

Implementation of Agglomerative Hierarchical Clustering Based on The Classification of Food Ingredients Content of Nutritional Substances

Syabdan Dalimunthe¹, Anggi Hanafiah²

Departement of Computer Engineering, Politeknik Caltex Riau¹

Departement of Informatic Engineering, Universitas Islam Riau²

syabdan@mahasiswa.pcr.ac.id¹, anggihanafiah@eng.uir.ac.id²

Article Info

Article history:

Received May 14, 2021

Revised May 25, 2021

Accepted August 4, 2021

Keyword:

Agglomerative hierarchical Clustering
Average Linkage
Nutrients
Silhouette Index

ABSTRACT

Health is something very precious. Maintaining health can be done in many ways, one of them by keeping your diet. The correct diet will keep your immune system so that it can avoid various diseases. The proper diet will also put the body in a balanced nutrition state, which all need to be nourished. Nutrient requirements include calories, protein, fat, carbohydrates, calcium, phosphorus, iron, vitamin A, vitamin B, and vitamin C with a mass of 100 grams each. To facilitate the search for nutrients needed, then build a system that can categorize food based on its nutritional status and calculate the average value of nutrients in agglomerative hierarchical clustering using average linkage. Calculation of intermediate linkage methods produces data that has some similarities to the data sought nutrients that can be seen from its index, so precise data are in each group. Of the 6 categories of foodstuffs and 10 nutrient contents obtained, the cluster results were selected in the staple food category for testing in cluster 5, obtaining an SI value of 1, in the animal side dish category, in cluster 3, obtaining an SI value of 1, in the test vegetable side dish category. in cluster 4 obtained an SI of 0.846. In the vegetable category, the test in cluster 8 obtained an SI value of 9, in the fruit category, the test in cluster 3 obtained an SI value of 1, and in the egg, fruit and test categories in cluster 4 obtained an SI value of 1.

© This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Corresponding Author:

Syabdan Dalimunthe
Applied Master of Computer Engineering Program
Politeknik Caltex Riau
Jl. Umban Sari (Patin) No. 1 Rumbai, Pekanbaru, Riau
Email: pcr@pcr.ac.id

1. INTRODUCTION

A well-chosen daily diet will provide all the nutrients needed for the normal functioning of the body. Conversely, if the food is not appropriately chosen, the body will experience a deficiency of certain essential nutrients.

Nutrients in food can be divided into two parts based on the amount contained therein, namely micronutrients (micronutrients) and macronutrients (macronutrients). Macronutrients are nutrients that contribute a lot of energy to the body. The term macronutrient describes chemicals that provide calories for energy, including carbohydrates, protein, and fat. The body requires these

nutrients in large quantities. Meanwhile, micronutrients are substances such as vitamins and minerals that are essential for healthy growth and development. Although micronutrients are only needed in small amounts, micronutrient deficiencies can also cause serious problems. Examples of micronutrients include vitamin A, folic acid, iodine, iron, and zinc. Deficiencies in these nutrients can have severe consequences for children, pregnant women, and women of childbearing age.

The human body needs macronutrients and micronutrients. However, the food contained in the surrounding environment has different types and nutritional content. People only know the nutrients in the food. For example, rice contains carbohydrates. Even though the ones that contain carbohydrates are not only rice, there are sweet potatoes, wheat, sago, and others. People find it challenging to find foods that have similar nutritional content in large quantities. Besides, people affected by certain diseases, people who are undergoing a diet program, and particular sports require the consumption of special nutrients. Most of them find it difficult to choose foods with the similarity of the nutrients they need.

Based on this, data mining and clustering techniques can help provide solutions to these problems. Clustering is a method that can be used to make it easier to find data in the database. Clustering is a technique in data mining for grouping data. The agglomerative hierarchical clustering method is used to group data. In this method, food ingredients will be grouped based on nutritional attributes. The process of grouping foodstuffs is done by calculating the distance matrix using the euclidean distance. Where each data item is a cluster. After getting the distance between the data, the next step is to find each distance matrix's average value using the average linkage clustering method. Where is the average value of the comparison between the two distance matrices that will be taken to find the next iteration to obtain the desired number of clusters.

