:: IJEEIT ::
(International Journal of Electrical Engineering and Information Technology)
Volume 03 Number 02 September 2020

ISSN : 2615-2096 (ONLINE)
ISSN : 2615-2088 (PRINTED)

GIS implementation and classterization of potential blood donors using the agglomerative hierarchical clustering method

Pratama Ryan Harnanda
Faculty of Computer Science, Narotama University, Indonesia
pratama.ryan@mhs.fasilkom.narotama.ac.id

Tresna Maulana Fahrudin
Faculty of Computer Science, Narotama University, Indonesia
tresna.maulana@narotama.ac.id

Natalia Damastuti
Faculty of Computer Science, Narotama University, Indonesia
natalia.damastuti@narotama.ac.id

ABSTRACT

The blood needs of PMI (Indonesian Red Cross) in the Surabaya City area are sometimes erratic, the problem occurs because the amount of blood demand continues to increase while the blood supply is running low. As the main objective of this research, data mining was applied to able to cluster the blood donor data in UTD-PMI Surabaya City Center which was to determine both potential and no potential donors and also visualize the pattern of donor distribution in Geographic Information System (GIS). Agglomerative Hierarchical Clustering was applied to obtain the clustering result from the existing of 8757 donors. The experiment result shown that the cluster quality was quite good which reached 0.6065410 using Silhouette Coefficient. We concluded the one interesting analysis that private male employees with blood type O, and live in the eastern part of Surabaya City are the most potential donors.

Keywords: Blood Donor, Clustering, Agglomerative Hierarchical Clustering, Data Mining, Geographic Information System.

:: IJEEIT ::
(International Journal of Electrical Engineering and Information Technology)
Volume 03 Number 02 September 2020

ISSN : 2615-2096 (ONLINE)
ISSN : 2615-2088 (PRINTED)

1. INTRODUCTION

The Indonesian Red Cross (PMI) is an independent and neutral organization in the Indonesian state, for its activities covering the social and humanitarian field. In carrying out all its activities, PMI always adheres to the seven principles of the International Red Cross and Red Crescent, namely humanity, volunteerism, neutrality, equality, independence, unity and universality (Raufun et al., 2019). In its implementation, the Indonesian Red Cross also does not make distinctions but rather prioritizes objects to victims who desperately need immediate help for the safety of their souls.

Blood donation is one of the humanitarian activities that aims to assist and assist community members who need blood, blood donation activities are organized and managed by the Indonesian Red Cross. Blood supply is often not constant, it happens because the number of donors is always uncertain or fluctuates, so it will be a problem when the amount of blood demand increases while the blood supply is running low (Atmaja et al., 2018). PMI in the City of Surabaya always conducts socialization aimed at raising public awareness to conduct regular blood donations, by disseminating information thoroughly to all elements of society of all ages, professions and regions. This method was deemed ineffective because each element of society who had donated had different characteristics to receive the information presented.

It is hoped that through the existing donor data at the Surabaya City Center PMI, the clustering process can be carried out using the Agglomerative Hierarchical Clustering (AHC) method and can be implemented into the visualization of the Geographical Information System (GIS) which is useful as a visualization of potential donor distribution patterns by determining the region. from donors. So that it can be focused on where the dissemination of information must be done to be more effic

2. REASEARCH METHOD

In this study, there are several stages and methods that can be used as materials to solve problems in the study. The system design in this study is shown in Figure 1:

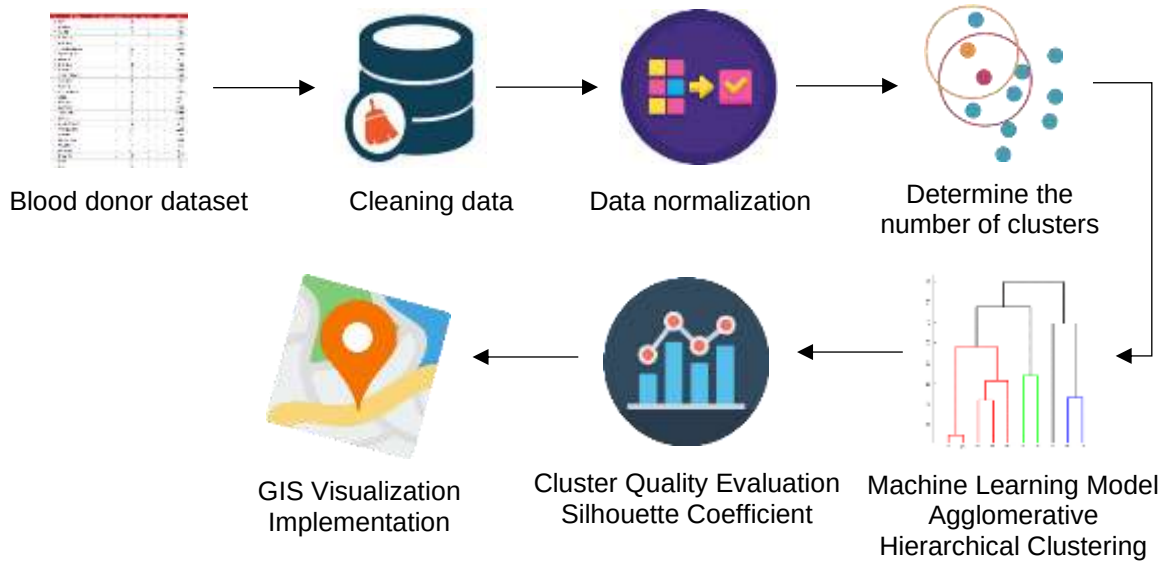

Figure 1. Research system design

ISSN : 2615-2096 (ONLINE)
ISSN : 2615-2088 (PRINTED)

2.1 Data mining

Data mining is a series of processes to explore added value in the form of information that has not been known manually from a database (Atmaja et al., 2018). Data Mining Method has been known since 1990 (Iqbal, 2019). Basically Data Mining is at the core of the Knowledge Discovery in Database (KDD) process which involves algorithms to explore data, develop models and find previously unknown patterns.

2.2 Clustering

Grouping or can be called clustering is a method for finding and grouping data that has (similarity) or similar characteristics between one data and another. There are two types of data clustering, namely partitional clustering and hierarchical clustering (Dani et al., 2019). With partitional clustering, data objects are divided into sub-set clusters that do not overlap so that each data object is in one sub-set. Meanwhile, hierarchical clustering is a nested cluster arranged as a hierarchical tree.

2.3 Min-max normalization

Min-Max Normalization is a normalization method by performing linear transformations of the original data so that it can produce a balance of value comparisons between data before and after processing (A. Nasution. H. Khotimah. N. Chamidah, 2019). This method can use equation (1).

$$x' = \frac{(x - x_{Min})}{(x_{Max} - x_{Min})} \quad (1)$$

Information :

x '= normalized data value

x = actual data value

xmin = smallest data value

xmax = largest data value

2.4 Agglomerative hierarchical clustering method

Agglomerative Hierarchical Clustering is a clustering method that can group objects in data into a hierarchy (Bachtiar et al., 2017) . In this method, there are two types of grouping, namely Agglomerative (bottom-up) and Divisive (top-down). Agglomerative Hierarchical Clustering is a bottom-up clustering method which combines several clusters into a single cluster.

The process starts from each data as a cluster, then recursively scatter looking for the closest group as a pair which will then be combined into a larger group (Suhirman & Wintolo, 2019) . The process will continue to be repeated until it appears to form a hierarchy. The following is an equation for calculating the distance between data contained in the Agglomerative Hierarchical Clustering method using the distance matrix formula (2) and (3).

Euclidean Distance:

$$D(x_2, x_1) = \sqrt{\sum_{j=i}^{n} |x_{2j} - x_{1j}|^2} \quad (2)$$

Manhattan Distance:

$$D = \sum_{k=0}^{n} |x_{2j} - x_{1j}| \quad (3)$$

ISSN : 2615-2096 (ONLINE)
ISSN : 2615-2088 (PRINTED)

There are three hierarchical grouping methods in the Agglomerative Hierarchical Clustering method using equations (4), (5) and (6).

