University of Sfax

***

The Higher Institute of Computer Science and Multimedia of Sfax

Summer internship report

Introduced at:

The Higher Institute of Computer Science and Multimedia of Sfax

Performed at:

Run Way Tek

By:

Abir Aloulou

Specialty:

License in Big Data and Data Analysis

2021/2022

I. Preface and Acknowledgement:

First, I would like to thank **Mrs. Feten Besbes** the company manager of Run Way Tek for giving me the opportunity to do an internship for August month within the company, for being such a patient supervisor and for her continued guidance. Your encouraging words and thoughtful detailed feedback have been very important to me.

This opportunity I had, was a great chance for personal development and for learning.

Furthermore, I express my thanks to the **ISIMS administration** and all the **Teaching Staff** especially **Mrs. Wafa Naifer** who have been providing me with guidance and sharing experiences before and during the practical training.

Finally, my biggest thanks goes to both my parents and sisters who have helped me in terms of encouragement and financial support.

Without them all I might have problems achieving this exercise properly.

Table of contents

II. Abstract:

Market basket analysis is becoming a very influential tool for decision-making today in retail businesses. It is important to determine the placement of goods, designing sales promotions to improve customer satisfaction and hence the profit of the retail organisation.

The application of data mining can help to analyse data obtained from transactions in the information system so that it can explore the patterns of customer shopping habits in the retail company.

This study is concerned to compare the use of only Apriori algorithm method and the use of Apriori algorithm method along with the model Profset. It investigates for the most effective method to get a better market basket analysis.

Our analysis shows the difference that model Profset can make in the profit of retail organisation.

Keywords: Data mining, association rules, Apriori algorithm, model Profset, Market Basket Analysis.

III. Company description:

1. Identity:

RUNWAYTEK is a Tunisian services and software solutions company created in 2017 and specialized in the development of embedded software. Its mission is to support customers and partners in the development of efficient and innovative solutions mainly for telecommunications.

2. Branch of activity:

The main contributions of RUNWAYTEK are:

- Learning by doing: learning to manage to solve problems on your own.
- Self-training: in web and mobile development, machine learning and SCRUM methodology.
- Communication / Group work: the essential point put forward by the program is to prepare the "trainee" for professional integration.
- Personal development.

3. Company contact details:

The contact details of the company are:

- Company manager: Besbes Faten
- Address: avenue Mohamed Karray, 3000, Sfax
- Phone: +216 70 032 054
- Email : feten.besbes@runwaytek.com

4. The company's activities:

- Consulting service in the field of embedded and telecommunications.
- Solution development.
- Specialized website development.

IV. General introduction:

1. Problem statement:

When you go to the supermarket, usually the first thing you do is grab a shopping cart. As you move up and down the aisles, you will pick up certain items and place them in your shopping cart. Most of these items may correspond to a shopping list that was prepared ahead of time, but other items may have been selected spontaneously. That’s why we need to find a way so the retail businesses will understand more their customer choice and better organize store layouts. All we need here is to find the best method for market basket analysis.

2. Project Objectives:

This project aims to analyze the data of market basket.

This project’s purpose is to find the best method for market basket analysis from researches done through the last years and test it.

3. Tasks:

First part: Reading posted articles that gives different methods to market basket analysis.

Second part: Comparing the methods used and choosing the better one

Third part: Code the chosen method and test it.

Final part: Making a conclusion.

V. Research:

1. Data Analysis:

Data analysis is a process of inspecting, cleansing, transforming and modelling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.[1]

There are several methods and techniques to perform analysis depending on the industry and the aim of the analysis.

As a result, data analysis will drive success to your marketing strategies, allow you to identify new potential customers.

2. Data Mining:

Data mining is a process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. [2]

Data mining is a particular data analysis technique that focuses on statistical modelling and knowledge discovery for predictive rather than purely descriptive purposes.

3. Market Basket Analysis:

Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves analyzing large data sets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together. [3]

4. Association Rules:

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. [4]

5. Apriori Algorithm:

The Apriori algorithm is used for mining frequent item sets and devising association rules from a transactional database. The parameters “support” and “confidence” are used. Support refers to items frequency of occurrence; confidence is a conditional probability.

