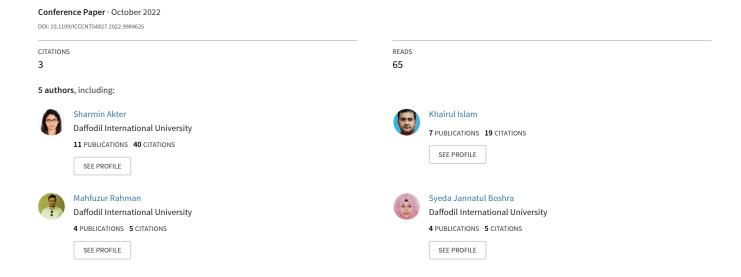
People Thoughts Prediction Using Machine Learning on Online Shopping in Bangladesh During COVID-19 Pandemic



People Thoughts Prediction Using Machine Learning on Online Shopping in Bangladesh During COVID-19 Pandemic

Sharmin Akter
Dept. of CSE
Daffodil International University
Dhaka, Bangladesh
sharminakter.cse@diu.edu.bd

MD. Khairul Islam

Dept. of CSE

Daffodil International University

Dhaka, Bangladesh
khairul15-9838@diu.edu.bd

MD. Nabir Hossain

Dept. of CSE

Daffodil International University

Dhaka, Bangladesh

nabir15-9679@diu.edu.bd

Mahfuzur Rahman

Dept. of CSE

Daffodil International University

Dhaka, Bangladesh

mahfuzur15-10030@diu.edu.bd

Syeda Jannatul Boshra

Dept. of CSE

Daffodil International University

Dhaka, Bangladesh
syeda15-10032@diu.edu.bd

Abstract—E-commerce business has become a prominent entity of global retail as online transaction saves time and cost at the same time. COVID-19 pandemic and lockdown accelerated the growth of e-commerce. The new e-business companies are booming rapidly whereas insincerity in customer concern is noticeable. As a result, the purchasers are facing numerous problems while buying online. The main objective of this study is to predict preferences on online shopping of buyers and based on that analysis, the pattern can be observed. While doing the study, we used some popular Supervised machine learning algorithms such as Decision Tree (DT), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Naïve Bayes (NB) algorithm. Amongst those, best accuracy was delivered by the Decision Tree algorithm. The output clearly demonstrates that, people are more likely to participate in online shopping if the obstacles could be alleviate which means, buyers are still not satisfied and confident about the online platform. Hopefully, the result of this study can be a great asset for improving the Ecommerce sector of Bangladesh if it is optimized wisely.

Index Terms—Online shopping, Decision Tree, Machine Learning

I. Introduction

Internet has brought tremendous change in the field of shopping. Previously it was common for people to buy something from a physical shop. Now-a-days the way of buying goods has changed and people are getting used to shop through online [1]. More precisely, because of the recent pandemic situation, the number of online customers has been increased. Lockdown has also changed the view of purchasing products. People can't go outside and that's why online shopping is the only way to buy necessary stuffs [2]. People having busy schedule are more likely to shop

from online sectors because it saves both time and energy [3]. Multiple vendors are investing in online businesses and online shopping is becoming more competitive day by day. Still there are limitations about identifying fraud websites, false products, payment complications etc. Companies and production houses are concentrating more on the quality of their products. To satisfy customers, online based companies need to know the mindset, behavior and demand of the customers to improve their field of business. As online shopping is becoming popular rapidly, it is important to analyze the facts, reasons, satisfaction level, trust issues, behavior, future and overall situation of this up growing percentage of consumers. Researchers are discussed different types of analysis based on E-commerce such as Predict customer behavior [4], online adaptation [5], product return policy [6] and the satisfaction level on online shopping [7]. Most of this work are based on specific shopping data. Like data collected from some company's website or collected from shopping mall or specific topics survey. But this not clear idea for overall online shopping. To remove this barrier this paper aims to find out the present situation and future potential of overall online shopping system. This paper is based on a core real time data set which have been retrieved from online shoppers. We conduct an online survey from different profession online shopper and used the Supervised Machine learning algorithm [8] Decision Tree predicting user prefer in online shopping.

This paper is constructed as follows. Phase II discusses the background study. Phase III expand on methodology. Phase IV highlights on performance measurement and phase V on experimental result. Finally, phase VI concludes the paper.

II. BACKGROUND STUDY

Numerous researchers have discussed on online shopping in different perspective.

