Extracting hidden patterns from dates' product data using a machine learning technique

Mohammed Abdullah Al-Hagery
Computer Science Department, College of Computer, Qassim University, Buraydah, Saudi Arabia

Article Info

Article history:

Received Apr 12, 2019
Revised May 9, 2019
Accepted Jun 17, 2019

Keywords:

Association rules
Data analysis
Data mining
Dates product
Features extraction
Machine learning

ABSTRACT

Mining in data is an important step for knowledge discovery, which leads to extract new patterns from datasets. It is a widespread methodology that has the capability to help ministries, companies, and experts for diving into the data to find important insights and patterns to help them take suitable decisions. The farmers and marketers of the date product in the production regions lack to discover the most important characteristics of dates types from the economically, healthy, and the type of consumers point of view to achieve the highest profits by choosing the best types and the most consumed. The research objective is to extract interesting patterns from the dates' product dataset, using Machine Learning, based on association rules generation. This, in turn, will support the farmers, and marketers to discover new features related to the production, consumption, and marketing processes. This research used a real dataset collected from KSA, Qassim region, which is the first region of cultivation of palm, that produces the best types of dates in the Arab region. The data preprocessed and analyzed by the Apriori algorithm. The results show important features and insights related to the health benefits of dates, production, its consumption, consumers types, and marketing. Consequently, these results can be employed, for instance, to encourage individuals to consume dates for their nutritional value and their important health benefits. Furthermore, the results encourage producers to focus on the production of preferable types and to improve the marketing policies of the other types.

Corresponding Author:

Mohammed Abdullah Al-Hagery,
Department of Computer Science,
College of Computer, Qassim University,
Buraydah, Arab Saudi.
Email: hajry@qu.edu.sa

1. INTRODUCTION

The dates' data contain hidden patterns as valuable knowledge, including the most produced types, the most consumed types, the undesirable types, etc. The Qassim region in KSA is one of the most producers places of dates in the Gulf region and in the world besides Iraq. Currently, the industry of dates becomes is an important food industry [1]. Dates are considering the most popular kinds in KSA and in the Gulf region [2-3], many research works concentrated on the management, marketing, and traceability of date's product [4-6], during the processes of sell and export this product. And till right now, there is not any study employing data mining or Machine Learning techniques to benefit from the features and characteristics hidden in the dates' data. The research problem focuses on the weaknesses that are related to the dates product in regards to the random production and marketing and the un-ability to discover the most important characteristics of this product from the economically, healthy, and the type of consumers point of view (male, female, level of age, etc.) to achieve the highest profits by identifying and choosing the best types and the most consumed based on the analysis of the real datasets of this product by the machine learning tools.

There are several insights and hidden features can be discovered from the analysis of the product data and may help for producing the most important types then marketing them in proper ways. Till right now, these features are not available, such as consumer impression about this product, the best and the worst types, the identify the reasons for the consumption of some types and the abandonment of other types. There are also some difficulties in terms of production increase, which far exceeds domestic demand, whereas there is a clear weakness in the process of marketing the undesirable types. By using the dates' data and recent technology to benefit from these data. The required features relevant to the production and consumption process can be discovered. The importance of this paper comes from the importance of this product globally and especially in KSA, as one of the most important regions in palm growing and date production in the world.

This paper aims to extract interesting patterns from the dates' product dataset, using the Machine Learning (Apriori algorithm), by extracting the most important types of knowledge (unseen patterns as new features), which consists of a set of association rules. The proposed solution comprises many tasks including information and data collection, divide data into samples, data preprocessing, mining process, rules generation, validation of results, and accuracy improvement. Based on the application of the Apriori algorithm [7-9] and the establishment of a set of association rules to help decision-makers in that domain. The association rules usually used to find strong relations and important characteristics from the data [10-12]. This paper is organized as follows; Section 2 introduces the literature review. Section 3 explains the research methodology and results generation, in addition, section 4 describes the validation of the results, while section 5 explains rules filtration and section 6 includes the discussion of the results. Finally, section 7 demonstrates the conclusions. Finally, section 8 illustrates future works.

2. LITERATURE REVIEW

Currently, world companies and organizations are drowning in data but starving for knowledge. Data can be found as numerical values, records, figures, text documents, structures that are more complex, and etc. The complex data may appear in various forms; multimedia data, spatial data, and hypertext. To take complete advantage of data, we can retrieve and analyse it by different methods. These methods are complex and not enough for that purpose. It requires strong tools to discover patterns from raw data. With the massive amount of data placed in files, databases, and data warehouses, it is progressively imperative to utilize effective and powerful tools for data analysis and extraction of interesting patterns to help the decision-makers. This can be accomplished using Data Mining. Data Mining contains effective tools with great mechanisms to help miners focus to find the most important patterns from data using the Machine Learning algorithms. The Machine Learning deals with algorithm development as software that can learn and extract hidden patterns or features or relations from datasets. The Machine Learning algorithms adjust to changes and enhance performance according to the learning and training process. The Data Mining role is the application of Machine Learning algorithms on data for various purposes, such as prediction, classification, clustering, and extraction of association rules.

