Identification of Consumer Buying Patterns using KNN in E-Commerce Applications

[1]**Wasim Yasin**
Department of Computer Science and Engineering
CHRIST(Deemed to be University)
Bangalore, Karnataka, India
wasim.yasin@mtech.christuniversity.in

[2]**Karthikeyan Harimoorthy**
Department of Computer Science and Engineering
CHRIST(Deemed to be University)
Bangalore, Karnataka, India
Karthikeyanh.a@gmail.com

[3]**Divya Vetriveeran**
Department of Computer Science and Engineering
CHRIST(Deemed to be University)
Bangalore, Karnataka, India
Vdivya.3793@gmail.com

2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN) | 979-8-3503-2284-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICPCSN58827.2023.00036

***Abstract*- In recent days, with the advancement of technologies, people use electronic medium to carry out their businesses. E-commerce is a process of allowing people to buy and sell products online using electronic medium. E-commerce has a wide range of customer base as well. The data generated through transaction helps the enterprises to develop the marketing strategy. The growth of this e-commerce application depends on several factors. Some of the factors are follows 1) Customer demand, 2) Analyzing buying pattern of the users, 3) Customer retention, 4) dynamic pricing etc. It is very difficult to analyze the buying pattern of customers as there is a wide range of customer base in the online platform. To overcome this problem, this research study discusses about the challenges and issues in e-commerce applications, also identifies and analyses the buying patterns of customer using various machine learning techniques. From the implementation it is identified that, KNN algorithm performed well while comparing it with various other machine learning algorithms. Performances of these algorithms have been analyzed using various matrices. For analyzing, the model is tested using e-commerce dataset (Amazon dataset downloaded from Kaggle.com). From the analysis it found that KNN algorithm computes and predicts better compared to other machine learning algorithms either Naïve Bayes, or Random Forest, or Logistic Regression etc.**

Keywords: Machine Learning, Consumer behavior analysis, K-Nearest Neighbor, buying behavior, E- Commerce applications

I. INTRODUCTION

In recent days, due to advancement in technologies and internet [1], many people prefer to purchase various products in online mode [2]. E-commerce applications generate large amount of data; the customers browse through different items before baking a final decision of buying a product. As such enormous data are being generated from the prospective customer; and these data contain customer details, about their transactions, recently viewed products etc. During COVID-19 pandemic, e-commerce helped people to buy their essential needs through online mode. As per the leading daily newspaper *The Times of India*'s report on22, Jan 2022, E-commerce retail trade increased drastically from 14 % to 17 % from the year 2019 to 2020 respectively.

It is also seen that there are lot opportunities in the e-commerce sector and it is ever expanding. This can be justified as per *India brand Equity Foundation* report which states that India's Equity market may reach US$ 350 billion by 2030.

In order to increase the sales and identify the customer needs, there is a need of analyzing the E-commerce customers' transaction details data. Using traditional techniques, it is very difficult to analyze and identify the underlying patterns that may exist on it. Thus to analyze there is need of machine learning algorithms [3], to identify the customer needs and product recommendation for the users (Hopkins et al., 2022).

Altogether, this research work is organized in the following manner, Section 2 discusses about the literature review, section 3 discusses about the model that is proposed, and implementation results are being discussed in section 4 while section 5 discusses briefly about the conclusion and the future work.

II. LITERATURE REVIEW

Many authors have used several machine learning algorithms to analyze e-commerce data to understand the customer buying behavior.

Kashvi et al. (2019) has suggested that K-Nearest Neighbor is a powerful tool in machine learning algorithm as it is very effective in classification and in regression, as well [4]. Although, it is highly used for classification, it generally groups the data into clusters or subsets and classifies the data based how similar the data are based on the previously inputted training datasets. The KNN model is effective but there lie many weaknesses while classifying data. The authors give a humble effort to remove the weaknesses of the model, thus trying to modify the model to a great extent.