Dr. Neetu Narwal [9] analyzed sentiment on the reviews of customers on newly launched mobile phone Moto G Plus from a website. They collected total 1560 reviews which are consist of 18720 words. They used R language and its library package syuzhet and getNrcSentiment function. And calculate 8 different emotions based on this reviews word. Their calculation showed that 62% emotions is positive and the other 38% emotions are negative.

Norhaslinda Kamaruddin, Siti Azizah Abas, Abdul Wahab[10] analyzed sentiment based on reviews from comment section of products. They showed that sentiment analysis used as an alternative solution to measure the positive, negative, and neutral feedback of the past reviews. They used supervised Machine Learning Approach to rank the product based on the sentiment category. The performance between 65% to 95% accuracy.

Sven F. Crone, Didier Soopramanien [11] predicted online shopping adoption of consumers. They collect data from 685 users by online shopping channels. Mainly they find out three types users based on their giving data. They are people browse online and buy from online, people browse online but buy from store and people who don't use internet for shopping. They used Logistic Regression & Neural Network for classification users. The Neural Network give the best accuracy both of them.

Katerina Kalaidopoulou, Stratos Triantafyllou, Anastasia Griva, Katerina Pramatari [12] identified the customer satisfaction patterns/profiles from a survey data which collected from 120 Greek-e-shops from 18 industries. They collect data two step from a user. They used unsupervised machine learning approach in form of cluster analysis and the algorithm used Expectation-Maximization (EM) for perform business & technical term.

Dr. K. Maheswari, P. Packia Amutha Priya [13] classified the customers on Regular, Occasions, Festival Seasons, Offer, Window shopping (Never purchase, just watch and visit, rarely purchase), Recent customers based on their data. They used Support Vector Machine (SVM) algorithm and observed that according to dataset 15 to 24 ages people purchased higher percentage than the other ages people in year between 2007-2014.

R. Venkatesan, A. Sabari [14] reviewed some papers of recommendation systems in the e-commerce field with data mining algorithms and conducted a survey among the young generation based on Indian various E-commerce websites. They found that, 21-25 & 26-30 age group people buy their product with most frequent manner, majority of people visit the e-commerce site consider TV ads, price of the product, offer deals and they attract on a product to read the review comment who buy in recent.

Rana Alaa El-Deen Ahmeda, M.Elemam.Shehaba, Shereen Morsya, Nermeen Mekawiea [15] studied online shopping data and predicted the user attitude and behavior. They applied several numbers of data mining algorithm. The decision table & Filtered Classifier gave the best accuracy in their data set.

Kartika Purwandari, Join W. C. Sigalingging, Muhammad Fhadli, Shinta Nur Arizky, Bens Pardamean [16] studied the customer opinion based on some questionnaires. They collected data from a family restaurant. They analyzed data with K-means Clustering, Spectral Clustering (SC), and Agglomerate Clustering (AC) algorithms. K-means Clustering produced the better output.

Kangning Wei, Jinghua Huang, Shaohong Fu [17] discussed personalized recommendation techniques in recommender system through summarizing and analyzing personalized recommendation research. Mainly they include three aspects such the input of recommender system, the typical methods of various recommendation techniques, based on current research application situation. Collaborative filtering approach technology used by collaborative algorithm and sorted it into two types such Heuristic-based method, Model based method.

D. Kalaivani, T. Arunkumar [18] used A multi process prediction model which is consist of Linear Regression, Analysis of variance analysis(ANOVA) and logistic regression for analysis customer behavior. They categorized buyer behavior in different segments and founded that, the Drop Out people are lower class income people and and they don't have enough time and ability to use internet. Finally, Never buys people are more alert about online privacy for that reason they were not eager to purchase.

Abhishek Kanani, Sanket Chodavadiya, Prajapati Maharshi, Prof. Shanti Verma [19] worked on customer satisfaction about e-commerce service and focused on customers age group. Collected 520 user responses based on age group, service type and satisfaction level. Founded that there is no dependency between age group and satisfaction level of Indian e-consumer. Most attended people are range is 18 to 28 and e-shopping and e-ticket are the most popular online services.

D. Kalaivani, P. Sumathi [20] proposed a model to get the idea of customer behavior named FBPCA (Factor based principle component analysis) method. This method and decision tree prediction model is refined and proposed factor for customer analysis. They used Decision Tree, Back propagation Neural Network, Support Vector machine (SVM), Rule based Classification and Bayesian Classification for analysis and Classification. Analyzing data from decision rules they used Decision Tree(DT) and building the tree they used the recursive splitting method.