2. RESEARCH METHOD

2.1. Related research

The literature study in this study is a literature study, where the literature referred to according to previous researchers is as follows:

Based on research conducted by [1] regarding the grouping of food and beverages at PT. Indomarco Palembang. PT. Indomarco Palembang is a company engaged in the distribution of food and beverages. Not only PT. Indomarco Palembang, there are still quite a lot of other companies engaged in similar fields. This, of course, creates business competition between companies. From these problems, the solution obtained is to apply data mining to make decisions in the business world to develop business and to see the sales that consumers are most interested in, especially food and beverage sales. Based on this research, it can be concluded that one of the methods contained in data mining used in this study is the clustering method. The result of this research is that the application built can help the company as an illustration in making decisions in order to get product sales patterns. Originality in this study with this research is the problem under study. Namely, this study examines the problem of grouping food ingredients based on the nutrients contained therein, while the method used is different from this research, namely Agglomerative Hierarchical Clustering with the Average Linkage method.

Based on research conducted by [2] regarding the grouping of types of food based on the nutrients contained therein. From these problems, the solution obtained is to classify the types of food into the number of micronutrients (carbohydrates and calcium) and macronutrients (protein and fat). The method used is the Discriminant Analysis method. Data taken from the List of Food Ingredients Composition (DKBM) of Indonesia includes 50 types of food and 4 variables of types of nutrients. The 4 types of nutrition are Protein, and Fat, carbohydrates, Calcium. Based on this problem, it can be concluded that dividing 4 variables into 2 groups, namely micronutrients and macronutrients in food. The results of the data analysis will show that the type of food is the type of food that contains a lot of micronutrients or contains more macronutrients. The originality of this study compared to that of the study is the method used. This study, using the Agglomerative Hierarchical Clustering algorithm with the Average Linkage method.

Based on research conducted by [3] about classifying food ingredients based on nutrient content in food. Based on these problems, the solution taken is to classify the food material data consisting of food ingredients with nutritional content as a source of combustion substances, a source of regulatory substances, and a source of building substances. From these problems, it can be concluded that each foodstuff group is grouped into three groups using the fuzzy c-means method. Information regarding foodstuff groups can be used as a reference for selecting food ingredients according to the required nutritional content. The originality of this study compared to that of the study is the problem under study. In this study, the types of foodstuffs were classified into 4 groupings, namely high-micronutrient high macronutrient, high-micronutrient high macronutrient, low macronutrient high-micronutrient, low macronutrient low-micronutrient. In addition, this study also differs in methods, namely using the Agglomerative Hierarchical Clustering algorithm with the Average Linkage method.

2.3. Average Linkage Clustering

Average linkage clustering is a method in which the similarity rules between clusters are based on the average distance of all objects in a cluster with all other objects in other clusters [12]. In the average linkage method, the distance between two clusters is defined as the average distance between all object pairs, where one of the pair members comes from each cluster.

The following is the formula for calculating distance using the euclidean distance method in the following equation:

$$d_{AB} = \sqrt{\sum_{i=1}^n |x_i - y_i|^2} \quad (1)$$

Measurement of the distance between two clusters in the average linkage uses the average proximity formula as in equation as follows:

$$d(U, V) = \frac{1}{n_u \times n_v} \sum d(U, V); d(U, V) \in D \quad (2)$$

Note: Nu and NV, respectively, is the amount of data in clusters U and V.

For all the distances that have been obtained, if the clusters U and V are candidates for the two clusters to be combined while W is the other clusters that are not candidates, then for all the distances that have been calculated in equation (2.2), the smallest of them is chosen. . The formulation can be seen in the equation as follows:

$$U \cap (V, W) = \min\{d(U, V), d(U, W)\}; d(U, V), D(U, W) \in D \quad (3)$$

2.4. Silhouette Index

The silhouette index (SI) can be used to validate either a single data cluster (one cluster out of a number of clusters) or even an entire cluster. This method is most widely used to validate clusters that combine cohesion and separation values. To calculate the SI value of an ith data, there are 2 components, namely ai and bi. ai is the average distance of the I data to all other data in one cluster, while bi is obtained by calculating the average distance of the I data to all data from other clusters not in one cluster with the ith data, then take the smallest [12].

Here's a formula for calculating cohesion:

$$cohesion(C_i) = \sum_{x \in C_i} proximity(x, C_i) \quad (4)$$

The following is the formula for separation between two clusters:

$$separation(C_i, C_j) = proximity(C_i, C_j) \quad (5)$$

Here's the formula for calculating a_i^j :

$$a_i^j = \frac{1}{m_j - 1} \sum_{\substack{r=1 \\ r \neq i}}^{m_j} d(x_i^j, x_r^j), \quad i = 1, 2, \dots, m_j \quad (6)$$

$d(x_i^j, x_r^j)$ is the distance between the i th data and the r -th data in one cluster j , while m_j is the amount of data in the j th cluster.

