OPTIMIZATION OF THE FUZZY C-MEANS CLUSTER CENTER FOR CREDIT DATA GROUPING USING GENETIC ALGORITHMS

Dicky Apdilah[1], Oris Krianto Sulaiman[2], Indah Purnama Sari[3]
[1]Department of Informatic Engineering, Universitas Asahan, Indonesia
[2]Department of Informatic Engineering, Universitas Islam Sumatera Utara, Indonesia
[3]Department of Information Technology, Universitas Muhammadiyah Sumatera Utara, Indonesia

ABSTRACT

Data grouping can be used in the marketing strategy of a product. The process of grouping data that previously behaved differently into groups that now behave more uniformly. As with the grouping of creditworthiness assessment data, this data grouping is needed to obtain the dominant values that will characterize each group or segment. The clustering method is quite widely used to overcome problems related to data segmentation. Clustering is a grouping method based on a measure of proximity, the more accurately the clusters are formed, the clearer the similarity of customer behavior patterns will be. Thus, companies can determine marketing strategies more precisely, based on customer behavior patterns. One of the clustering methods that can be used to group data is Fuzzy C-Means (FCM), which is a method of grouping data determined by the degree of membership. Optimization by presenting a Genetic Algorithm to obtain test data cluster results regarding the grouping of credit data. The purpose of this study is to examine the application of the Genetic Algorithm in fuzzy clustering, especially Fuzzy C-Means, and to examine the extent to which the Genetic Algorithm can improve the performance of Fuzzy C-Means in optimizing cluster centers to obtain grouping of customer data which will later be used for assessing creditworthiness.

Keyword : Data Grouping, Clustering, Fuzzy C-Means, Genetic Algorithms

Corresponding Author:
Dicky Apdilah,
Department of Informatic Engineering,
Universitas Asahan,
Jalan Jend. Ahmad Yani, Kisaran Naga, Kec. Kota Kisaran Timur, Kisaran, Sumatera Utara 21216, Indonesia.
Email: dickyapdi1404@gmail.com

1. INTRODUCTION

Data segmentation is a grouping of data that can later be selected as a target for marketing a product. Data segmentation is also the process of grouping data that previously behaved differently into several market groups that now behave more uniformly. The grouping of data in companies should be further analyzed and utilized using the Clustering method. This analysis provides many outputs that companies can take advantage of. All of these outputs can be used by the company to win the competition or increase the company's income and turnover. As is the case with the credit worthiness assessment data grouping consisting of 1,000 records and 20 variables plus a target variable or response variable. This data grouping is needed to get the dominant values that will become the characteristics of each group or segment.

The clustering method is quite widely used to overcome problems related to data segmentation. This clustering is a grouping method based on a measure of proximity (similarity) where clusters do not have to be exactly the same but are groupings based on the closeness of a characteristic of an existing data sample, one of which is by using the ecluidean distance formula. The application of the clustering method can be used to group customers who have similarities in customer shopping behavior. The more accurately the clusters are formed, the clearer the similarity of customer behavior patterns will be. Thus, companies can determine marketing strategies more precisely, based on customer behavior patterns that have been obtained from the cluster processes that have been formed.

One of the clustering methods that can be used to group data is Fuzzy C-Means (FCM), which is a data grouping technique where the existence of each data point in a group (cluster) is determined by

the degree of membership. By repairing the cluster center and the membership value of each data repeatedly, it will be obtained that the cluster center is heading to the right location [1]. The initial value of the Fuzzy C-Means cluster center point is generated randomly so that a local optimum occurs, where the next process depends on the initial value generated randomly, here Genetic Algorithm will be used to optimize the cluster center value.

Optimization by presenting the Genetic Algorithm will obtain the results of the test data clusters. Genetic Algorithms are very appropriate for solving complex optimization problems that are difficult to solve using conventional methods. Genetic algorithm (GA) as an optimization technique can be applied to optimization-based clustering. In the GA approach for fuzzy clustering the fitness function is taken from the minimized objective function. In the Genetic Algorithm approach, in each generation, chromosomes are evaluated based on the value of the fitness function, to find the cluster center is done by evolving the cluster center matrix (selection, crossover and mutation) through the fitness function using the objective function contained in Fuzzy C-Means.

