Adaptive kNN based Intention Prediction

Dhaarani S
Department of Computer Science and Engineering
dhaarani2816@gmail.com

Pavithra L
Department of Electronics and Communication Engineering
pavithragowni@gmail.com

Keerthana B
Department of Computer Applications
keerthikec@gmail.com

Dr. V. Latha Jothi
Department of Computer Science and Engineering
lathajothi.s@gmail.com

Abstract- E- shopping most preferably called as Online shopping has become a trend-setter in today's business sector especially during this medical pandemic (COVID- 19). Due to the fear of this life sucking disease, most of the people around the world have changed their lifestyle from physical shopping to online shopping. People find it easy to surf the products that they are looking for in the shopping site and use their Net-banking (or) Google Pay linked with their bank account to make the payment. But thinking from the vendor's point of view, the job is not so easy as we think. The vendor has to keep records and analyze the action of every customer opening the site. Based on the database that holds the activity of every user in the site, the vendor has to provide product suggestions for the customers when they visit the site next time. A classification algorithm named Adaptive kNN is implemented to provide the relevant recommendations to the customer so that they find it easy to complete their shopping as quickly as possible. The paper is reported in a comparative manner of two algorithms namely Random Forest and Adaptive kNN. Alongside, ensemble techniques were also used to evaluate the performance measure of both the algorithms. For the dataset that we applied, the proposed algorithm shines out with an accuracy of about 91.87% in the line of customer intention prediction

Keywords – E- shopping, Online Business, Product Recommendation, Accuracy, Classification, Ensemble Techniques.

I. INTRODUCTION

Having a look at the past, internet was not much popular among people. But now, in today's techno world, internet and smart phone has become a basic need in human life to survive. From going browsing centers to seeing the whole world in our single hand via smart phones, internet has given our lives a rapid change over. Especially in the concept of Online shopping, people find it comfort to purchase all the products sitting at one place than roaming the whole city to get the best products. Online shopping has technically arisen from the word Business to Consumer communication. The use of smartphones and the internet has become a mandatory need in one's life to survive. The increase in use of internet has eventually given way to many online businesses. People either go for browsers or smart phones to surf the products they need and make their payments either with debit card or Net banking services. The visit to online shopping sites has almost raised to about 22 billion in June 2020 from 16.07 billion visits in January 2020 [1]. As of 2020, online shopping sites has the biggest share of online purchases across the world and Amazon stands the leading online retail website [2].

Convenience in using the site, bulk number of products, easy exchange of items if damaged or misplaced, and the reviews given for each product that is purchased stand behind the success of online shopping [13]. How the customer feels while using the site is very much important for increasing online sales. If the customer is not comfortable in accessing the site, then it greatly affects the marketing of that particular online shopping site.

Predicting the purchase intention of the buyers is quite challenging because there will be no individual conversation between the vendor and the buyer [3]. To improve marketing and sales, understanding the customer's behavior and providing relevant suggestions is much more important. These product recommendations can be done efficiently by

examining the purchase history of each customer as well as the database containing the action of each customer while using the site.
The vendor maintains the database of every customer visiting the site. The database contains the record of the pages that the customer has just made window shopping, the pages where the purchase was made, the pages whose links are copied by the customers to share with others, and the products that the buyer adds to the shopping bag. The Purchase history database includes [12] the product's unique ID, name, quantity, reviews of that particular product, mode and time of payment, shipping charges, and the delivery date.
In the field of machine learning and data mining, purchase prediction with the help of empirical data of buyers has become an emerging area of research.