6. Model Profset:

The key idea of the PROFSET model is that when evaluating the business value of a product, one should not only look at the individual profits generated by that product (the naïve approach), but one must also take in to account the profits due to cross-selling effects with other products in the assortment. Therefore, to evaluate product profitability, it is essential to look at frequent sets rather than at individual product items since the former represent frequently co-occurring product combinations in the market baskets of the customer. [5]

VI. The work done:

1. Discovering part:

After getting to know the necessary information in the research part, we came to the part of discovering and reading the articles and the methods used in each one.

2. Comparison:

This is a comparison between the two articles we've read:

Name of the article	***Data Mining Based Store Layout Architecture for Supermarket. [9]***	***Analysis of Shopping Cart in Retail Companies Using Apriori Algorithm Method and Model Profset. [8]***
About	Its purpose is to find a technique that can help in marketing and sales. Five parameters induce the transition from recreational shopping to purchase oriented shopping: • Attractive store layout. • Navigational aids. • Sales people contact. • In-store events. • Environmental design characteristics. They believed that there are three parameters that engender transition from purchase-oriented shopping to recreational shopping. • Retail crowding. • Time pressure.	Its purpose is to find a solution for cost optimization in retail businesses. This research discussed the algorithm used in these three studies: • The study of Handojo [6]. • The study by Asana [7]. • The research on data mining made by Tom Brijs using the Profest Model. More specifically, this research is about merging the previous studies. It is expected to improve layout good goods.

	• Contextual factors. So they managed to make a technique that can cover cross-sales and related products.	
Method used	This system develops a relational database and uses Apriori algorithm techniques as methodologies for the store layout. This method requires two files: the first one contains transactions during a specific period and the second one contains product data. This method goes from transforming text files into MS SQL Server within OBDC environment to loading the text formatted into a relational database for querying, then to choosing the specific categories to construct correlation matrix. This approach allows supermarkets to cluster products around meaningful purchase opportunities related to use association.	This system uses association rules, Apriori algorithm techniques and the Profset model to maximize cross-selling opportunities. This method requires two files: the first one contains transactions during a specific period and the second one contains product data. This method goes from the association search that show the conditions of attribute values, to the classification and prediction which is a process of finding models, to the clustering which maximize the intraclass similarity. Then, it's time to use the Profset model to maximize the business profit. This approach allows retail businesses to cluster products around meaningful purchase opportunities related to use association and maximize cross-selling opportunities, which optimize the retail business profit.
Steps they took	This project contains the following steps: 1. Store information about categories of the products in data input model. 2. Use Apriori algorithm in order to understand the market basket.	This project contains the following steps: 1. Get the transactions from the database. 2. Use Apriori algorithm in order to understand the market basket.

	3. Apply the Association Rule to find sets of products that are frequently bought together. 4. Data modeling using Clementine to generate results for various situations.	3. Apply the Association Rule to find sets of products that are frequently bought together. 4. Use the Profset model to get the item sets with the most profit.
Advantages	• The speed of application of the Apriori algorithm. • They used a data warehouse, which is the newest form of decision support system. • This technique helps in marketing and sales. • Help the costumer by organizing store layout	• Product optimization is carried out using the Profest model. • The Profest model evaluates the profit margins generated per product. • The application of Apriori algorithm in the data mining technique is very efficient. • The Apriori algorithm accelerate the process of forming the combination of sales items. • Help the costumer by organizing store layout and optimize the business retail profit in the same time.
Disadvantages	• Do not maximize the profit of the retail business.	

3. Choosing part:

Now that we have read the articles and compared them, we need to choose the better method between the two.

We choose the 2nd article "***Analysis of Shopping Cart in Retail Companies Using Apriori Algorithm Method and Model Profset.***" [8] which works with Algorithm Apriori and Model Profset.

Why we chose that method?

The 1st method use only the Apriori algorithm which is speed. The technique of this method helps in marketing and sales but only benefits the costumer.

Otherwise, the method we chose use the Apriori algorithm which accelerate the process of forming the combination of sales items and also use the Profset model which evaluates the profit margins generated per product. Which means, this method benefits both the costumer and the retail business and this is why we chose it.

