International Journal of Advanced Trends in Computer Science and Engineering

E-Commerce Product Recommendations using XGBoost with User Clusters and Clickstream

Anditika Maulida Purnamasari[1], Antoni Wibowo[2]

[1] Computer Science Department, BINUS Graduate Program - Master of Computer Science Bina Nusantara University, Jakarta, Indonesia 11480, anditika.purnamasari@binus.ac.id

[2] Computer Science Department, BINUS Graduate Program - Master of Computer Science Bina Nusantara University, Jakarta, Indonesia 11480, anwibowo@binus.ac.id

ABSTRACT

E-commerce companies are competing to be the best by finding users' interests and displaying products more effectively to get users to visit websites for long periods of time. The XGBoost method did not have optimal accuracy in providing product recommendations. This study aims to improve the product recommendation system using the XGBoost method by combining it with the K-Means method and clickstream data. As the results, these two contributions were able to improve the accuracy of the product recommendation system with an accuracy value of 95.72%, a precision value of 95.93%, a recall value of 95.72%, and an f-1 score of 95.80%.

Key words : Clickstream, K-Means, product recommendation, XGBoost.

1. INTRODUCTION

Along with technological developments, e-commerce has come to simplify the shopping process. In searching for products, users generally have to open various categories and product variations to get the desired product. However, this can be difficult for users because too much information has to be processed. As a result, users can switch to other e-commerce due to inefficient product data processing.

In the end, e-commerce companies are competing to be the best by always innovating to meet the desires and needs of users so that their companies remain the first choice. As a result, challenges arise to find user interests and display products more effectively so that users continue to visit the website for a long period of time [1]-[2].

The recommendation system can analyze the use of prospective buyer data, products, and user interactions with the product to find the relationship between them. The results received will be displayed as recommendations [3]. Product recommendations can help users to handle information overload and provide personalized service. In addition, the recommendation system can increase product sales and build user loyalty [4].

Some researchers have conducted research on product recommendations and datasets used such as product purchases [5]-[11], carts [7], and clickstream [5]-[11]. Clickstream contains information about user behavior seen from all mouse activities when the user is surfing the internet. Clickstream data captures various kinds of information in a complete, timely, and accurate manner that covers user activities such as browsing routes and purchased products.

Contextual information can help create a recommendation system, especially for some e-commerce, where it is not enough to only consider users and products, such as travel package recommendations and personal content [12]. Contextual information, such as age, gender, and location, can help product recommendations solve common problems, such as cold starts when a new user appears in the database.

User data will be grouped into several sections using one of the clustering methods, namely K-Means. It is used because users in certain groups tend to have the same interests [13]. Demographic data is used in the K-Means method because users in certain groups tend to have the same interests. For example, women ages 21 to 30 buy beauty products more often, while women ages 31 to 40 buy kitchen utensils more often.

Then, the user group data will be processed with transaction data and clickstream. [14] Clickstream is data that contains a user's browsing history on a website, from opening to closing a website and and this data is widely used by researchers to analyze various things [15]-[19]. Several studies state that the more often users open a product, the higher the user's interest in that product [7], [9], [20]. Those data will be processed using the XGBoost method to determine the accuracy of product recommendations. The XGBoost method is used because this method has the highest level of accuracy when it is used to analyze this dataset compared to other methods, such as Random Forest and Gradient Boost.

The remainder of this paper is organized as follows. In section 2, a brief literature review along with a statement of the problem is provided. a brief description of the data set used for the experiment; The research methodology was introduced in section 3. Analysis of the experiment and the results obtained are presented in section 4 and it is discussed in section 5. We conclude our findings in this paper in section 6.

2. RELATED WORKS

In addition to using purchase history and clickstream data, product recommendations can use data to analyze products that have been placed in a cart but are deleted so they are not purchased by the user. The product will be recommended to users if the user still sees the product several times after it is removed from the cart. Just like the previous research, the dataset used in this study are clickstream, product purchase, and cart. The dataset will be added with user demographic data to increase the level of accuracy of the product recommendations [7].