Here's a formula for calculating b_i^j :

$$b_i^j = \min_{\substack{n=1, \dots, k \\ n \neq j}} \left\{ \frac{1}{m_n} \sum_{\substack{r=1 \\ r \neq i}}^{m_n} d(x_i^j, x_r^n) \right\}, \quad i = 1, 2, \dots, m_n \quad (7)$$

To get the i th Silhouette Index (SI) data, use the following equation:

$$SI = \frac{b_i^j - a_i^j}{\max\{a_i^j, b_i^j\}} \quad (8)$$

The value of a_i measures how dissimilar a data is to the cluster it follows. The smaller the value, the more precisely the data is in the cluster. A large b_i value indicates how bad the data is against other clusters. The SI values obtained are in the range $[-1, +1]$. The SI value that is close to 1 indicates that the data is increasingly right in the cluster. A negative SI value ($a_i > b_i$) indicates that the data is not right in the cluster (because it is closer to another cluster). SI value 0 (or close to 0) means that the data is positioned on the border between the two clusters.

The SI value of a cluster is obtained by calculating the average SI value of all data that joins the cluster, as in the following equation:

$$SI = \frac{1}{m_j} \sum_{i=1}^{m_j} SI_i^j \quad (9)$$

While the global SI value is obtained by calculating the average SI value of all clusters as in the following equation:

$$SI = \frac{1}{k} \sum_{j=1}^k SI_j \quad (10)$$

3. RESULTS AND ANALYSIS

3.1. System Analysis

In this stage, a grouping application will be designed and built using existing techniques in data mining. This application will classify food material data based on the content of certain nutrients. Grouping to find information in food ingredients based on these nutrients uses clustering techniques in data mining using agglomerative hierarchical clustering algorithms.

This grouping of information in foodstuff data will help make it easier for people to find information related to nutrient content and food ingredients as needed. This application is used to find out information on the relationship between existing food material data with food material data or nutrient content and find out information on existing food data groups based on certain criteria. An overview of the data mining process can be seen as follows:

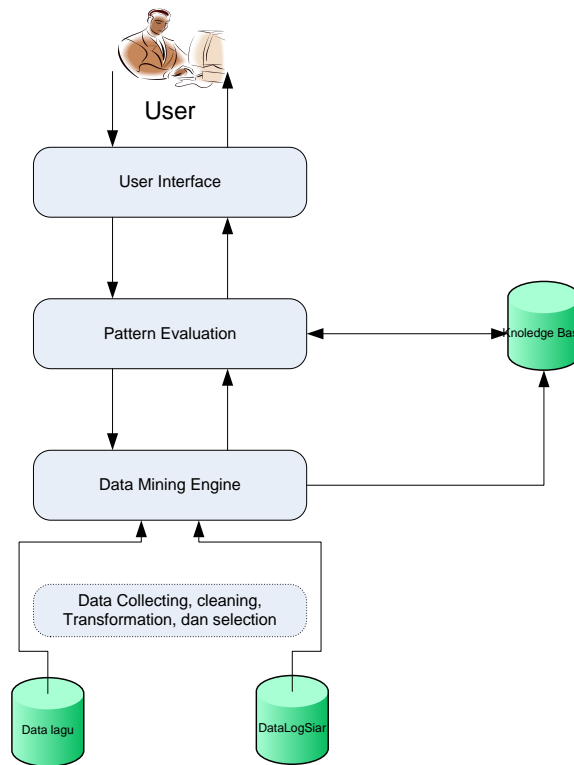


Fig 1. Process Flow in Data Mining

3.2. Implementation

1. System Implementation in Society

The implementation of the system is by making a questionnaire with 4 (four) questions and 10 (ten) correspondents from various circles of society. The fourth question in question is as follows:

- a. What do you think about the appearance of this system design?
- b. Are the input and output display easy to understand?
- c. Is the system easy to use?
- d. Does this system make it easier for you to find substitute food ingredients similar to what you want?

