# Data Mining Model for Predicting Customer Purchase Behavior in e-Commerce Context

Orieb Abu Alghanam, Sumaya N. Al-Khatib, Mohammad O. Hiari
Al-Ahliyya Amman University
Amman, Jordan

***Abstract*—Nowadays e-commerce environment plays an important role to exchange commodity knowledge between consumers commonly with others. Accurately predicting customer purchase patterns in the e-commerce market is one of the critical applications of data mining. In order to achieve high profit in e-commerce, the relationship between customer and merchandise are very important. Moreover, many e-commerce websites increase rapidly and instantly and competition has become just a mouse-click away. That is why the importance of staying in the business, and improving the profit needs to accurately predict purchase behavior and target their customers with personalized services according to their preferences. In this paper, a data mining model has been proposed to enhance the accuracy of predicting and to find association rules for frequent item sets. Also, K-means clustering algorithm has been used to reduce the size of the dataset in order to enhance the runtime for the proposed model. The proposed model has used four different classifiers which are C4.5, J48, CS-MC4, and MLR. Also, Apriori algorithm to provide recommendations for items based on previous purchases. The proposed model has been tested on Northwind trader's dataset and the results archives accuracy equal 95.2% when the number of clusters were 8.**

***Keywords*—Apriori PT algorithm; C4.5; CS-MC4; Data mining; decision tree; e-commerce; K-means**

## I. INTRODUCTION

The technique of examining data from a different category is known as data mining [1]. This data contains important information, also in data mining additional knowledge will be extracted. Also, it is a helpful strategy for extracting and detecting patterns in huge data sets that incorporate methods from machine learning, statistics, and database system [2, 3].

Nowadays corporate organization is attempting to adopt a digital marketing strategy and competitive markets in order to gain worldwide commercial benefits. on the other hand, to get such competitive advantages, e-commerce businesses must first comprehend their customers' sentiments, thoughts, and seasons in relation to their products and services [4].

Competitive economy and customer repurchasing behavior are critical to a company's existence. Deeper marketing tactics and managerial decisions can be made with a better grasp of customers and their preferences [5]. A typical online retail store has thousands of transactions in its database, and it serves hundreds, if not thousands, of customers per day. Manipulation and processing of this data in various ways to provide a model with increased prediction accuracy allow for the extraction of novel knowledge that aids in one-to-one marketing, personalization, increased sales, and customer retention [6]. Network marketing has become a significant marketing technique, and as internet technology has advanced, many companies have built online stores to give customers purchasing materials. Because of the numerous benefits of e-commerce, the number of people who engage in online trade, as well as the volume of transactions, has significantly expanded [7].

The difficulty in data mining applications is identifying valid, relevant, and intelligible information from raw and sparse data by mining frequent patterns for knowledge discovery [8]. One of the most important applications of data mining in the e-commerce sector is accurately anticipating client purchase habits because the number of e-commerce websites (both customer and merchandise) grows swiftly and instantly, and competition is only a mouse click away. To stay in business, providers must be able to reliably forecast customer buying behavior and target them with customized services based on their preferences.

Machine learning (ML) techniques are one of the most techniques that are used as data mining techniques. Also, using ML to develop the learner model based on previous experiences and get new knowledge when the size of data becomes huge. ML has been used in many fields such as security [9, 10] medical field, e-commerce field and others.

e-Commerce data is referred to as "Big Data" therefore dealing with this data to extract the knowledge is considered a challenge [11]. In addition to size, using analytical approaches and solutions to extract patterns in hidden relationships in order to make better decisions and get new knowledge makes it more complicated. Furthermore, choosing suitable algorithm to get the best pattern and extract the knowledge to improve the performance is also not easy.

The data mining applications have problems in the mining of recurrent patterns seeking knowledge discovery in order to identify valid, useful and understandable information out of raw and sparse data.

Applying a data mining model to enhance the accuracy of prediction in the context of e-commerce and dealing with big data to extract the knowledge at a reasonable time is an important task. Also, presenting suggestions for associated item set using a prior PT algorithm to help the customer is a desirable task.

The major contributions in this paper are as follows:

*1)* Applying data mining algorithms in such a way to provide a model that predicts customers' next purchase and recommends it to them.

*2)* Providing a comparison between different decision tree classification algorithms to choose the best classification algorithm for a product recommendation system based on a set of considered parameters.

*3)* Clustering the data to enhance the runtime and using Apriori PT association rule algorithm to extract the set of items.

The rest of the paper is organized as follows: Section 2 presents related works and Section 3 presents data collection while Section 4 gives the suggested system framework and major contribution. Moreover, the experiment results are shown in Section 5, and the conclusion is presented in Section 6.

## II. Related Work

The rise in popularity of social media has ushered in a new era for e-commerce, transforming online shopping. Several studies have been proposed to enhance the performance for the prediction in e-commerce [12]. Also, some of them used to predict the customer opinion based on the comments [4]. This section presents different classification algorithm that has been used in the literature for data miming or for classifying.