Table 1: Related Works

Authors (Year)	Method	Performance Measure
Bansal, Gupta, Sinsha (2017)	CART, Regression Tree, Apriori algorithm	Recall, Precision
Dixit dan Gupta (2019)	K-Means + CPC, PC, CV + Support, Confidence, Lift	Recall, Precision, F1
Dixit dan Gupta (2019)	KNN + CPC, PC, CV + ET, ANN, DT	Recall, Precision
Dixit, Gupta, dan Jain (2018)	Random Forest, Gradient Boost + CPC, PSS	Recall, Precision, F1
Prasetya (2017)	KNN + CF, Content Based, Cosine Similarity	Recall, Precision, F1
Rawat, Goyal, dan Singh (2017)	PSS, Rough Set Leader Clustering + DT, ANN, LR, GB, RF	Recall, Precision, F1
Xue, Xiao, Lin, Mu (2015)	CF + Bipartite Network	Recall, Precision, F1

Table 1 shows some studies related to this research. Some machine learning algorithms that can be used such as (1) Naïve Bayes [8]; (2) K-Nearest Neighbor used to search for top-n recommendations [21]; (3) a priori that was used to search for best-selling products and recommend in future purchases [5]; and (4) Random Forests, Gradient Boosts, Artificial Neural Networks, and Decision Trees that were used to analyze products that have been viewed but not yet purchased [7].

[7] Dixit, Gupta, and Jain compare four algorithms, namely Random Forest, Gradient Boost, Artificial Neural Network, and Decision Tree, to find out which algorithm has the greatest influence on the level of accuracy of product recommendations. Of the four algorithms, XGBoost is the best algorithm to be used in this research dataset. In addition, this method is also not widely used, especially in product recommendations.

Olivetti's research laboratory developed an active badge system to adapt applications to the presence of people using contextual data by sensing the locality of the mobile user [22]-[23]. [24] Dey developed a prototype application that can help conference participants and presenters decide which activities to attend with the aim of helping users decide which activities to attend, providing awareness about coworkers' activities, to improve interaction between users and the environment, helping users take notes on presentations, and assist in retrieval of conference information. after the conference ended. [25] Dey also developed a context tool to support rapid prototyping of several types of applications using contextual data.

This study used precision, recall, and f-1 score as performance measures to measure the accuracy of the product recommendation system. This method was used because there are so many studies that use this method as a performance measure for their product recommendation system [5]-[7], [20]-[21], [26].

3. RESEARCH METHODOLOGY
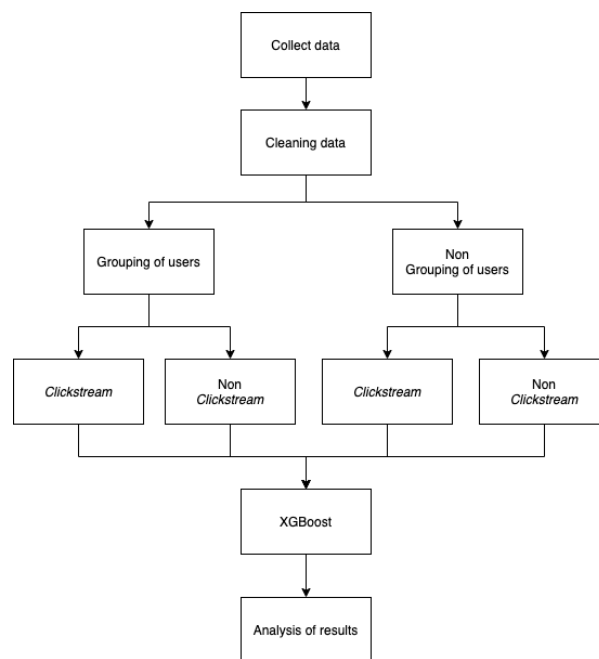
Figure 1: Research Methodology

Figure 1 shows the steps of the research carried out in this study.