The most common types of association rules algorithms are the frequent itemset mining and mining association rules. Three classes of these algorithms discussed and compared; Apriority algorithm, FP-growth algorithm, and Eclat algorithm [13]; the Eclat algorithm is suitable for Big data sets and the Apriori algorithm and the FP-growth are better for small data sets, that's why we use it in this research. A typical example of using the association rules is to discover which items in a supermarket are normally put together in the basket market for a specific customer. Various approaches are employed for the association rules extraction [9]. In Data Mining, the datasets can be employed to compare and select the best methods such as classifiers and predictors for improving Data Mining techniques and algorithms [14]. One of the common Data Mining algorithms is the Apriori algorithm that is used for frequent patterns analysis and extraction of association rules. This algorithm usually used to generate all significant association rules between items in a database. Currently, many organizations/companies are using Data Mining task and Machine Learning on a regular basis. Some of these companies include; retail stores, schools, banks, and insurance companies. Many of these organizations combine Data Mining with such things as pattern recognition, statistics, and software tools. Data Mining used to find interesting patterns and relations that would otherwise be difficult to find. It allows data owners to study and understand their customer's behaviour and make smart marketing decisions [15], for their products and services.

The Data Mining always aims at the analysis of historical datasets from different perspectives [16-18], to sum up, the data in new ways that are both clear and useful to increase revenue, cut costs, or both for the data owner [2]. It becomes common in both the private and public sector [19, 20] to satisfy various needs using various applications that are employed in a local and global society to enhance the services and procedures. Therefore, there is an increasing request for mining about interesting patterns in datasets. The process of analyzing such data is a really computationally very complex process when using traditional methods [21]. In addition to what previously discussed, there are many research works provided as contributions in this field of study, some are focusing on the data analysis [22-25] and others are concentrating on the development and refinement of the algorithm [12, 26, 25, 16]. This is because the Data Mining is a multidisciplinary field with a wide and diverse application developed for data analysis. In fact, there exist non-slight gaps between knowledge discovery fundamentals and domain applications. A few of the application domains include; the analysis of product data, educational data, retail industry, spatial-temporal data, and medical data [26]. Furthermore, there are more related contributions are similar to this research, for instance, Cornelis studied and analyse the association rules problem relevant to positive and negative values for Big Data [27], likewise, Mahmood et al. concentrated on proposing an algorithm for discovering positive and negative association rules among frequent and infrequent item sets. The identified associations among medical test results using Data Mining algorithms [8]. Association rule generally comprises of a set of antecedent parts that lead to a consequent part with a certain confidence. Pazzani and Billsus see the list of subjects of books customers suggest for as transactions, which enable them to find groups of association rules for concerns that frequently appeared together as part of a customer's interests [28]. Also, Osadchiy et al. proposed an algorithm that recognized a model of collective preferences independently of the customer's interests. This requires a simple system of ratings, the performance of that algorithm evaluated by a large dataset of various transactions of real dietary recalls. It has demonstrated that the execution based on pairwise association rules achieves better for the defined task [29]. In fact, our research concentrates on a different idea, where it depends on the generation of association rules using a different kind of data consequently discover other types of knowledge.

Other research work provided a valuable community service, where Vasavi, used Data Mining algorithms for Hidden Patterns extraction from Road Accident dataset of highways that pass through Krishna district Indian for (2013), as a heterogeneous data collected from police stations. The objective was to find the shared features between accidents. The data analyzed using Machine Learning algorithms and the results generated are sets of association rules by Apriori algorithm [30], as well, Sene et al. worked on association rules but for analyzing a different database describing in-flight medical incidents to extract interesting knowledge from that data [7]. Miholca et al. investigated the problem of incremental relational of association rule mining. They proposed a new method named "Incremental Relational Association Rule Mining (IRARM)" for incrementally uncovering interesting relational association rules within a dynamic dataset during updates. A number of experiments carried out in order to show that the proposed method generates the results more rapidly than the execution of the Data Mining algorithms, on the extended dataset [31]. An additional approach presented for mining generalized association rules. An algorithm developed to scan the database one time only and use transaction dataset to compute the support of generalized item set faster than other similar algorithms [32]. Vidhate and Kulkarni proposed an efficient algorithm to a set of data collected from different shops to find a set of frequent items [33], on the other hand, Fernandez-bassso et al. proposed a parallelization algorithm for association rule extraction using Big Data technologies, which uses an efficient algorithm to address the problems related to the massive amounts of data [9].

Sadh and Shukla proposed a mining-based optimization technique for rule generation based on the Apriori algorithm and ant colony optimization approach. They applied the Apriori algorithm [34], on the other hand, Prajapati et al. identified consistent and inconsistent association rules from sales using a distributed datasets [21]. A modified form of the frequent itemset mining method presented using an improved formula for generating valid candidates by decreasing the number of invalid candidates. During the generation process of association rule sets, the confidence and support measures were applied [12]. The produced frequent k-item set is specified to the association rule generator to create all possible rules [35]. Rajeswari et al. proposed a modified fuzzy algorithm for Apriori rare Item sets mining to detect the outliers that represent weak student depend on the heap space usage [36]. An additional approach was proposed to extract a set of association rules based on medical data, the objective is to select the best mining algorithm of association rules according to multiple-criteria decision analysis [37]. In this paper, our approach is concentrating on the analysis of dates' data in order to find interesting patterns within the extracted association rules. These patterns are strongly relevant to the production and consummation of the date's product.

