Tuijin Jishu/Journal of Propulsion Technology
ISSN:1001-4055
Vol. 44 No. 4 (2023)

Consumer Online Behavior Analysis and CRISP-DM Process Model Development

[1]Amit Manglik, [2]Dr. JN Singh, [3]Prof. (Dr.) Munish Kumar Tiwari

[1]Research Scholar, Computer Science and Engineering Department, Galgotias University, Greater Noida, UP,

[2]Professor, Computer Science and Engineering Department, Galgotias University, Greater Noida, UP,

[3]Professor, Institute of Management, Commerce, and Economics, Shri Ramswaroop Memorial University, Barabanki, UP

Abstract:

Understanding the behavior of a company's customers forms the cornerstone of modern marketing. Data analysis is penetrated by cutting-edge artificial intelligence techniques like data mining and machine learning. These techniques can be used in a variety of businesses and for online sales of any type of commodity in big volumes. They are frequently employed in the sale of clothing, computers, and electronics. However, they can also be used in the B2B as well as B2C sales of seeds, agricultural goods, or agricultural machinery in the agricultural sector. A selling company might gain competitive advantages or higher profits by combining the right offers and customer knowledge. Our study's major objective is to create a CRISP-DM process model that will let small firms examine the behavior of their online clients. We analyze the online sales data using machine learning techniques including clustering, decision trees, and association rules mining to achieve the main objective. The usage of the proposed model in the area of online sales in the agriculture sector is discussed after the proposed model has been evaluated.

Keywords: *Consumer behaviour, data analysis, association rules, classification, online shopping.*

I. Introduction:

The acquired information and knowledge are extremely valuable in today's world of continuous change. The amount of data generated is also growing as a result of a high level of computerization in every aspect of our life, which is then further stored and monitored. Today, digital data analysis is more important than ever for the automotive, energy, engineering, and agricultural sectors. The objective is to extract information from a large volume of data that can be efficiently translated into knowledge that is crucial for comprehending a business's or company's trend. The gathering of such data frequently occurs to support an entity's choice to continue with the business process. Knowledge discovery in databases (KDD) has been approached in a variety of ways by different authors (Brachman and Anand, 1996; Fayyad, Piatetsky-Shapiro, and Smyth, 1996; Klösgen and Zytkow, 2002; Mannila, 1997; Simoudis, 1996), but they all share a few commonalities. The Enterprise Standard Data Mining Process (CRISP-DM), which was developed in 1996 as part of an EU project, outlines the fundamentals of data mining (Smart Vision Europe, 2015). The KDD process includes data mining (DM) techniques like business intelligence (BI), the outputs of which are used to enhance business decision-making through the creation of reports.

The sale of things online is one area that has a lot of potential and allows for effective real-time data collection. When it comes to online sales, the main objective is to guarantee client pleasure, reduce expenses, increase revenue, or streamline procedures. You can examine customer purchase behavior by using the data you've collected and the right machine learning techniques.In order to enhance corporate decision-making in the field of business processes, our contribution intends to provide a CRISP-DM process model. Using specific data from

the online sales of electronic components of the global company SOS Electronic, we will apply the proposed CRISP-DM process model to create marketing letters that will increase the efficacy of the marketing strategy and, hopefully, increase sales of these components. The major goal is to look into how big association rules and customer size relate to each other in terms of order volume and payment amount during the online ordering procedure. Due in large part to its high success rate, clustering is frequently utilized in segmentation, as several writers have noted (Safri, Arifudin, and Muslim 2018; Prashar, Vijay, and Parsad 2018; Suchacka, Skolimowska-Kulig, and Potempa 2015; Keller, Gray, and Givens, 1985). The primary principle of clustering is to maximize the differences between components from various clusters, while attempting to make elements within a cluster as similar as feasible. The benefit is that, in contrast to other classification and prediction methods based on the initial, regularly unaltered training set, the training set is regularly updated upon the inclusion of new features.