Based on the discussion above, this study proposes Optimization of the Fuzzy C-Means Cluster Center for Grouping Credit Data using a Genetic Algorithm. The Fuzzy C-Means method was chosen because it can determine the number of clusters to be formed. The purpose of this study is to examine the application of Genetic Algorithms in fuzzy clustering, especially Fuzzy C-Means, and to examine the extent to which Genetic Algorithms can improve the performance of Fuzzy C-Means in optimizing cluster centers in classification problems for assessing creditworthiness. So that this information can later be followed up as material for consideration for decision making and it is also possible to carry out a marketing approach that is in accordance with the dominant characteristics of the groups formed.

2. RESEARCH METHOD/MATERIAL AND METHOD/LETERATURE REVIEW

A. State of the Art

Several previous studies related to the title raised:

Research entitled Student Grouping Using Adaptive Genetic Algorithm by Putu Indah Ciptayani, Kadek Cahya Dewi, I Wayan Budi Sentana, 2016. This study describes learning in groups by forming several groups of students, the best combination of students in a group will give the best results in learning. To get the best combination, it is necessary to evaluate all existing solutions. Adaptive genetic algorithm is used in this paper to find the best group formation. Adaptive population size, crossover probability, and mutation rate are applied in this paper based on the fitness achieved in each generation. This paper presents how adaptive genetic algorithms find solutions to these problems [1].

Research entitled Application of Genetic Algorithms to Maximize Profits in Hijab Production by Samaher and Wayan Firdaus Mahmudy in 2015. This research is related to the maximum profit from the production process in the industry. However, it is said that income is limited by the availability of production materials and investment funds. The head of the production department must determine the amount of each type of product (hijab) to gain a temporary advantage by considering various production constraints. This study proposes a genetic algorithm to obtain the appropriate production quantity. Computational experiments were carried out to get the best parameters for the Genetic Algorithm. Using the best parameters, the proposed algorithm produces a combination of product types that must be produced with maximum profit and with the lowest difficulty [2].

B. Test Data

This research will later use test data that is widely used in classification problems for assessing creditworthiness called the German Credit Dataset, this data set was donated by Prof. Hofman from Hamburg University, Germany. This dataset consists of 1000 records and 20 variables plus a target variable or response variable, of which 13 variables are of the categorical type and the remaining 7 variables are of the numeric type. The German Credit Dataset can be downloaded at the UCI Machine Learning Repository.

The case study raised in this research is about market segmentation. The definition of marketing provides a more detailed definition by saying that market segmentation is the process of dividing markets that previously behaved heterogeneously into several market groups that now behave more uniformly [1].

Determination of market segmentation is based on several criteria including: Demographic segmentation is based on population characteristics that can be measured by age, gender, income, education, and occupation. Psychographic Segmentation is the process of grouping people in terms of attitudes, espoused values, and lifestyles. Behavioral segmentation focuses on whether people will buy and use a product or not, as well as how often and how much they use. Consumers can be categorized according to usage levels, for example, heavy, medium, light users and non-users. Benefit segmentation focuses on the value equation [3].