3. RESEARCH METHOD

The overall steps of the methodology are shown in Figure 1. It comprises dataset gathering, data preprocessing, mining process, knowledge generation & representation, and accuracy improvement. These steps are explained in the following sub-sections:

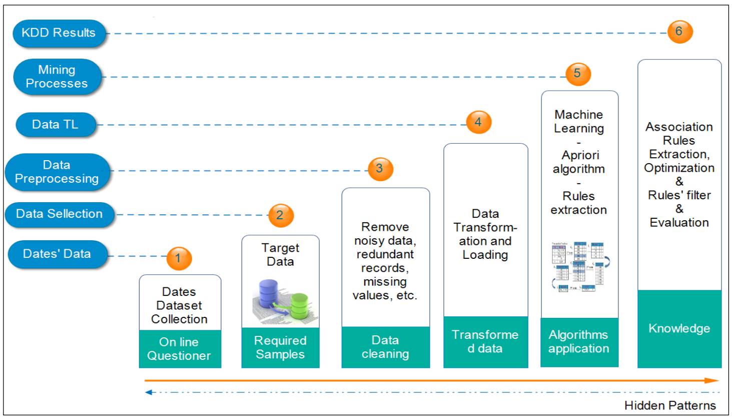

Figure 1. Methodology layouts

3.1. Information and Data Collection

Important information collected by interviews with a number of people and the collected data was by an online questioner designed, evaluated, and distributed to a sample of consumers, producers, marketers, and product manager. It distributed to a sample of 640 people. The collected dataset attributes presented in Table 1. Some values of the collected data are incorrect, and others are incomplete, it contains missing values, this reason leads to a cleaning. After the cleaning process, we got 499 records as a total number of instances employed in this research.

Table 1. The dataset attributes

I	Attribute	Description	Available values for this attribute
1	MCK	Annual Consumption in Kilo Gram (KG)	1-10 Kg, 10-20 Kg, …more-than-50-kg
2	NDC	Number of Daily Consumption	One-time, Two-times, Rarely, None
3	ASR	Spending Rate/family per year	50-100, 100-200, 200-300, 300-400, 400-500, 500-600, 600-800, 800-1000, more-than-1000
4	DBM	Dates as a Basic Meal	Yes, No
5	VF	Value of Food	Yes, No, To-some-extent, don't know
6	CR	Consumption Reason	Healthy, Social customs
7	CG	Consumer's Gender	Female, Male
8	MCT	Most Consumption Type	Ajwa, Barhi, Garawea, Hulwa, Khalas, Nabtat-Ali, Nabtat-Seif,
9	US	Undesirable Species	Rushodia, Ruthana, Shagra, Sukkari, Wannana, Ajwa, Barhi, Garawea,
10	MPT	Most Produced Types	Hulwa, Maktomi, Nabtat-Ali, Nabtat-Rashed, Nabtat-Seif, Nabtat-Sultan, Om-hmam, Rushodia, Ruthana, Shagra, Sukkari, Wannana
11	CA	Consumer's Age	All Ages, All Ages, Elderly, Children, Young

3.2. Dataset Samples

The dataset divided into four samples. The first consists of two attributes, the second includes four attributes, and the third & fourth contain five attributes. Some samples are overlapped. The output is a set of rules reflexing some features of the date's product relevant to production and consummation processes.

3.3. Data Preprocessing

The Pre-processing task is the basic step in knowledge discovery using machine learning [38], it includes various tasks [39]; remove inconsistent data, noisy data, attributes coding, transformation, and loading [40]. This, in turn, will improve the data quality and the accuracy of the results. The Apriori algorithm was selected as a useful rule-based technique in order to discover strong hidden patterns as a set of rules.

3.4. Data Analysis and Rules Generation

Association rules method is considered one of the important functionality of Data Mining, it includes three types; multilevel association rules, multidimensional association rules, and quantitative association rules. This research is using the multilevel association rules, the results of the analysis of four samples are demonstrated as follows:

– Rule Set 1 (Rs_1):

This sample consists of six rules based on two attributes {The Most Produced Types and Undesirable Species}. The results generated when the minimum support and the Minimum metric confidence were 0.1. The extracted rules are five rules, as follows:

a. Undesirable Species=Ajwa 59 → The Most Produced Types=Sukkari 444 conf:(0.93)
b. Undesirable Species=Garawea 52 → The Most Produced Types=Sukkari 444 conf:(0.92)
c. Undesirable Species=Om-hmam 100 → The Most Produced Types=Sukkari 444 conf:(0.94)
d. The Most Produced Types=Sukkari 444 → Undesirable Species=Om-hmam 84 conf:(0.19)
e. The Most Produced Types=Sukkari 444 → Undesirable Species=Ajwa 55 conf:(0.12)
f. The Most Produced Types=Sukkari 444 → Undesirable Species=Garawea 48 conf:(0.11)

– Rule Set 2 (Rs2):

It consists of a set of 5 rules generated based on four attributes, which are {Most Consumption Type, Monthly Consumption in KG, Number of daily consumptions, and Consumption Reason}. The minimum support value was 0.3 and the minimum metric confidence was 0.1. These rules are as follows:

a. Number of daily consumptions=One-time 218 → Most Consumption Type=Sukkari 189 conf:(0.87)
b. Consumption Reason=Social traditions 217 → Most Consumption Type=Sukkari 187 conf:(0.86)
c. Consumption Reason=Healthy 282 → Most Consumption Type=Sukkari 240 conf:(0.85)
d. Monthly Consumption in KG=1-10-kg 247 → Most Consumption Type=Sukkari 209 conf:(0.85)
e. Most Consumption Type=Sukkari 427 → Consumption Reason=Healthy 240 conf:(0.56)

– Rule Set 3 (Rs3):