The most popular clustering technique is thought to be the k-means method, which MacQueen (1967) initially employed.Prashar, Vijay, and Parsad (2018) evaluate k-means, neural network prognostic ability, and linear discriminant analysis techniques to lessen the susceptibility of online retailers to demand in the market. Statistical proof of the approaches' forecast accuracy was shown at the study's conclusion, with the k-nearest neighbor method demonstrating an accuracy of 79.1%. Suchack, Skolimowska-Kulig and Potempa (2015) looked into the topic of distinguishing two user sessions in an online store: a shopping session and a browsing session. They determined that the k-nearest neighbor approach, which uses Euclidean distance, was the most successful in terms of both shopping and overall forecasts by comparing the outcomes of other strategies. Among the fundamental machine learning classification techniques are decision trees (Quinlan 1986), which allow data to be categorized based on decisions made in response to specific tests. The most widely used type of classifier representation is this one. In addition to concentrating on criteria statistics specifically for regression and classification trees, Komprdová et al. (2012) characterize the CART algorithm as one of the most well-known algorithms for building decision trees, which is also a fundamental representation of binary trees. Because more binary trees can be created by altering the CART tree's rules, it clarifies the fundamentals of decision tree construction. Based on historical consumer behavior data from the seller's website, the decision tree classification method is most frequently used to assess online customer behavior.

Decision trees and the program Weka are presented by Sun, Cárdenas, and Harrill (2016) as a novel method and instrument that pinpoints crucial characteristics that influence the caliber of the client experience when accessing a travel agency website. Raj and Singh (2016) use decision tree techniques to divide consumers into three groups: frequently shoppers, frequent customers, and less frequent customers, in order to examine how the demographic environment influences the frequency of online transactions.Products acquired through online or in-store sales could be related. An association rule (AR) is established when the existence of one product influences the existence of another in the same purchase. This relationship is expressed as an implication. The authors of this idea were Agrawal, Imielinski, and Swami (1993). Association rules mining (ARM) is the practice of looking through consumers' shopping carts to identify ARs that need to fulfill specific requirements in order to be considered significant. Language studies (Adamov, 2018), electronic transaction security (Askari, Md, and Hussain 2020), medicine (Buczak et al., 2015; Soni et al., 2011; Luo et al., 2013), but also frequently in customer analysis (Kaur and Kang, 2016; Guo, Wang, and Li, 2017) are just a few of the research and analysis fields that currently use ARM. Initially presented by Agrawal et al. (1994), the Apriori method is one of the most widely applied techniques for ARM.

Since then, improvements have been made to the Apriori approach to accelerate the ARM process. For example, Yuan (2017) has suggested minimizing the number of database scans required or Wu et al. (2009) has suggested reducing the amount of operation required in order to enhance operational efficiency. Apriori algorithm adaptation to real-time online consumer behavior analysis for particular enterprises was the goal of further revisions (Kaur and Kang 2016; Guo, Wang, and Li, 2017). Alfian et al. (2019), for instance, suggest utilizing ARM to analyze consumer behavior in real-time for online commerce. They track clients, product browsing history, and transaction data from digital tagging using the suggested method for analysis. ARM is more frequently used to track the purchasing process in a brick and mortar store and for specific goods, despite the

fact that it has been mentioned in numerous studies that evaluate consumer behavior during the purchasing process (Avcilar and Yakut 2014; Chen et al. 2015). We have not discovered any publications on the research of the sale of agricultural commodities by this method, and it is not frequently employed in the study of the online purchase process. This has grown to be the main driving force for our investigation.

Although decision-tree (CART), ARM (Apriori), and clustering (k-means) techniques are widely used, we did not come across a combination of them and utilize them in the research of online customer behavior. Nonetheless, other writers have suggested or employed comparable combinations to examine customer behavior. Kunjachan, Hareesh, and Sreedevi (2018) analyze a lot of data—online sales data—using k-means, Apriori, and Eclat techniques combined. They suggest using this technology to mine hidden data links and analyze consumer behavior more easily. In order to develop a model for client segmentation in an online setting, Ma, Haiying, and Dong Gang (2011) used a methodology based on the combination of ARM and decision trees. Helping managers comprehend clients, assess the market, and make business management judgments was the advantage. Our objective is to demonstrate a blend of machine learning techniques for interpreting consumer behavior and a customized offer of a set of products. Our goal is to demonstrate that the provided method may be applied to the selling of different agricultural goods and products in the agricultural industry. We could not find many examples of these techniques being used in the agriculture sector when we looked through the current literature. The sales of gadgets, personal computers, and apparel are the main industries where their usage can be observed. An overview of data mining approaches in agriculture is given by Gandhi and Armstrong (2016), who cite methods such support vector machines, neural networks, and Bayesian networks. Santosh Kumar and Balakrishnan (2019) directly recommend products to customers in the agriculture sector by utilizing the Apriori algorithm. The analysis of agricultural data also frequently makes use of clustering (Shedthi et al. 2017; Zhao et al., 2009). Nevertheless, none of these studies concentrate on applying these techniques to real-time customer behavior analysis.