Single Linkage (Closest Distance)

$$d_{uv} = \min\{d_{uv}\}, d_{uv} \in D \quad (4)$$

Complete Linkage (Farthest Distance)

$$d_{uv} = \max\{d_{uv}\}, d_{uv} \in D \quad (5)$$

Average Linkage (Average Distance)

$$d_{uv} = \text{average}\{d_{uv}\}, d_{uv} \in D \quad (6)$$

2.5 Evaluation of the silhoutte coefficient

Silhoutte Coefficient (SC) figures used in testing the quality of the clusters formed (Wardani et al., 2019). This method can be used as a test method in conducting research, especially in the clustering method. Calculation of the Silhoutte Coefficient value is through equation (7).

$$s(i) = \frac{b(i)-a(i)}{\max((b(i),a(i))} \quad (7)$$

Information :

Si = silhouette coefficient

ai = the average distance between the ith object and all objects in the same group

bi = the average distance of the ith object with all objects in different groups

The results of the Silhoutte Coefficient calculation have a range between -1 to 1, it can be said to be good if it is positive (Dani et al., 2019) and it is said to be bad if it is negative.

2.6 Geographical information system

Geographical Information System is a computer technology-based system that is used as data storage that can manipulate geographic information. In addition, GIS can present information or data in graphical form using maps as an interface (Teknik et al., 2016). The several stages carried out in this research are :

1. Data Collection
   Collecting data in this study using donor data obtained at the Indonesian Red Cross Blood Transfusion Unit (UTD-PMI) Surabaya City Center, totaling 8757 donor data. The data has an extension of * .xlsx with six supporting attributes as shown in Table 1, the data used as a dataset ranges from 2013 to 2018.

Table 1. Donor data samples

Gender	Blood group	Amount	Age	Profession	Region
Male	O	43	49	Private employees	North
Female	A	9	25	Student	South
Male	AB	56	63	Entrepreneur	West
Male	O	38	54	Government employees	West
Male	A	22	65	Entrepreneur	South
Male	A	90	40	Government employees	North
Male	B	89	45	Private employees	South
Female	O	5	34	Private employees	Center
Female	B	36	43	Civil servants	East

ISSN : 2615-2096 (ONLINE)
ISSN : 2615-2088 (PRINTED)

Male	AB	29	51	Army / Police	Center
Male	B	47	40	Student	North
Male	A	97	68	Government employees	West
Female	O	15	34	Private employees	South
Male	AB	31	35	Army / Police	Center
Female	A	4	26	Student	East
Male	A	18	38	Army / Police	West
Male	B	9	27	Government employees	South
Female	O	11	31	Private employees	South
Male	AB	23	40	Private employees	East
Female	AB	34	42	Entrepreneur	North
Female	B	27	38	Government employees	North
Female	AB	6	24	Private employees	West
Male	O	33	41	Student	South
Male	A	28	32	Army / Police	Center
Female	AB	12	26	Student	West
Female	A	5	25	Student	East
Male	A	16	35	Army / Police	West
Female	B	26	35	Government employees	North
Male	AB	54	61	Entrepreneur	West

2. Pre-Processing Data
   At this stage, it is a data pre-processing stage where there are two processes, namely data cleaning and data normalization.
   a. Cleaning Data
      Data cleaning includes several operations including identification of data entry without data and entry of lost data.
   b. Data Normalization
      Data normalization is carried out with the aim of making all data variables in the study within the same value range, so that it will be able to minimize the differences between research variables. In normalizing the data in this study using Min-Max Normalization can be seen from equation (1).
3. Data Processing (AHC)
   Data processing in this study was carried out using the Agglomerative Hierarchical Clustering method to obtain clusters from donor data that would produce potential and non-potential donors, by utilizing matrix calculations between Euclidean Distance data distances in equation (2) with the AHC Average Linkage process. (average distance) in equation (6).
4. Cluster Validation Test
   The next thing is to test the validity of the cluster, with the aim of seeing the goodness or quality of the results of the cluster analysis. In this study using the Silhoutte Coefficient (SC) method as cluster validation, which is in equation (7).
5. GIS implementation
   The final stage is the implementation stage into the Geographical Information System (GIS), which aims to find the location of the potential distribution of blood donors by utilizing the region / region attribute data of donors in the city of Surabaya.