The related data mining algorithm has been presented for e-commerce but from different perspectives. On the other hand, the proposed approaches have differed from the contribution of this paper such as the objectives and the datasets and the way that the proposed model is designed. This section presents the data mining algorithms and how it has been used in the context of e-commerce.

### A. K-means

The k-means algorithm is a data mining technique that splits entities into K groups based on attributes or features, where K is a positive integer number [13]. In order to group data, the sum of squares of distances between data and the respective cluster centroid is minimized. K-mean clustering is used to organize data into categories. Fig. 1 shows the K-means algorithm when the number of clusters has been selected to be 5.

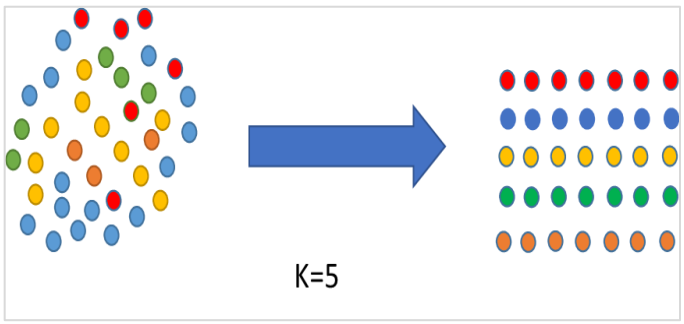


Fig. 1. Dataset Attributes and Types.

Anitha and Patil applied the Recency, Frequency and Monetary model (RFM) and deploy the principles of dataset segmentation using the K-Means Algorithm. This model objective is to employ business intelligence (BI) in recognizing potential customers by providing timely data that is relevant to the retail industry's business units. The used data was based on systematic research and scientific applications in the analysis of sales history and consumer purchasing behavior. The data, carefully selected and organized as a result of this scientific research, not only increase business sales and profits but also provides intelligent insights for predicting consumer purchasing behavior and related patterns. They also used the KMeans clustering algorithm Silhouette Analysis to evaluate the clusters by varying the number of clusters. Based on the Silhouette Score, they concluded that they could analyze the up-to-dateness of the sale, the frequency of the sale, and the money of the sale and find the best solution. [14].

Mulyawan and et al analyzed the behavior of customer shopping by made web shopping. Analyzing and comprehending clients purchasing patterns might assist web shopping in determining what they are seeking. Based on a "transaction" data set, they used Frequency and monetary (FM) analysis. They then divided the clients into groups based on how frequently they bought, how much they bought, and how much the acquired item was worth. They also clustered customers based on their transactions using the K-means method. The study indicates that the K-means algorithm suggestion products were successfully achieved and shown on the customer page [15].

### B. Classification by Decision Tree Induction

Databases have plenty of interesting hidden information that can be intelligently used for decision-making. Decision trees are known as classification trees and they are used in machine learning due to their ability to handle both discrete and continuous data in big databases and their easy implementation.

Redouan ABAKOUY and et al employed a learning model for predicting the "clicks" and "conversions" of targeted marketing emails. They compared algorithms of regression and classification for predicting whether an email sent will be opened, clicked, or converted by the intended recipient or not. The features gathered from the emails and client profiles were used to create the model. They compared categorization approaches for predicting whether an email sent to a possible recipient will be opened or not. They are the SVMs classifier and the C4.5 Decision Tree classifier. In all the cases, the Decision Tree classifier results outperform the SVM. [16].

For e-commerce logistics businesses to manage enormous client bases and develop long-term and profitable connections, Luk and et al presented an intelligent customer identification model (ICIM). This ICIM comprises a historical view and analysis of all existing or potential consumers. That model aided in the accurate identification of actual consumer needs, as well as the classification of new clients in the future in the shortest period possible. The ICIM combines the k-means clustering technique and the C4.5 classification algorithm to extract important hidden knowledge from both continuous and discrete variables [17].

*1) C4.5 Decision tree:* C4.5 is an improved version of the greedy, top-down, recursive, divide-and-conquer ID3 algorithm; the improvement in the algorithm included its ability to handle continuous variables, prune the tree after being created and its ability to deal with missing values. C4.5 rules are then constructed by greedily prune conditions from each rule if this decreases its estimated error.

*2) Improved J48 decision tree classification algorithm:* The J48 algorithm is a well-known machine learning algorithm that is based on the J.R. Quilan C4.5 algorithm [18]. In this paper, the algorithm is evaluated against C4.5 for verification purposes. With this technique, a tree is built to model the categorization process using this technique. Once the tree has been constructed, it is applied to each tuple in the database, yielding categorization for that tuple.

*3) CS-MC4 Decision tree algorithm:* The main goal of decision tree induction algorithms is to increase accuracy while minimizing costs. The m-estimate smoothed probability estimation process, which is a generalization of the Laplace estimate [19], is used in the cost sensitive decision tree algorithm. This approach decreases the expected loss by detecting the best prediction within leaves using a misclassification cost matrix.

Table I presents a comparison between different approaches that have been proposed for e-commerce. The comparisons have been done in terms of the algorithm that is used, the datasets and the experiment results. On the other hand, in this paper different data mining model has been proposed that aims to enhance the prediction in addition to apply the prior algorithm to generate a rule for association items.

