Rivista Italiana di Filosofia Analitica Junior Vol 14, No. 1 (2023)
ISSN: 2037-4445

The Studies of the Shopping Behavior of Male and Female Using Dissimilar Clustering Algorithm with Data Mining Tool

Teena Vats[1]

[1]Guru Gobind Singh Indraprastha University, Delhi, India

tina.summer27@gmail.com

(0000-0002-3378-3051)

Dr. Kavita Mittal[2]

[2]Jagannath University NCR, Haryana, India

kavitamittal.it@gmail.com

(0000-0002-2967-0804)

Abstract:

Data mining is a way to deal with mine prized stowed away information, examples and relationshipfrom large and sparse datasets. This way continues through in excess of a couple of methodologies for example order, bunching and connection and so on Grouping is an essential insights mining approach which team equivalent realities protests all in all in a gathering. In this get some answers concerning appraisal is performed with five selective grouping strategies the utilization of five extraordinary datasets. Correlation used to be completed on the foundation of diverse examination boundaries. By conventional results it is inferred that simplek-Mean algorithms are ideal, least difficult, created quality groups and has extreme in general execution among all unique four calculations. Execution of EM calculation is most noticeably awful among all other fourcalculations as it required some investment to deliver off base outcomes. Canopy algorithms and simple k-mean clustering take the same time to build a model. But the parallel k-mean is considered as a best performance algorithm to build the model. In this paper the researcher analysis's the shopping behaviour of male and female. This study assessment and impacts make higher insight for group analyst to work on current systems and furthermore to investigate additional methodologies and to exhort another clustering method.

Keywords: Data Mining, Clustering, simple k-Mean Clustering, EM Clustering, Canopy Clustering, Cobweb, Parallel k-mean**.**

1. INTRODUCTION

Data mining sometimes called data or knowledge discovery process. Data mining discovers knowledge or information that you never knew was present in your data. Usually, the uncovered hidden knowledge manifests itself as relationship or patterns relationship may be between two or more different objects along with time dimension. Relationship may be between the attributes of the same object. Pattern discovery is another outcome of data mining operations. Data *m*ining refers to extracting or mining knowledge from large amount of data .datamining is knowledge mining from databases knowledge extraction, data/pattern analysis, data archaeology and data dredging. Data mining popular name is knowledge discovery in databases or KDD.data mining as an analogyimagines a very wide and very deep pit densely packed with some important material. We use set of sophisticateddrilling tool to dig and unravel the contents.[7]