II. Methods and techniques:

We discovered that clustering is the most widely used technique for customers to examine an online purchase process based on a review of recent research in this field. Because it enables you to establish relevant client groups and select the best marketing management for them, this approach makes sense. One can also presume that distinct association rules will apply to various customer clusters. As a result, we logically propose the following study hypothesis:

H1: Every associated customer group has its own set of association rules.

In our investigation, we noted the frequency with which a buyer of the relevant things simultaneously purchased another. We were only interested in the relationship under analysis under the presumption that the presence of such pairs (in our example) would be sufficient to satisfy the predefined conditions. The ability of the business to do online sales was a must when choosing a research site, and the most reliable data was preferred. The dataset includes information from the company that distributes electronic components' online orders over the course of a year. We acquired information on 185,706 purchases made from 4,111 businesses in a single year.

The dataset has the following information in each row:

The id of the purchased products,

The id of the company that made the purchase,

The buy date,

The quantity of things purchased, and the price per unit of goods.

Identification numbers have taken the place of customer and item names in order to maintain data anonymity. Because the R software environment is freely available and has numerous packages that enable us to do every step of our study, we completed the entire work process there. Naturally, we selected the number of orders placed and the amount of money paid from the dataset as our criterion for selection. From the perspective of developing a marketing strategy, this choice best captures the client and is also feasible from the dataset's

perspective. First, we determined how many purchases each client had made overall and how much they had spent overall on online purchases from the chosen retailer over the course of the year. After that, we eliminated the outliers from these numbers that would have distorted the outcome, and we might still investigate these outliers further. This is accomplished by using the boxplot(Data)$out command (Chambers, 2017; Becker, 2018; Murrell, 2018), which outputs show the outliers that are present in the Data table.

Utilizing the robustHD package and the standardize() command, normalize the data in the second phase to get it ready for clustering. We can determine the ideal number of clusters using the NbClust program (Charrad and Ghazzali, 2014). It analyzes the findings, suggests the ideal number of clusters, and compares the optimal number of clusters found using twenty-four different techniques. The original, non-normalized total amount and purchase data are included in the matching results that are put to the table after this is utilized for cluster analysis of k-means utilizing the kmeans() (Hartigan and Wong, 1979) command.

Using the rpart (Breiman, 2017) package, the decision tree and associated categorization rules are generated in the third stage. The decision tree's goal characteristics will be the column cluster that represents the value of the burst to which the client belongs, while the input attributes to be tested will be the number of purchases and total amount. This stage will yield a model that an organization may use to divide up new clients into preexisting clusters.

Using the arulez and arulezviz programs, we will mine the association rules for each cluster independently in the last stage. Hahsler and Gruen programmed the Apriori function into the R environment using the original (Agrawal, Imieliński, and Swami, 1993) and more contemporary (Lepping, 2018; Borgelt and Kruse, 2002; Borgelt, 2003) designs of this method. The Customer Data table is first divided into as many tables as the number of clusters that occur, and only the entries in each table that correspond to their clusters are recorded. Next, a list of transactions (shopping carts) will be generated, with each transaction including a set of commodities that were bought. We will next mine the ARs in the clusters under the assumption that a single purchase will be made up of the items that a single customer purchases in a single day. Using the itemset, frequency, support, confidence, and lift indicators, the resulting ARs will be assessed.