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	A11	6	A34	A43	1169	A65	A75	4	A93	A101	4	A121	67	A143	A152	2	A173	1	A192	A201	1
2	A12	48	A32	A43	5951	A61	A73	2	A92	A101	2	A121	22	A143	A152	1	A173	1	A191	A201	2
3	A14	12	A34	A46	2096	A61	A74	2	A93	A101	3	A121	49	A143	A152	1	A172	2	A191	A201	1
4	A11	42	A32	A42	7882	A61	A74	2	A93	A103	4	A122	45	A143	A153	1	A173	2	A191	A201	1
5	A11	24	A33	A40	4870	A61	A73	3	A93	A101	4	A124	53	A143	A153	2	A173	2	A191	A201	2
6	A14	36	A32	A46	9055	A65	A73	2	A93	A101	4	A124	35	A143	A153	1	A172	2	A192	A201	1
7	A14	24	A32	A42	2835	A63	A75	3	A93	A101	4	A122	53	A143	A152	1	A173	1	A191	A201	1
8	A12	36	A32	A41	6948	A61	A73	2	A93	A101	2	A123	35	A143	A151	1	A174	1	A192	A201	1
9	A14	12	A32	A43	3059	A64	A74	2	A91	A101	4	A121	61	A143	A152	1	A172	1	A191	A201	1
10	A12	30	A34	A40	5234	A61	A71	4	A94	A101	2	A123	28	A143	A152	2	A174	1	A191	A201	2
11	A12	12	A32	A40	1295	A61	A72	3	A92	A101	1	A123	25	A143	A151	1	A173	1	A191	A201	2
12	A11	48	A32	A49	4308	A61	A72	3	A92	A101	4	A122	24	A143	A151	1	A173	1	A191	A201	2
13	A12	12	A32	A43	1567	A61	A73	1	A92	A101	1	A123	22	A143	A152	1	A173	1	A192	A201	1
14	A11	24	A34	A40	1199	A61	A75	4	A93	A101	4	A123	60	A143	A152	2	A172	1	A191	A201	2
15	A11	15	A32	A40	1403	A61	A73	2	A92	A101	4	A123	28	A143	A151	1	A173	1	A191	A201	1
16	A11	24	A32	A43	1282	A62	A73	4	A92	A101	2	A123	32	A143	A152	1	A172	1	A191	A201	2
17	A14	24	A34	A43	2424	A65	A75	4	A93	A101	4	A122	53	A143	A152	2	A173	1	A191	A201	1
18	A11	30	A30	A49	8072	A65	A72	2	A93	A101	3	A123	25	A141	A152	3	A173	1	A191	A201	1
19	A12	24	A32	A41	12579	A61	A75	4	A92	A101	2	A124	44	A143	A153	1	A174	1	A192	A201	2
20	A14	24	A32	A43	3430	A63	A75	3	A93	A101	2	A123	31	A143	A152	1	A173	2	A192	A201	1
21	A14	9	A34	A40	2134	A61	A73	4	A93	A101	4	A123	48	A143	A152	3	A173	1	A192	A201	1
22	A11	6	A32	A43	2647	A63	A73	2	A93	A101	3	A121	44	A143	A151	1	A173	2	A191	A201	1
23	A11	10	A34	A40	2241	A61	A72	1	A93	A101	3	A121	48	A143	A151	2	A172	2	A191	A202	1

Figure 1. Test Data

C. Analysis Flow

Research is a systematic, controlled, empirical and critical investigation of a hypothesis proposal about certain relationships between phenomena. Research here aims to contribute to science knowledge in solving problems using appropriate methods. The stages of research methods that will be carried out by the author for making this research, including the following:

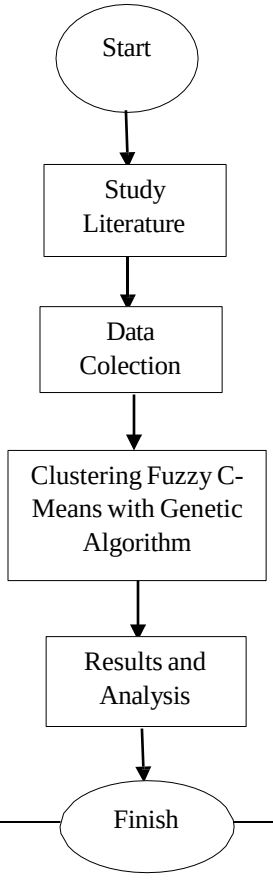

Figure 2. Analysis Flow

The stages of the research analysis flow that will be carried out include the following:

1. The benefits of a literature study can provide a comprehensive picture of the extent of progress studies related to the research to be taken. In this study references were obtained from journals, research report articles, and books related to Fuzzy C-Means research for data grouping.
2. Data Collection. Based on the form and nature, the research data used is mixed, namely qualitative data (in the form of words/sentences) and quantitative data (in the form of numbers). Quantitative data can be grouped based on how to get it, namely discrete data and continuum data. Based on By its nature, quantitative data consists of nominal data, ordinal data, interval data and ratio data.
3. The method proposed in this research is a combination of Genetic Algorithm and Fuzzy C-Means methods Which. Where in this study will try to form clusters on mixed-type data namely the determination of the provision of credit that is optimized with the Genetic Algorithm, thereby optimizing the results clusters.
4. Results and analysis stages are evaluation stages that measure the results obtained from the results of previous studies. The expected end result of research regarding application the Fuzzy C-Means method in grouping data is to obtain credit data clusters.