a) Data collection. The data used in this research is user activity data for three months in one of e-commerce in Indonesia, which consists of clickstream data, product purchases, and carts. The dataset contains 242,888 entries with clickstream data from 86,498 visitors and 50 products. Clickstream data is taken based on visitors who have purchased 50 selected products in a predetermined period.

b) Data cleaning. This stage is carried out to obtain quality data by replacing or removing null data, data outliers, and object or string data types.

c) Grouping of users. Users are grouped using one of the clustering methods, namely K-Means. User grouping uses

user demographic data consisting of gender, age and place of residence.

d) Clickstream data. This data is used to increase the accuracy of product recommendations based on user browsing data while visiting the website. The data used include add to cart, device, click, and impression data.

e) Implementation of the XGBoost method. This method is used to predict product preferences using all existing attributes.

f) Analysis of results. Analyzing the results of the implementation of methods and algorithms to find out the answer to each problem identification from this study using several performance measurement methods, such as Precision, Recall, and F1-Score.

3.1 K-Means

K-Means is an iterative clustering algorithm that divides a data set into a certain number of groups by minimizing the total distance from each point to a group of center points. The K-Means algorithm consists of two phases, namely calculating k data center points and grouping each point into a group of points that have the closest center point. Figure 2 shows one of the most frequently used methods of determining the distance to the nearest center point is Euclidean. After the grouping is complete, this method recalculates the new center point of each group based on the previous center point to get the optimal center point. The center point for each group is the point where the total distance from all objects in the group is minimized [27].
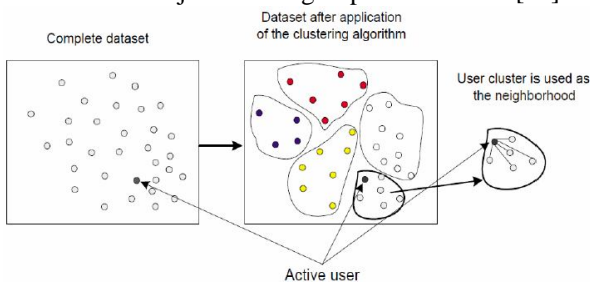
Figure 2: Formation of the environment of clustered partitions

The customer context preference matrix is performed by minimizing the sum of squares of the distance between the center of mass of the group and the associated data points [2].

$$Minimize \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - c_j)^2 \qquad (1)$$

Where K is the number of groups, N is the number of data points, C is the center of mass of group j, and $(x_{ij}-c_j)$ is the distance between the data points and the center of mass where the data is grouped. The main input of K-means grouping is the number of groups obtained with the concept of minimizing within a cluster sum of square (WCSS). The image below is a scree plot plotting the number of groups on the X and WCSS axes for each group on the Y axis.
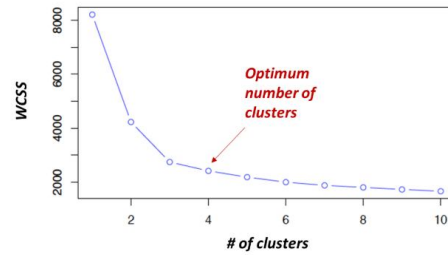
Figure 3: Scree Plot

As the number of groups increases, WCSS will continue to decline. Figure 3 shows the settlement of the WCSS was initially steep and then the rate of descent slowed down resulting in an angular plot. The number of groups in the elbow formation usually gives an indication of the optimal number of groups.

3.2 XGBoost

XGBoost is short for eXtreme Gradient Boosing. This method is an implementation of the Gradient-Boosted Decision Tree (GBDT) which can support various objective functions, such as regression, classification and ranking [28]. Assembly algorithms create and combine several weak but complementary classification methods to produce strong recommendations. This combination can be made by bagging (random forest) and boosting.