The itemset is the collection of products sold online to SOS Electronic. The absolute abundance of product A in an itemset is represented by frequency (A), while the relative abundance of product A occurrences in an itemset is represented by support (A). Equation 1 illustrates the concept of confidence $(A \tilde{=} B)$, which represents the ratio of purchases containing both A and B products to those containing only product B.

Confidence (A=>B) = Frequency (A, B)/ Frequency (A) (1)

Thus, it establishes the probability that a purchase containing product A will include product B. But let's look at an example where product A is only used in 5% of purchases and product B is used in all purchases . Because the occurrence of Product B in the same transaction is unaffected by the existence of Product A in the buy , the predictive value of this AR would be minimal even though the confidence $(A \tilde{=} B)$ would be equal to 1. Because of this, the Lift $(A \tilde{=} B)$ indicator is used to measure the strength of AR by determining the impact that the purchase of product A has on the purchase of product B (see to equation 2).

Lift (A=>B) = Support (A,B)/ Support (A) * Support (B) (2)

Lift (A=>B) would equal 1 in our case, where Product B was in every transaction, indicating that the incidence of Product B in the same buy is unaffected by the existence of Product A in the purchase. The more product A influences product B in the same purchase, the higher the Lift $(A \tilde{=} B)$ and the more product A is negatively impacted by product B in the same purchase, the lower the Lift value is below one. The Apriori technique looks for ARs that satisfy the user-selected parameters of min_confidence and min_support (or absolute value min_frequency). We left the arulez package's default values of min_frequency = 7 and min_confidence = 1 for association rules mining. A table expressing the association rules and the corresponding values for count (frequency), lift, confidence, and support will be the end product.

III. Findings and conversation

The following components make up the analysis results:

Client groups determined by the total amount spent and the total number of purchases made on products;

Classification rules as a textual notation of particular branches of the decision-making tree; decision tree as a model for allocating consumers to clusters from the preceding point;

Association rules that are mined independently for each active cluster.

The NbClust() command was used to determine the ideal number of bursts, and the outcome showed that three clusters is the ideal number.Next, we used the kmeans() tool to divide the consumers into three clusters.

Table 1 shows the absolute and relative (roundedto 2 decimal places) abundance of customers in clusters.

Absolute abundance	3237	547	327
Relative abundance	78.74%	13.31%	7.95%

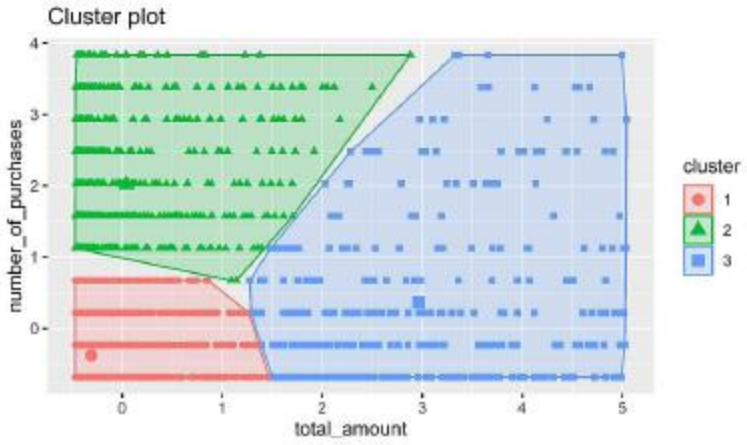

Figure 1: Clusters visualization.

Figure 1 displays a graphical depiction of the client assignment to the clusters. According to the graphical evaluation, customers with a low number of purchases overall make up Cluster 1, customers with a higher number of purchases overall make up Cluster 2, and customers with a higher number of purchases overall make up Cluster 1 across the whole range of the number of purchases. Since clients who made inexpensive, little transactions are included in Cluster 1, which has the largest absolute abundance. We are unable to ascertain the number of purchases or total amount that serve as the boundaries between distinct clusters, nevertheless, based on Table 1 and Figure 1. Because of its binary, we produce a CART decision tree for simplicity in order to determine these bounds.