3. RESULTS AND DISCUSSION

A. Cluster Analysis

In the cluster analysis the cluster validity method used is [4]:

a. Correlation
   Using 2 types of matrices, namely proximity matrix and incidence matrix
   1) Proximity matrix is a matrix that contains the distance between objects
   2) Incidence matrix is a binary matrix that indicates cluster membership
      I. 0 if members of different clusters
      II. 1 if members of the same cluster

b. Cohesion and Separation
   Cohession is a cluster validity that calculates the intra-cluster variance (WSS).

$$WSS = \sum_{i=1}^{k} \sum_{x \in Ci} (x - mi)^2 \quad (1)$$

   While Separation is a cluster validity that calculates the inter-cluster variance (BSS).

$$BSS = \sum_{i=1}^{k} |Ci|(m - mi)^2 \quad (2)$$

c. Silhouette Coefficient
   Using a combination of the two basic principles of cohesion and separation.

$$s = \begin{cases} 1 - \frac{a}{b}, & jika\ a < b \\ \frac{b}{a} - 1, & jika\ a \geq b \end{cases} \quad (3)$$

   Where:
   a = average distance to objects in one cluster
   b = minimum average distance to objects in different clusters

d. Dunn Index
   Has the premise that a good cluster is one that has a small diameter and a large distance to other clusters.

$$D = \min_{i=1\ldots nc} \left(\min_{j=i+1\ldots nc} \left(\frac{d(ci,cj)}{\max_{k=1\ldots nc}(diam(ck))} \right) \right) \quad (4)$$

   Where:
   1) d(ci, cj) is the distance to other clusters
   2) silence(ck) is the distance to fellow clusters

e. Davies-Bouldin Index
Having basic principles in calculating similarity between clusters

$$DB = \left(\frac{1}{nc}\right) \sum_{i=1}^{nc} \left(\max_{\substack{j=1\ldots nc \\ i=j}} (Rij) \right) \quad (5)$$

where value :

$$Rij = \frac{si+sj}{dij} \quad (6)$$

a. si is the average object distance of all cluster i to its center
b. dij is the distance between cluster center i and cluster center j

B. System Implementation

System implementation is carried out after the system design stage. As for the results of system implementation about Optimization of Fuzzy C-Means Cluster Centers for Clustering Credit Data Using Algorithms Genetics can be seen in figure 3.

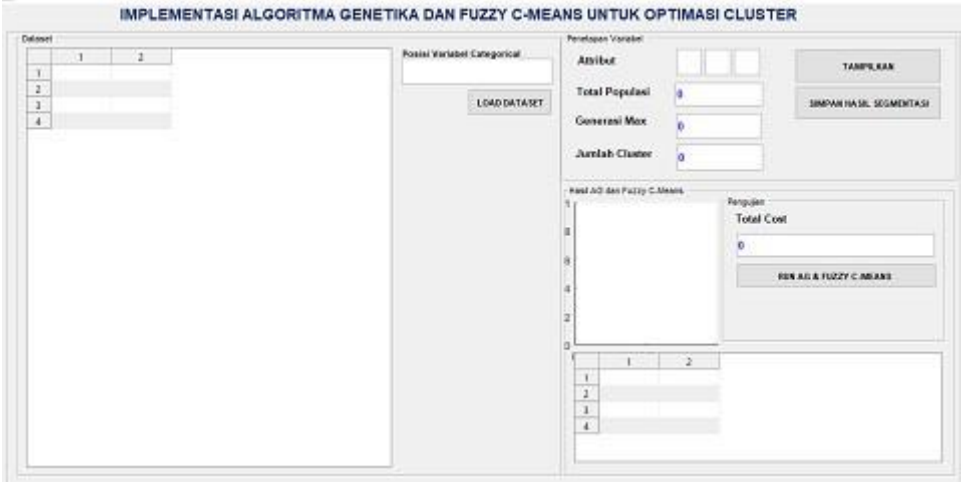

Figure 3. Main Page

Figure 4 shows the input dataset in the form of credit data, where the input data is of the data type mixed namely numeric and categorical. A categorical data position will appear to indicate the data position the.