Wenchao et al. (2019) has given a very good idea as to how to use the big data as well as Machine learning concepts to come up with a conclusion related to medical health records [5]. For its simplicity, the KNN algorithm has been used to deduce the datasets but it has been noticed that when the sample size is very large and feature attributes are huge, there occurs a robust declination in the efficiency of the

979-8-3503-2284-2/23/$31.00 ©2023 IEEE
DOI 10.1109/ICPCSN58827.2023.00036

algorithm. This study attempts to improvise on the KNN algorithm by removing the shortcomings that were there in the traditional K-Nearest Neighbor algorithm based on de-noising the cluster and the density cropping methods. This has been done by clustering speedily while maintaining the accuracy of the search speed of the KNN algorithm.

Shichao et al. (2017) has used different k parameters for the k values that the KNN algorithm takes and has used different test samples for finding the outcomes in different applications, i.e. classification, regression and imputations of the missing data [6]. This has been done by reconstructing the sparse coefficient matrix between test samples and training data. There is also the use of Locality Preserving Projection regularization process, to keep intact the local structure of data even though keeping a firm eye on the accuracy of the model. 20 real datasets were being used to come up with inference that their algorithm is better suited than the previous K-Nearest Neighbor algorithm based on regression, classification and imputation of the missing value.

Yang Li et al. (2007) has suggested that network intrusion is very hard to detect, thus devising the Transductive Confidence Machines for K-Nearest Neighbors machine learning algorithm. This method detects the high anomalies persisting in the high detection rate wherein low false positives can be ascertained and detected [7].

Ni Li et al. (2016) has suggested that when there exists no prior knowledge of the distribution of the data, then K-Nearest Neighbor algorithm is the best option that suits the model. To cope up with the prediction problem, an improved KNN algorithm was proposed and implemented. The effectiveness of the algorithm was ascertained with proper testing of the algorithm with real world data[8]. Table 1 summarizes the literature review of using machine learning algorithms in various fields.

Table 1 Summary of Literature Review

Sl. No.	Literature Title & Year of Publishing	Name & Size of the Dataset	Algorithm Used	Parameter Considered
1	Using KNN Algorithm for Classification of Textual Documents - 2017 [9]	Two articles were being taken for analysis. There are two websites whose data are being taken.	KNN algorithm was being used.	The value of k ranged from 1 to 50. Accuracy factor was considered.
2	Efficient kNN Classification With Different Numbers of Nearest Neighbors- 2018 [10]	20 datasets have been used.	KNN-based model was being used.	AD K Nearest Neighbour was being used.
3	An amalgam KNN to predict diabetes mellitus- 2013 [11]	Pima Indian diabetes dataset was being used.	KNN classifier was being used.	Amalgam KNN was being used.
4	Efficient kNN classification algorithm for big data - 2016. [12]	Several real datasets from UCI and medical imaging datasets were being used.	KNN classifier was being used.	LC-K-Nearest Neighbour and RC-K-Nearest Neighbour were being used.

III. The Proposed Model

It is observed that there is a buying pattern for the customers for a particular product which will be analyzed below using various machine learning classification algorithms.

K-NN algorithm is a classification algorithm that uses to classify test data based on the distance metrics, i.e. a test sample is classified as Class-1 if there are more number of Class-1 training samples closer to the test sample compared to other Classes training samples. In this proposed model k value is been randomly chosen ranging from 2 to 8.

In KNN algorithm we select the k entries in our database which are closest to the new sample. The k has to be a positive integer. The same algorithm has been graphically described in the figure 1. The most common classification entries are found. The new samples are being given this classification.