We created a CART decision tree with an infinite supply of clients at each stage, as seen in Figure 2, using the software rpart. Using the created decision tree, the accuracy of classifying customers into distinct clusters is 99,37%. We also produced a C5.0 decision tree with a greater accuracy (99.93%) for comparison. However, C5.0 decision tree has ten branches, making it more difficult to employ going forward when compared to the four branches of CART decision tree, which is why we stuck with CART decision tree.

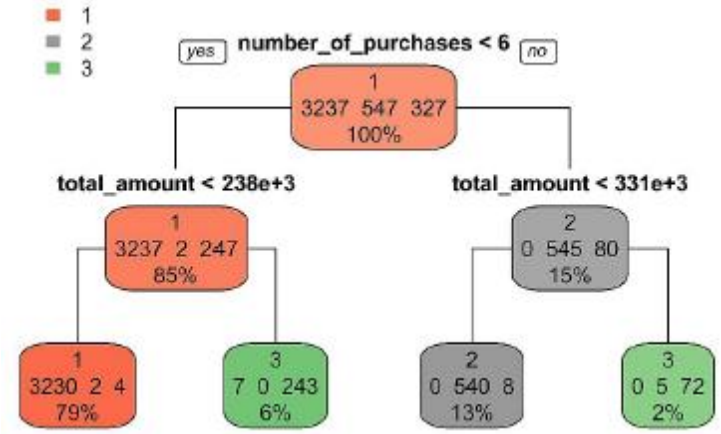

Figure 2: Decision tree.

We obtain classification rules based on the created decision tree displayed in Table 2 by using the same software rpart. Final client assignments to the clusters indicated by the accompanying listing are represented by lines ending in the symbol *. Examine all relatives in the unsuitable cluster.

For instance: 5)

where: number_of_purchases>=16.5 72 13 3 (0.18 0 0.81) *

Step(5) is the number of the decision tree branch (step no. 5 in this case); test (number_of_purchases >=16.5);

Total (72) is the total number of customers that entered the test;

Unfit (13) is the number of consumers that the model assigned to the cluster differently. As the decision tree flows in the example, k-means assigned 13 consumers to the opposite cluster;

Cluster (3) is the customer's cluster number if he passes the test; relatives (0.18 0 0.81) is the cluster-by-cluster relative abundance of unfit customers. In the case, k-means placed 18% of the customers to cluster 1, while a decision tree assigned them to cluster 3.

Table 3 represented mined association rules for cluster 1 and its values for each mined

1) root 4111 874 1 (0.7873996595 0.1330576502 0.0795426903)

2) number_of_purchases< 5.5 3486 249 1 (0.9285714286 0.0005737235 0.0708548480)

3) total_amount< 238490.7 3236 6 1 (0.9981458591 0.0006180470 0.0012360939) *

4) total_amount>=238490.7 250 7 3 (0.0280000000 0.0000000000 0.9720000000) *

5) number_of_purchases>=5.5 625 80 2 (0.0000000000 0.8720000000 0.1280000000)

6) total_amount< 330888.3 548 8 2 (0.0000000000 0.9854014599 0.0145985401) *

7) total_amount>=330888.3 77 5 3 (0.0000000000 0.0649350649 0.9350649351) *

Rule of association. The implication input is in the lhs (left hand side) column, while the implication output is in the rhs (right hand side) column. Table 3's first association rule can be understood as follows: "The rule has been applied in 13 cases, and if the customer bought the product 75125, he also bought the product 75127 with 100% probability (expressed by confidence = 1)."

The number of association rules that k-means with the same settings mined for each cluster is:

Cluster 1 = 92 association rules;

Cluster 2 = 16 association rules;

Cluster 3 = 12 association rules.

We looked for intersections between the ARs and compared the values of the ARs in the event that a match occurred. Only C2 ∩ C3 = 1 AR was the intersection of ARs across clusters with the same lhs and rhs in both clusters. AR was shown to have varying values of indicators (lift, confidence, and support) between clusters, even when they intersected. We draw the conclusion that each connected customer group has particular association rules based on this, and for that reason, we do not reject the hypothesis.Business management can now group clients into clusters based on the results to have a better knowledge of their organizational structure. Clusters can be used to evaluate consumer baskets and forecast consumer behavior by adhering to association principles. Following the decision tree or categorization criteria, the client would first be assigned to the cluster, after which they might abide by the cluster's association rules. To swiftly allocate new consumers to the clusters, we created a model using a CART decision tree because of its simplicity. Nonetheless, based on his preferences, the user can select from a variety of decision tree forms for analysis.