TABLE I. COMPARISON BETWEEN DIFFERENT DATA MINING APPROACHES

| Reference | Dataset | Proposed approach | Results |
|---|---|---|---|
| [20] | Amazon DVD musical product:- .Net Crawler, 9555 reviews | Hybrid approaches | Precision: - 0.89, Recall: - 0.84, F-Measure: - 0.86 |
| [21] | Review of cellphone & accessories:- 21600 reviews [22] | Linear support vector machine | Accuracy: - 93.52% |
| [11] | UCI Machine Learning Repository. | Decision Tree | Accuracy: - 95% |
| [23] | Data in [24] | C4.5 | Accuracy: - 86.59% |
| | | Random forest | Accuracy: - 86.78 |

### C. Apriori PT Association Rule Algorithm

Yuanzhu and et al present a study that implements the Apriori algorithm and C5.0 which are considered as association rules; also, decision tree techniques for data mining [25]. It has been used to help managers or decision maker people to extract knowledge ‘from’ and 'about' customers in order to determine their preferences, allowing enterprises to develop the correct goods and achieve a competitive advantage.

The findings show that the knowledge-based approach is effective, and the returned knowledge is represented as a set of rules that can be used to identify relevant patterns for both new product development and marketing tactics.

The original Apriori algorithm was proposed by Agarwal and Srikant in 1994 [26]. Apriori is constructed to operate on transactional databases; the algorithm determines item sets in the database that are subsets of at least one transaction. Apriori PT is an enhancement of this algorithm that works well with big data by:

Step 1 - find all frequent elements that have support more than the minimum support needed. Step 2 - the set of frequent elements to build association rules with a high enough level of confidence.

Pruning -using the fact that any subset of a frequent item set should be frequent.

## III. DATA COLLECTION

The real lifetime large transactional data set from Kaggle to test the purposed model. The northwind.mdb sample transaction database has been used to test the robustness of the proposed model. The dataset passed through preprocessing phases to meet the requirements for each selected algorithm such as data type conversion from continuous to discrete was also handled for prediction purposes.

Northwind originally consisted of many tables with relations between them, each table consists of many details. In this paper, processing and understanding the dataset have been done. Furthermore, 2155 product sales on 8 product categories of 77 different item types for 91 different customers have been taken into consideration. Also, the demographic variable such models, gender and customer job have been taken. Table II presents the dataset details while Fig. 2 contains further dataset details.

TABLE II. CHARACTERISTICS NORTHWIND DATABASES

| Attribute | Category | Information |
|---|---|---|
| Customer ID | Discrete | 89 values |
| Gender | Discrete | 2 values |
| Customer Job | Discrete | 12 values |
| Order ID | Continue | - |
| Category Name | Discrete | 8 values |
| Product Name | Discrete | 77 values |
| Unit price | Continue | - |
| Quantity | Continue | - |
| Discount | Continue | - |
| Extended Price | Continue | - |