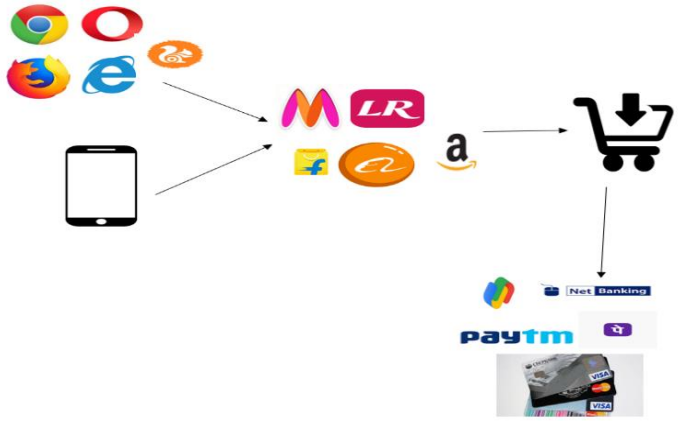

Figure 1. The flow of online shopping

In this paper, experimentation is done with the help of a classification algorithm alongside an ensemble approach to boost up prediction accuracy.
The upcoming topics of this paper is organized as Literature Survey in Section 2, Proposed methodology in Section 3, Experimentation in Section 4, Result Analysis and Conclusion in Section 5 and 6.

II. RELATED WORK

Humphrey Sheil et.al proposed a twin tasking approach of user classification and content ranking in e-commerce done using a non-restrictive dataset. Gradient Boosted Machine (GBM) is a straightforward idea to train as it ceaselessly extends an ensemble of classification and regression trees (CART) to foster judgment on unseen data [5]. New trees are continually added throughout the training to stronger the objective function and for correcting the flaws made by the initial trees. The features are calculated and are saved in LIBSVM (label: value) format which is used by GBM to mould a forest of CART trees. In all the experiments done, GBM performed compatibly well that the outcome attained depends greatly on feature engineering than algorithm improvements.

Aida Mustapha conducted experiments to help the businesses in understanding their customers much better so that marketing can be increased [6]. Naive Bayes algorithm was implied to categorize the customers based on the variables in the dataset. The case study was carried on java based open-source software named RapidMiner and the datasets are downloaded from UCI Repository. It has been encountered that during data preparation, interpretation and evaluation of the model, the analysis is quite time consuming and crucial.

Jitendra Kumar Jaiswal used Random forest algorithm to analyze the selection of feature subset [7]. The work is implemented using R language with Borusta and randomForest packages. The dataset used is taken from UCI Machine Learning Repository (Chronic_kidney_disease data). The packages used depends on wrapper algorithms and top-down technique is implemented to evaluate the feature. Though the random forest algorithm performs well on classification basis, it consumes huge time for larger datasets.

Fatemah Safara analyzed two approaches for predicting the consumer behavior namely statistical approach and machine learning approach [9]. Correlation among features is evaluated in statistical approach whereas in machine learning technique, a predictive model is constructed. The dataset used for both the approaches is taken from DigiKala online shopping site. Weka 3.8.1 software is used for machine learning way of classification and Python

Anaconda for correlation analysis. However, the prediction models are constructed without the features that gets affected from COVID- 19.

K. Maheswari discussed about SVM method of classifying the customer behavior that uses multi-dimensional hyperplane. The experiment was done using an inventory dataset with six attributes in R software with e1071 package. The customers are clustered based on their frequency in visiting the site, number of items purchased, product id and so on... After applying SVM classifier [8] to the dataset, the result is generated using the histogram graph. More in- depth analysis of the algorithm and comparison with other methods would have made this study much precise.

Michael Shekasta came up with the Purchase Intent Session- bAsed algorithm to know the buying intention of customers who has not made any purchase before. The algorithm performs well [4] on imbalanced datasets and also has better performance. A proprietary dataset containing events and item catalogue is taken into account that provides recommendations for both registered and guest users. The approach is super effective when goes in combination with classic user- item recommendation systems but less effective in case of small datasets.

II. PROPOSED ALGORITHM

The study is designed in a way to analyze and compare the prediction accuracy of two classic classification technique. First and foremost, the pre-processing of data must be done to use the dataset. Pre- processing of data is the common step that was carried out in both the algorithms. Removing unwanted and missed data, replacing data into standard format is the process of data pre-processing. The remaining steps mentioned in Fig.2 are explained in the algorithms below.