It includes seven rules based on the analysis of five attributes {Spending Rate, Consumer Age, Value of Food, Consumer Gender, and Dates as a Basic Meal}. The minimum support value was 0.05 and the minimum metric confidence was 0.9. The best rules extracted are:

a. Spending Rate=more-than-1000-SR Consumer Age=Elderly Consumer Gender=Male Dates as a Basic Meal=Yes 31 → Value of Food=Yes 29 conf:(0.94)
b. Spending Rate=800-1000-SR Consumer Gender=Female 30 → Value of Food=Yes 28 conf:(0.93)
c. Spending Rate=400-500-SR Consumer Age=Elderly 28 → Value of Food=Yes 26 conf:(0.93)
d. Spending Rate=100-200-SR 41 → Value of Food=Yes 38 conf:(0.93)
e. Spending Rate=more-than-1000-SR Consumer Age=Elderly Dates as a Basic Meal=Yes 40 → Value of Food=Yes 37 conf:(0.93)
f. Spending Rate=400-500-SR Dates as a Basic Meal=Yes 30 → Value of Food=Yes 33 conf:(0.92)
g. Spending Rate=more-than-1000-SR Consumer Age=Elderly Consumer Gender=Male 46 → Value of Food=Yes 42 conf:(0.91)

– Rule Set 4(Rs4):

This set of rules generated based on five attributes overlapped with the previous sets, including {Undesirable Species, The Most Produced Types, Most Consumption Type, Monthly Consumption in KG, and Number of daily Consumption}. The minimum support was 0.1 and the minimum metric confidence was 0.89. The output is a set of 14 rules as follows:

a. Undesirable Species=Ajwa Most Consumption Type=Sukkari 52 → The Most Produced Types=Sukkari 51 conf:(0.98)
b. Most Consumption Type=Sukkari Monthly Consumption in KG=20-30-kg 61 → The Most Produced Types=Sukkari 59 conf:(0.97)
c. Most Consumption Type=Sukkari Number of daily consumptions=Rarely 64 → The Most Produced Types=Sukkari 61 conf:(0.95)
d. Most Consumption Type=Sukkari Monthly Consumption in KG=1-10-kg Number of daily consumptions=One-time 104 → The Most Produced Types=Sukkari 98 conf:(0.94)
e. Monthly Consumption in KG=20-30-kg 69 → The Most Produced Types=Sukkari 65 conf:(0.94)
f. Most Consumption Type=Sukkari Monthly Consumption in KG=1-10-kg 209 → The Most Produced Types=Sukkari 196 conf:(0.94)
g. Undesirable Species=Ajwa 59 → The Most Produced Types=Sukkari 55 conf:(0.93)
h. Most Consumption Type=Sukkari Number of daily consumptions=One-time 189 → The Most Produced Types=Sukkari 176 conf:(0.93)
i. Most Consumption Type=Sukkari 427 → The Most Produced Types=Sukkari 397 conf:(0.93)
j. Monthly Consumption in KG=1-10-kg Number of daily consumptions=One-time 125 → The Most Produced Types=Sukkari 116 conf:(0.93)
k. Undesirable Species=Ajwa The Most Produced Types=Sukkari 55 → Most Consumption Type=Sukkari 51 conf:(0.93)
l. Most Consumption Type=Sukkari Number of daily consumption=Two-times 119 → The Most Produced Types=Sukkari 110 conf:(0.92)

m. Undesirable Species=Om-hmam The Most Produced Types=Sukkari 84 → Most Consumption Type=Sukkari 77 conf:(0.92)
n. Monthly Consumption in KG=1-10-kg Number of daily consumption=Rarely 58 → The Most Produced Types=Sukkari 53 *conf:(0.91)*

3.4. Measuring Support and Confidence

In this step, the Support and Confidence measures applied to validate the outputs. Appendix A contains all generated rules with the ranking values of these measures. The values show the importance of each rule amongst other rules. The formulas of Support and Confidence are given in Formula (1) and (2), respectively [41-42]. The association rules format can be written as "IF" part = antecedent "THEN" part = consequent. The whole dataset applied once, but the final rules were limited and covering all Dates' types partially, that is the justification of divided the attributes into 4 samples and generate a big set of rules some of them were weak and the others were strong, then the filtration process.

$$supp(x \rightarrow y)=supp(x \cup y)=P(x \cap y)supp(x \rightarrow y)=supp(x \cup y)=P(x \cap y) \quad (1)$$

$$conf(x \rightarrow y)=supp(x \rightarrow y)/supp(x)=supp(x \cup y)/supp(x)=P(x \cap y)/P(x)=P(y|x) \quad (2)$$

4. RULES VALIDATION

To validate the generated rules, the frequent item generates strong association rules must satisfy minimum support and minimum confidence [42]. The minimum confidence of a rule is a user-defined value and an association rule is strong if it has supported greater than the minimum support value and confidence greater than the minimum confidence value [43]. All the generated rules are shown in Table 2 contains 32 rules. All of them have support and confidence values greater than the minimum support and minimum confidence values.