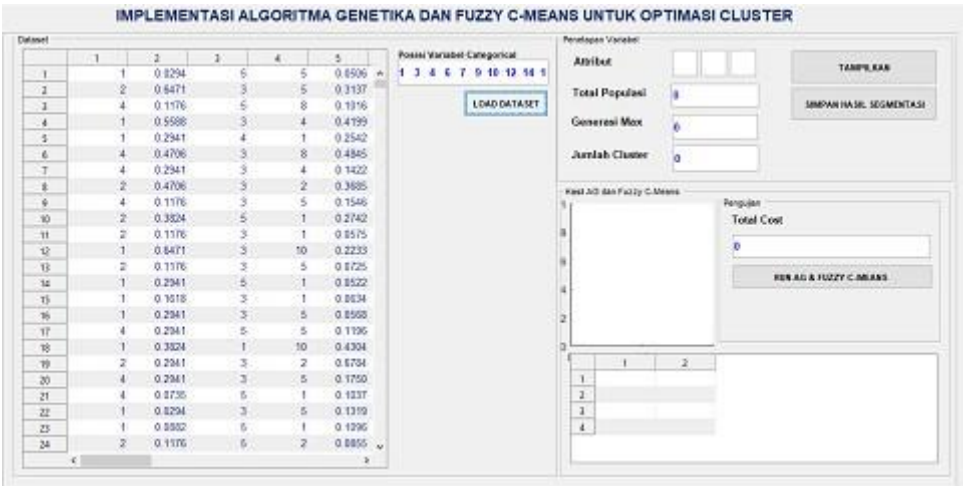

Figure 4. Dataset input

Figure 5 is a display of variable assignment by inputting 3 attributes, where attributes

is the position of the selected data column from 1 to 20 columns. By filling in the total free filled population with numbers that will later affect the length of the process and Generation Max which represents the total population which will produce chromosomes and finally fill in the column for the number of clusters specified, namely 4 clusters.

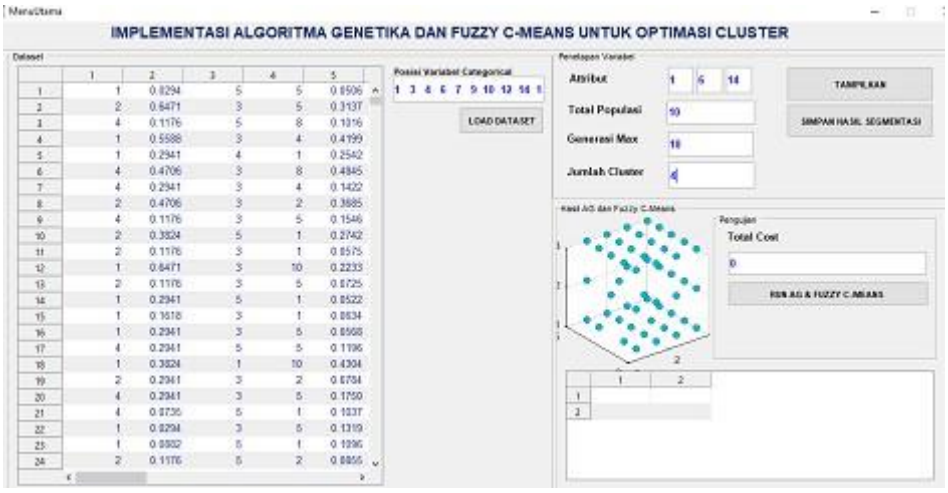

Figure 5. Variable Assignment

After the variable determination process, three-dimensional graphic results will be obtained to describe the clusters formed. From the graph, a table of cluster centers formed from clusters that have been generated with the total cost which is the total distance between the clusters formed will also appear.