4. Implementation and coding:

Just after choosing the method, we need try it and see how does it works.

4.1. Learning Python basics:

Python is a powerful high-level, object-oriented programming language. It has simple easy-to-use syntax and when it comes to text processing Python is a very strong choice. The data analysis and parsing required for machine learning go well with Python, and its libraries.

4.2. Used tool:

Here we are using Python with PyCharm which is an integrated development environment used in computer programming (specifically for the Python language).

4.3. Python libraries used:

- **mlxtend**

Mlxtend (machine learning extensions) is a Python library of useful tools for the day-to-day data science tasks.

- **pandas**

Pandas is a software library for data manipulation and analysis.

4.4. Steps of coding:

We worked in this part on the example used in the article. We considered the transactions he gave us and started working from there.

First of all, we import the libraries we need for the process. Then, we insert the datasets we are going to work in which we find the transactions and the price of each product.

- **Apriori algorithm developing steps:**
  1. Instantiate transaction encoder: we transform the dataset into an array format suitable for typical machine learning APIs using the method .fit(), the TransactionEncoder() and via transform() method.
  2. Convert the result to a DataFrame.
  3. Compute frequent items using the Apriori algorithm: show each item sets and its support resulted by the Apriori() method. (here we fix the min support and the max length of item sets)
  4. Compute all association rules for frequent item sets with association_rules() method.
- **Profset model developing steps:**
  5. Get the basic combination of 2-itemset (xi):

     #: Prepare the information we will need later: putting each column from the prices data in a list, get the number of transactions and the number of item sets etc.

5.1. Get a list containing the item sets and its support (from the result of the 3rd step).

5.2. Take from the previous list the 3 item sets (xi) who has the highest support.

6. Calculate the gross profit (M(xi)) for each taken item sets using the formula:

   M(xi) = (sp(pi) – pp(pi)) + (sp(pj) – pp(pj))

   With: pi: product i, pj: product j, sp: selling price, pp: purchase price.

7. Get the combination (Zi) with products from the 5th step(xi):

   7.1. Get the list of products we are going to work on.

   7.2. Get combinations (Zi) from the previous products with max length 3 for each combination.

8. Calculate the total profit of each combination Zi using the formula:

   Z = M(x1)Px1 + M(x2)Px2 + M(x3)Px3 – C1Q1 – C2Q2 – C3Q3 – C4Q4

   With: M(xi): gross profit calculated in the 6th step,

   Pxi: 1 if the combination xi exists in Z and 0 if it doesn't,

   Ci: Storage and handling fees for product i,

   Qi: 1 if the product i exists in Z and 0 if it doesn't.

   8.1. Prepare a function to confirm the existence of the combination xi in the combination Zi (Pxi): return 1 if it exists and 0 if it does not.

   8.2. Prepare a function to confirm the existence of a product pi in the combination Zi (Qi): return 1 if it exists and 0 if it does not.

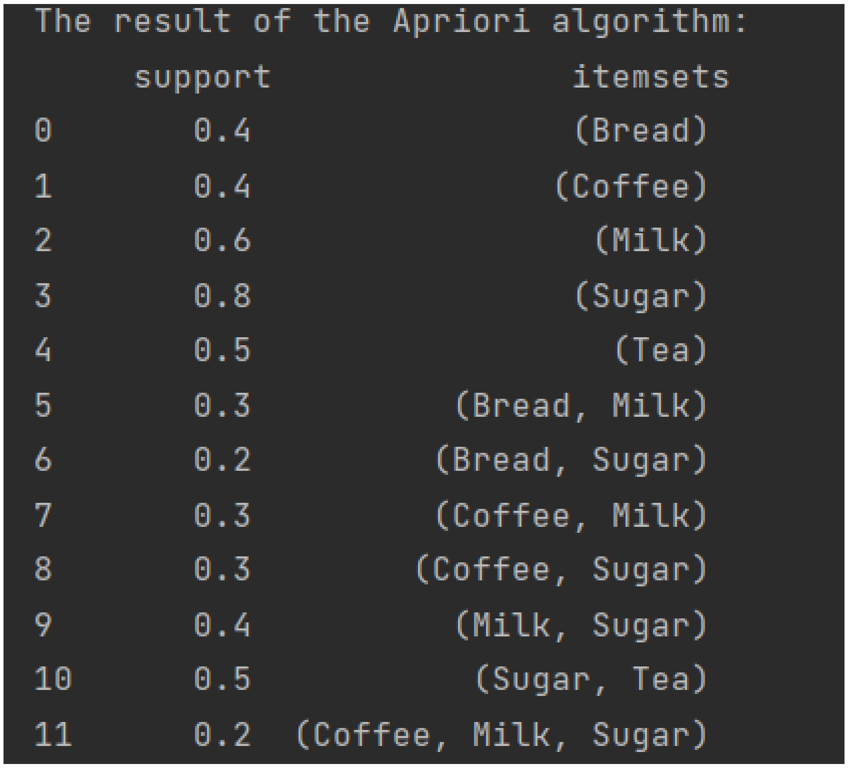
   8.3. Prepare a function to count the storage and the handling fees of each combination Zi: Return the storage and the handling fees.