The boosting gradient is constructed sequentially. The new weak learners are built to be maximally correlated with the negative gradient with each iteration. Decision trees allow making predictions on the output variable based on a series of rules arranged in a tree structure consisting of a series of node split points. The specific value will be obtained at the last node or leaf. Tree learning algorithms do not require linear features or linear interactions between features. In addition, XGBoost is a gradient enhancement method that can accelerate tree development [29].

XGBoost combines multiple weak learners with iterative results that reduce errors in previous iterations [7]. Unlike the Random Forest, weak students at XGBoost do not have to only use a Decision Tree, but can combine it with an Artificial Neural Network. XGBoost prediction uses equation (2) which is calculated using the sklearn.ensemble python library using the same variables as RF [30].

$$XGB = \frac{1}{n} \sum_{i=1}^{n} ModelResult(i) \qquad (2)$$

3.3 Analysis of Result

Evaluations can be measured using the recall measure and the precision given by Equations.

$$recall = \frac{\sum_{i \in N} |Hid\,(i) \cap Top_{N(i)}|}{\sum_{i \in x} Hid\,(i)} \qquad (3)$$

$$precision = \frac{\sum_{i \in x} |Hid(i) \cap Top\_N(i)|}{N, |X|} \qquad (4)$$

where Hid (i) is the number of products hidden for i users.
N is the number of product recommendations for each user.
Top_N is a top N product recommendation for i users.
X is the number of users who have hidden products.

The combination of F1 steps is used for the evaluation purposes given by Eq. (6) [22]. Size of F1 with a high value shows better performance than the proposed method.

$$F - score = \frac{2 \times Recall \times Precision}{Recall + Precision} \qquad (5)$$

Recall value is a measure of the ability of the prediction model to select products from the list of product preferences used; precision is a measure of accuracy obtained; and the F-Score value is used to obtain the average value of precision and recall.

4. RESULT

Currently, the XGBoost method used in the product recommendation system is not optimal. This study still maintains the use of the XGBoost method compared to other methods because several studies have proven that the XGBoost method has a higher level of accuracy compared to other methods.
Table 2 shows the comparison value of XGBoost (XGB) accuracy with Random Forest (RF), Gradient Boosting (GB), KNN, and Decision Tree (DT). The results prove that the XGBoost method has the highest level of accuracy compared to other methods. Therefore, this study still uses XGBoost.

Table 2: Comparison of XGBoost accuracy with other methods

	XGB	RF	GB	KNN	DT
Accuracy	89,44	89,07	89,37	88,20	86,00
Precision	88,30	88,38	88,30	87,26	86,07
Recall	89,44	89,07	89,37	88,20	86,00
F1-Score	88,23	86,75	87,72	87,60	86,04
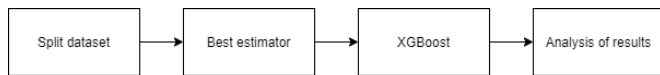
Figure 4: XGBoost Workflow

a) Split dataset. The dataset is divided into two parts, train data and test data. The buy column on each train data and test data is the main objective of the recommendation system (label). The data were divided randomly with 80% train data and 20% test data. Train data is data that will be used as learning for the XGBoost algorithm, while test data is data that will be matched for accuracy with train

data that has been processed by XGBoost.
b) Best estimator. The best estimator search process becomes a reference for the best parameters in the XGBoost process. These parameters are used to strengthen the accuracy of the XGBoost algorithm. The best estimator method is used to get the most out of the XGBoost model. This method can be used using existing libraries. The parameters used in this study are learning_rate, max_depth, n_estimators, and min_child_weight. Twenty percent of the total data train is tested by the best estimator ten times using the value of the randomly provided parameter to get the maximum result from the random value.
c) XGBoost. At this stage, the train and test data are processed using the XGBoost method using the best estimator.
d) Results analysis. After the XGBoost process is carried out, the accuracy results can be issued by displaying the precision, recall, and F1-score values. So that the results of the accuracy of the XGBoost process can be seen.
In this study, the XGBoost method was tested for its accuracy by adding clickstream data and the K-Means method. Here are the details of each test:

4.1 XGBoost, K-Means, and Clickstream

Table 3 shows the results of confusion matrix from the application of the XGBoost, K-Means, and clickstream methods. In tests carried out 26738 times, there were 3374 true positives, 22220 true negatives, 403 false positives, and 741 false negatives. In this condition, the accuracy is 95.72%.

Table 3: Confusion Matrix of XGBoost + K-Means + Clickstream

n = 26738	prediction: 0	prediction: 1
actual: 0	3374	403
actual: 1	741	22220
Accuracy	95,72%	

Table 4 shows the precision, recall, and f1-score results from the application of the XGBoost, K-Means, and clickstream methods. Where the precision value is 95.93%, recall 95.72%, and f1-score 95.80%.

Table 4: Result of XGBoost + K-Means + Clickstream

	Value (%)
Accuracy	95,72
Precision	95,93
Recall	95,72
F1-Score	95,80

4.2 XGBoost, K-Means, and Non Clickstream

Table 5 shows the results of confusion matrix from the application of the XGBoost, K-Means, and non-clickstream methods. In tests carried out 26738 times, there were 1619 true positives, 22309 true negative times, 2158 false positives,

and 652 false negatives. In this condition, the accuracy is 89.49%.

Table 5: Confusion Matrix of XGBoost+K-Means+Non Clickstream

n = 26738	prediction: 0	prediction: 1
actual: 0	1619	2158
actual: 1	652	22309
Accuracy	89,49%	

Table 6 shows the precision, recall, and f1-score results from the application of the XGBoost, K-Means, and non-clickstream methods. Where the precision value is 88.37%, the recall is 89.49%, and the f1-score is 88.35%.

Table 6: Result of XGBoost + K-Means + Non Clickstream

	Value (%)
Accuracy	89,49
Precision	88,37
Recall	89,49
F1-Score	88,35

4.3 XGBoost, Non K-Means, and Clickstream

Table 7 shows the results of confusion matrix from the application of the XGBoost, non K-Means, and clickstream methods. In tests carried out 26738 times, there were 3364 true positives, 22226 true negative times, 413 false positives, and 735 false negatives. In this condition, the accuracy is 95.71%.

Table 7: Confusion Matrix of XGBoost+Non K-Means+Clickstream

n = 26738	prediction: 0	prediction: 1
actual: 0	3364	413
actual: 1	735	22226
Accuracy	95,71%	

Table 8 shows the precision, recall, and f1-score results from the application of the XGBoost, non K-Means, and clickstream methods. Where the precision value is 95.90%, recall is 95.71%, and f1-score is 95.78%.

Table 8: Result of XGBoost + Non K-Means + Clickstream

	Value (%)
Accuracy	95,71
Precision	95,90
Recall	95,71
F1-Score	95,78

4.4 XGBoost, Non K-Means, and Non Clickstream

Table 9 shows the results of confusion matrix from the application of the XGBoost, non K-Means, and non clickstream methods. In tests carried out 26738 times, there were 3364 true positives, 22226 true negative times, 413 false

positives, and 735 false negatives. In this condition, the accuracy is 89.44%.

Table 9: Confusion Matrix of XGBoost

n = 26738	prediction: 0	prediction: 1
actual: 0	1582	2195
actual: 1	628	22333
Accuracy	89,44%	

Table 10 shows the precision, recall, and f1-score results from the application of the XGBoost, non K-Means, and non-clickstream methods. Where the precision value is 88.30%, recall is 89.44%, and the f1-score is 88.23%.