3. RESULTS AND DISCUSSION

3.1 Data discretization formation

The initial stage in establishing potential and non-potential donor clusters in this study is to determine the data variables used in the formation of the cluster model in the form of gender, goal, number, age and profession. By making changes to categorical data to be discrete as shown in Table 2 and discretization as follows:

a. Gender: Men = 1, Women = 0
b. Goals: A = 1, AB = 2, B = 3, O = 4

ISSN : 2615-2096 (ONLINE)
ISSN : 2615-2088 (PRINTED)

c. Profession: government employees= 1, private = 2, army / police= 3, students = 4, farmers / factory workers= 5, housewife= 6, self-employed = 7 and others = 8

Table 2. The results of the formation of discrete data

Gender	Blood group	Amount	Age	Profession
1	3	54	46	2
1	1	20	43	7
1	4	33	50	1
0	3	14	26	4
0	1	38	53	6
1	4	15	58	7
1	2	52	46	1
1	4	70	51	2
0	3	21	35	8
1	1	44	53	3
0	2	34	40	1
0	4	38	50	2
1	3	22	36	5
0	4	46	39	6
1	3	51	47	2
1	4	21	35	2
0	1	33	37	3
0	2	44	58	1
1	2	38	53	8
1	4	38	51	7
1	3	22	35	5
1	4	22	36	6
0	1	47	53	6
1	2	14	28	7
0	3	20	43	4
0	4	33	50	3
0	2	17	28	1

3.2 Normalization of data (pre-processing)

There are lots of data that have different value ranges so that it is required to carry out the normalization process. The donor dataset is transformed using the min-max normalization method by processing the minimum and maximum values of each attribute. The range used in this method is 0 to 1 as shown in Table 3.

Table 3. Results of normalization of donor datasets

Gender	Blood group	Amount	Age	Profession
0.006431	0.025723	0.926031	0.372985	0.051446
0.010148	0.030443	0.872703	0.48709	0.010148
0.017705	0.070821	0.106232	0.9915	0.017705
0.010556	0.042222	0.781114	0.62278	0.010556
0.009419	0.037677	0.875995	0.480384	0.018839
0.008021	0.024063	0.922424	0.385012	0.016042
0.01175	0.029967	0.799005	0.599254	0.01175
0.010277	0.010277	0.894061	0.441892	0.071936
0.008447	0.033788	0.920725	0.388563	0.008447
0.008114	0.008114	0.916913	0.397599	0.032457
0.007491	0.029964	0.91389	0.404509	0.014982
0.008339	0.025016	0.900563	0.433604	0.016677
0.022299	0.089198	0.156096	0.981175	0.066898

ISSN : 2615-2096 (ONLINE)
ISSN : 2615-2088 (PRINTED)

0.00872	0.017441	0.915644	0.401139	0.017441
0.011881	0.047525	0.843566	0.534655	0.011881
0.006751	0.006751	0.938418	0.344312	0.027005
0.008743	0.034972	0.935506	0.349722	0.034972
0.009726	0.038905	0.885082	0.45713	0.077809
0.006162	0.024647	0.948919	0.314253	0.012324
0.008334	0.033336	0.883395	0.466699	0.025002
0.013846	0.041539	0.706157	0.706157	0.027692
0.007865	0.031461	0.880907	0.471915	0.01573
0.023088	0.092351	0.392494	0.900426	0.161615
0.006679	0.006679	0.921763	0.387408	0.013359
0.011183	0.044733	0.816369	0.57034	0.078282
0.008098	0.008098	0.882714	0.469701	0.008098
0.007519	0.030077	0.849672	0.526346	0.007519
0.007475	0.007475	0.881998	0.470897	0.014949
0.006386	0.025543	0.900393	0.434232	0.006386
0.011203	0.033608	0.862613	0.504125	0.022406

3.3 Establishment of a donor data clustering model

At the donor data processing stage, the first thing to do is determine the estimated number of data clusters to be formed and calculate the distance between data by ensuring that the dataset used is discrete data. The Euclidean Distance method is a method of calculating the distance between data used in this study. The next stage is to carry out the cluster formation process. Cluster formation carried out in this study uses Agglomerative Hierarchical Clustering - Average Linkage, namely by determining the closeness between two groups of the average distance between two data from a different cluster.