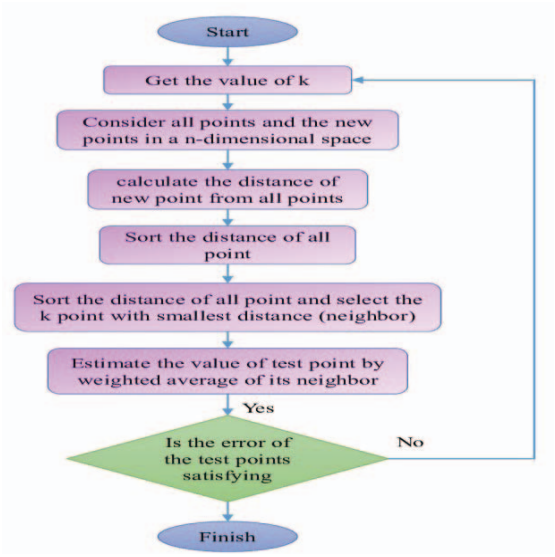

Figure 1: Flowchart of the K Nearest Neighbour algorithm

(***Source***: *Modeling viscosity of crude oil using k-nearest neighbor algorithm by Mohammad Reza Mahdiani, Ehsan Khamehchi, Sassan Hajirezaie, Abdolhossein Hemmati – Sarapardeh)*

A. How K-Nearest Neighbour (KNN) algorithm works

The working principle of K-nearest neighbours (KNN) algorithm can be understood by the following steps:

Step 1: The training as well as the test datasets have to be uploaded.
Step 2: The value of K i.e. the nearest data points has to be ascertained.
Step 3: For each point in the test data we have to do the following :

3.1: Calculating the distance between the test data and each row of the training data has to be done.
3.2: Sorting the distances in the ascending order based on their value.
3.3: The top K rows will be selected from the sorted array.
3.4: The test points will be assigned a class based on the most frequent class of the underlying rows.

Step 4: End.

In the above algorithm, to calculate the distance of new point from all the points Euclidean distance has been used the equation of which is described in the equation 1.

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots . + \quad (q_n - p_n)^2} \quad (1)$$

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

In the above algorithm, the calculated distances are sorted in ascending order based on distance values.

A dataset has been taken where the purchase decision of a consumer is reflected in the last column of the table based on the age and salary of the consumer. There are enormous datasets available in the database but the work will be done on the purchase decision made by a customer based on the columns like age and salary of the customer. There are 400 data, which are being tested for the various classification algorithms.

With the help of KNN algorithm the model is tested to see if the model rightly depicts whether the purchase decision is being made compared to the actual purchase decision. The prediction will be purely based on the training dataset that will be used. Four hundred customers' data are being taken and based on their decision of purchasing an item the dataset has been taken. 75% data are being taken into account in the training set while 25% are being taken into consideration into the test set.

B. Outcome of the implementation

The KNN algorithm is being applied to come up with the decision making process.

Out of the 400 customers, 100 customers are taken into consideration in the test set, as because 25% of the data are in the test set.

After running the KNN algorithm it is observed that:

- 93 times the algorithm predicted correctly whether the customer will buy the product or not.
- 4 times the algorithm wrongly predicted that the customer won't buy the product while he/she bought the product.
- 3 times the algorithm wrongly predicted that the customer will buy the product while he/she didn't buy the product.

Thus the accuracy of the model is 93%, which is by far much appreciated.

The training set data has been graphically described in the figure 2. From the figure it is clearly seen when the customers are going to buy the product with a clear indication of which age group of salaried persons are going to buy the product.

The test set data has been graphically described in the figure 3. From the figure it describes ascertain those individuals who are going to buy the product with a clear indication of their age group and their salary. 93 times out of 100 have seen that the model depicts the clear picture of those individuals who are going to buy the product.

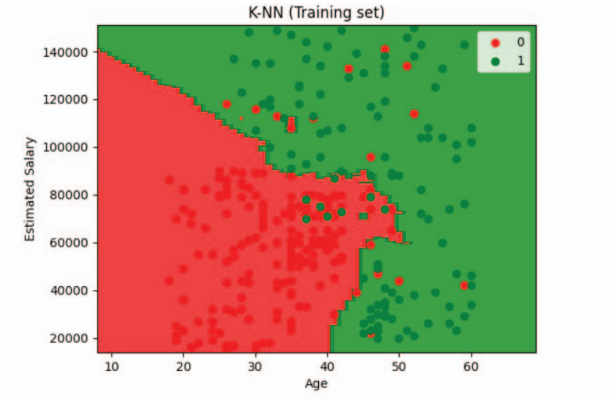

Figure 2: Graphical representation of the training set data

The figure 2 shows how the points are being plotted after analyzing the customer buying behavior using KNN algorithm for the training set. It shows a sparse distribution but correct in most of the case.