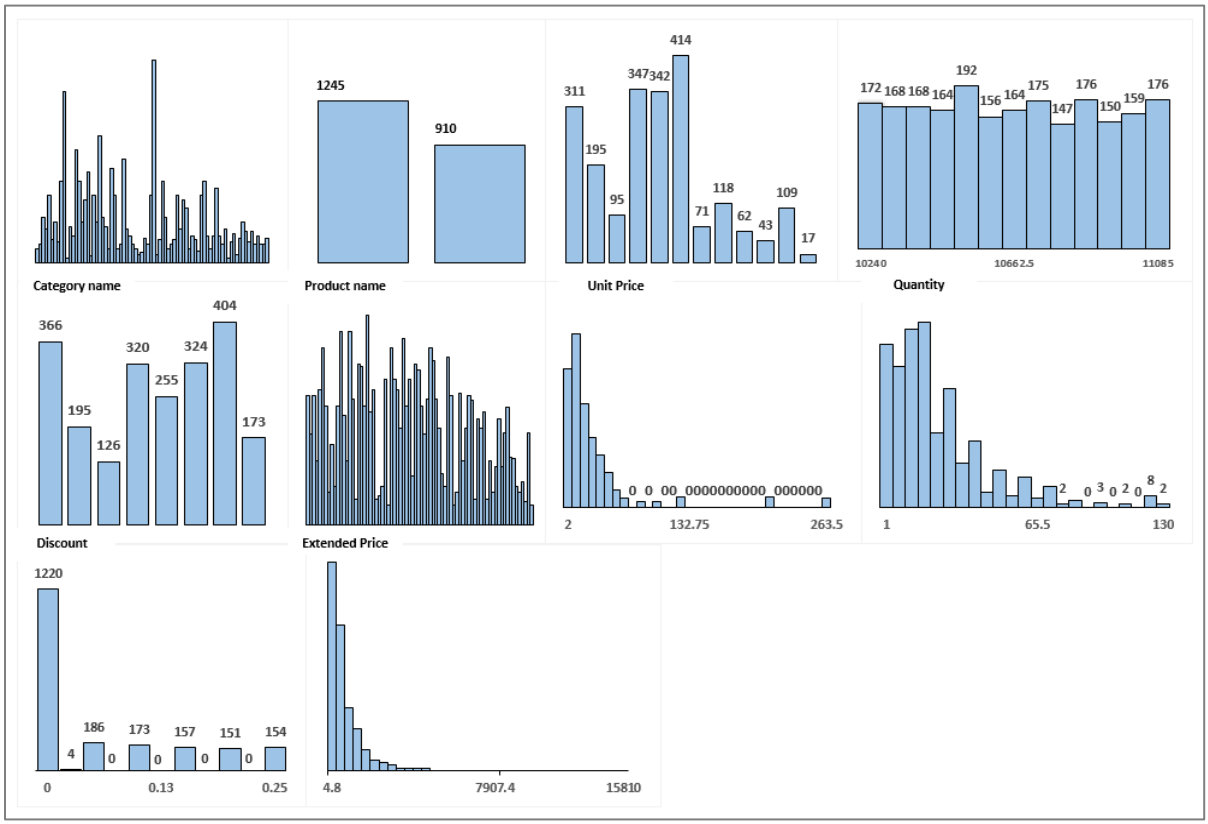


Fig. 2. Dataset Attributes and Types.

## IV. System Framework

Fig. 3 presents the proposed model which went through a group of phases. In the first phase a normalization and duplicate removal have been done, then k-means clustering was performed in the second phase. In the next phase, the data was split into training and testing. Moreover, the modeling stage has been built based on four algorithms which are C4.5, J48, CS-MC4, and MLR. Also, a prior PT algorithm has been applied to get the association rules.

The proposed system firstly starts by deciding how many K clusters need to split the dataset. The centroid or center of these clusters has been randomly chosen to start the calculation for the whole data. Moreover, to divide the dataset into clusters this will be done based on the distance between each object and the centroid to categorize the objects based on the minimum distance (closest centroid). Table III presents the characteristics of the K-means clustering algorithm and the number of clusters that are generated in the proposed model.

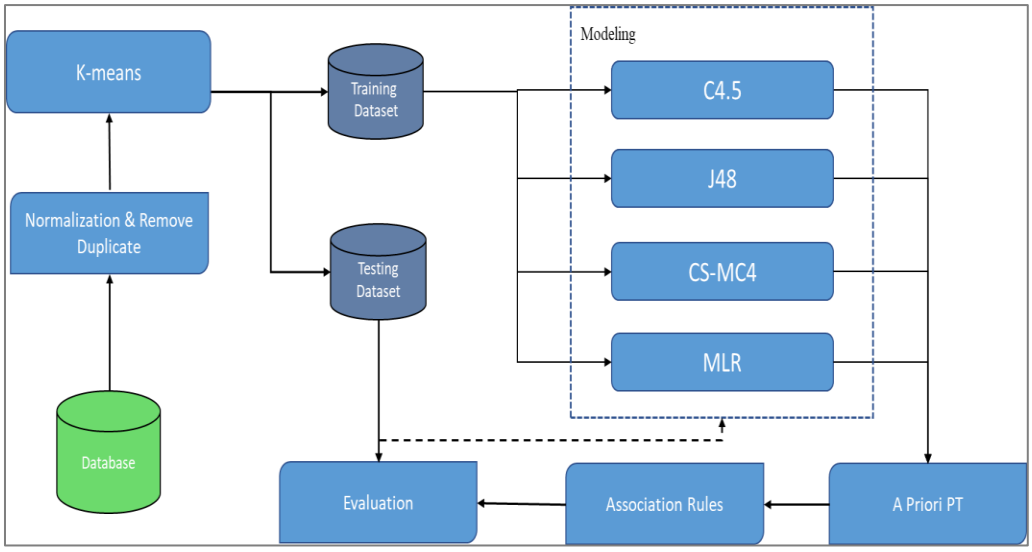


Fig. 3. The Proposed Data Mining Model.

TABLE III. K-MEANS CLUSTERING PARAMETERS

| Name | Description | Value |
|---|---|---|
| **Distance function** | The distance function to use for instances comparison. | Fixed - Euclidean distance |
| **Number of clusters** | The number of clusters to be created. | Dynamic – starting with 2 clusters and finished at 12 K= 2, 8, 10,12 |
| **Max Iterations** | Set the maximum number of iterations. | Fixed - 500 |
| **Seed** | The random number seed to be used. | Fixed - 10 |

In this paper, the dataset has been clustered into different clusters which are 2,8,10,12. Furthermore, the maximum iteration that is used is 500 while the fixed Euclidean distance is used for instances comparison between the records. Also, the random number called seed that is used is 10.

The Euclidean distance or the Manhattan distance is used to cluster data when utilizing the K means technique as shown in equation 1. If the Manhattan distance is employed, the component-wise median rather than the mean is used to calculate the centroids [27].

$$d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (1)$$

Classification's fundamental goal is to accurately anticipate the target class for each record. Its training procedure seeks to uncover correlations between predictor and target variables.