The work can be used as a CRISP-DM process model, which is appropriate for smaller organizations that haven't yet examined customer behavior in online sales and doesn't require additional application expenditures. Real-time consumer behavior prediction is possible with association rules. The consumer is immediately assigned to one of the clusters with associated ARs after logging into the website. All notable ARs with the product on the lhs side are searched when a customer adds an item to his shopping cart. To increase the chances of success, customized advertising can be made using ARs that have the highest level of confidence and support.The degree of validity of the AR and its frequency in the original data are discussed by confidence and support. This kind of targeted advertising can boost revenue for businesses. KPIs, such as the following, can be used to track how well the proposed CRISP-DM process model is being implemented: KPI 1: The percentage of revenues and costs related to the implementation of the proposed CRISP-DM process model; KPI 2: The quantity of recommended goods that customers purchase after seeing targeted advertising; and KPI 3: The total revenue from recommended goods.

It is possible to analyze frequently occurring consumer baskets using the generated data by the ARs. The examination of the connections between certain items is fundamental. An organization can examine why a customer chooses to purchase product B over product A in spite of the fact that product B has a number of alternatives. The findings of such a study might be put to many uses, such as providing items A and B together in a discounted package or acting as a catalyst to confirm and enhance the caliber of product B's substitute. Process adjustments based on these analyses may result in cost savings or increased revenue for the company. For example, commodities A and B might be packaged together in a single box to save money on additional packaging. Management should concentrate on the ARs with the highest Lift value, which assesses the strength of the relationship between the goods in each AR, when choosing association rules in such an analysis. KPIs comparable to the previously discussed consumer behavior prediction can be used to assess the effectiveness of the adjustments based on shopping cart analysis.

Our findings demonstrate that a company's sales earnings can be raised by making the appropriate offers to the appropriate clients. Our goal is to use this set of techniques in a less common setting, such the agro-business industry.

Conclusion:

With profit as the main benefit, consumer behavior analysis makes it feasible to supply information for business decision-making that helps achieve corporate objectives. Either raising revenue or cutting expenses within specific corporate operations can accomplish this. In this work, we develop and offer a low-cost process model, CRISP-DM, which, when applied and used appropriately, permits both. The foundation of the suggested CRISP-DM process model is the use of the k-means approach to divide current consumers into discrete groups; the decision tree method to create a model for segmenting potential customers; and the Apriori method to analyze the shopping cart. We explore potential applications for the data as well as metrics for gauging the effectiveness of the various options. The article can also be used as a quick summary of what is currently known about machine learning techniques and how they are being applied in scientific research. While our study was

conducted using data from an electronics company, we believe that the suggested CRISP-DM model can also be applied in the agricultural sector, where it would enable the analysis of consumer behavior in the selling of seeds, agricultural products, or equipment useful for the agro-industry. In contrast with the previously mentioned studies (Gandhi and Armstrong, 2016; Kumar and Balakrishnan, 2019; Shedthi et al., 2017; Zhao et al., 2009), we present a model that combines a number of data mining techniques and enables real-time consumer behavior analysis of agricultural product sales in an online sales environment.

We have abstracted the CRISP-DM process model from customers, and we will concentrate on extreme values in our future study. We are going to analyze customers with a modern machine learning method RFM (recency, frequency and monetary value), which assesses the importance of customers based on how often they make purchases when they made their last purchase, and at what value they made during the reporting period. The model presented in this work, along with its other proposed adjustments, will be put to the test in an online agricultural sales environment. By dividing up the consumer base according to certain factors, you may eventually improve the association rules' performance and accomplish your corporate objectives.

References

[1] Adamov, Abzetdin Z. (2018) “Mining Term Association Rules from Unstructured Text in Azerbaijani Language”, *In 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*, IEEE, pp. 1-4. DOI 10.1109/ICAICT.2018.8747143.