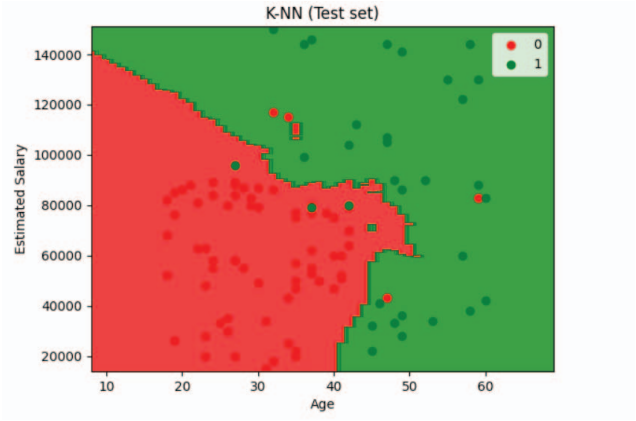

Figure 3: Graphical representation of the test set data

The above figure shows how the points are being plotted after analyzing the customer buying behavior using KNN algorithm for the test set. It shows a sparse distribution but correct in most of the case.

IV. RESULTS AND DISCUSSIONS:

A comparison of the KNN algorithm is being made with various classification algorithms along with setting various values for the value of K in K-Nearest Neighbor algorithm and then a comparison chart is being prepared.

The same dataset is being used to compare the outcome of the implementation of various classification algorithms. It is evaluated in terms precision recall, f1-score, accuracy and results are represented in figure 4, figure 5, figure 6 and figure 7 respectively.

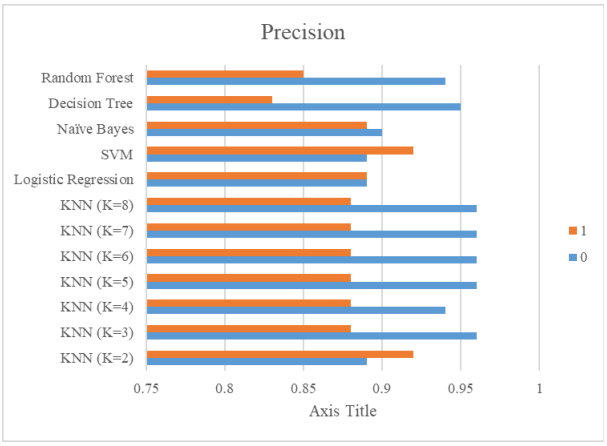

Figure 4: Precision comparison of the classification algorithms

The precision shows how correctly identifies, whether a product will be bought by a customer or not. Higher the precision higher is the accuracy of the model.

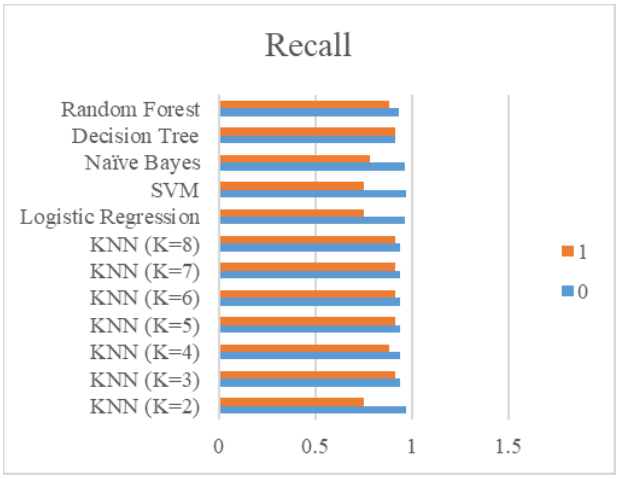

Figure 5: Recall comparison of the classification algorithms

The recall shows how correctly identifies whether a product will be bought by a customer or not. Higher the recall higher is the accuracy of the model.

