']]
34	chicken	[]
35	sugar	[['sugar', 'whole milk', '3.0%'], ['sugar', 'rolls/buns', '3.0%']]
36	fruit/vege	[['fruit/vegetable juice', 'brown bread', '3.0%']]
37	newspape	[]
38	packaged	[]
39	specialty	[]
40	butter mil	[]
41	pastry	[['pastry', 'whole milk', '3.0%'], ['pastry', 'soda', '3.0%']]

Fig X: The frequent relations that have been obtained

After apriori is implemented, the 1-hot-encoded data is turned into a frequent itemset as shown in the table above. This is then connected with the windows of the front end to make it viewable for the user.

5.2 Impact on Environment

This project is a software solution for solving a business problem and it does not impact the environment negatively in any way. The solution is light and resource-friendly at the software's current state. It can be considered a prototype for systems that might be used in warehouses everywhere.

5.3 Life Cycle Analysis

The software that has been developed in the project is simple and light-weight. It does not require much resources when compared to other management systems. It is fully front end and only connects with the dataset files in the back end. The phases that were involved here are:

1) Planning - To plan the project and the steps that would be required to implement it properly.
2) Analysis - Analyzing and preparing the data took a significant amount of time. The algorithms and methods have been tested and created however. This does not leave much room for further issues when preparing the data for a new dataset, as long as it is in the correct format.
3) Design - Creating the views and windows to display the tables and other functionalities to the user, which also includes a log in screen.
4) Implementation - Joining the processed data output with the front end and making it fully functioning for the user.
5) Maintenance - The software is easy to maintain and small bugs and issues that may occur in the foreseeable future have been dealt with.

5.4 Solution Verification

The project was initially set up with the goal of predicting customer behavior and using that

for creating associations between products that were not known before. In absence of a real dataset, an open dataset from Kaggle was used in the project. Later after data-cleaning and managing inconsistencies, datasets were used to create associations between different products. And for further use-cases, another implemented functionality was to suggest users what items to restock. For verifying whether the system's result works, a real supershop dataset is needed. After using the dataset, a comparison can be shown about whether the super shop's sales increase after using customer behavior patterns or not.

5.5 Tools Used

The project was implemented by using a particular set of tools in detail. These consisted of frameworks, libraries and datasets that were implemented as per requirement. The tools included and were not restricted to:

- Python - Data Mining to provide with frequent itemsets, namely the mlxtend and pandas library was used
- React JS - For creating a front-end user interface that provides suggestions
- CSV files - to store and retrieve data results for processing and viewing for the users

---

CONCLUSION

This project used sales data to predict customer behavior and buying patterns using which associations or connections were made between different products. This was achieved using certain data-mining techniques named association rule mining and apriori algorithm which falls under Market Basket Analysis domain. After getting the association results, several insights were shown to suggest to the customer what products are frequently bought together, and what to restock. All of the insights were represented through a web application for the ease of user's use.

For further improvement, a server with higher computation power will be required to deploy the backend portion which takes a significant time to create the associations from sample data.

REFERENCE

---

1. Warnia Nengsih. (2015). A Comparative Study on Market Basket Analysis and Apriori Association Technique. 3rd International Conference on Information and Communication Technology (ICoICT).
2. Manpreet Kaur, Shivani Kang. (2016). Market Basket Analysis: Identify the changing trends of market data using association rule mining. ScienceDirect.
3. Mauricio A. Valle, Gonzalo A. Ruz, Rodrigo Morrás. (2018). Market basket analysis: Complementing association rules with minimum spanning trees. ScienceDirect.
4. Fachrul Kurniawan, Binti Umayah, Jihad Hammad, Supeno Mardi Susiki Nugroho, Mochammad Hariadi. (2018). Market Basket Analysis to Identify Customer Behaviors by Way of Transaction Data. Knowledge Engineer and Data Science (KEDS)

5. Suprianto Panjaltan et al. (2019). Implementation of Apriori Algorithm for Analysis of Consumer Purchase Patterns. Journal of Physics: Conference Series
6. Iston Dwija Utama, Ivan Diryana Sudirman, Doni Purnama Alamsyah. (2020). Optimizing the Products Combination by Using Data Mining Market Basket Analysis Approach. Journal of Critical Reviews.