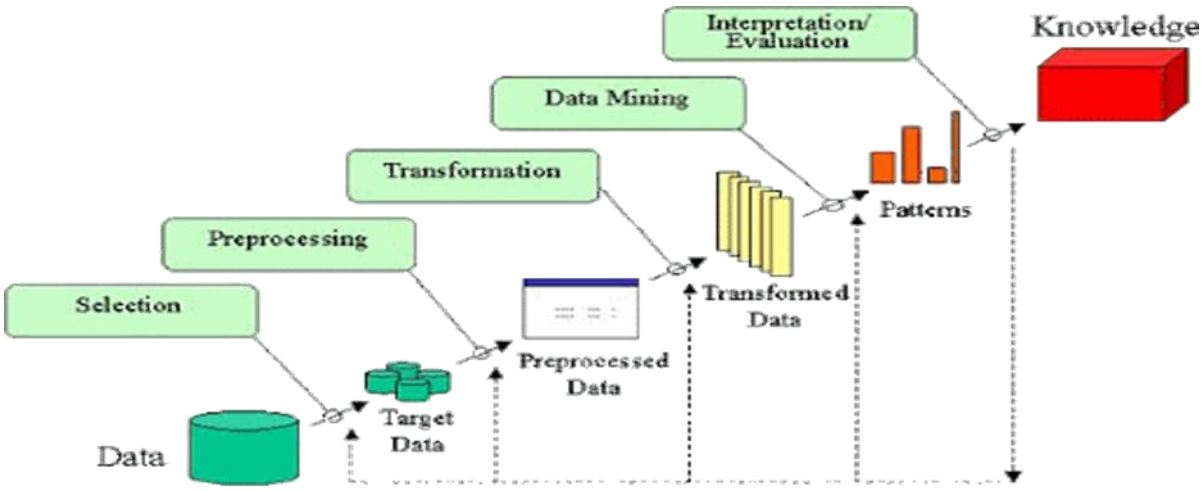


Fig.1.Knowledge discovery process

DATA PREPROCESSING

Information can be introduced from a document in dissimilar establishments: ARFF, CSV, C4.5, similar. Data can similarly be examined from a URL or from a SQL data set (utilizing JDBC). Pre-preparing apparatuses trendyWEKA are named strainers. WEKA contains filters for: Discretization, correction, approaching, quality choice, changing and consolidating ascribes. Information that given to the Weka device might be unstructured or no quality information so information pre-handling is significant. In each field, the information assortment is the significant factor however in the event that the data is insignificant, the immense issue might happen. Those issuesare missing qualities, unimaginable information mix, out of reach esteems. Because of these issues, it might create surprising flaw result. Information arrangement and separating steps can take extensive measure of handling time. Information pre-handling incorporates cleaning, standardization, change, highlight extraction and determination and so on. the result of information pre-handling is the last preparing set. [7, 5]

Clustering

WEKA contains clusterers for discovering gatherings of comparative occurrences in a dataset. Executed plans are: k-Means, EM, Cobweb, X-implies, and Farthest First. Groups can be envisioned and looked at to true bunches (whenever given). Assessment depends on log likelihood, if grouping plan delivers a likelihood conveyance.

Clustering Algorithms: while there are various clustering algorithms but for analysis the data these clustering algorithms are used by the researcher these are canopy, coweb, em, simple k-mean, parallel k-mean Clustering algorithms for analysis the data availability of algorithms is vast but the researcher select these algorithms becausethese algorithms are the major clustering algorithm and every algorithm has its own benefits and drawback. Clustering analysis is applied by using Weka tool, all algorithms used by researcher are available in Weka. [9,12]

Canopy: this Clustering method is unsupervised learning, it's fast and simple, fast as compare to other lustringalgorithms. It's used for grouping objects into clusters. Every object is represented as a point in a multidimensional feature. This algorithm run in pending stage until the original set is blank, gather a set of Canopies, all contain one or more points. [2,14]

Input: A set S of data points $\boldsymbol{x}_i$, thresholds T_1 and T_2
$\mathcal{C} \leftarrow \emptyset$ ▷ $\mathcal{C}$ is a set of canopies
$\Sigma \leftarrow S$ ▷ Σ is a set of center candidates
while $\Sigma \neq \emptyset$ **do**
  $\boldsymbol{c} \leftarrow$ get a point from Σ at random ▷ $\boldsymbol{c}$ is a center
  $C \leftarrow \emptyset$
  for $\boldsymbol{x} \in S$ **do**
    if $d(\boldsymbol{x}, \boldsymbol{c}) \leq T_1$ **then**
      $C \leftarrow C \cup \{\boldsymbol{x}\}$ ▷ a canopy C includes a point $\boldsymbol{x}$
    end
    if $d(\boldsymbol{x}, \boldsymbol{c}) \leq T_2$ **then**
      $\Sigma \leftarrow \Sigma - \{\boldsymbol{x}\}$ ▷ remove $\boldsymbol{x}$ from the candidates
    end
  end
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{C\}$
end
return $\mathcal{C}$

Algorithm Canopy clustering

Cobweb: this method calculated by the incrementally observations into the classification tree. every node in that classification tree representing as a class (concept) and is categorized by a probabilistic concept that recapitulate the attribute-value distributions of the object classify below the node. [5,11]