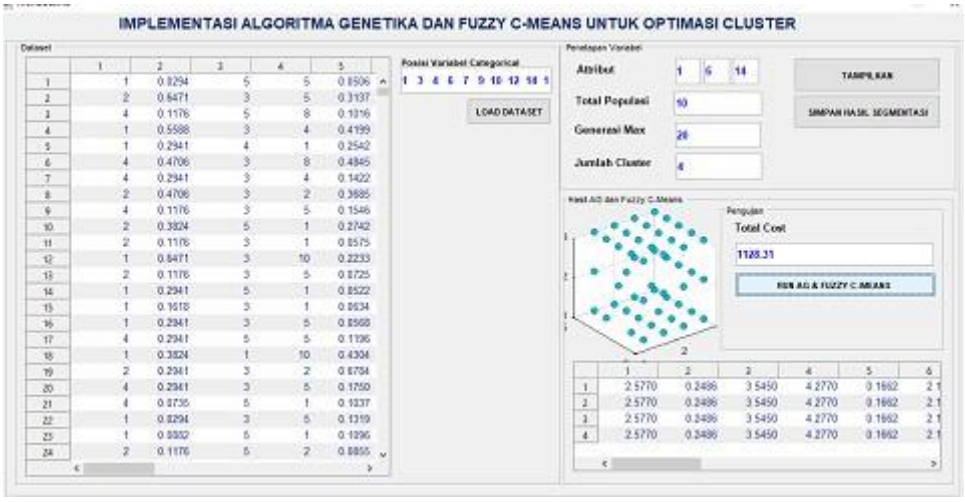

Figure 6. Process Results of AG and Fuzzy C-Means

4. CONCLUSION

The conclusion that can be drawn from this study is that the Fuzzy C-Means method with Genetic Algorithms has been successfully applied to produce cluster center optimization regarding credit data grouping, where this test data is widely used in classification problems for assessing creditworthiness called the German Credit Dataset. This system can assist companies in making decisions and also allows for a marketing approach that is in accordance with the dominant characteristics of the formed groups.

REFERENCES

[1] Sari, I.P., Al-Khowarizmi, A., & Batubara, I.H (2021). Cluster Analysis Using K-Means Algorithm and Fuzzy C-Means Clustering For Grouping Students' Abilities In Online Learning Process*. Journal of Computer Science, Information Technology and Telecommunication Engineering*, 2(1), 139-144.

[2] Erwin, 2009. Analisis Market Basket Dengan Algoritma Apriori dan FPGrowth. *Jurnal Generic*, vol. 4.

[3] Sari, I.P., Batubara, I.H., & Al-Khowarizmi, A (2021). Sensitivity Of Obtaining Errors In The Combination Of Fuzzy And Neural Networks For Conducting Student Assessment On E-Learning*. International Journal of Economic, Technology and Social Sciences (Injects)*, 2(1), 331-338.

[4] Sari, I.P., Fahroza, M.F., Mufit, M.I., & Qathrunad, I.F (2021). Implementation of Dijkstra's Algorithm to Determine the Shortest Route in a City. *Journal of Computer Science, Information Technology and Telecommunication Engineering, 2(1)*, 134-138.

[5] Batubara, I.H., Saragih, S., Syahputra, E., Armanto, D., Sari, I.P., Lubis,B.S., & Siregar, E.F.S (2022). Mapping Research Developments on Mathematics Communication: Bibliometric Study by VosViewer. *AL-ISHLAH: Jurnal Pendidikan* 14(3), 2637-2648.

[6] Sari, I.P., Al-Khowarizmi, A.K., & Batubara, I.H. (2021). Analisa Sistem Kendali Pemanfaatan Raspberry Pi sebagai Server Web untuk Pengontrol Arus Listrik Jarak Jauh. InfoTekJar: *Jurnal Nasional Informatika dan Teknologi Jaringan,* 6 (1), 99-103.

[7] Hariani.,P.P, Sari.,I.P, & Batubara., I.H. (2021). Implementasi e-Financial Report BUMDes. *IHSAN: JURNAL PENGABDIAN MASYARAKAT,* 3 (2), 169-177.

[8] Sari, I.P., Basri, Mhd., Ramadhani, F., & Manurung, A.A. (2023). Penerapan Palang Pintu Otomatis Jarak Jauh Berbasis RFID di Perumahan. Blend Sains Jurnal Teknik, 2(1), 16-25.

[9] Batubara, I.H., & Sari, I.P. (2021). Penggunaan software geogebra untuk meningkatkan kemampuan pemecahan masalah matematis mahasiswa. *Scenario (Seminar of Social Sciences Engineering and Humaniora)*, 398-406

[10] Sari, I.P., & Batubara, I.H. (2020). Aplikasi Berbasis Teknologi Raspberry Pi Dalam Manajemen Kehadiran Siswa Berbasis Pengenalan Wajah*. JMP-DMT* 1(4), 6.

[11] Sari, I.P., Al-Khowarizmi, A.K., Ramadhani, F., & Sulaiman, O.K. (2023). Implementation of the Selection Sort Algorithm to Sort Data in PHP Programming Language. *Journal of Computer Science, Information Technology and Telecommunication Engineering,* 4(1).

[12] Batubara, I.H., Sari,I.P., Hariani, P.P., Saragih, M., Novita, A., Lubis, B.S., & Siregar, E.F.S. (2021). Pelatihan Software Geogebra untuk Meningkatkan Kualitas Pembelajaran Matematika SMP Free Methodist 2. *Martabe: Jurnal Pengabdian Kepada Masyarakat,* 4(3), 854-859.