Previous researchers worked with specific demography group, products or shops. In our study, we will predict the preference of overall online shoppers and compare some attributes of online shopping in COVID-19 pandemic.

III. METHODOLOGY USED

The key objective of this research paper is to determine the thought of people about online shopping on covid pandemic situation or after pandemic situation. The number of purchases in online are increasing day by day. By this time people demand and their purchase thinking are changing day by day. The proposed methodology of this research work is given below:

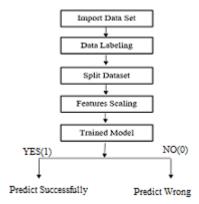


Fig. 1. Proposed Methodology

We see that from the figure 1, after collecting the data we make fit data set for applying machine learning algorithm. Four machine learning algorithms applied for prediction. They are Decision Tree, K-Nearest Neighbors, Support Vector Machine and Naive Bayes. The final predicted outcome is people prefer online shopping or not. The result "YES" stands for prefer online shopping and "NO" stands for don't prefer online shopping.

A. Data Set

Covid-19 pandemic situation changes everything in our daily life routine[21]. These changes happen all over the country also in Bangladesh. Gathered people are not allowed in shopping mall and sometime people try to avoid public place. For that reason, buying product in online are increasing day by day and new online business are starting based on public demand. At the same time the quality-of-service decrease in different online shopping. Sometime people don't get their product in exact time, they don't receive same product which they ordered and so on. This type of problem plays bad impact in online shopping or E-commerce sector. Some company

who's originally serve the people, yet the lose their customer for the bed impact of other company. For the measure of this impact, we conducted an online survey and collected data via Google Form.

We gave a short description of our research work in the google form for the people who are taking part in our research work and didn't't take any personal data. We confirmed that their records would be kept mystery. We gathered data in Covid-19 lockdown period and it was impossible for us to reach physically to the people. Finally, We accrued 510 real-time records from Bangladesh through online survey using social media together with Facebook, Linked In, E-mail etc. The majority of portion of data came from university and college student. We also collected data from other professing people such Engineer, Teacher, Govt Employee. We used a questionnaires pattern to collect data. The gender ratio is 0.65:0.35(Men: Women) and 63% of people are student, 12% of people are Engineer, 5% of people are teacher and they are all Bangladeshi people.

A total 19 question asked in survey via google form and these questions consist of three categories. They are YES or No type, Multiple options and submit text. We contacted few experts while making the questionnaires and our predicting level are discussed with them.

B. Data Cleaning

We pre-processed data set after collecting data. Initially we have 19 columns and some columns contain same kind of information. After feature scaling we kept total 10 columns for final work. Here we keep 9 columns for input and 1 column for output. In our data set there were two types of data they are numerical and categorical. We converted categorical data to numeric using Label Encoder. The final output column (Prefer online shopping or not) contains two values they are YES and NO. YES, stands for 1 and No stands for 0. The data cleaning flow are given below in Figure 2:

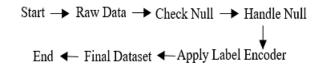


Fig. 2. Data Processing Process

C. Identify Prefer level in online shopping

Nowadays people are more interested in online shopping. Current pandemic is one of the reasons, on the other hand online shopping save the time and people need not to go the shopping mall. In contrast people faced various problem when they buy product from different online shop. One of the common problems is people don't get exact product what they saw in advertisement and another is payment method. Though people like buy product in online but they want to pay cash on delivery.

The key objective of this research paper is to predict the preferred label of people in online shopping in perspective in Bangladesh. So, the class of the target attribute is 0 and 1. Where 0 for NO and 1 for YES.

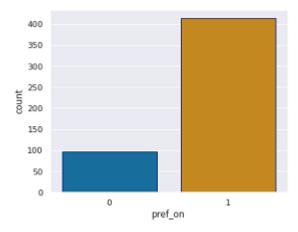


Fig. 3. Prefer Level Ratio

IV. PERFORMANCE MEASUREMENT

Classification algorithms are widely used in data mining and it can be performed both structured and unstructured data. The main aim of classification algorithm is to identify the category or class by analyzing the training data set. In our data set we applied four different classification algorithms which are Decision Tree, K- Nearest Neighbors, Naive Bayes and Support Vector Machine. To find the appraisal of our trained algorithm we used some parameter as follows:

A. Accuracy

Accuracy is a ratio of correctly predicted items to the total items. Accuracy is the most important performance parameter. To calculate the accuracy, we used this formula:

$$Accuracy = (TP + TN)/Total$$
 Number of Items (1)

B. Precision

Precision is a ratio of appropriately positive predict items and total positive predict items. To calculate the precision, we used this formula:

$$Precision = TP/(TP + FP)$$
 (2)

C. Recall

Recall is the value of appropriately positive predict items by the classification algorithms. This is also called TP rate. To calculate the Recall, we used this formula:

$$Recall = TP/(TP + FN) \tag{3}$$

D. F1 Score

F1 score the means value of precision and Recall. To calculate the Recall, we used this formula:

F1 Score =
$$2/\{(1/Precision) + (1/Recall)\}$$
 (4)

Here full form of some Matrix of Performance and their exact meaning:

- TP = True Positive: The exact result is Positive and classifier predict Positive.
- FP = False Positive: The exact result is Negative but classifier predict positive.
- FN = False Negative: The exact result is positive but classifier predict Negative.
- TN = True Negative: The exact result is Negative and classifier predict Negative.

V. EXPERIMENT & RESULT

A. Experimental Setup

We have divided the data set into two parts. One training and the other is testing with the ratio of 7:3(training: testing). To remove the less important column we implemented feature scaling and keep to 9 columns to predict the final result and data set scaled by Standard Scalar. While doing the analysis the data set researcher got some observation about the most important thing and payment method based on data set who's are prefer or not prefer online shopping. People, who prefer online shopping "Product Quality" and "Reliability" are most important thing. Almost 200 people said that "Product Quality" is the most concern topics while they buy a product in online. On the other hand, almost 125 people said that "Reliability" is the most concern topics and 75 people said that "Price" is most important for online shopping and others are concern about all. Here in X-axis(most_important) 0 stands for All, 1 for Price, 2 for Product Quality, 3 for Reliability. The graphical representation of most_important and prefer type(pref_on) of online shopping are given below:

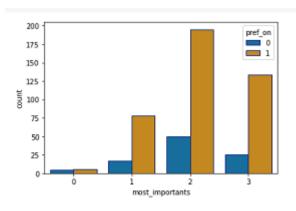


Fig. 4. Graphical Representation of most important and prefer type of Online Shopping

Another observation is about payment method. Though people prefer online shopping but most of them want to pay cash on delivery. It means that, they don't want to trust the online shop without getting product because some fake online shop collect money from customers but don't give any products or give invalid product. Almost 325 people they prefer online shopping but they want to pay cash on delivery and 100 people are agreed to pay via online payment. Here in X-axis(payment_method) 0 stands for cash on delivery and 1 for online payment. The graphical representation of payment method and prefer type(pref_on) of online shopping are given below:

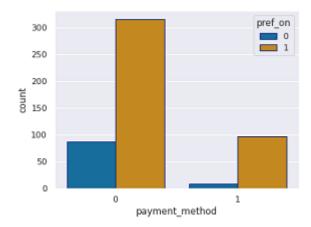


Fig. 5. Graphical Representation of Payment Method and Prefer Type of Online Shopping

The researchers applied multiple machine learning algorithm for the classification[22]. The algorithms are Decision Tree, K- Nearest Neighbors, Naive Bayes and Support Vector Machine. The Decision Tree algorithm gave the best accuracy among them. We calculate the Confusion Matrix to check how data they predict correctly. The confusion matrix of different algorithms is given below:

TABLE I
COMPARATIVE HEAT MAP OF DIFFERENT ALGORITHM

Algorithms	Heat Map			
	Predicted Level/True Level	YES	NO	
DT	YES	113	7	
	NO	14	19	
KNN	YES	119	1	
	NO	24	9	
SVM	YES	117	3	
	NO	22	11	
NB	YES	104	16	
	NO	15	18	

The graphical representation of Decision Tree Heat Map is given below:

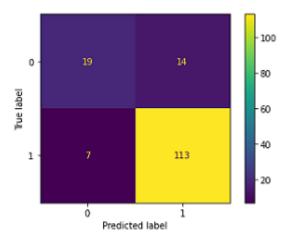


Fig. 6. Graphical Representation of Decision Tree

The graph represents and summarize the calculate the right predict and wrong predict of the Decision Tree algorithm. Different algorithm is used to trained the model and Decision Tree gave the best output among them.