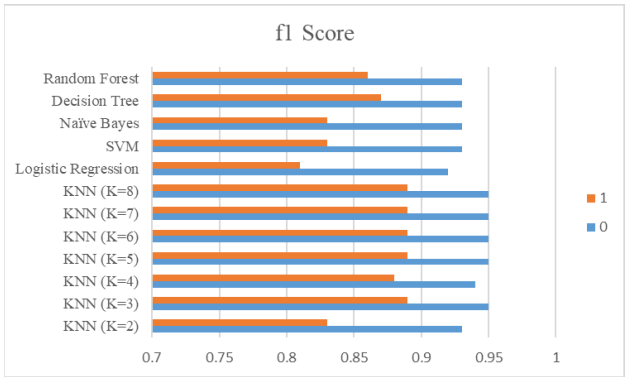

Figure 6: f1 score comparison of the classification algorithms

The f1 score determines how correctly the model has given the accurate results and represented in figure 6.

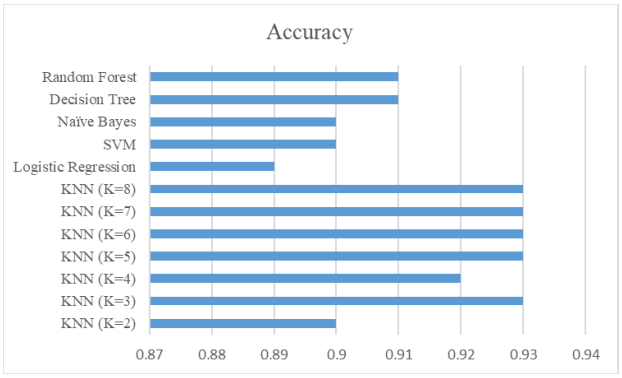

Figure 7: Accuracy comparison of the classification algorithms

Figure 7 shows how accurate the classification models are in determining whether the product will be bought by the customer or not. Higher the accuracy higher is the fruitfulness of the model.

Thus, by looking at the graph above it is clear that KNN algorithm best suites the dataset that is being used as the accuracy level is 93%. The accuracy level remains in this case 93% for KNN algorithm while keeping K=3,5,6,7,8 whereas accuracy level of predicting the purchasing decision drops to 92% and 90% for K=4 and K=2 respectively.

While it is being observed compared to KNN algorithm the other classification algorithms like Random Forest [13], Decision Tree [14], Naïve Bayes [15], Support Vector Machine [16], Logistic Regression [17] gives a low precision, recall, f1-score and accuracy level. Thus, the model selected i.e. KNN is a powerful tool to come up with a prediction.

Conclusion:

It is observed from the above implementation of the K-Nearest Neighbour model, that the accuracy of correct predictions is almost 93% in this above dataset. The working of the KNN algorithm can be further be enhanced by adding pre-processing stage.

It is observed that 93 times out of 100 the model predicted correctly, thus KNN model can be used to predict the customer buying behaviour. With the help of buying pattern data any e-commerce site can use the KNN algorithm to come up with the conclusion of whether a customer is going to buy the product or not.

The e-commerce sites are already availing the use of varied machine learning techniques and will definitely be beneficial while using the KNN algorithm for marking the prediction of those customers who are going to buy the product and thus can pitch their products to these customers to lure them to buy the product.

REFERENCES

[1] Zeyu Wang, Mingyu Li, Jia Lu, Xin Cheng, Business Innovation based on artificial intelligence and Blockchain technology, Information Processing & Management, Volume 59, Issue 1, 2022, 102759, ISSN 0306-4573. Available from: https://doi.org/10.1016/j.ipm.2021.102759

[2] Sabina Lissitsa, Ofrit Kol, Generation X vs. Generation Y – A decade of online shopping, Journal of Retailing and Consumer Services, Volume 31, 2016, Pages 304-312, ISSN 0969-6989, https://doi.org/10.1016/j.jretconser.2016.04.015.