Classification algorithms [28, 29] differ in the strategies employed to identify these associations, which are further summarized in a model then applied to a record (test data) where the class label is unknown.

The modeling stage is built based on four algorithms which are C4.5, J48, CS-MC4, and MLR. Each algorithm has been applied on the whole clustered dataset, then substituting error rates for each. Moreover, unbiased error rate estimation ‘10 folds cross-validation was used to evaluate each learning algorithm. Table IV presents the parameters that have been used by C4.5 algorithm and the splitting ratio for the dataset.

In the proposed model the weighted total of the error estimates for all of the subtree's leaves has been used to get the error estimate. The upper bound of the error estimate for a node is derived as shown in equation 2, where f represents the error on the training data and N is the number of instances covered by the leaf.

$$e = \left(f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right) / \left(1 + \frac{z^2}{N}\right) \quad (2)$$

TABLE IV. C4.5 DECISION TREE PARAMETERS

| | |
|---|---|
| **Minimum Size of Leaves** | 5 |
| **Confidence level - lower values incur heavier pruning** | 25% |
| **Cross Validation** | 10 folds |
| **Train-Test** | 80% (train) 20% (test) |

The parameters for J48 algorithm that have been used in the proposed model are presented in Table V. more details about these parameters in [30]. The default Number of Folds has been used which is 3 and the seed was 1.

TABLE V. J48 TREE PARAMETERS

| | |
|---|---|
| **Collapse Tree** | yes |
| **Confidence Factor** | 25% |
| **Min. No. of Objects** | 2 |
| **Number of Folds** | 3 |
| **Seed** | 1 |
| **Use MDL Correction** | True |
| **Unpruned** | False |
| **Cross Validation** | 10 folds |
| **Subtree Raising** | True |

Table VI shows the parameters applied to implement CSMT4 algorithm.

TABLE VI. CS-MC4 CLASSIFICATION TREE SUPERVISED PARAMETRS

| | |
|---|---|
| **Minimum Size of Leaves** | 5 |
| **Lambda** | 3 |
| **Cross Validation** | 10 folds |

In the next step, Apriori PT Christian Borgelt’s as shown in Fig. 4 was applied, which is a highly effective association rule generator, it can handle large datasets quickly. Further processing has been done before using Apriori algorithm. The processed data consisted of 830 transactions and 77 attributes. The "item types” was set to ‘yes' for each item purchased by each transaction and 'no' otherwise. Also, the main support was set at 0.1 (10%), the min confidence min as 85%, the max cardinal of the item was set as 4 (Max Card Item sets).

**$I_k$: frequent item set of size K**

**Li= {frequent item}**

**For (K=1; $I_k$ !=0 ; K++) do begin**

**$C_k$ +1= candidates generated**

**From $I_k$;**

**do**

**Increment the count of all**

**Candidates in $C_k$+1**

**$I_k$ +1= Candidates in $C_k$ +1 with**

**Min_support**

**End**

**Return: K, $I_k$**

Fig. 4. Input Data for Apriori PT Algorithm and Algorithm Pseudocode.

The association rules were generated based on the proposed model. Here is a sample set of generating rules which are related to the attribute number. If attribute 39 and attribute 77 are combined, then the attribute for 46 should represent the product reality-lifetime and unique ids this means if a customer purchased item number 39 and item number 77 then item 46 most probably will be bought. Thus, the proposed model recommends this item to that customer. Finally, to test the usability of the proposed model we applied the model to the real-life time usability of big Dataset and the model showed high robustness.

One of the strong features of C4.5 algorithm is its ability to handle discrete and continuous data types, this feature was used and the algorithm was implemented on clustered data, with 96.9 % accuracy.

## V. Experiments and Results

This section presents the experimental results for the proposed model. Moreover, each of the above-mentioned decision trees is executed for each learning method on the whole dataset. After that, we substituted error rates for each, then we use unbiased error rate estimation '10 folds cross-validation' to evaluate each learning algorithm.

### A. The Performance Measurement

In this paper, the accuracy and the error rate have been used to measure the performance of the proposed model and measure the performance of each classifier.

*1) Accuracy:* It is referring to the proportion of valid predictions (including true positives and true negatives) among the total number of cases analyzed is the accuracy [31]. Classified by the classifier as shown in equation 3:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

*2) Error-Rate:* It's also known as the Misclassification rate, and it's calculated as 1-Acc (M), where Acc (M) represents M's accuracy, as given in equation. 4:

$$Error\ Rate = 1 - \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Table VII shows the experiment results for all classifiers in terms of the accuracy and error rate for each classifier on different clusters size. The experiment results have been done on the dataset based on clustering it into a different cluster. K-means algorithm has been used to reduce the size of the dataset and enhance the speed and performance of the classifiers.it can be noticed that when the size of the cluster was 10 the best accuracy has been reached compared with other clusters size.

Moreover, C4.5 outperforms the other classifiers in terms of accuracy in all clusters size. On the other hand, CS-MC4 has reached the lowest accuracy compared with the other classifiers in all clusters size.