B. Result

The researcher used 70% of data for training and 30% of data for testing to calculate to model performance. We used "train_test_split" for dividing the training and testing data for model. After training the model the "Decision Tree' Algorithm gave the best accuracy of 86.25%. Different performance result has been observed, the following table are representing the value of performance of algorithms:

TABLE II

OVERALL PERFORMANCE TABLE OF VARIOUS ALGORITHM

Algorithms	Classes	Precision	Recall	F1-Score	Support
DT	0	0.73	0.58	0.64	33
	1	0.89	0.94	0.91	120
KNN	0	0.90	0.27	0.42	33
	1	0.83	0.99	0.90	120
SVM	0	0.79	0.33	0.47	33
	1	0.84	0.97	0.90	120
NB	0	0.53	0.55	0.54	33
	1	0.87	0.87	0.87	120

As we have found that our model trained by 70% of data and tested by 30% of data. Different types of training and testing accuracy are observed in the data model. The training and testing accuracy of different algorithms are given below:

TABLE III
ACCURACY TABLE OF VARIOUS ALGORITHM

Algorithms	Training and Testing Accuracy		
	Training Accuracy	Testing Accuracy	
DT	95.79%	85.62%	
KNN	86.55%	83.66%	
SVM	88.51%	83.66%	
NB	82.63%	79.73%	

The decision tree algorithm showed the best accuracy of 85.62%. Below is the display of the accuracy result in bar chart:

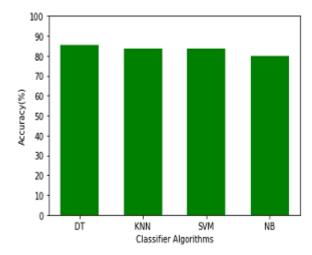


Fig. 7. Accuracy percentage of various algorithms

We used Scikit-learn library in python for graphical representation and the other calculation in data model.

VI. CONCLUSION

E- commerce sector is one of the growing sectors in digital world also in Bangladesh. The participation of people in this sector both as a customer and business owner increasing day by day. In addition, the Covid-19 pandemic plays very important role in this sector. Though most of the people are interested online shopping but people are not satisfied some characteristics of online shopping and which topics is very important in online shoppers we already showed in section IV. The limitation of our research work is we cannot collect big amount of data because of Covid-19 pandemic, which might be gave more accurate result. In future we will analyze the people's sensibility in online shopping. This research work will help the business owners to improve their services and who want to start online business. Two specific point we mentioned in section IV that what is more priority when people buy a product in online and people prefer online shopping but they unwilling to pay online. We believed that this paper will help business owner to alert these topics.

REFERENCES

- [1] Bijoy, Md Hasan Imam, Maksuda Akter Rubi, Shanjida Chowdury, and Siddiqur Rahman. "Growth and Trustworthiness of Online Shopping & Business during the Pandemic: A Case Study on Bangladesh." Journal of Sales, Service and Marketing Research (e-ISSN: 2582-7804) (2022): 6-16.
- [2] Md. Hasan Imam Bijoy, Masud Rabbani, Md. Ishrak Islam Zarif, Md. Mahbubur Rahman, Md. Rakibul Hasan, and Tridip Bhowmik , "A stupendous statistics on the pandemic impact on all sectors of Bangladesh", AIP Conference Proceedings 2451, 020006 (2022) https://doi.org/10.1063/5.0095349
- [3] Mosa, Mona & Agami, Nedaa & khayat, Ghada & Kholief, M. (2020). A Literature Review of Data Mining Techniques for Enhancing Digital Customer Engagement. International Journal of Enterprise Information Systems. 16. Article 5. 10.4018/IJEIS.2020100105.