[3] S. Ray, "A Quick Review of Machine Learning Algorithms," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 35-39, doi: 10.1109/COMITCon.2019.8862451.

[4] K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747

[5] W. Xing and Y. Bei, "Medical Health Big Data Classification Based on KNN Classification Algorithm," in IEEE Access, vol. 8, pp. 28808-28819, 2020, doi: 10.1109/ACCESS.2019.2955754

[6] S. Zhang, X. Li, M. Zong, X. Zhu and R. Wang, "Efficient kNN Classification With Different Numbers of Nearest Neighbors," in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1774-1785, May 2018, doi: 10.1109/TNNLS.2017.2673241

[7] Yang Li, Binxing Fang, Li Guo, and You Chen. 2007. Network anomaly detection based on TCM-KNN algorithm. In Proceedings of the 2nd ACM symposium on Information, computer and communications security (ASIACCS '07). Association for Computing Machinery, New York, NY, USA, 13–19. https://doi.org/10.1145/1229285.1229292

[8] Li, N., Kong, H., Ma, Y. et al. Human performance modeling for manufacturing based on an improved KNN algorithm. Int J Adv Manuf Technol 84, 473–483 (2016). https://doi.org/10.1007/s00170-016-8418-6

[9] A. Moldagulova and R. B. Sulaiman, "Using KNN algorithm for classification of textual documents," 2017 8th International Conference on Information Technology (ICIT), Amman, Jordan, 2017, pp. 665-671, doi: 10.1109/ICITECH.2017.8079924.

[10] S. Zhang, X. Li, M. Zong, X. Zhu and R. Wang, "Efficient kNN Classification With Different Numbers of Nearest Neighbors," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 5, pp. 1774-1785, May 2018, doi: 10.1109/TNNLS.2017.2673241.

[11] M. NirmalaDevi, S. A. alias Balamurugan and U. V. Swathi, "An amalgam KNN to predict diabetes mellitus," 2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), Tirunelveli, India, 2013, pp. 691-695, doi: 10.1109/ICE-CCN.2013.6528591.

[12] Zhenyun Deng, Xiaoshu Zhu, Debo Cheng, Ming Zong, Shichao Zhang, Efficient kNN classification algorithm for big data, Neurocomputing, Volume 195, 2016, Pages 143-148, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2015.08.112.

[13] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," in IEEE Access, vol. 7, pp. 60134-60149, 2019, doi: 10.1109/ACCESS.2019.2914999.

[14] H. Cui, D. Huang, Y. Fang, L. Liu and C. Huang, "Webshell Detection Based on Random Forest–Gradient Boosting Decision Tree Algorithm," 2018

IEEE Third International Conference on Data Science in Cyberspace (DSC), Guangzhou, China, 2018, pp. 153-160, doi: 10.1109/DSC.2018.00030.

[15] J. Zhang, C. Chen, Y. Xiang, W. Zhou and Y. Xiang, "Internet Traffic Classification by Aggregating Correlated Naive Bayes Predictions," in IEEE Transactions on Information Forensics and Security, vol. 8, no. 1, pp. 5-15, Jan. 2013, doi: 10.1109/TIFS.2012.2223675.

[16] A. Soualhi, K. Medjaher and N. Zerhouni, "Bearing Health Monitoring Based on Hilbert–Huang Transform, Support Vector Machine, and Regression," in IEEE Transactions on Instrumentation and Measurement, vol. 64, no. 1, pp. 52-62, Jan. 2015, doi: 10.1109/TIM.2014.2330494.

[17] D. Prasetio and D. Harlili, "Predicting football match results with logistic regression," 2016 International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA), Penang, Malaysia, 2016, pp. 1-5, doi: 10.1109/ICAICTA.2016.7803111.