The results turn out that using 10 clusters gave petter results for the C4.5 algorithm that is reach 96.9% accuracy while when 8 clusters have been used MLR reach 89%. C4.5 can handle discrete and continuous data types; it is used this strong point feature and implemented the algorithm on clustered data.

When j48 induction tree algorithm has been applied to '10' k-means clustering dataset the prediction accuracy was 93.8%. The testing has been performed in the 10-fold cross validation. After that, the results are then used to generate decision rules.

TABLE VII. The Comparisons between Different Algorithms for Different Clusters Numbers

| **K-means** | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **# of clusters** | **K=2** | | | | **K=8** | | | | **K=10** | | | | **K=12** | | | |
| **Algorithm** | **C4.5** | **J48** | **CS-MC4** | **MLR** | **C4.5** | **J48** | **CS-MC4** | **MLR** | **C4.5** | **J48** | **CS-MC4** | **MLR** | **C4.5** | **J48** | **CS-MC4** | **MLR** |
| **Accuracy%** | 86.5 | 82.4 | 62.7 | 70.6 | 95.2 | 92.2 | 81.7 | 89.3 | 69.9 | 93.8 | 83.4 | 91 | 93.1 | 90 | 79 | 87.1 |
| **Error Rate %** | 13.5 | 17.6 | 37.3 | 29.4 | 5 | 7.8 | 18.3 | 10.7 | 3.1 | 6.2 | 16.6 | 9 | 6.9 | 10 | 21 | 12.9 |

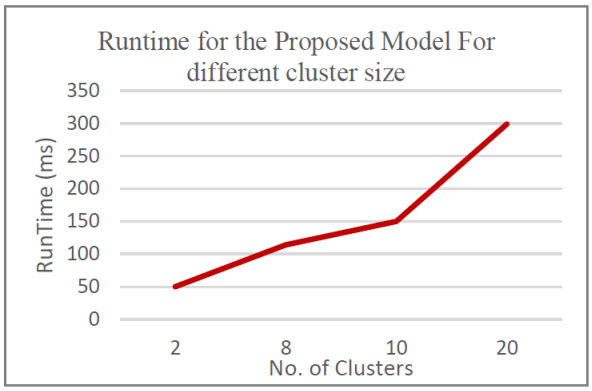


Fig. 5. The Runtime for the Proposed Model with different Clusters.

Fig. 5 shows the results for the runtime that is needed when the different size of clusters has been used. It can be noticed that when the number of clusters increases the runtime is increasing. It can be noticed that when the data has clusters into any two clusters the runtime that has been taken was 50msWhile when the number of clusters becomes 12 the runtime reached approximately 7 times greater than when the data was 2 clusters.

Table VIII illustrated the size of the decision tree in terms of the number of leaves and number of nodes that are generated by C4.5 algorithm, J48 algorithm and CS-MC4 algorithm for the dataset in terms of the number of nodes and the number of leaves.

TABLE VIII. NUMBER OF NODES AND LEAVES FOR EACH CLASSIFIER

| Algorithm | C4.5 | J48 | CS-MC4 |
|---|---|---|---|
| **No. of nodes** | 35 | 57 | 83 |
| **No. of Leaves** | 18 | 23 | 45 |

Fig. 6, Fig. 7, Fig. 8 and Fig. 9 represent the accuracy for different algorithms based on different cluster sizes. The clusters that have been selected were two, eight, ten and twelve respectively. It can be noticed that when the cluster size was 10 all classifiers reach a better accuracy compared with other cluster sizes. Also, the results for C4.5 algorithm outperforms the other algorithms in all cluster size.

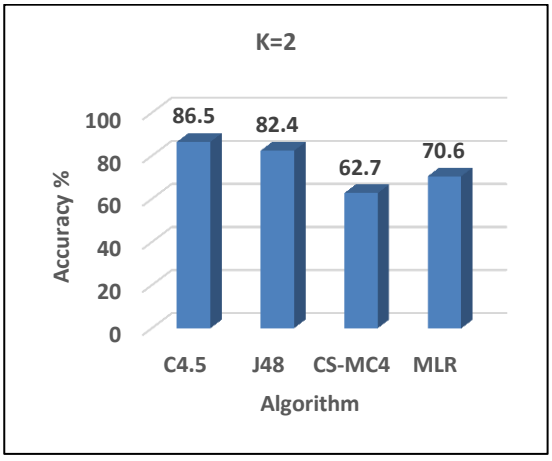


Fig. 6. The Accuracy for different Classifiers when k=2.

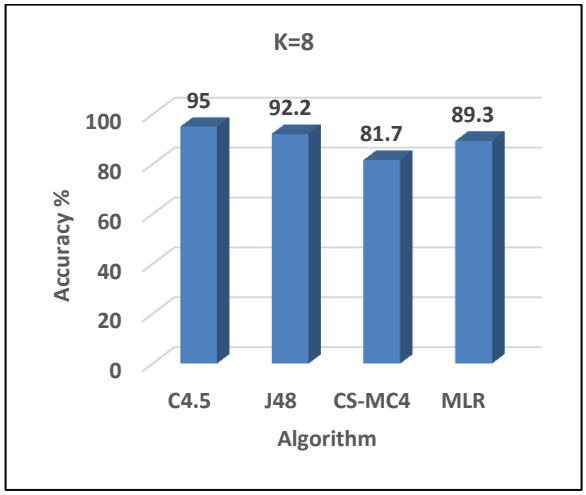


Fig. 7. The Accuracy for different Classifiers when k=8.