Table 10: Result of XGBoost

	Value (%)
Accuracy	89,44
Precision	88,30
Recall	89,44
F1-Score	88,23

5. DISCUSSION

Based on the results of the implementation and evaluation described in section 4, the K-Means and clickstream methods were proven to increase the accuracy rate of product recommendations on the XGBoost method with an accuracy value of 95.72%, a precision value of 95.93%, a recall value of 95.72. %, and an f-1 score of 95.80%. The percentage addition between the XGBoost method and the combination of the XGBoost, K-Means, clickstream methods was quite significant, where the accuracy value increases 0.07%, the precision value increases 0.09%, the recall value increases 0.07%, and the F1-Score value increases 0.09%.

Table 11: Result of All Methods

	XGB+KM+C	XGB+KM	XGB+C	XGB
Accuracy	95,72	89,49	95,71	89,44
Precision	95,93	88,37	95,90	88,30
Recall	95,72	89,49	95,71	89,44
F1-Score	95,80	88,35	95,78	88,23

6. CONCLUSION

Based on the results of the implementation and evaluation described in section 4, it can be concluded that the K-Means method and clickstream data can increase the accuracy of product recommendations. The combination of these methods were able to improve the accuracy of the product recommendation system with an accuracy value of 95.72%, a precision value of 95.93%, a recall value of 95.72%, and an f-1 score of 95.80%.

REFERENCES

1. Lee, J. S., & Lee, J. C. **Context awareness by case-based reasoning in a music recommendation**

system, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4836 LNCS, 2007, pp. 45–58.
https://doi.org/10.1007/978-3-540-76772-5_4

2. Qin, J., Fu, W., Gao, H., & Zheng, W. X. **Distributed k-Means Algorithm and Fuzzy c-Means Algorithm for Sensor Networks Based on Multiagent Consensus Theory**, IEEE Transactions on Cybernetics, 2017, pp. 772–783.

3. Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. **Explaining the user experience of recommender systems**, User Modeling and User-Adapted Interaction, 2012, pp. 441–504.

4. Mobasher, B. **Data Mining for Web Personalization,** Lecture Notes in Computer Science: The Adaptive Web, 2007, pp. 90–135.
https://doi.org/10.1007/978-3-540-72079-9_3

5. Bansal, S., Gupta, C., & Sinha, A. **Clickstream & behavioral analysis with context awareness for e-commercial applications**, 2017 10th International Conference on Contemporary Computing, 2018, pp. 1–6.

6. Xue, W., Xiao, B., & Mu, L. **Intelligent Mining on Purchase Information and Recommendation System for E-Commerce**, 2015 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, 2015, pp. 611-615.

7. Dixit, V. S., Gupta, S., & Jain, P. **A Propound Hybrid Approach for Personalized Online Product Recommendations**, Applied Artificial Intelligence, 2018.

8. Madke, P. N. B., Kulkarni, Y., Mule, R., & Lakade, A. **User Profile Based Behavior Identificaton Using Data Mining Technique**, International Research Journal of Engineering and Technology (IRJET), 2018, pp. 326–331.

9. Su, Q., & Chen, L. **A method for discovering clusters of e-commerce interest patterns using click-stream data**, Electronic Commerce Research and Applications, 2015, pp. 1–13.

10. Utku, A., Aydogan, E., Mutlu, B., & Akcayol, M. A. **A New Recommender System Based on Multiple Parameters and Extended User Behavior Analysis**, Proceedings of the 9th International Conference on Information Management and Engineering - ICIME 2017, 2017, pp. 133–138.

11. Xiao Y., & Ezeife, C.I. **E-Commerce Product Recommendation Using Historical Purchases and Clickstream Data**, In: Ordonez C., Bellatreche L. (eds) Big Data Analytics and Knowledge Discovery. DaWaK 2018. Lecture Notes in Computer Science, vol 11031, 2018.