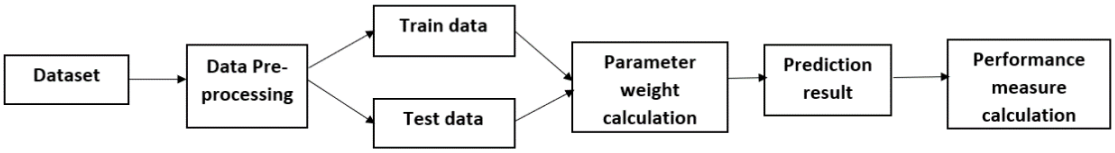

Figure 2. Study Framework

3.1 Random Forest Algorithm –

Random Forest algorithm involves a collection of trees that finally produces a correlated decision tree based on prediction. Each tree in the Random Forest is considered to be an ensemble model [10]. Every tree produces a result and the final output is brought up by the voting technique. Since all the trees join as a group and produce the final result, the efficiency of Random Forest is based on the strength of each tree and the correlation between them [11]. Number of attributes a node has and the total number of trees stands the basic parameters of this algorithm. Overall rating of Random Forest algorithm can be evaluated by considering the accuracy rate. Working of Random Forest can be split up as two phase- Creating Random Forest by combining N trees and then calculating predictions for each tree [14]. Working steps in Random Forest algorithm is as follows:

Step 1: From the training set, select the data points and build a decision tree for each data point.

Step 2: Repeat step 1 for all the data points.

Step 3: Find predictions for each tree and update the tree with new data.

3.2. Adaptive kNN Algorithm –

In oftentimes, kNN algorithm is applied for text categorization. The algorithm depends on standard vector space framework. Even distribution of training samples among different classes is the assumption of classic kNN algorithm. Text similarity is first evaluated using which the text to be tested is determined. k texts which are similar to the text that needs to be classified is identified and then compared with the training data. Though kNN algorithm

is quite simple and the implementation process is also easy, the selection of k and the similarity measure decides the success of the algorithm. An improved kNN algorithm named Adaptive kNN is proposed that has the focus on identifying the suitable k value to get a possible class label for each test sample. To increase the expected accuracy rate, Adaptive kNN dynamically choses k for each point. To tell this in theoretical manner, the value of k must be identified for each point so that expected accuracy would be exp_acc(p,k)= max{exp_acc(p,k) | k € Range}.
Initially, the values of opt_k and max_acc is initialized so that during first iteration, they can be rewritten with the corresponding values. For a point p, k value with highest expected accuracy rate for the corresponding iteration is stored in opt_k variable and the value of highest expected accuracy is stored in max_acc variable. For every 'k' value, the candidate accuracy rate is computed and compared with the highest expected accuracy rate that was obtained during previous iteration. The algorithm repeats this process till k € Range and at the final iteration, the k value will be assigned with highest expected accuracy. The final phase of the algorithm would be applying the obtained highest k value to the classic kNN algorithm and as a result, an appropriate class will be assigned.

IV. DATASET AND EXPERIMENTATION

In our experiment, purchase history of a supermarket has been taken as the dataset. The dataset was collected for nearly 3 months period that consists of nearly 1000 records of the users who has made their purchase through online and the summary of the dataset is shown in Fig 3. Each row in the data file describes the session data of one customer. The dataset involves invoice ID, Branch name, City, Type of Customer, Gender, Product category, Price per unit, Quantity, Tax of 5%, Total amount, Date and time of purchase, Payment method and Rating and Gross income. The dataset has to be pre-processed first to replace all the categorical values with numeric values. Data filtering and data tagging is the process involved in data pre-processing stage. Data filtering is done to filter out the cases that are needed to perform the calculation from the entire dataset. In data tagging, unwanted, missed and uneven data are removed and the similar data are sorted out and grouped with a name tag.

Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	Rating	gross income
750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	01-05-2019	13:08	Paytm	9.1	26.1415
226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.82	80.22	03-08-2019	10:29	Google Pay	9.6	3.82
631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	03-03-2019	13:23	PhonePe	7.4	16.2155
123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.288	489.048	1/27/2019	20:33	Debit Card	8.4	23.288
373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	02-08-2019	10:37	Debit Card	5.3	30.2085
699-14-3026	C	Naypyitaw	Normal	Male	Electronic accessories	85.39	7	29.8865	627.6165	3/25/2019	18:30	Debit Card	4.1	29.8865
355-53-5943	A	Yangon	Member	Female	Electronic accessories	68.84	6	20.652	433.692	2/25/2019	14:36	Google Pay	5.8	20.652
315-22-5665	C	Naypyitaw	Normal	Female	Home and lifestyle	73.56	10	36.78	772.38	2/24/2019	11:38	Paytm	8	36.78
665-32-9167	A	Yangon	Member	Female	Health and beauty	36.26	2	3.626	76.146	01-10-2019	17:15	Paytm	7.2	3.626
692-92-5582	B	Mandalay	Member	Female	Food and beverages	54.84	3	8.226	172.746	2/20/2019	13:27	Google Pay	5.9	8.226
351-62-0822	B	Mandalay	Member	Female	Fashion accessories	14.48	4	2.896	60.816	02-06-2019	18:07	Google Pay	4.5	2.896
529-56-3974	B	Mandalay	Member	Male	Electronic accessories	25.51	4	5.102	107.142	03-09-2019	17:03	Paytm	6.8	5.102
365-64-0515	A	Yangon	Normal	Female	Electronic accessories	46.95	5	11.7375	246.4875	02-12-2019	10:25	Debit Card	7.1	11.7375
252-56-2699	A	Yangon	Normal	Male	Food and beverages	43.19	10	21.595	453.495	02-07-2019	16:48	Debit Card	8.2	21.595
829-34-3910	A	Yangon	Normal	Female	Health and beauty	71.38	10	35.69	749.49	3/29/2019	19:21	Debit Card	5.7	35.69
299-46-1805	B	Mandalay	Member	Female	Sports and travel	93.72	6	28.116	590.436	1/15/2019	16:19	Google Pay	4.5	28.116
656-95-9349	A	Yangon	Member	Female	Health and beauty	68.93	7	24.1255	506.6355	03-11-2019	11:03	PhonePe	4.6	24.1255
765-26-6951	A	Yangon	Normal	Male	Sports and travel	72.61	6	21.783	457.443	01-01-2019	10:39	PhonePe	6.9	21.783
329-62-1586	A	Yangon	Normal	Male	Food and beverages	54.67	3	8.2005	172.2105	1/21/2019	18:00	Google Pay	8.6	8.2005
319-50-3348	B	Mandalay	Normal	Female	Home and lifestyle	40.3	2	4.03	84.63	03-11-2019	15:30	Google Pay	4.4	4.03
300-71-4605	C	Naypyitaw	Member	Male	Electronic accessories	86.04	5	21.51	451.71	2/25/2019	11:24	Google Pay	4.8	21.51
371-85-5789	B	Mandalay	Normal	Male	Health and beauty	87.98	3	13.197	277.137	03-05-2019	10:40	Google Pay	5.1	13.197
273-16-6619	B	Mandalay	Normal	Male	Home and lifestyle	33.2	2	3.32	69.72	3/15/2019	12:20	Debit Card	4.4	3.32
636-48-8204	A	Yangon	Normal	Male	Electronic accessories	34.56	5	8.64	181.44	2/17/2019	11:15	Google Pay	9.9	8.64
549-59-1358	A	Yangon	Member	Male	Sports and travel	88.63	3	13.2945	279.1845	03-02-2019	17:36	Google Pay	6	13.2945
227-03-5010	A	Yangon	Member	Female	Home and lifestyle	52.59	8	21.036	441.756	3/22/2019	19:20	Google Pay	8.5	21.036

Figure 3. Summary of the dataset

Classification algorithms are then applied on the dataset to predict the performance of each algorithm and that is the reason of pre-processing the dataset in the first step. To train the classification models, 85% of the total data points is taken into account and for calculating the performance of the models, 15% of the data points is taken. The experiments were carried out on a machine that has Intel Core i3 processor with 8GB RAM. MATLAB software with Java Programming Language is used to implement this study. For each of the model, evaluation metrics involving accuracy, precision and recall have been calculated for comparison. Test datafile is the dataset that is used for evaluating the performance. Data in the test datafile should be properly labeled because those labels will be compared with the labels that are predicted from classification.