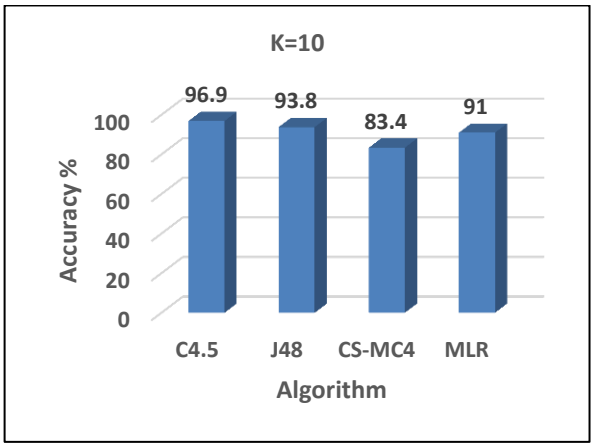


Fig. 8. The Accuracy for different Classifiers when k=10.

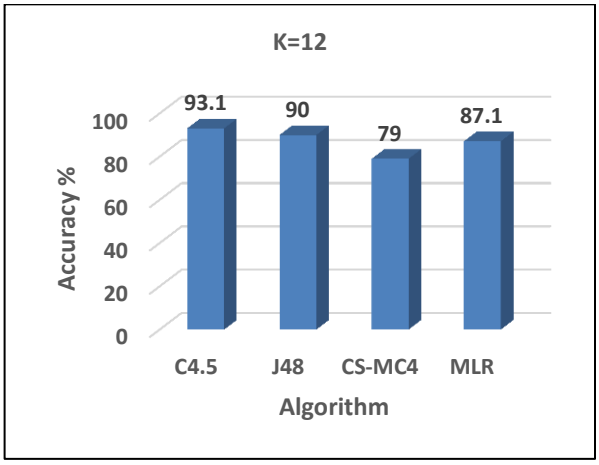


Fig. 9. The Accuracy for different Classifiers when k=12.

## VI. CONCLUSION

In conclusion, this paper achieved higher accuracy for predicting in the proposed data mining model. Also, it gives an indicator for the suitable size of the clusters that should be selected for northwind.mdb. Moreover, the proposed model suggests the most association items that are related to each

other. It amid to understand the purchase behavior to predict customer next purchase based on a set of selected parameters when Apriori PT algorithm has been used. On the other hand, the proposed model aimed to enhance the prediction for a huge database. The experimental results show that J48 and C4.5 algorithms produce high accuracy measurements compared with other algorithms.

In this paper, Apriori PT is applied for a fast and powerful association rules generation in e-commerce customer purchasing field. Moreover, data clustering has provided a good performance, such as the run time of the proposed model or the accuracy. Clustering the dataset does not affect the value of the data. Finally, the proposed model achieved 95.2% accuracy when the number of clusters was assigned to be 8 for C4.5 algorithm. On the other hand, the CS-MC4 algorithm achieved the lowest accuracy when the number of clusters was 2 it reached 62.7%.

## REFERENCES

[1] S. F. Abdullah, A. F. N. A. Rahman, Z. A. Abas, and W. H. M. Saad, "Fingerprint gender classification using univariate decision tree (j48)," Network (MLPNN), vol. 96, no. 95.27, pp. 95-95, 2016.

[2] T. Reutterer, M. Thomas, and N. Schröder, "Leveraging purchase regularity for predicting customer behavior the easy way," International Journal of Research in Marketing, vol. 38, no. 1, pp. 194-215, 2021.

[3] R. Heldt, C. S. Schmitt, and F. B. Luce, "Predicting customer value per product: From RFM to RFM/P," Journal of Business Research, vol. 127, pp. 444-453, 2021.

[4] A. Moazzam, Y. Farwa, H. Mushtaq, A. Sarwar, A. Idrees, S. Tabassum, BaburHayyat, and K. Ur Rehman, "Customer Opinion Mining by Comments Classification using Machine Learning," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 5, pp. 385-393, 2021.

[5] K. Maheswari, and P. P. A. Priya, "Predicting customer behavior in online shopping using SVM classifier," in 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), pp. 1-5, IEEE, 2017.

[6] X. Dou, "Online purchase behavior prediction and analysis using ensemble learning," in 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), pp. 532-536, IEEE, 2020.

[7] R. Heldt, C. S. Silveira, and F. B. Luce, "Predicting customer value per product: From RFM to RFM/P," Journal of Business Research, vol. 127, pp. 444-453, 2021.

[8] A. Dogan, and D. Birant, "Machine learning and data mining in manufacturing," Expert Systems with Applications, vol. 166, pp. 114060, 2021.