[2] Alfian, G., Ijaz, M. F., Syafrudin, M., Syaekhoni, A., Fitriyani, N. L. and Rhee, J. (2019) “Customer Behavior Analysis Using Real-Time Data Processing: A Case Study of Digital Signage-Based Online Stores”, *Asia Pacific Journal of Marketing and Logistics*, Vol. 31, No. 1, pp. 265-290. ISSN 1355-5855. DOI 10.1108/APJML-03-2018-0088.

[3] Askari, S., Md. S. and Hussain, Md. A. (2020) “E-Transactional Fraud Detection Using Fuzzy Association Rule Mining”, *Proceedings of the 2nd International Conference on Information Systems & Management Science (ISMS) 2019*, Tripura University, Agartala, Tripura, India, 6 p.

[4] Avcilar, M. Y. and Emre, Y. (2014) “Association Rules in Data Mining: An Application on a Clothing and Accessory Specialty Store”, *Canadian Social Science*, Vol. 10, No. 3, pp. 75-83. E-ISSN 1923-6697, ISSN 1712-8056.

[5] Becker, R. A. (2018) "*The New S Language: A Programming Environment for Data Analysis and Graphics*", CRC Press. ISBN 053409192X. DOI 10.1201/9781351074988.

[6] Borgelt, Ch. and Kruse, R. (2002) “Induction of Association Rules: Apriori Implementation”, In: Härdle W., Rönz B. (eds) *Compstat*, Physica, Heidelberg. E-ISBN 978-3-642-57489-4. DOI 10.1007/978-3-642-57489-4_59.

[7] Borgelt, Ch. (2003) “Efficient Implementations of Apriori and Eclat”, In *Proc. 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations*, FIMI 2003, Melbourne, FL, CEUR Workshop Proceedings 90.

[8] Brachman, R. J. and Anand, T. (1996) “The Process of Knowledge Discovery in Databases”, In *Advances in Knowledge Discovery and Data Mining*, pp. 37-57. ISBN 9780262560979.

[9] Breiman, L. (2017) "Classification and Regression Trees", Routledge. ISBN 1138469521. DOI 10.1201/9781315139470.

[10] Buczak, A. L., Baugher, B., Guven, E., Ramac-Thomas, L. C., Elbert, Y., Babin, S. M. and Lewis, S. H. (2015) “Fuzzy Association Rule Mining and Classification for the Prediction of Malaria in South Korea”, *BMC Medical Informatics and Decision Making*, Vol. 15, No. 1, pp. 47. ISSN 1472-6947. DOI 10.1186/s12911-015-0170-6.

[11] Chambers, J. M. (2017) "*Graphical Methods for Data Analysis*", Chapman and Hall/CRC, 410 p. ISBN 9781315893204.

[12] Charrad, M. and Ghazzali, N., Boiteau, V. And Niknafs, A. (2014) "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set", *Journal of Statistical Software*. ISSN 1548-7660.

[13] Chen, Ch.-Ch., Huang, T.-Ch., Park, J. J. and Yen, N. Y. (2015) "Real-Time Smartphone Sensing and Recommendations towards Context-Awareness Shopping", *Multimedia Systems*, Vol. 21, No. 1, pp. 61-72. E-ISSN 1432-1882, ISSN 0942-4962. DOI 10.1007/s00530-013-0348-7.

[14] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) "The KDD Process for Extracting Useful Knowledge from Volumes of Data", *Communications of the ACM*, Vol. 39, No. 11, pp. 27-34. E-ISSN 1557-7317, ISSN 0001-0782. DOI 10.1145/240455.240464.

[15] Gandhi, N. and Armstrong, L. J. (2016) December. A review of the application of data mining techniques for decision making in agriculture. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, IEEE, pp. 1-6. DOI 10.1109/IC3I.2016.7917925.

[16] Guo, Y., Wang, M. and Li, X. (2017) "Application of an Improved Apriori Algorithm in a Mobile E-Commerce Recommendation System", *Industrial Management & Data Systems*, Vol. 117, No. 2, pp. 287-303. ISSN 0263-5577. DOI 10.1108/IMDS-03-2016-0094.