Input: The current node *N* of the concept hierarchy.
An unclassified (attribute-value) instance *I*.
Results: A concept hierarchy that classifies the instance.
Top-Level Call: Cobweb(Top-node,*I*).
Variables: *C*, *P*, *Q*, and *R* are nodes in the hierarchy.
U, *V*, *W*, and *X* are clustering scores

```
Cobweb(N,I)
    If N is a terminal node Then
        Create-new-terminals(N,I).
        Incorporate(N,I).
    Else
        Incorporate(N,I).
        For each child C of node N
            Compute the score for placing I in C.
        Let P be the node with the highest score W.
        Let R be the node with the second highest score.
        Let X be the score for placing I in a new node Q.
        Let Y be the score for merging P and R into one node.
        Let Z be the score for splitting P into its children.
        If W is the best score Then
            Cobweb(P,I)      ; place I in category P.
        Else If X is the best score Then
            Initialize Q's probabilities using I's values; place I by itself.
        Else If Y is the best score Then
            Let O be Merge(P,R,N).
            Cobweb(O,I).
        Else If Z is the best score Then
            Split(P,N).
            Cobweb(N,I).
```

Algorithm Cobweb Clustering

EM: EM is an iterative method which interchange among two steps first is expectation and second is maximization. For clustering, in EM a finite of Gaussian mixtures model is prepared after that a approximation aset of constraint iteratively awaiting a desired junction value is reach. Iteratively refine

the constraint with E and M steps. [10,12]

1. Initialize a guess for the parameters, call it $\hat{\theta}_0$
2. (**Expectation**) On step j, compute

$$Q(\theta, \hat{\theta}_j) = \mathbb{E}\left[\ell(\theta; X, \Delta)|X, \hat{\theta}_j\right]$$

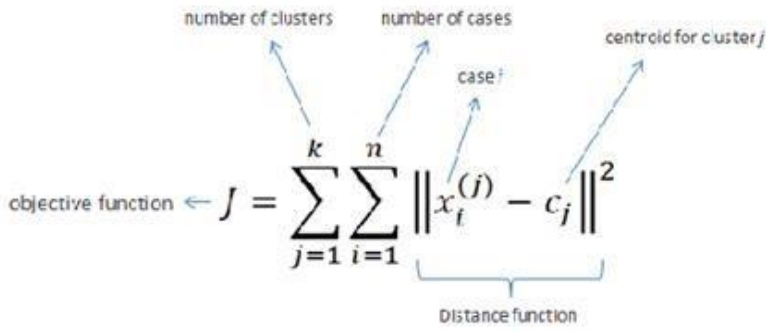
   regarded as a function of θ
3. (**Maximization**) Maximize $Q(\theta, \hat{\theta}_j)$ with regards to θ and call the result $\hat{\theta}_{j+1}$, the new estimate.
4. Loop through steps 2 and 3 until convergence

K-mean: k-mean perform in 3 steps, [2,13]

1. Identify sum of clusters *K*.

2. Reset centroids by initial scuffle the dataset and at that time at random select *K* data points for the centroidslacking alternative.

3. Wait repeating till there is no alter to the centroids. i.e. arrangement of data facts to clusters be situated altering.



Parallel k-mean Clustering: Clustering is a method that divides the data into groups of similar items. In a same manner parallel k-mean algorithms are designed in a mode that each P active node is responsible for conductn*/*P data facts. Scholar runs that program on a Linux Cluster with a supreme of eight nodes using message-passingencoding classical. Steps of Parallel K-Mean Algorithm. [3,6]

Parallel k-means(*DS*,*k*)
1. Select initial *k* centers
2. Repeat
3. parfor *i*=1 to *n*
4. Calculates the distances between the current poin and *k* centers
5. Endparfor
6. Assign each point to its nearest cluster
7. Calculate the new *k* centers
8. Until stable *k* centers reached

Related work:

1. sanjay garg and ramesh Chandra jain [2] in this paper the author discusses various clustering algorithms like k- mean-k-means and k-medoid algorithms. But according to the author k-mean

clustering algorithms is best but h-k-means algorithm is provide over performing as compare two others so that h-k-means algorithms provide best quality results. running time of k-mean is low. The author takes the conflicts of interest from that paper to understand the k-mean method to solve the problem.