12. Gupta, S., Ojha, U., & Dixit, V. S. **Personalized web recommendations analyzing sequential behaviour using implicit data streams: A survey**, 8th International Conference on Computing, Communications and Networking Technologies, 2017.
https://doi.org/10.1109/ICCCNT.2017.8203976

13. Kant, S., & Mahara, T. K-**means Atkinson clustering approach for collaborative filtering-based**

recommendation system, International Journal of Intelligent Systems Design and Computing, 2017.

14. Koehn, D., Lessmann, S., & Schaal, M. **Predicting Online Shopping Behaviour from Clickstream Data using Deep Learning**, Expert Systems with Applications, 2020.

15. Li, Q., Baker, R., & Warschauer, M. **Using clickstream data to measure, understand, and support self-regulated learning in online courses**, The Internet and Higher Education, 2020.
https://doi.org/10.1016/j.iheduc.2020.100727

16. Iwanaga, J., Nishimura, N., Sukegawa, N., & Takano, Y. **Improving collaborative filtering recommendations by estimating user preferences from clickstream data**, Electronic Commerce Research and Applications, 2019.

17. Kumar, A., Salo, J., & Li, H. **Stages of User Engagement on Social Commerce Platforms: Analysis with the Navigational Clickstream Data**, International Journal of Electronic Commerce, 2019, pp. 179–211.

18. Chen, Y., & Yao, S. **Sequential search with refinement: Model and application with click-stream data**, Management Science, 2017, pp. 4345-4365.

19. Obiedat, R. **A Comparative Study of Different Data Mining Algorithms with Different Oversampling Techniques in Predicting Online Shopper Behavior**, International Journal of Advanced Trends in Computer Science and Engineering, 2020.
https://doi.org/10.30534/ijatcse/2020/164932020

20. Dixit, V. S., & Gupta, S. **Personalized Recommender Agent for E-Commerce Products Based on Data Mining Techniques**, In Intelligent Systems, Technologies and Applications, 2019, pp. 77-90.

21. Prasetya, C. S. E. **Sistem Rekomendasi Pada E-Commerce Menggunakan K-Nearest Neighbor**, Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK), 2017, pp. 2355–7699.

22. Want, R., Hopper, A., Falcao, V., Gibbons, J. **The Active Badge Location System**, ACM Transactions on Information Systems 10, 1992, pp. 91–102.

23. Ward, A., Jones, A., Hopper, A. **A New Location Technique for the Active Office**, IEEE Personal Communications 4, 1997, pp. 42–47.
https://doi.org/10.1109/98.626982

24. Dey, A.K., Futakawa, M., Salber, D., Abowd, G.D. **The Conference Assistant: Combining Context-Awareness with Wearable Computing**, In: Proceedings Symposium on Wearable Computers, 1999, pp. 21–28.

25. Dey, A.K., Salber, D., Abowd, G.D. **A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications**, Human-Computer Interaction 16, 2001, pp. 97–166.

26. Rawat, M., Goyal, N., & Singh, S. **Advancement of recommender system based on clickstream data using gradient boosting and random forest classifiers,** In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), July 2017, pp. 1-6.

27. Dhanachandra, N., Manglem, K., & Chanu, Y. J. **Image segmentation using K-means clustering algorithm and subtractive clustering algorithm**, Procedia Computer Science, 2015, pp. 764-771.

28. Enireddy, V., Gunda, K., Kalyani, N. L., & Prakash, K. B. **Nature Inspired Binary Grey Wolf Optimizer for Feature Selection in the DETECTION of NEURODEGENERATIVE (PARKINSON) DISEASE**, International Journal of Advanced Trends in Computer Science and Engineering, 2020. https://doi.org/10.30534/ijatcse/2020/222932020

29. Torlay, L., Perrone-Bertolotti, M., Thomas, E., & Baciu, M. **Machine learning–XGBoost analysis of language networks to classify patients with epilepsy**, Brain informatics, 2017, pp. 159.

30. Breiman, L. **Random Forest**, Machine Learning, 2001, pp. 5-32.