   8.4. Calculate the combinations Zi using the previous functions.

9. Show the result of the Profset model: this gives the combination Zi, which has the highest total profit.

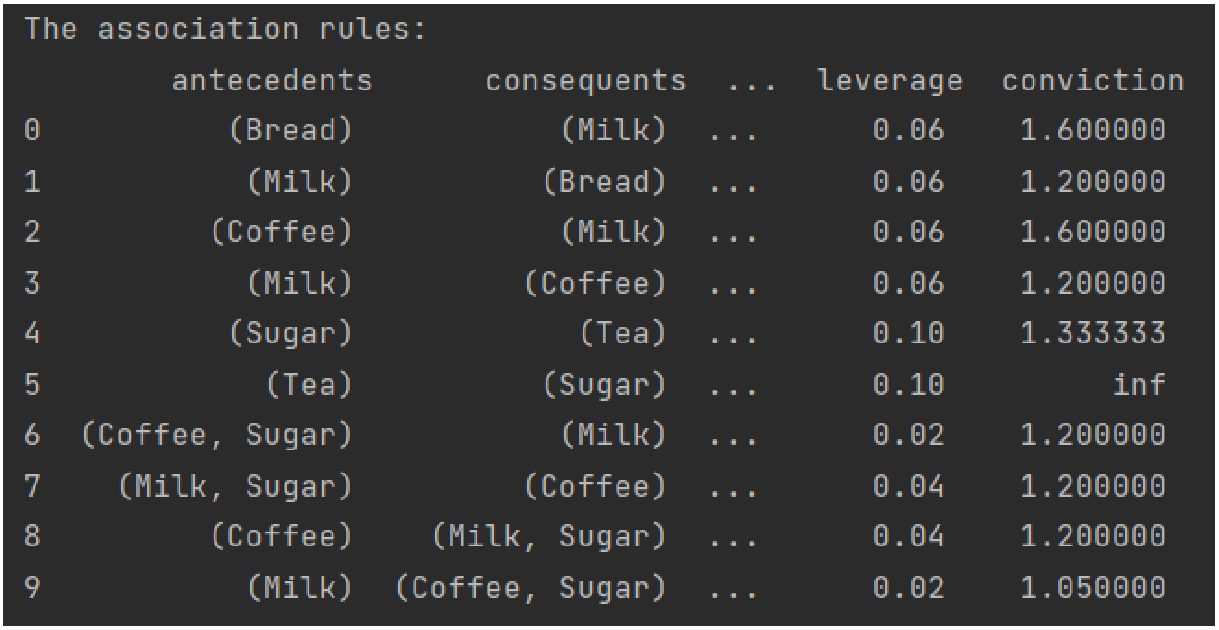
4.5. Running results:

- ### Apriori algorithm result:



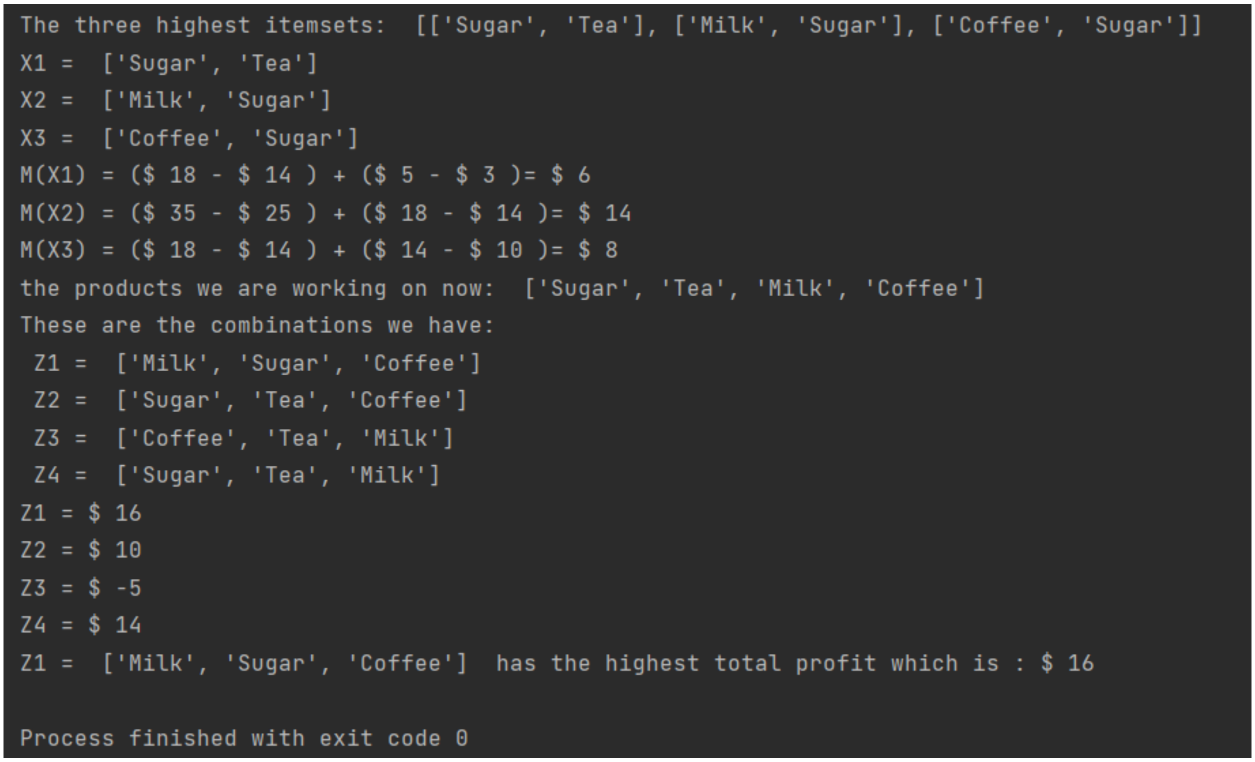
```
The result of the Apriori algorithm:
    support               itemsets
0       0.4                (Bread)
1       0.4               (Coffee)
2       0.6                 (Milk)
3       0.8                (Sugar)
4       0.5                  (Tea)
5       0.3          (Bread, Milk)
6       0.2         (Bread, Sugar)
7       0.3         (Coffee, Milk)
8       0.3        (Coffee, Sugar)
9       0.4          (Milk, Sugar)
10      0.5           (Sugar, Tea)
11      0.2  (Coffee, Milk, Sugar)
```

- ### Association rules result:



```
The association rules:
       antecedents      consequents  ...  leverage  conviction
0          (Bread)           (Milk)  ...      0.06    1.600000
1           (Milk)          (Bread)  ...      0.06    1.200000
2         (Coffee)           (Milk)  ...      0.06    1.600000
3           (Milk)         (Coffee)  ...      0.06    1.200000
4          (Sugar)            (Tea)  ...      0.10    1.333333
5            (Tea)          (Sugar)  ...      0.10         inf
6  (Coffee, Sugar)           (Milk)  ...      0.02    1.200000
7    (Milk, Sugar)         (Coffee)  ...      0.04    1.200000
8         (Coffee)    (Milk, Sugar)  ...      0.04    1.200000
9           (Milk)  (Coffee, Sugar)  ...      0.02    1.050000
```