2. Pavel Turcinek, Jiri Stastny, Arnost Motycka [3] in this paper the author says that the applications of cluster analysis are suitable for all kinds of attribute but the all methods are not suitable of all kind of data. The author uses various type of clustering algorithms on super market data to analysis the customer behaviour. The author also says that association rule also applies on that data to predict the customer behaviour. The author takes the conflicts of interest from that paper to understand the k-mean system to break the association problem

3. Fazilah Othman, Rosni Abdullah, Nur'Aini Abdul Rashid, and Rosalina Abdul Salam [4] according to the author the main aim of this paper is partition the data into similar groups and make a unwanted data group. parallelK-means algorithm used in this paper. Parallel K-mean algorithm is responsible for p partition and this p partitionis handling for n/p data points. According to author this algorithm is suitable for large data set to improve accuracythat algorithm must be used. The author takes the conflicts of interest from that paper to understand the parallel k- mean algorithms and solve the partition problem in that paper.

4. Basma Jumaa Saleh and Ahmed Yousif Falih Saedi [5] in this paper author predict the heard disease by using data mining algorithms in weka tool. After reading various paper the researcher conclude that accuracies based on chooses numbers of features for testing from the data set. Result concludes on the basis of these algorithms J48, SMO, Naïve Bayes, MLP, Bayes Net REPTREE, K-star. The author takes the conflicts of interest from thatpaper to understand the various algorithms that run in weka because weka tool used in that paper to analysis the data.

5. Waseem Ahmed ,2 Muhammad Rafiq Kakar [6] author says that parallel k-mean algorithm is an effective and feasible solution to remove the percentage of failure in results in various area. The author takes the conflicts of interest from that paper to understand the k-mean system to understand the percentage problem that is used in thatpaper.

6. Vijayakumar. M 1, R. Porkodi [8] in this paper the author says that when the researcher uses both the algorithmsApriori and Predictive Apriori provide better results. In this paper author analysis, the whole sale data by using above algorithms. The author takes the conflicts of interest from that paper to understand the apriori system to break the association problem.

7. Mrs.Yogita Bhapkar and Dr. Ajit More [9] in this paper The Researcher use variation of k-mean algorithms onbank data in weka tool. In this paper dendrogram forms are used to represent the result of hierarchical clustering but this foam suitable for the small data sets hard to big data sets. The author takes the conflicts of interest from that paper to understand the k-mean system to solve bank data problem by using weka tool.

Data set Description: researcher use the primary data, conduct by questionnaire. Questionnaire contains various attribute but conclude all attribute is not possible so that researcher select some selected attribute. Researcher includes 400 samples from the whole population. This sample contain 5 attribute like User-ID, Gender, Age, Estimated Salary, Purchased value.

Data	Customer behaviour
Instances	400
Attribute	5
Type of attribute	Numeric
Name of attribute	User_id,gender,age,estimated salary and purchasedvalue