[9] O. AbuAlghanam, L. Albdour, and O. Adwan, "Multimodal Biometric Fusion Online Handwritten Signature Verification using Neural Network and Support Vector Machine," International Journal of Innovative Computing, Information and Control, vol. 7, no. 5, pp. 1691-1703, 2021.

[10] S. N. Mohanty, E. L. Lydia, M. Elhoseny, M. M. G. Al Otaibi, and K. Shankar, "Deep learning with LSTM based distributed data mining model for energy efficient wireless sensor networks," Physical Communication, vol. 40, pp. 101097, 2020.

[11] E. F. Zineb, N. RAFALIA, and J. ABOUCHABAKA, "An Intelligent Approach for Data Analysis and Decision Making in Big Data: A Case Study on E-commerce Industry," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 7, pp. 723-736, 2021.

[12] Y. Fu, M. Yang, and D. Han, "Interactive Marketing E-Commerce Recommendation System Driven by Big Data Technology," Scientific Programming, vol. 2021, 2021.

[13] T. Mitchell, Machine Learning. McGraw hill Burr Ridge, 1997.

[14] P. Anitha, and M. M. Patil, "RFM model for customer purchase behavior using K-Means algorith," Journal of Kings Saud University-Computer and Information Sciences, 2019.

[15] B. Mulyawan, M. V. Christanti, and R. Wenas, "Recommendation Product Based on Customer Categorization with K-Means Clustering Method," IOP Conference Series: Materials Science and Engineering, vol. 508, no. 1, pp. 012123, 2019.

[16] R. Abakouy, E. M. En-naimi, A. E. Haddadi, and E. Lotfi, "Data-driven marketing: how machine learning will improve decision-making for marketers," in proceedings of the 4th international conference on Smart City Applications, pp. 1-5, 2019.

[17] C. C. Luk, K. L. Choy, and H. Y. Lam, "Design of an intelligent customer identification model in e-Commerce logistics industry," MATEC Web of Conferencs, vol. 255, pp. 04003, 2019.

[18] Y. Zhan, K. H. Tan, and B. Huo, "Bridging customer knowledge to innovative product development: a data mining approach," International Journal of Production Research, vol. 57, no. 20, pp. 6335-6350, 2019.

[19] S. M. Karst, R. M. Ziels, R. H. Kirkegaard, E. A. Sørensen, D. McDonald, Q. Zhu, R. Knight, and M. Albertsen, "High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing," Nature methods, vol. 18, no. 2, pp. 165-169, 2021.

[20] U. A. Chauhan, M. T. Afzal, A. Shahid, M. Abdar, M. E. Basiri, and X. Zhou, "A comprehensive analysis of adverb types for mining user sentiments on amazon product reviews," World Wide Web, vol. 23, no. 3, pp. 1811–1829, 2020.

[21] T. U Haque, N. N. Saber, and F. M. Shah, "Sentiment Analysis on Large Scale Amazon Product Reviews," in 2018 IEEE international conference on innovative research and development (ICIRD), pp. 1-6, IEEE, 2018.

[22] R. He, and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one- class collaborative filtering," in proceedings of the 25th international conference on world wide web, pp. 507-517, 2016.

[23] K. Baati, and M. Mohsil, "Real-time prediction of online shoppers' purchasing intention using random forest," in IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 43-51, Springer, Cham, 2020.

[24] C. O. Sakar, S. O. Polat, M. Katircioglu, Y. Kastro, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks," Neural Computing and Applications, vol. 31, no. 110, pp. 6893-6908, 2019.

[25] D. Cirqueira, M. Hofer, D. Nedbal, M. Helfert, and M. Bezbradica, "Customer purchase behavior prediction in e-commerce: A conceptual framework and research agenda," in International Workshop on New Frontiers in Mining Complex Patterns, pp. 119-136, Springer, Cham, 2019.

[26] M. H. W. Ho, and H. F. Chung, "Customer engagement, customer equity and repurchase intention in mobile apps," Journal of business research, vol. 121, pp.13-21, 2020.

[27] S. Yang, and G. M. Allenby, "Modeling interdependent consumer preferences," Journal of Marketing Research, vol. 40, no. 3, pp. 282-294, 2003.

[28] J. Qiu, Z. Lin, and Y. Li, "Predicting customer purchase behavior in the e-commerce context," Electronic commerce research, vol. 15, no. 4, pp. 427-452, 2015.

[29] S. Moon, and G. J. Russell, "Predicting product purchase from inferred customer similarity: An autologistic model approach," Management Science, vol. 54, no. 1, pp. 71-82, 2008.

[30] K. Kang, and J. Michalak, "Enhanced version of AdaBoostM1 with J48 Tree learning method," arXiv preprint arXiv:1802.03522, 2018.

[31] N. Kavha, and S. Karthikeyan, "Customer Buying Behavior Analysis: A clustered Closed Frequent Itemsets for Transactional Database," International Journal of Computer Science Issues (IJCSI), vol. 10, no. 3, pp. 113, 2013.