The results obtained from the three processes will display a dendrogram visualization of the results of clustering potential and non-potential donor data as illustrated in Figure 2.

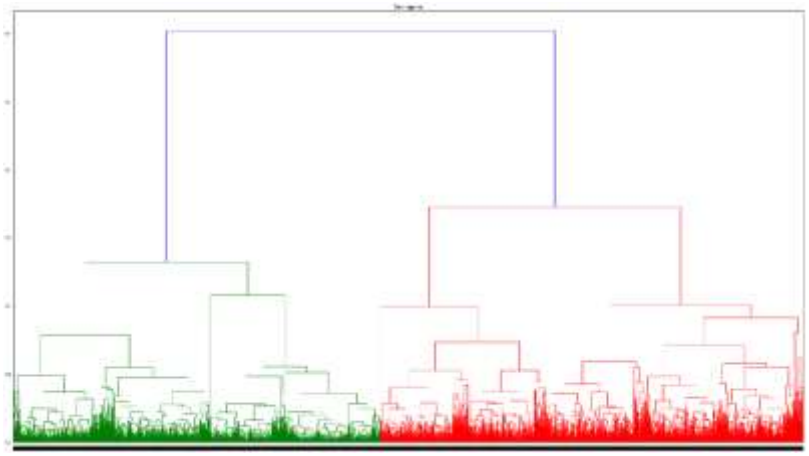

Figure 2. Dendrogram of potential and non-potential donor data clusters

From the cluster forming, it can also be represented in the form of an array by showing the number labels 1 and 0 as shown in Figure 3, provided that the number 1 is a label for potential donors and for number 0 the label of potential donors.

```
array([1, 1, 0, ..., 0, 0, 0], dtype=int64)
```

Figure 3. Array numbers resulting from the formation Cluster

ISSN : 2615-2096 (ONLINE)
ISSN : 2615-2088 (PRINTED)

Furthermore, the array can be displayed in a form like Table 4 by adding cluster attributes to the processed donor dataset, aiming to find out which patterns in the donor dataset labeled 1 and 0. By doing this process it will be easier to find out the status of the donor.

Table 4. Labels of the results of donor cluster formation

Gender	Blood group	Amount	Age	Profession	Region	cluster
Male	B	65	39	Private employees	West	1
Male	AB	44	54	Government employees	South	1
Male	B	27	45	Private employees	West	0
Female	A	39	33	Entrepreneur	East	1
Female	O	16	40	Other	East	0
Male	B	33	28	Student	North	1
Female	AB	9	44	Housewife	Center	0
Male	B	45	40	Army / Police	North	1
Female	O	37	30	Student	West	1
Male	A	43	55	Private employees	South	1
Male	AB	12	33	Army / Police	South	0
Male	B	56	50	Government employees	Center	1
Female	AB	41	59	Government employees	East	1
Female	O	8	29	Student	North	0
Male	A	40	47	Private employees	Center	1
Female	B	32	44	Government employees	North	1
Male	AB	45	40	Army / Police	West	1
Male	AB	22	45	Private employees	South	1
Male	O	7	30	Private employees	West	0
Female	A	16	33	Other	East	0
Female	B	58	52	Government employees	West	1
Male	B	42	39	Private employees	East	1
Female	O	11	29	Student	South	0
Male	AB	40	36	Entrepreneur	Center	1
Male	O	39	45	Army / Police	East	1
Female	O	9	44	Housewife	North	0
Female	A	40	37	Private employees	West	1
Male	B	47	40	Army / Police	South	1
Male	B	30	29	Student	East	1
Male	A	12	45	Other	East	0

3.4 Evaluation of cluster results

As the final stage of donor data processing by calculating the accuracy of cluster validity using the Silhoutte Coefficient (SC), the results show that the level of accuracy obtained reaches a value of 0.6065410 as shown in Figure 4.