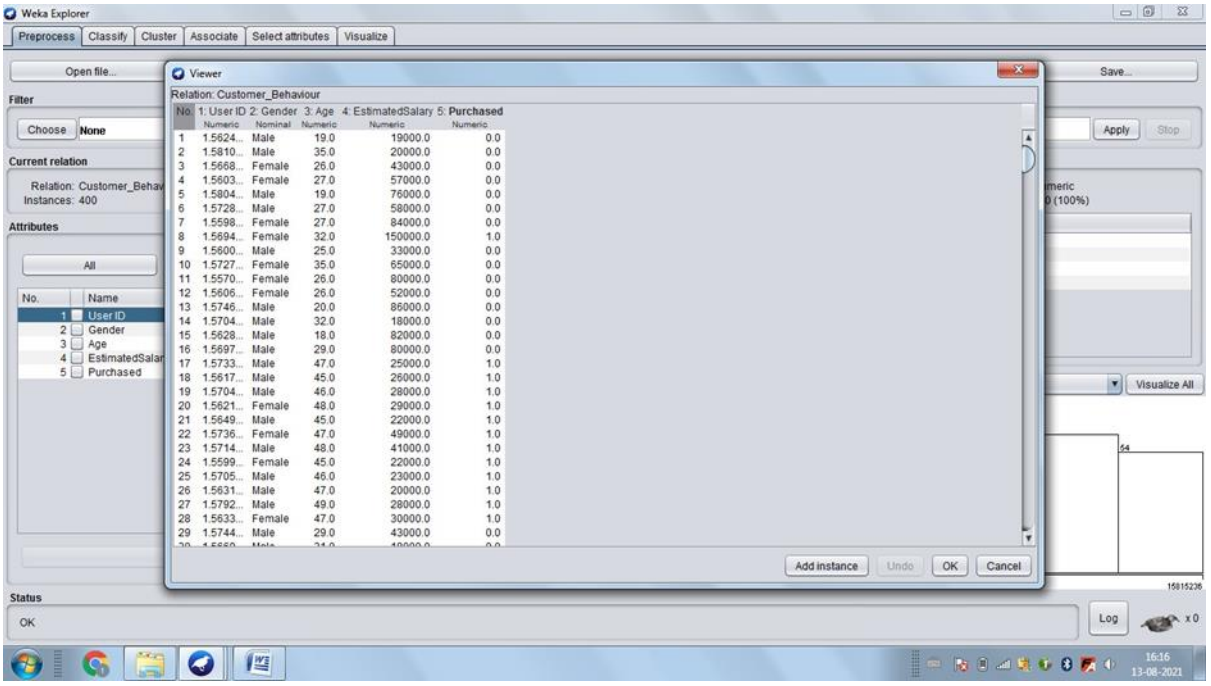


Fig.2. customer data in Weka

Data in figure 2 show the data that is used for analysis in Weka. That data taken from the Kaggle repository. Thisdata contains five attribute customer id, gender, age, estimate salary, purchased value. This data contains records of 400 customers that purchase goods in November 2019 to May 2020.kaggle provide data in various format likeCSV and ARFF. Research finds that data in CSV format then change it in ARFF format for analysis that data andtake out the results. [1,9]

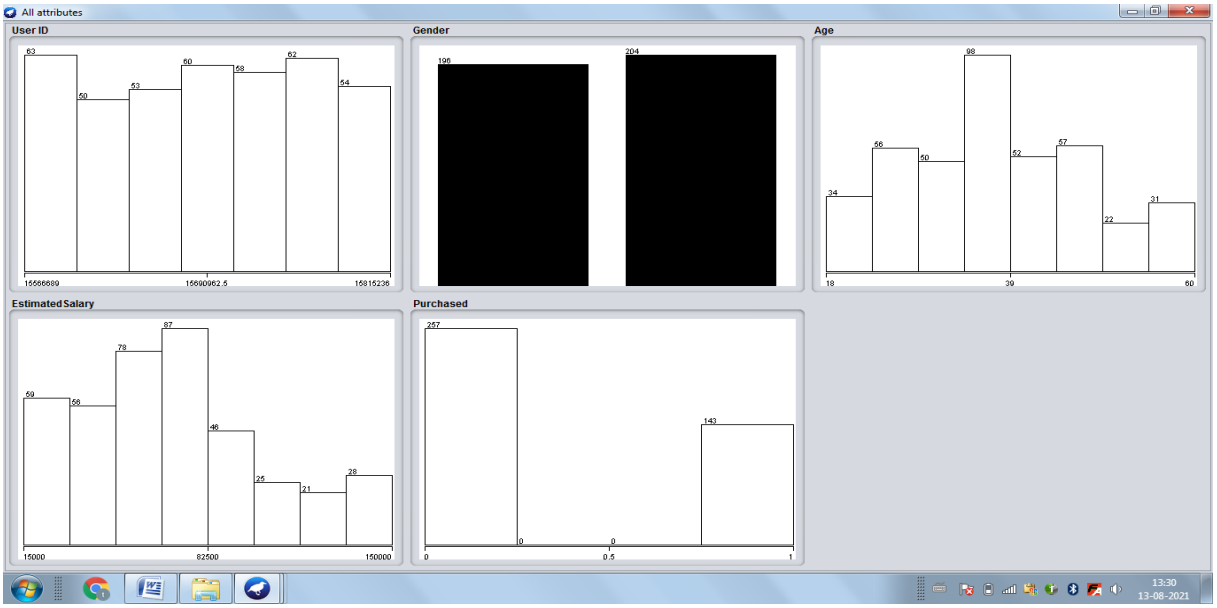


Fig.3. Visualization

Results calculated on basis of Canopy algo, Cobweb algo, EM algo, Simple k-mean and parallel k-mean algo.