Table 2. All generated rules

Index	Rank	Rule Sets Components
9	522	Consumption Reason=Healthy 282 → Most Consumption Type=Sukkari 240 conf:(0.85)
8	404	Consumption Reason=Social traditions 217 → Most Consumption Type=Sukkari 187 conf:(0.86)
10	456	Monthly Consumption in KG=1-10-kg 247 → Most Consumption Type=Sukkari 209 conf:(0.85)
28	241	Monthly Consumption in KG=1-10-kg Number of daily consumptions=One-time 125 → The Most Produced Types=Sukkari 116 conf:(0.93)
32	111	Monthly Consumption in KG=1-10-kg Number of daily consumption=Rarely 58 → The Most Produced Types=Sukkari 53 conf:(0.91)
23	134	Monthly Consumption in KG=20-30-kg 69 → The Most Produced Types=Sukkari 65 conf:(0.94)
11	667	Most Consumption Type=Sukkari 427 → Consumption Reason=Healthy 240 conf:(0.56)
27	824	Most Consumption Type=Sukkari 427 → The Most Produced Types=Sukkari 397 conf:(0.93)
24	378	Most Consumption Type=Sukkari Monthly Consumption in KG=1-10-kg 209 → The Most Produced Types=Sukkari 196 conf:(0.94)
22	202	Most Consumption Type=Sukkari Monthly Consumption in KG=1-10-kg Number of daily consumptions=One-time 104 → The Most Produced Types=Sukkari 98 conf:(0.94)
20	120	Most Consumption Type=Sukkari Monthly Consumption in KG=20-30-kg 61 → The Most Produced Types=Sukkari 59 conf:(0.97)
26	365	Most Consumption Type=Sukkari Number of daily consumptions=One-time 189 → The Most Produced Types=Sukkari 176 conf:(0.93)
21	125	Most Consumption Type=Sukkari Number of daily consumptions=Rarely 64 → The Most Produced Types=Sukkari 61 conf:(0.95)
30	229	Most Consumption Type=Sukkari Number of daily consumptions=Two-times 119 → The Most Produced Types=Sukkari 110 conf:(0.92)
7	407	Number of daily consumptions=One-time 218 → Most Consumption Type=Sukkari 189 conf:(0.87)
15	79	Spending Rate=100-200-SR 41 → Value of Food=Yes 38 conf:(0.93)
14	54	Spending Rate=400-500-SR Consumer Age=Elderly 28 → Value of Food=Yes 26 conf:(0.93)
17	63	Spending Rate=400-500-SR Dates as a Basic Meal=Yes 30 → Value of Food=Yes 33 conf:(0.92)
13	58	Spending Rate=800-1000-SR Consumer Gender=Female 30 → Value of Food=Yes 28 conf:(0.93)
18	88	Spending Rate=more-than-1000-SR Consumer Age=Elderly Consumer Gender=Male 46 → Value of Food=Yes 42 conf:(0.91)
12	60	Spending Rate=more-than-1000-SR Consumer Age=Elderly Consumer Gender=Male Dates as a Basic Meal=Yes 31 → Value of Food=Yes 29 conf:(0.94)
16	77	Spending Rate=more-than-1000-SR Consumer Age=Elderly Dates as a Basic Meal=Yes 40 → Value of Food=Yes 37 conf:(0.93)
5	499	The Most Produced Types=Sukkari 444 → Undesirable Species=Ajwa 55 conf:(0.12)
6	492	The Most Produced Types=Sukkari 444 → Undesirable Species=Garawea 48 conf:(0.11)
4	528	The Most Produced Types=Sukkari 444 → Undesirable Species=Om-hmam 84 conf:(0.19)

Index	Rank	Rule Sets Components
1	503	Undesirable Species=Ajwa 59 → The Most Produced Types=Sukkari 444 conf:(0.93)
25	114	Undesirable Species=Ajwa 59 → The Most Produced Types=Sukkari 55 conf:(0.93)
19	103	Undesirable Species=Ajwa Most Consumption Type=Sukkari 52 → The Most Produced Types=Sukkari 51 conf:(0.98)
29	106	Undesirable Species=Ajwa The Most Produced Types=Sukkari 55 → Most Consumption Type=Sukkari 51 conf:(0.93)
2	496	Undesirable Species=Garawea 52 → The Most Produced Types=Sukkari 444 conf:(0.92)
3	544	Undesirable Species=Om-hmam 100 → The Most Produced Types=Sukkari 444 conf:(0.94)
31	161	Undesirable Species=Om-hmam The Most Produced Types=Sukkari 84 → Most Consumption Type=Sukkari 77 conf:(0.92)

5. RULES FILTRATION

The values of support measure normalized to a small range by dividing each value over 499, (the dataset size), to be compatible with the values of Confidence as a primary step to finding the rank values. The Support and Confidence values used to calculate the rank of each rule, according to Formula (3), after that the rules filtered by removing the redundancy and removing the rules that have lower ranks (lower quality). The next step is the selection of the rules that have the highest quality/highest ranks. Figure 2 demonstrates the final results for all generated rules and their ranks. The rules shown above the value 1.6 in the Y-axis in the chart, the highest points in this figure represent the highest ranks. These rules are shown in Figure 2, it includes the following set of rules {1, 2, 3, 7, 8, 9, 10, 12, 16, 18, 24, 26, 27}. This set contains the best rules, where it found that there are 13 rules have the highest ranks, it covers all dates' types included in the research; it represents the final results as in Table 3.

$$\text{Rank} = (\text{Sup-of Consequent}/\text{Ds}) + (\text{Sup-of Antecedent}/\text{Ds}) + \text{Confidence} \qquad (3)$$

where Ds is the dataset size =499.