- [4] Abdul Rahman, Nur. (2020). Preliminary Studies on Predicting Customer Purchase Behaviour in Online Retail Business. International Journal of Advanced Trends in Computer Science and Engineering. 9. 408-412. 10.30534/ijatcse/2020/5891.42020.
- [5] Pramanik, Anik, Md Bijoy, Hasan Imam, and Md Rahman. "Depression-Level Prediction During COVID-19 Pandemic Among the People of Bangladesh Using Ensemble Technique: MIRF Stacking and MIRF Voting." In Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021, pp. 71-87. Springer, Singapore, 2022.
- [6] Safari, Ehram & Pourhashemi, Seyedeh & Gharahkhani, Mohsen. (2020). A New Model for Predicting the Probability of Product Return in Online Shopping.
- [7] Bijoy, Md Hasan Imam, Sumiya Alam Akhi, Md Ali Ashraf Nayeem, Md Mahbubur Rahman, and Md Jueal Mia. "Prediction of internet user satisfaction levels in Bangladesh using data mining and analysis of influential factors." Bulletin of Electrical Engineering and Informatics 11, no. 2 (2022): 926-935.
- [8] M. R. Hasan, M. H. Imam Bijoy, S. A. Khushbu, S. Akter and S. A. Hossain, "Supervised Method Pursued For Overall Impact of Online Class During Lockdown in Bangladesh," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-5, doi: 10.1109/ICCCNT51525.2021.9579827.
- [9] Dr. Neetu Narwal, "A STUDY OF PRODUCT SENTIMENT ANALY-SIS BASED ON REVIEWER'S COMMENTS", Jamshedpur Research Review (UGC Approved International Research Journal) (ISSN: 2320-2750) Year V Vol. V Issue XXIV, OCTOBER –NOVEMBER 2017
- [10] Norhaslinda Kamaruddin, Siti Azizah Abas, Abdul Wahab, "Comparative Study on Sentiment Analysis Approach for Online Shopping Review", Turkish Journal of Computer and Mathematics Education Vol.12 No.3 (2021), 1358-1370, doi: https://doi.org/10.17762/turcomat.v12i3.907.
- [11] Crone, Sven & Soopramanien, Didier. (2005). Predicting Customer Online Shopping Adoption - an Evaluation of Data Mining and Market Modelling Approaches.. Proceedings of the 2005 International Conference on Data Mining, DMIN'05. 215-221.
- [12] Kalaidopoulou, Katerina & Triantafyllou, Stratos & Griva, Anastasia & Pramatari, Katerina. (2017). Identifying Customer Satisfaction Patterns via Data Mining: The Case of Greek E-Shops.
- [13] K. Maheswari and P. P. A. Priya, "Predicting customer behavior in online shopping using SVM classifier," 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), 2017, pp. 1-5, doi: 10.1109/ITCOSP.2017.8303085.
- [14] Venkatesan, R., and A. Sabari. "ISSUES IN VARIOUS RECOM-MENDER SYSTEM IN E-COMMERCE-A SURVEY." Journal of Critical Reviews 7, no. 7 (2020): 604-608.
- [15] R. A. E. -D. Ahmeda, M. E. Shehaba, S. Morsya and N. Mekawiea, "Performance Study of Classification Algorithms for Consumer Online Shopping Attitudes and Behavior Using Data Mining," 2015 Fifth International Conference on Communication Systems and Network Technologies, 2015, pp. 1344-1349, doi: 10.1109/CSNT.2015.50.
- [16] K. Purwandari, J. W. C. Sigalingging, M. Fhadli, S. N. Arizky and B. Pardamean, "Data Mining for Predicting Customer Satisfaction Using Clustering Techniques," 2020 International Conference on Information Management and Technology (ICIMTech), 2020, pp. 223-227, doi: 10.1109/ICIMTech50083.2020.9211272.
- [17] K. Wei, J. Huang and S. Fu, "A Survey of E-Commerce Recommender Systems," 2007 International Conference on Service Systems and Service Management, 2007, pp. 1-5, doi: 10.1109/ICSSSM.2007.4280214.
- [18] Kalaivani, D. & Thangavelu, Arunkumar. (2018). Multi process prediction model for customer behaviour analysis. International Journal of Web Based Communities. 14. 54. 10.1504/IJWBC.2018.090918.
- [19] Kanani, Abhishek & Chodavadiya, Sanket & Maharshi, Prajapati & Verma, Shanti. (2020). Predict online customer satisfaction level on the basis of e-commerce services and age group.
- [20] Kalaivani, D., Sumathi, P. Factor based prediction model for customer behavior analysis. Int J Syst Assur Eng Manag 10, 519–524 (2019). https://doi.org/10.1007/s13198-018-0739-4
- [21] M. Hasan, M. H. I. Bijoy and S. A. Akhi, "Refute the Decision of Auto-Promotion and Real Facts of Digital Online Classes During the Pandemic in Bangladesh," 2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI), 2020, pp. 1-6, doi: 10.1109/ICATMRI51801.2020.9398326.
- [22] M. H. Imam Bijoy, A. Pramanik, M. S. Rahman, M. Hasan, S. A. Akhi and M. M. Rahman, "MKRF Stacking-Voting: A Data

Mining Technique for Predicting Educational Satisfaction Level of Bangladeshis Student During Pandemic," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), 2022, pp. 1-7, doi: 10.1109/I2CT54291.2022.9824357.