Clustering Algo	Time taken to build a model	No. Of cluster	Purchased percentage (%) by male	Purchased percentage (%) by female
Canopy	0.02	6	49	51
Cobweb	0.47	458	48	49
EM	1.14	7	33	37
Simple k-mean	0.02	2	33	37
Parallel k-mean	0.01	2	31	35

Table.1.

On the basis of results the researcher conclude that higher shopping done by the female customer but window shopping done by the male customer. So that researcher concludes the customer behaviour on the basis of purchased value 0(window shopping) show no purchased and 1(shopping) show purchased in above data. Resultcalculated on the basis of the clustering algorithms.

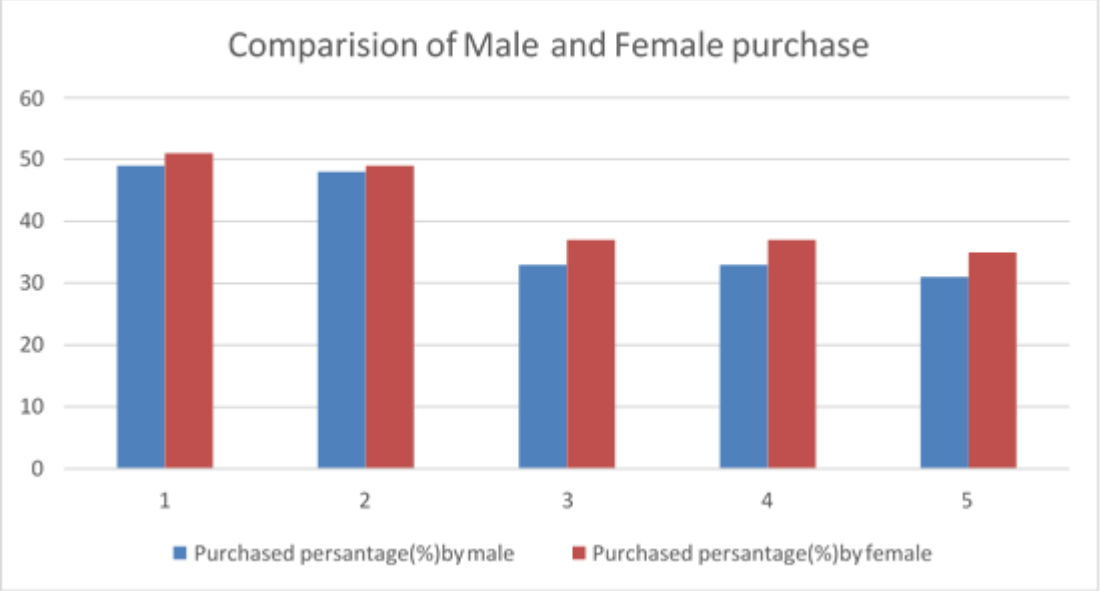


Fig.4. male and female purchasing

5. Conclusion

In this research, correlation of five diverse grouping calculations which incorporates ('Canopy', EM', 'Cobweb', 'Parallel k-mean', 'simple k-Means') has been performed utilizing five distinctive datasets. Calculations are thoughtabout based on time taken to build the model, number of clusters created, purchased percentage (%) by male, Purchased percentage (%) by female. From results it is reasoned that presentation of Parallel k-Mean calculationis best among any remaining four calculations to build a model. Execution of EM calculation is most noticeably awful among any remaining four calculations as it required some investment to create wrong outcomes. Various levelled calculation is touchy to measure of information. It is ideal for little datasets yet on gigantic datasets it requires some investment as contrast with different calculations. Execution of two algorithms k-mean and Canopyis same to build the model. Performance of cobweb isn't so much effective. In general end can be made that parallel k-Mean calculation is most straightforward, created quality groups and has elite among any remaining four calculations. Researcher additionally analysed that study outcomes with current results and established that our outcomes are very practical and exact as there was minor difference between them. Scholar projected researchand outcomes improve understanding that more shopping is done by woman as compare to man. But the more window shopping is done by man as compare to female according to data. Trendy Future, inspection and examination of other grouping policies will be performed and

outcomes will be thought about with present resultsfor better tolerant and study.