Table 3. The Final set of rules

I	Rank	Rules
1	503	Undesirable Species=Ajwa 59 → The Most Produced Types=Sukkari 444 conf:(0.93)
2	496	Undesirable Species=Garawea 52 → The Most Produced Types=Sukkari 444 conf:(0.92)
3	544	Undesirable Species=Om-hmam 100 → The Most Produced Types=Sukkari 444 conf:(0.94)
7	407	Number of daily consumptions=One-time 218 → Most Consumption Type=Sukkari 189 conf:(0.87)
8	404	Consumption Reason=Social traditions 217 → Most Consumption Type=Sukkari 187 conf:(0.86)
9	522	Consumption Reason=Healthy 282 → Most Consumption Type=Sukkari 240 conf:(0.85)
10	456	Monthly Consumption in KG=1-10-kg 247 → Most Consumption Type=Sukkari 209 conf:(0.85)
16	77	Spending Rate=more-than-1000-SR Consumer Age=Elderly Dates as a Basic Meal=Yes 40 → Value of Food=Yes 37 conf:(0.93)
18	88	Spending Rate=more-than-1000-SR Consumer Age=Elderly Consumer Gender=Male 46 → Value of Food=Yes 42 conf:(0.91)
24	378	Most Consumption Type=Sukkari Monthly Consumption in KG=1-10-kg 209 → The Most Produced Types=Sukkari 196 conf:(0.94)
26	365	Most Consumption Type=Sukkari Number of daily consumptions=One-time 189 → The Most Produced Types=Sukkari 176 conf:(0.93)
27	824	Most Consumption Type=Sukkari 427 → The Most Produced Types=Sukkari 397 conf:(0.93)

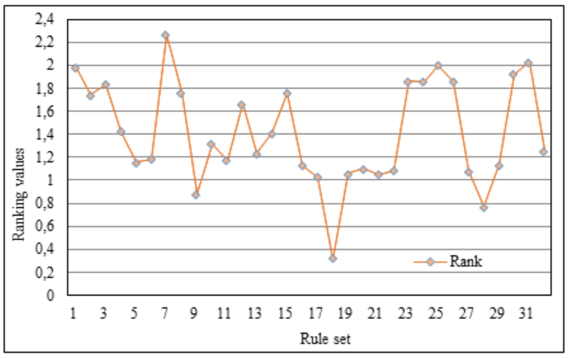

Figure 2. Rules ranking

6. RESULTS DISCUSSION

The results obtained as an association rule set, filtered by selecting the high ranks rules according to the two measures of support and confidence. This process includes; exclude the rules that are partially covering the required cases and the redundant rules. The final results include the minimum number of these rules, that cover the most cases of dates' types included in this research. The initial set of association rules extracted based on the ranking values (the highest values) amongst all generated rules that include 32 rules, although, all these rules, covering all required cases, it reduced to a small set containing 13 very strong rules. These rules provide the required knowledge given by the initial set; it will be discussed in the following points:

a. People are eating dates in the KSA because they are believing that it is a healthy meal in the first class then the second reason is eating dates as social traditions, especially with relatives and guests as shown in rule number 8 and 9 in Appendix A and in Table 3, their ranks are (522 and 404), respectively.
b. The results show that the most consumption type of dates is "Sukkari", this type is the desirable type by the consumers. All rules show this except the rules 16 and 18, the ranks of these two rules are low (88, 77), see Appendix A.
c. The amount of dates monthly consumed in Kg for each family is between 1 to 10-kg, this result supported by rule number10 and 24 with a ranking value equal to (456, 378).
d. The most produced type of dates in KSA is "Sukkari", based on the results illustrated by the rules number 1, 2, 3, 24, 26, and 27 that have a high rank based on the support and confidence values (503, 696, 544, 378, 365, and 824).
e. Likewise, the results demonstrate the number of daily consumption times is only one time as illustrated in the rules 7 and 26 with a ranking value equal to (407, 365).
f. As well as it found that the spending rate for each family per year is more-than 1000 Saudi Riyal, and the most consumer's age is "Elderly", as noticed in the rules number 16 and 18, corresponding to the following values of ranks (77, 88). In addition, these two rules show that the dates are representing an important food value.
g. Moreover, the rule number 18, illustrates that the consumer's type mostly is Male. The two rules (16 and 18) are giving patterns that are more important.
h. Also, the results showed that the dates consumed by KSA consumers as a basic meal, this obtained by the results of rule number 16, which has a ranking value is (77). Finally, the analyzed results have shown the most undesirable species of dates are respectively "Om-hmam" comes in the first place, secondly "Ajwa" after that is "Garawea". A set of rules are supporting this approach, includes rule number 1, 2, and 3, which have the following ranks (503, 496, 544), respectively.

7. CONCLUSIONS

The research results provide a type of contributions, represent the cooperation and interaction between the agricultural and information technology fields, for serving the community in KSA, Gulf region, and other countries around the world that are producing the dates. The research concentrated on dates' data analysis for serving the marketing and the production process of dates' product, and to understand the consumer Interests. The research results provide important knowledge, through the employment of the Data Mining in dates' data. The research concentrated on the extraction of new features and insights to improve the marketing and the production process of dates' product and to understand the consumer interests. Based on the extracted sets of association rules shown above, and the discussions carried out on those rules, we can reach the following conclusions:

a. The highest quality date's product type is "Sukkari" because it is the most produced and the most consumed type.
b. The undesirable types of dates are "Om-hmam", "Ajwa", and "Garawea", this, in turn, leads us to the following facts:
c. The focusing on a specific type of dates "Sukkari", which is the most consumption and the most productive type may be as a result of a strong marketing policy for this type and weak marketing policies of the other types or maybe is an inherited culture from the ancestors to their sons and grandchildren in this country over past decades.
d. Many producers have a great interest in cultivating specific species of dates although there are many other species, and this is probably due to the benefits they get.
e. Prices may have an important role to play in the process of buying or abstaining from a certain type of reason in the fame of that product (cheap or expensive).
f. Undesirable species of dates can be marketed in modern ways such as using the social media various platforms, Television channels, in addition to using other traditional methods, exploiting social and

cultural events and other occasions to distribute free samples of other species to create a strong consumption culture based on real experience.