Different classification algorithms and ensemble techniques have been applied to the data and the performance of each approach is eventually analyzed. Accuracy, Precision, Recall stands the basic evaluation metric of analyzing the performance of any classification algorithm. The following content gives the explanation about each of the evaluation metrics.

For a given input data, the binary classifier generates either Yes/ No and 1/0 as the output. Prediction of "Yes" and "1" is considered to be positive while "No" and "0" is considered to be negative. Four outcomes generated as the result of binary classification is used to form the confusion matrix.

1) Test cases that are positive at the base and are correctly evaluated as positive is said to be the True Positive rate.

2) Test cases that are negative at the base and are shown as positive during classification is the False Positive rate.

3) Test cases that are negative at the base and are correctly identified as negative is said to be the True Negative rate.

4) Test cases that are actually positive and are evaluated as negative during classification is the False Negative rate.

To determine how well the algorithm predicts the data points correctly out of all the data points, accuracy stands one of the best methods. Adding true predictions of the dataset and dividing it by the total dataset gives the accuracy percentage of the classification. 1- ERR i.e., 1- Error Rate is another way of evaluating accuracy. Fig 4 pictures the accuracy rate for both the algorithms.

$$\text{Accuracy} = \text{Sum of true predictions} / \text{Sum of input predictions} \quad (1)$$

Dividing the true predictions of positive classes to the sum of all the predictions of the positive class gives the precision value of the algorithm.

$$\text{Precision} = \text{True Positive predictions} / \text{Sum of true and false positive predictions} \quad (2)$$

Among all the actually positive cases, how many of them were correctly predicted as true predictions gives the value of Recall.

$$\text{Recall} = \text{True Positive predictions} / \text{Sum of actual positive cases} \quad (3)$$

V. RESULT

Performance indicators has been calculated for the two classification algorithms namely Random Forest and Adaptive kNN which is pictured in the below figures. Results are generated for the individual classifiers and also for their ensembles like bagging. The classification results for consumer behavior are pictured in the Fig 4.

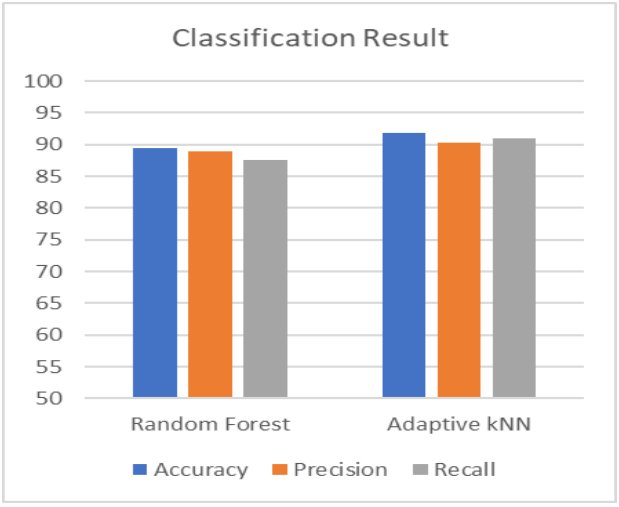

Figure 4. Result of individual classifiers

Among the two classifiers analyzed, Adaptive kNN stands the highest in terms of accuracy and also for other evaluation metrics like recall and precision. In the Bagging concept of supervised learning, the learners are trained in parallel in order to enhance the accuracy of weak classifiers which is shown in Fig 5. In the Figures 5, 6 and 7, Blue color line indicates the level of Adaptive kNN and red line indicates the level of Random Forest algorithm.