The manuscript in part or full has not been submitted or published anywhere. The data is data secondary dataand Data has been taken from Kaggle repository.

This research paper doesn't have any grant from any company.

"Conflict of Interest: The authors declare that they have no conflict of interest."

Reference

1. Sharma, S., Ahmed, A., Naseem, M., & Sharma, S. (2021). Machine Learning for Solving a Plethora ofInternet of Things Problems. In *Intelligent Communication and Automation Systems* (pp. 255-274). CRCPress.
2. Garg, S. (2006). Variation of k-mean algorithm: a study for high-dimensional large data set. *InformationTechnology Journal*, *5*, 1132-1135.
3. Turcinek, P. A. V. E. L., Stastny, J. I. R. I., & Motycka, A. R. N. O. S. T. (2012). Usage of cluster analysisin consumer behaviour research. In *Proceedings of the 12th WSEAS International Conference on Applied Informatics and Communications (AIC '12)* (pp. 172-177).
4. Othman, F., Abdullah, R., & Salam, R. A. (2004, December). Parallel k-means clustering algorithm on DNA dataset. In *International Conference on Parallel and Distributed Computing: Applications and Technologies* (pp. 248-251). Springer, Berlin, Heidelberg.
5. Saleh, B. J., Saedi, A. Y. F., Al-aqbi, A. T. Q., & Salman, L. A. (2020). A review paper: analysis of wekadata mining techniques for heart disease prediction system. *Library Philosophy and Practice*, 1-17.
6. Akhtar, M. N., Ahmed, W., Kakar, M. R., Bakar, E. A., Othman, A. R., & Bueno, M. (2020). Implementation of Parallel K-Means Algorithm to Estimate Adhesion Failure in Warm Mix Asphalt. *Advances in Civil Engineering*, *2020*.
7. Ponyiam, P., & Arch-int, S. (2018, September). Customer Behavior Analysis Using Data Mining Techniques. In *2018 International Seminar on Application for Technology of Information and Communication* (pp. 549-554). IEEE.
8. Agapito, G., Calabrese, B., Guzzi, P. H., & Cannataro, M. (2019, January). A pipeline for mining association rules from large datasets of retailers invoices. In *Proceedings of the 2nd International Conference on Applications of Intelligent Systems* (pp. 1-6).
9. Bhapkar, M. Y., & More, A. Comparative analysis of clustering algorithms for the study of home loan applicants using WEKA tool.
10. Wang, Z., Gu, Q., Ning, Y., & Liu, H. (2015). High dimensional em algorithm: Statistical optimization and asymptotic normality. *Advances in neural information processing systems*, *28*, 2512.
11. Mulani, N., Pawar, A., Mulay, P., & Dani, A. (2015). Variant of COBWEB clustering for privacy preservation in cloud DB querying. *Procedia Computer Science*, *50*, 363-368.
12. Hussain Shah, S., Javed Iqbal, M., Bakhsh, M., & Iqbal, A. (2020). Analysis of Different Clustering Algorithms for Accurate Knowledge Extraction from Popular DataSets. *Information Sciences Letters*, *9*(1), 4.
13. Cleuziou, G. (2008, December). An extended version of the k-means method for overlapping clustering.In *2008 19th International Conference on Pattern Recognition* (pp. 1-4). IEEE.
14. Sagheer, N. S., & Yousif, S. A. (2021, March). Canopy with k-means clustering algorithm for big data analytics. In *AIP Conference Proceedings* (Vol. 2334, No. 1, p. 070006). AIP Publishing LLC.