g. Most of the consumers eat dates as a healthy meal and also as a basic meal in the first class, especially, amongst the elderly class.

h. Saudis buy dates with an average of more than SR 1,000 per family per year and most consumers are males.

8. FUTURE WORK

The research idea can be extended from different perspectives according to the following points:

– Study more features of the dates' product related to its health benefits.

– Collect inclusive data from various regions producing the dates, and increase the number of attributes used in the analysis.

– Improve the quality of the generated patterns using the Logical Analysis of Data (LAD) method.

– Using the same data to compare the accuracy of the results for various Machine Learning tools such as those included in Rapid Miner, IBM SPSS, Tanagra, python tools, KNIME, Orange, etc.

REFERENCES

[1] I. Pickrahn *et al.*, "Contamination incidents in the pre-analytical phase of forensic DNA analysis in Austria—Statistics of 17 years," *Forensic Sci. Int. Genet.*, vol. 31, no. 2, pp. 12–18, 2017.

[2] K. Anuradha and K. A. Kumar, "An E-Commerce application for Presuming Missing Items," vol. 4, no. 8, pp. 2636–2640, 2013.

[3] G. Kaur and L. Singh, "Data Mining : An Overview," vol. 4333, pp. 336–339, 2011.

[4] Y. M. Hwang, J. Moon, and S. Yoo, "Developing A RFID-based food traceability system in Korea Ginseng Industry: Focused on the business process reengineering," *Int. J. Control Autom.*, vol. 8, no. 4, pp. 397–406, 2015.

[5] L. Huang, Q. Luo, P. Yu, and G. Yu, "Designing and planning agricultural supply chain traceability system based on modern RFID technology," *Proc. 2011 Int. Conf. Mechatron. Sci. Electr. Eng. Comput. MEC 2011*, pp. 2112–2118, 2011.

[6] S. Hammami, A. Touir, and H. Alshede, "Analysis and Design of " Dates Traceability System "," vol. 4523, pp. 79–90.

[7] A. Sene, B. Kamsu-Foguem, and P. Rumeau, "Discovering frequent patterns for in-flight incidents," *Cogn. Syst. Res.*, vol. 49, pp. 97–113, 2018.

[8] S. Mahmood, M. Shahbaz, and A. Guergachi, "Negative and positive association rules mining from text using frequent and infrequent itemsets," *Sci. World J.*, vol. 2014, 2014.

[9] C. Fernandez-bassso, M. D. Ruiz, and M. J. Martin-bautista, "Extraction of Fuzzy association rules using Big Data technologies," vol. 11, no. 3, pp. 178–185, 2016.

[10] R. Agrawal, H. Road, and S. Jose, "Mining Association Rules between Sets of Items in Large Databases," no. May, pp. 1–10, 1993.

[11] Rakesh Agrawal and ramakrishnant, "Fast Algorithm for mining assocation Rules," pp. 1–32, 1993.

[12] J. Han, M. Kamber, and P. Jian, *Data Mining : Concepts and Techniques Third Edition*. 2011.

[13] H. Khanali, "A Survey on Improved Algorithms for Mining Association Rules," vol. 165, no. 9, p. 8887, 2017.

[14] M. Abdullah and H. Al-Hagery, "Classifiers' Accuracy Based on Breast Cancer Medical Data and Data Mining Techniques," *Int. J. Adv. Biotechnol. Res.*, vol. 7, no. 2, pp. 976–2612, 2016.

[15] S. P. Deshpande and V. M. Thakare, "Data Mining System and Applications: A Review," *Int. J. Distrib. Parallel Syst.*, vol. 1, no. 1, pp. 32–44, 2010.

[16] H. M. Al Shorman and Y. H. Jbara, "An Improved Association Rule Mining Algorithm Based on Apriori and Ant Colony approaches," vol. 07, no. 07, pp. 18–23, 2017.

[17] M. Kaya and R. Alhajj, "Genetic algorithm based framework for mining fuzzy association rules," vol. 152, pp. 587–601, 2005.

[18] Y. Tsay and J. Chiang, "CBAR : an efficient method for mining association rules," vol. 18, pp. 99–105, 2005.

[19] Oracle Manual, "Oracle® Data Mining Concepts, 11g Release 1 (11.1), B28129-04," *Doc. E16808-07, Oracle*, vol. 2, no. June, 2008.

[20] A. R. Kulkarni and D. S. D. Mundhe, "Data Mining Technique: An Implementation of Association Rule Mining in Healthcare," *Iarjset*, vol. 4, no. 7, pp. 62–65, 2017.

[21] D. J. Prajapati, S. Garg, and N. C. Chauhan, "Interesting Association Rule Mining with Consistent and Inconsistent Rule Detection from Big Sales Data in Distributed Environment," *Futur. Comput. Informatics J.*, vol. 2, no. 1, pp. 19–30, 2017.