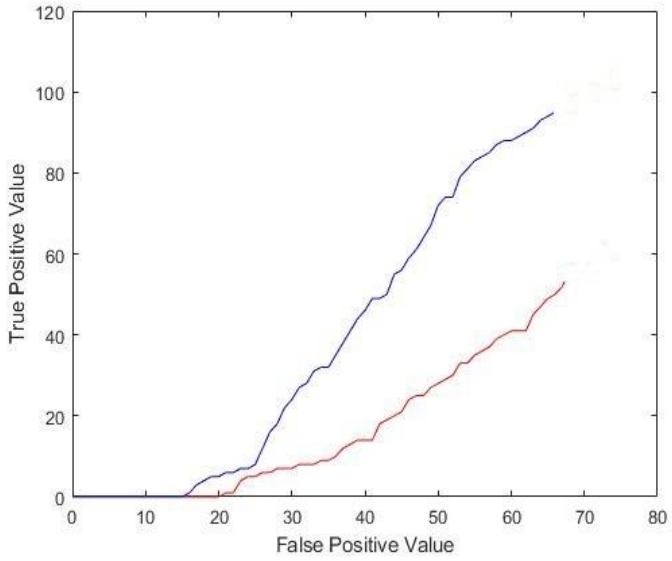

Figure 5. Accuracy level after Bagging

Precision and Recall results of the two classification algorithms are depicted in the Fig 6 and 7. On the whole, Adaptive kNN ensembled with Bagging gives the best prediction results.

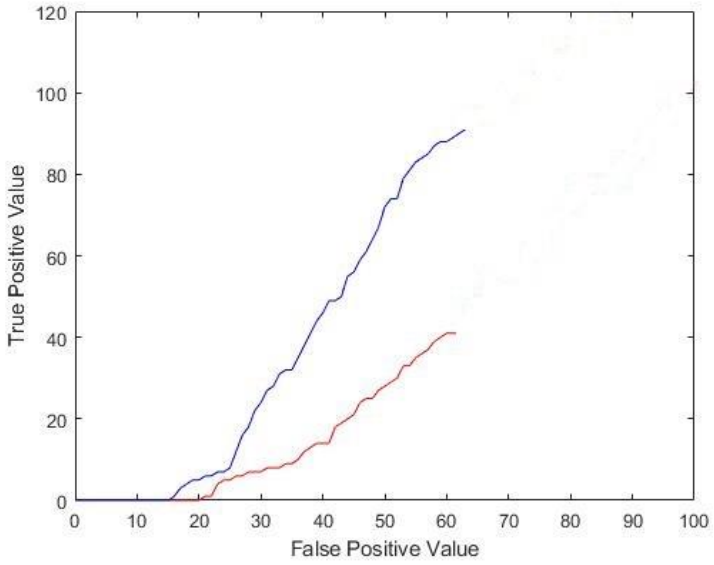

Figure 6. Precision level after Bagging

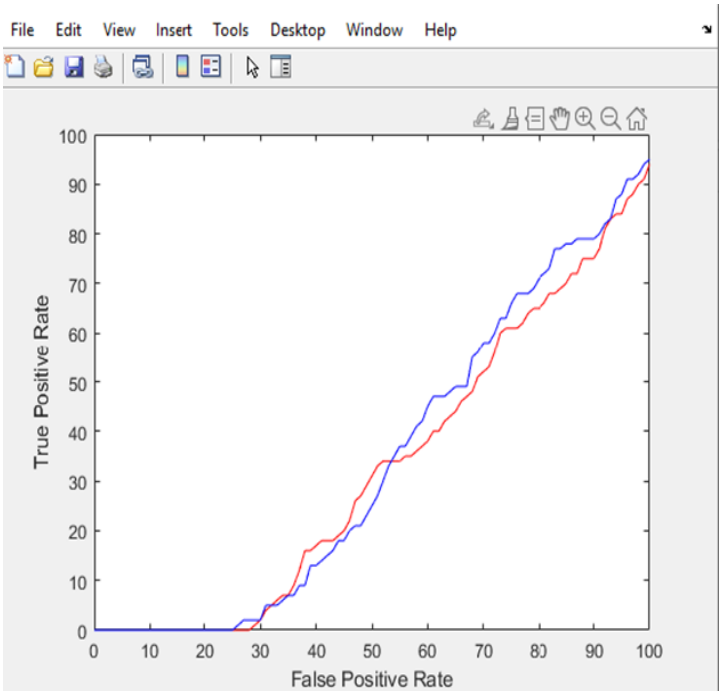

Figure 7. Recall level after Bagging

VI. CONCLUSION

In this paper, two different classification algorithms have been analyzed to identify the performance by using the session data of the customers in a supermarket. The main focus of the work is to predict the purchase intention of the clients visiting the site and evaluating their accuracy in percentage. From the experiments carried out so far, Adaptive kNN with Bagging stands the best in predicting the intension and also with the accuracy level of 91.87%. The next plan of the work will be using a very large dataset and also finding another efficient algorithm to get combined with Adaptive kNN to increase the accuracy level than that is achieved now.