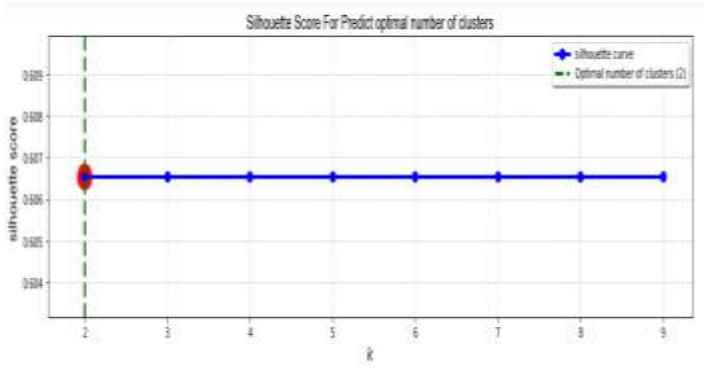

Figure 4.Graph result of silhouette coefficient value

ISSN : 2615-2096 (ONLINE)
ISSN : 2615-2088 (PRINTED)

3.5 Visualization of geographic information systems

The implementation process of the existing programs in this study is to visualize a Geographical Information System (GIS) by utilizing the region's attributes as a support for the distribution of donors. At this stage of the process the researchers listed five areas in the city of Surabaya, namely North, West, East, South and Center by utilizing the latitude and longitude of these areas.

The following Figure 5 is a login display in the distribution system for potential donors in the form of a website.

Figure 5. The pmi login page

The results of the GIS visualization are illustrated in Figure 6, it can be seen that the distribution of donors found on the map of the City of Surabaya produces five points of potential donor areas.

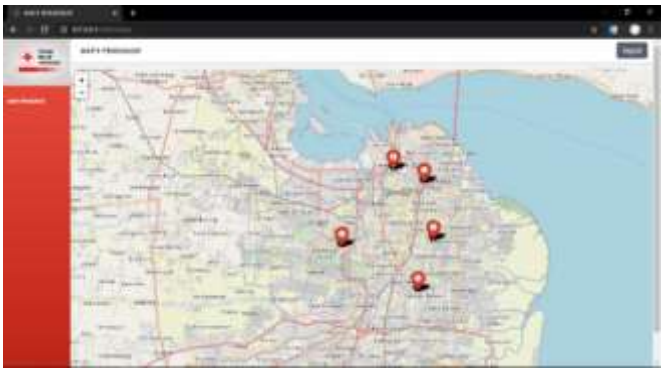

Figure 6. Map of the distribution of donors

Each region can also display insights in the form of a percentage of the number of potential and non-potential donors, gender, blood type, and occupation of the donor. In this discussion, an example is taken to display the percentage in the eastern part of the city of Surabaya.

Figure 7 shows the percentage of donors in the Eastern Region which is shown to have two parts, namely 49% for potential donors and 51% for non-potential donors.

Figure 7. Percentage of potential and non-potential donors

ISSN : 2615-2096 (ONLINE)
ISSN : 2615-2088 (PRINTED)

Figure 8 shows the percentage of blood donors based on gender in the Eastern Region, the information displayed is that potential donors are 5% for women and 95% for men, and 16% for women and 84% for men who are not potential donors.