[22] O. Maimon and L. Rokach, *Introduction to Knowledge Discovery and Data Mining*, vol. 2. 2009.

[23] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data*. 2014.

[24] I. H. Witten and E. Frank, "Data mining," *ACM SIGMOD Rec.*, vol. 31, no. 1, p. 76, 2002.

[25] S. Soni and J. Pillai, "Usage of nearest neighborhood, decision tree and Bayesian classification techniques in development of weight management counseling system," *Proc. - 1st Int. Conf. Emerg. Trends Eng. Technol. ICETET 2008*, pp. 691–694, 2008.

[26] M. A. Al-hagery, “Knowledge Discovery in the Data Sets of Hepatitis Disease for Diagnosis and Prediction to Support and Serve Community,” *Int. J. Comput. Electron. Res.*, vol. 4, no. 6, pp. 118–125, 2015.

[27] C. Cornelis, “Mining Positive and Negative Association Rules from Large Databases,” pp. 0–5, 2006.

[28] M. Pazzani and D. Billsus, “Content-Based Recommendation Systems The Adaptive Web,” *Adapt. Web*, pp. 325–326, 2007.

[29] T. Osadchiy, I. Poliakov, P. Olivier, M. Rowland, and E. Foster, “Recommender system based on pairwise association rules,” *Expert Syst. Appl.*, vol. 115, pp. 535–542, 2019.

[30] S. Vasavi, “Extracting hidden patterns within road accident data using machine learning techniques,” *Adv. Intell. Syst. Comput.*, vol. 625, pp. 13–22, 2018.

[31] D. Miholca, G. Czibula, L. Maria, D. Miholca, and L. Maria, “ScienceDirect A new incremental relational association rules mining approach A new incremental relational association rules mining approach,” *Procedia Comput. Sci.*, vol. 126, pp. 126–135, 2018.

[32] B. Vo and B. Le, “Fast Algorithm for Mining Generalized Association Rules,” vol. 2, no. 3, pp. 1–12, 2009.

[33] D. Vidhate, “To improve Association Rule Mining using New Technique : Multilevel Relationship Algorithm towards Cooperative Learning,” pp. 241–246, 2014.

[34] A. S. Sadh and N. Shukla, “Apriori and Ant Colony Optimization of Association Rules,” *Int. J. Adv. Comput. Res.*, vol. 3, no. 10, pp. 35–42, 2013.

[35] T. Ban, M. Eto, S. Guo, D. Inoue, K. Nakao, and R. Huang, “A study on association rule mining of darknet big data,” *Int. Jt. Conf. Neural Networks*, pp. 1–7, 2015.

[36] A.M. Rajeswar and C. Deisy Chelliah, “Outliers Detection on Educational Data using Fuzzy Association Rule Mining Outliers Detection on Educational Data using Fuzzy Association Rule Mining,” in *Int. Conf. on Adv. in Computer, Communication and information Science (ACCIS-14). Elsevier Publications*, 2014, no. July, pp. 1–9.

[37] M. Ramageri, “Data Mining Techniques and Applications,” *Indian J. Comput. Sci. Eng.*, vol. 1, no. 4, pp. 301–305, 2010.

[38] N. M. Samsudin, C. F. binti Mohd Foozy, N. Alias, P. Shamala, N. F. Othman, and W. I. S. Wan Din, “Youtube spam detection framework using naïve bayes and logistic regression,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, p. 1508, 2019.

[39] M. C. Babu and S. Pushpa, “Protecting sensitive information utilizing an efficient association representative rule concealing algorithm for imbalance dataset,” *Indones. J. Electr. Eng. Comput. Sci.*, vol. 15, no. 1, p. 527, 2019.

[40] T. R. S. Mary and S. Sebastian, “Predicting heart ailment in patients with varying number of features using data mining techniques,” *Int. J. Informatics Commun. Technol.*, vol. 8, no. 1, p. 56, 2019.

[41] R. Agrawal, T. Imieliński, and A. Swami, “Mining association rules between sets of items in large databases,” *ACM SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, 1993.

[42] C. Xiongying, “Research and Improvement of Apriori Algorithm for Association Rules,” pp. 0–3, 2010.

[43] H. Jnanamurthy, “Top Down Approach to find Maximal Frequent Item Sets using Subset Creation,” *Comput. Sci. Inf. Technol. (CS IT)*, pp. 445–452, 2012.

BIOGRAPHY OF AUTHOR

Mohammed Abdullah Al-Hagery: received his B.Sc in Computer Science from the University of Technology in Baghdad Iraq-1994. He got his MSc. in Computer Science from the University of Science and Technology Yemen-1998. Al-Hagery finished his Ph.D. in Computer Science and Information Technology, (Software Engineering) from the Faculty of Computer Science and IT, University of Putra Malaysia (UPM), November 2004. He was a head of the Computer Science Department at the College of Science and Engineering, USTY, Sana'a from 2004 to 2007. From 2007 to this date, he is a staff member at the College of Computer, Department of Computer Science, Qassim University, KSA. Dr. Al-Hagery was appointed a head of the Research Centre at the Computer College, and a council member of the Scientific Research Deanship Qassim University, KSA from September 2012 to October 2018. Currently, he is teaching the master degree students and a supervisor of four master thesis. He is a jury member of a number of PhD and master thesis, as an internal and external examiner in his field of his specialist.