[17] Hartigan, J. A. and Wong, M. A. (1979) "Algorithm AS 136: A k-Means Clustering Algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, Vol. 28, No. 1, pp. 100-108. E-ISSN 14679876, ISSN 00359254. DOI 10.2307/2346830.

[18] Kaur, M. and Kang, S. (2016) "Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining", *Procedia Computer Science*, Vol. 85, pp. 78-85. ISSN 1877-0509. DOI 10.1016/j.procs.2016.05.180.

[19] Keller, J. M, Gray, M. R. and Givens, J. A. (1985) "A Fuzzy K-Nearest Neighbor Algorithm", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-15, No. 4, pp. 580-585. ISSN 21682216. DOI 10.1109/TSMC.1985.6313426.

[20] Klösgen, W. and Zytkow, J. M. (2002) "The Knowledge Discovery Process", In *Handbook of Data Mining and Knowledge Discovery*, 10-21 p. ISBN 978-0-387-09823-4.

[21] Komprdová, K. (2012) "Rozhodovací stromy a lesy", Akademické nakladatelství CERM, 98 p., ISBN 978-80-7204-785-7.

[22] Kumar, M. S. and Balakrishnan, K. (2019) "Development of a Model Recommender System for Agriculture Using Apriori Algorithm", In *Cognitive Informatics and Soft Computing*, pp. 153-163, Springer, Singapore. ISBN 978-981-15-1451-7.

[23] Kunjachan, H., Hareesh, M. J. and Sreedevi, K. M. (2018) "Recommendation Using Frequent Itemset Mining in Big Data", In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 561–556. ISBN 9781538628430. DOI 10.1109/ICCONS.2018.8662905.

[24] Lepping, J. (2018) "*Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*". John Wiley and Sons Inc. E-ISSN 1942-4795.

[25] Luo, D., Xiao, Ch., Zheng, G., Sun, S., Wang, M., He, X. and Lu, A. (2013) "Searching Association Rules of Traditional Chinese Medicine on Ligusticum Wallichii by Text Mining", In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, IEEE., pp. 162-167. ISBN 978-1-4799-1309-1. DOI 10.1109/BIBM.2013.6732664.

[26] Ma, Haiying, and Dong Gang. (2011) "Customer Segmentation for B2C E-Commerce Websites Based on the Generalized Association Rules and Decision Tree", In 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), pp. 4600-4603, Piscataway, NJ : IEEE. ISBN 9781457705359. DOI 10.1109/AIMSEC.2011.6010255.

[27] MacQueen, J. (1967) “Some Methods for Classification and Analysis of Multivariate Observations”, In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Oakland, CA, USA. Vol. 1, pp. 281-297.

[28] Mannila, H. (1997) “Methods and Problems in Data Mining”, In *International Conference on Database Theory*, pp. 41-55. ISBN 3540622225. DOI 10.1007/3-540-62222-5_35.

[29] Murrell, P. (2018) "*R Graphics*", CRC Press.

[30] Parsad, Ch., Vijay, T. S. and Prashar, S. (2018) “Predicting Online Buying Behaviour-a Comparative Study Using Three Classifying Methods”, *International Journal of Business Innovation and Research*, Vol. 15, No. 1, pp. 62-78. E-ISSN 1751-0260, ISSN 1751-0252. DOI 10.1504/IJBIR.2018.10009022.

[31] Quinlan, J. R. (1986) “Induction of Decision Trees”, *Machine Learning*, Vol. 1, No. 1, pp. 81-106. E-ISSN 1573-0565, E-ISSN 0885-6125. DOI 10.1007/BF00116251.

[32] Rakesh, A. and Srikant, R. (1994) “Fast Algorithms for Mining Association Rules”, In *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487-99.

[33] Rakesh, A., Imieliński, T. and Swami, A. (1993) “Mining Association Rules between Sets of Items in Large Databases”, In *Acm Sigmod Record*, Vol. 22, pp. 207-16. DOI 10.1145/170036.170072.

[34] Sahil, R. and Singh, D. (2016) “Impact of Demographic Factors on Online Purchase Frequency - A Decision Tree Approach”, In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. ISBN 9789380544205.