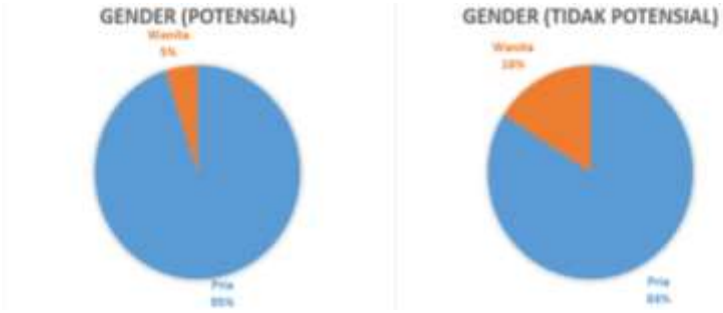

Figure 8. Percentage of potential and non-potential donors by gender\

Figure 9 shows the percentage of blood donors based on blood type in the Eastern Region, the information displayed is the potential donors of 23% for blood group A, 8% for blood group AB, 31% for blood group B and 38% for blood group O. While the donors which is not potential for 20% for blood group A, 7% for blood group AB, 31% for blood group B and 42% for blood group O.

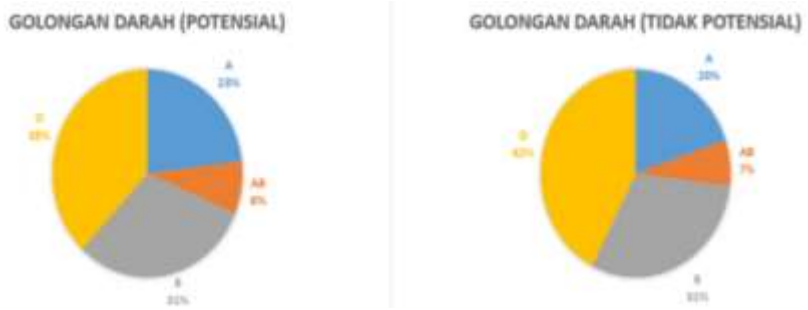

Figure 9. Percentage of potential and non-potential donors by blood group

Figure 10 shows the percentage of donor employment in the Eastern Region shown for potential donors of 30% for PNS, 46% for private sector, 2% for Army / Police, 8% for Self-employed, 1% for Housewives, 4% for Others and 9% for students, and 9% for civil servants, 65% for private sector, 7% for Army / Police, 6% for self-employed, 1% for housewives, 4% for Others and 8% for Student.

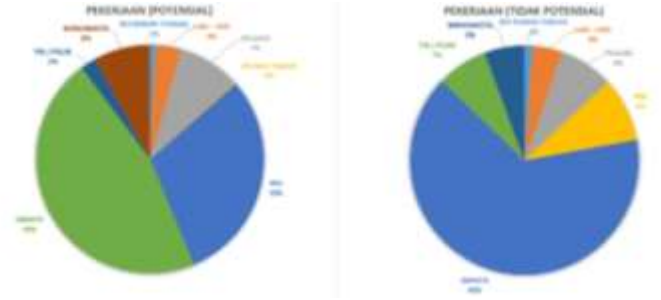

Figure 10. Percentage of potential and non-potential donors by occupation

ISSN : 2615-2096 (ONLINE)
ISSN : 2615-2088 (PRINTED)

4. CONCLUSIONS

From the results of the experiments conducted, it can be seen that the data used were 8,757 blood donors obtained from UTD-PMI Surabaya City Center. By applying the clustering method using Agglomerative Hierarchical Clustering, the results are quite good, namely the quality of the cluster reaches the Silhoutte Coefficient (SC) value of 0.6065410. The results of the clustering analysis of donor data that have been carried out can be drawn one interesting conclusion that male gender private employees with blood type O from the East Region of Surabaya City are the most potential donors.

5. REFERENCES

A. Nasution. H. Khotimah. N. Chamidah. (2019). *PERBANDINGAN NORMALISASI DATA UNTUK KLASIFIKASI WINE MENGGUNAKAN ALGORITMA K-NN*. *4*(1), 78–82.