From these questions, the results of the answers or responses from correspondents to the performance of the system based on the questions posed are as follows:

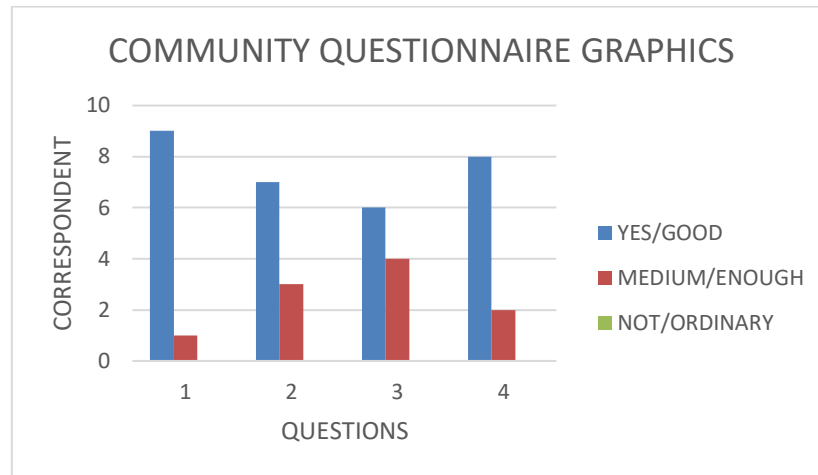


Fig 2. Graph of The Results of The Questionnaire From The Community

2. Implementation of Systems in the Health Sector

Implementation of the system is by making a questionnaire with 5 (five) questions and 10 (ten) correspondents, 5 of which are medical students and the other 5 are doctors. The five questions in question are as follows:

- a. What do you think about the appearance of this system design?
- b. Are the input and output display easy to understand?
- c. Is the system easy to use?
- d. Does this system make it easier for you to find food ingredients similar to what you want?
- e. Does the system provide the correct information on the group of foodstuffs and their nutritional content?

From these questions, the results of the answers or responses from correspondents to the performance of the system based on the questions posed are as follows:

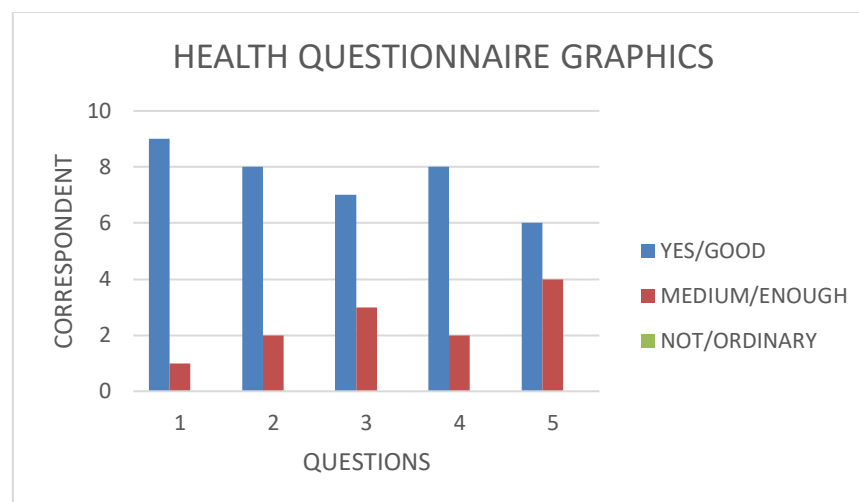


Fig 3. Graph of The Results of The Questionnaire From The Nutrition Sector

3.3. Test Result

In this study, there are 6 categories of food ingredients and 10 nutritional content. The testing is carried out by inputting the number of clusters you want to test, which will later calculate the validity of the clusters using the Silhouette Index (SI) method. The results of the calculation of the foodstuff clustering system can be seen in the table 1.

Table 1. Table of Silhouette Index Value Result

Foodstuff Category	Cluster Testing	<i>Silhouette Index</i>
Staple food	2	0.620
	3	0.547
	4	0.455
	5	1
	6	1
	7	1
	8	1
	Animal side dishes	2
3		1
4		1
5		1
6		1
7		1
8		1
Vegetable side dishes		2
	3	0.502
	4	0.846
	5	0.846
	6	0.846
	7	0.846
	8	0.846
	Vegetables	2
3		0.259
4		0.707
5		0.792
6		0.792
7		0.783
8		1
Fruits		2
	3	1
	4	1
	5	1
	6	1
	7	1
	8	1
	Eggs, milk, and milk products	2
3		0.980
4		1
5		1
6		1