REFERENCES

[1] Statistics and facts about global e-commerce. Retrieved from https://www.statista.com/topics/871/online-shopping, 2017.

[2] Statistics and facts about e-commerce in India. https://www.statista.com/topics/2454/e-commerce-in-india, 2017.

[3] Aghdaie, Mohammad Hasan, Sarfaraz Hashemkhani Zolfani, and Edmundas Kazimieras Zavadskas, "Synergies of data mining and multiple attribute decision making." *Procedia-Social and Behavioral Sciences 110: 767-776, 2014.*

[4] Michael Shekasta, et al., "New Item Consumption Prediction Using Deep Learning." , *arXiv preprint arXiv:1905.01686, 2019.*

[5] Humphrey Sheil and Omer Rana, "Classifying and Recommending Using Gradient Boosted Machines and Vector Space Models", *In Advances in Computational Intelligence Systems. UKCI 2017., Zhang Q Chao F., Schockaert S. (Ed.), 2017.*

[6] Aida Mustapha, Shazwani Mustapa, "A Classification Approach for Naive Bayes Of Online Retailers", *Acta Informatica Malaysia 1(1) (2017) 26-28, 2017.*

[7] Jaiswal, J. K., & Samikannu, R., "Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression", *2017 World Congress on Computing and Communication Technologies (WCCCT). doi:10.1109/wccct.2016.25, 2017.*

[8] Maheswari, K., & Priya, P. P. A, "Predicting customer behavior in online shopping using SVM classifier", *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS). doi:10.1109/itcosp.2017.8303085, 2017.*

[9] Fatemeh Safara, "A Computational Model to Predict Consumer Behaviour During COVID-19 Pandemic", *Computational Economics https://doi.org/10.1007/s10614-020-10069-3, 2020.*

[10] Breiman, Leo. "Random Forests.", *UC Berkeley TR567, 1999.*

[11] Ho, Tin Kam, "Random decision forest." *Proceedings of 3rd international conference on document analysis and recognition. Vol. 1. IEEE, 1995.*

[12] Rygielski, Chris, Jyun-Cheng Wang, and David C. Yen, "Data mining techniques for customer relationship management." *Technology in society 24.4: 483-502, 2002.*

[13] Anil Kumar and Manoj Kumar Dash. "Factor exploration and multi-criteria assessment method (AHP) of multi-generational consumer in electronic commerce.", *International Journal of Business Excellence 7.2: 767-776, 2014.*

[14] Yisen Wang, Shu-Tao Xia, Qingtao Tang, Jia Wu, and Xingquan Zhu, "A Novel Consistent Random Forest Framework: Bernoulli Random Forests", *IEEE Transactions on Neural Networks and Learning Systems, 2017.*

[15] Li, H., Jiang, H., Wang, D., & Han, B, "An Improved KNN Algorithm for Text Classification", *2018 Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC). doi:10.1109/imccc, 2018.*

[16] Sun, S., & Huang, R, "An adaptive k-nearest neighbor algorithm*", 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery. doi:10.1109/fskd.2010.5569740, 2010.*

[17] Kibanov, M., Becker, M., Mueller, J., Atzmueller, M., Hotho, A., & Stumme, G, "Adaptive kNN using expected accuracy for classification of geo-spatial data", *Proceedings of the 33rd Annual ACM Symposium on Applied Computing - SAC '18. doi:10.1145/3167132.3167226, 2018.*

[18] Rayhan Kabir, Rasif Ajwad, Faisal Bin Ashraf, "Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data", *2020.*