Atmaja, K. J., Anandita, I. B. G., & Dewi, N. K. C. (2018). Penerapan Data Mining Untuk Memprediksi Potensi Pendonor Darah Menjadi Pendonor Tetap Menggunakan Metode Decision Tree C.45. *S@Cies*, *7*(2), 101–108. https://doi.org/10.31598/sacies.v7i2.284

Bachtiar, A. M., Dharmayanti, D., & Hamzah, R. L. (2017). Penerapan Metode Hierarchical Agglomerative Clustering Untuk Segmentasi Pelanggan Potensial Di Jeger Jersey Indonesia. *Komputa : Jurnal Ilmiah Komputer Dan Informatika*, *6*(1), 35–42. https://doi.org/10.34010/komputa.v6i1.2475

Dani, A. T. R., Wahyuningsih, S., & Rizki, N. A. (2019). Penerapan Hierarchical Clustering Metode Agglomerative pada Data Runtun Waktu. *Jambura Journal of Mathematics*, *1*(2), 64–78. https://doi.org/10.34312/jjom.v1i2.2354

Iqbal, M. (2019). Klasterisasi Data Jamaah Umroh Pada Auliya Tour & Travel Menggunakan Metode K-Means Clustering. *JURTEKSI (Jurnal Teknologi Dan Sistem Informasi)*, *V*(2), 97–104.

Jo, T. (2020). Semantic string operation for specializing AHC algorithm for text clustering. *Annals of Mathematics and Artificial Intelligence*. https://doi.org/10.1007/s10472-019-09687-x

Li, Y., Liu, M., Wang, W., Zhang, Y., & He, Q. (2019). Acoustic Scene Clustering Using Joint Optimization of Deep Embedding Learning and Clustering Iteration. *IEEE Transactions on Multimedia*, *PP*(c), 1–1. https://doi.org/10.1109/tmm.2019.2947199

Magdalena, L., & Mulyasari, H. (2018). Rancangan Sistem Informasi PMI Dengan Mengintegrasikan Data Pendonor dan Stok Darah Antar Cabang PMI di Wilayah III Cirebon. *Sekolah Tinggi Manajemen Informatika Dan Komputer Cirebon*, 978–979.

Oey, E., Marpaung, A. B., & Idham Sofyan, M. (2019). Analysing salesmen itinerary with agglomerative hierarchical clustering and vehicle routing algorithm - A case study of a confectionery supplier in Indonesia. *International Journal of Industrial and Systems Engineering*, *31*(3), 287–303. https://doi.org/10.1504/IJISE.2019.098541

Rahayu, I. W., Atastina, I., Herdiani, A., Informatika, F., Telkom, U., Komunitas, D., & Sosial, J. (2018). Hierarchical Clustering Untuk Deteksi Komunitas Pada Media Sosial Facebook Analysis and Implementation of Agglomerative Hierarchical Clustering Algorithm for Community Detection in Social Media Facebook. *E-Proceeding of Engineering*, *5*(1), 1460–1468.

Raufun, L., Ode, W., Angraini, D., Prodi, D., Informatika, T., Dayanu, U., Baubau, I., & Tenggara, S. (2019). *Ketersediaan Darah Pada Palang Merah Indonesia Kabupaten Buton Berbasis Android*. *8*(1).

Suhirman, S., & Wintolo, H. (2019). System for Determining Public Health Level Using the Agglomerative Hierarchical Clustering Method. *Compiler*, *8*(1), 95. https://doi.org/10.28989/compiler.v8i1.425

Teknik, J., Teknik, F., Udayana, U., & Merah, P. (2016). Rancang Bangun Aplikasi Komunitas Donor Darah Berbasis Web Dan Android Yang Dilengkapi Layanan Informasi Geografis. *Jurnal Ilmiah Spektrum*, *3*(2), 77–83.

Triayudi, A., & Fitri, I. (2019). A new agglomerative hierarchical clustering to model student activity in online learning. *Telkomnika (Telecommunication Computing Electronics and Control)*, *17*(3), 1226–1235. https://doi.org/10.12928/TELKOMNIKA.V17I3.9425

Wardani, R. W., Setiawan, B. D., & Dewi, C. (2019). *Perbandingan Kualitas Hasil Klaster Algoritme K-Means dan Isodata pada Data Komposisi Bahan Makanan*. *3*(7), 6712–6720.

Zhou, S., Xu, Z., & Liu, F. (2017). Method for Determining the Optimal Number of Clusters Based on Agglomerative Hierarchical Clustering. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(12), 3007–3017. https://doi.org/10.1109/TNNLS.2016.2608001