- **Profset model results:**



```
The three highest itemsets:  [['Sugar', 'Tea'], ['Milk', 'Sugar'], ['Coffee', 'Sugar']]
X1 =  ['Sugar', 'Tea']
X2 =  ['Milk', 'Sugar']
X3 =  ['Coffee', 'Sugar']
M(X1) = ($ 18 - $ 14 ) + ($ 5 - $ 3 )= $ 6
M(X2) = ($ 35 - $ 25 ) + ($ 18 - $ 14 )= $ 14
M(X3) = ($ 18 - $ 14 ) + ($ 14 - $ 10 )= $ 8
the products we are working on now:  ['Sugar', 'Tea', 'Milk', 'Coffee']
These are the combinations we have:
 Z1 =  ['Milk', 'Sugar', 'Coffee']
 Z2 =  ['Sugar', 'Tea', 'Coffee']
 Z3 =  ['Coffee', 'Tea', 'Milk']
 Z4 =  ['Sugar', 'Tea', 'Milk']
Z1 = $ 16
Z2 = $ 10
Z3 = $ -5
Z4 = $ 14
Z1 =  ['Milk', 'Sugar', 'Coffee']  has the highest total profit which is : $ 16

Process finished with exit code 0
```

4.6. Conclusion:

After developing the method that came from the article ***“Analysis of Shopping Cart in Retail Companies Using Apriori Algorithm Method and Model Profset.”,*** we conclude that the use of the Model Profset and the Apriori Algorithm in retail businesses helps the select of the most interesting products from a product assortment based on their cross-selling potential given some retailer defined constraints. This method profits the retail businesses and also helps them better understand their costumer’s needs. Python has been very useful and helpful in this scientific research especially with its libraries.

VII. General conclusion:

In conclusion, I had the pleasure of interning at Run Way Tek company. The atmosphere at the Run Way Tek office was always welcoming which made me feel right at home. I can conclude that there have been a lot of practices I have learnt. The technical aspects of the work I have done are not flawless and could be improved upon provided with enough time.

My Internship at Run Way Tek was one of the most educational experiences I have ever encountered.

I learned the good side of not working with a team, which is mostly about managing time and self-motivation. In every project, there has to be multiple tasks and one month sure is not enough to do all the work, but with enough research done I managed to do more than what I expected.

I also had the opportunity to learn that Data Analysis is needed to grow your business and it provides insights that organizations need in order to make the right choices. Besides, I have improved my skills in Python language that I have learned in my university. Without forgetting that I have been taught the Algorithm Apriori and the Model Profset that helps in Market Basket Analysis.

Finally, this internship was a great experience that helped me know how to deal with a project and problems in a real work environment. I am glad I am leaving Run Way Tek with this amount of experiences, knowledge and lessons.

VIII. Preferences and bibliography:

[1]: https://en.wikipedia.org/wiki/Data_analysis

[2]: https://en.wikipedia.org/wiki/Data_mining

[3]:https://searchcustomerexperience.techtarget.com/definition/market-basket-analysis

[4]: https://en.wikipedia.org/wiki/Association_rule_learning

[5]: A Data Mining Framework for Optimal Product Selection in Retail Supermarket Data: The Generalized PROFSET Model, page3.

[6]: Handojo,2007. “Data Mining Application To Facilitate The Association Of Purchasing Goods With The Market Basket Analysis Method”

[7]: Asana,2013. “Shopping Basket Analysis With Apriori Algorithm In Retail Companies”

[8]: Permatasari, Putri Agung & Lie Jasa Linawati, 2020, "Analysis of Shopping Cart in Retail Companies Using Apriori Algorithm Method and Model Profset.", *International Journal of Engineering and Emerging Technology*, Vol.5, No.2, pp. 52-60.

[9]: Aishwarya Madan Mirajkar , Aishwarya Prafulla Sankpal ,
Priyanka Shashikant Koli , Rupali Anandrao Patil & Ajit Ratnakar Pradnyavant, 2016, “Data Mining Based Store Layout Architecture for Supermarket”, *International Research Journal of Engineering and Technology (IRJET)*, Vol.3, No.2, pp. 822-827.