[13] Sari., I.P, Batubara., I.P, Al-Khowarizmi., A, & PP Hariani. (2022). Perancangan Sistem Informasi Pengelolaan Arsip Digital Berbasis Web untuk Mengatur Sistem Kearsipan di SMK Tri Karya. *Wahana Jurnal Pengabdian kepada Masyarakat* 1 (1), 18-24.

[14] Batubara., I.H, Sari., I.P, EFS Siregar, & BS Lubis. (2021). Meningkatkan Kemampuan Penalaran Matematika Melalui Metode Penemuan Terpandu Berbantuan Software Autograph. *Seminar Nasional Teknologi Edukasi Sosial dan Humaniora* 1 (1), 699-705.

[15] Sari., I.P, A Syahputra, N Zaky, RU Sibuea, & Z Zakhir. (2022). Perancangan sistem aplikasi penjualan dan layanan jasa laundry sepatu berbasis website. *Blend sains jurnal teknik* 1 (1), 31-37.

[16] Sari., I.P, A Azzahrah, FQ Isnaini, L Nurkumala, & A Thamita. (2022). Perancangan sistem absensi pegawai kantoran secara online pada website berbasis HTML dan CSS. *Blend sains jurnal teknik* 1 (1), 8-15.

[17] Ramadhani., F, & Sari., I.P. (2021). Pemanfaatan Aplikasi Online dalam Digitalisasi Pasar Tradisional di Medan. *Prosiding Seminar Nasional Kewirausahaan* 2 (1), 806-811.

[18] Sari.,I.P, & Ramadhani., F. (2021). Pengaruh Teknologi Informasi Terhadap Kewirausahaan Pada Aplikasi Perancangan Jual Beli Jamu Berbasis WEB. *Prosiding Seminar Nasional Kewirausahaan* 2 (1), 874-878.

[19] Sari., I.P, A Jannah, AM Meuraxa, A Syahfitri, & R Omar. (2022). Perancangan Sistem Informasi Penginputan Database Mahasiswa Berbasis Web. *Hello World Jurnal Ilmu Komputer* 1 (2), 106-110.

[20] Hutasuhut, B.K., Sari, I.P., & Al-Khowarizmi, A (2023). Analysis the Effect of Digitalization and Technology on Web-Based Entrepreneurship*. Journal of Computer Science, Information Technology and Telecommunication Engineering* 4(1).

[21] Sari., I.P, & Batubara., I.H. (2021). Perancangan Sistem Informasi Laporan Keuangan Pada Apotek Menggunakan Algoritma K-NN. *Seminar Nasional Teknologi Edukasi dan Humaniora (SiNTESa)* 1 (2021 - ke 1.

[22] Ramadhani., F, A Satria, & Sari., I.P. (2022). Aplikasi Internet Berbasis Website sebagai E-Commerce Penjualan Komponen Sport Car. *Blend Sains Jurnal Teknik* 1 (2), 69-75.

[23] Sari., I.P, & Batubara., I.H. (2021). User Interface Information System for Using Account Services (Joint Account) WEB-Based*. International Journal of Economic, Technology and Social Sciences (Injects*), 462-469.

[24] PP Hariani, Sari., I.P, & Batubara., I.H. (2021). Android-Based Financial Statement Presentation Model*. JURNAL TARBIYAH* 28 (2), 1-16.

[25] Sari., I.P, Batubara., I.H, & M Basri. (2022). Implementasi Internet of Things Berbasis Website dalam Pemesanan Jasa Rumah Service Teknisi Komputer dan Jaringan Komputer. *Blend Sains Jurnal Teknik* 1 (2), 157-163.

[26] Ramadhani, F., Satria, A., & Sari, I.P (2023). Implementasi Metode Fuzzy K-Nearest Neighbor dalam Klasifikasi Penyakit Demam Berdarah. *Hello World Jurnal Ilmu Komputer* 2(2), 58-62.
[27] Sari., I.P, Al-Khowarizmi., A, & Batubara., I.H. (2021). Implementasi Aplikasi Mobile Learning Sistem Manajemen Soal dan Ujian Berbasis Web Pada Platform Android. *IHSAN: JURNAL PENGABDIAN MASYARAKAT* 3 (2), 178-183.
[28] Batubara, I.H., Saragih, S., Simamora, E., Napitupulu, E.E., Sari, I.P. (2022). Analysis of Student's Mathematical Communication Skills through Problem Based Learning Models Assisted by Augmented Reality. *Budapest International Research and Critics Institute-Journal (BIRCI-Journal),* 5(1), 1024-1037.