	7	1
	8	1

In this test, the number of clusters was tested as many as 2 clusters to 8 clusters. From the test results, the number of clusters in the staple food category, the highest silhouette index value 1 was found in clusters 5, 6, 7, and 8, while the lowest silhouette index value was 0.455 exist in 4 total clusters. For the category of animal side dishes, the highest silhouette index value was 1 in clusters 3, 4, 5, 6, 7, and 8, while the lowest silhouette index value in the category of animal side dishes was 0.989 in 2 clusters. Then for the category of vegetable side dishes, the highest silhouette index value was 0.846, while the lowest silhouette index value was 0.502, which was located in 3 total clusters. In the vegetable foodstuff category, the highest silhouette index value is 1, located in 8 total clusters, while the lowest silhouette index value is 0.259, located in 3 total clusters. In the fruit foodstuff category, the highest silhouette index value is 1, which is located in 3, 4, 5, 6, 7, and 8 of the number of clusters, while the lowest value of the silhouette index lies in 2 total clusters in that category. Furthermore, for the category of eggs, milk, and milk, the highest silhouette index value is 1, which is located at 4, 5, 6, 7, and 8 in the number of clusters, while the lowest silhouette index value is 0.705, which is located in 2 total clusters in the category of eggs, milk. and milk yields. The results of the calculation of the validity of this cluster will later select the number of clusters with the highest SI value, which is 1 or close to 1. The number of selected clusters is the highest SI value which can be seen in table 2.

Table 2. Table of Number of Clusters

Foodstuff Category	Cluster Testing	<i>Silhouette Index</i>
Staple food	5	1
Animal side dishes	3	1
Vegetable side dishes	4	0.846
Vegetables	8	1
Fruits	3	1
Eggs, milk, and milk products	4	1

From the results of testing using a black box, the system that has been done, the following conclusions are obtained:

1. Food material data can be grouped based on similarity in nutritional content.

2. The best silhouette index (SI) value is close to 1 (one) or 1 (one), meaning that the data for the food being sought have similar nutrients in the cluster. In the category of staple food, the best SI value was in 5 clusters, in the category of animal side dishes the best SI value was in 3 clusters, in the category of vegetable side dishes the best SI value was in 4 clusters, in the vegetable category the best SI value was in 8 clusters. Whereas in the fruit category, the best SI value was in 3 clusters, and in the category of eggs, milk, and milk, the best SI value was in 4 clusters.

From table 2, we can see the silhouette index value and the number of selected clusters which will later become the cluster values that will be studied in foodstuff data. The selected silhouette index value is the highest value of the silhouette index value from all clusters in one category. If the largest silhouette index value in one category is the same, for example, the staple food category, the largest silhouette index values are 5, 6, 7, and 8, Then the smallest number of clusters is chosen, namely, the number of clusters 5 (five) because the greater the number of clusters, the less data there is even one data in one cluster, Therefore, the smallest number of clusters was chosen with the largest silhouette index value.

4. CONCLUSION



From the design and manufacture of food clustering systems based on their nutritional content, it can be concluded that The implementation of the Agglomerative Hierarchical Clustering Average Linkage algorithm for grouping food ingredients has been successfully designed and implemented in the form of a web application that can provide detailed information on food ingredients needed based on nutritional content. The best silhouette index (SI) value is close to 1 (one) or 1 (one), meaning that the data for the food being sought have similar nutrients in the cluster. In the category of staple food, the best SI value was in 5 clusters, in the category of animal side dishes, the best SI value was in 3 clusters, in the category of vegetable side dishes, the best SI value was in 4 clusters, in the vegetable category the best SI value was in 8 clusters. Whereas in the fruit category, the best SI value was in 3 clusters, and in the category of eggs, milk, and milk, the best SI value was in 4 clusters.

REFERENCES

- [1] Sutrisno, Afriyudi, and Widiyanto, "Penerapan Data Mining Pada Penjualan Menggunakan Metode Clustering Study Kasus Pt . Indomarco," *Penerapan Data Min. Pada Penjualan Menggunakan Metod. Clust.*, vol. Vol.x No.x, no. Data Mining, pp. 1–11, 2013, [Online]. Available: [http://eprints.binadarma.ac.id/78/1/Penerapan data mining pada penjualan menggunakan metode clustering study kasus pt. Indomarco Palembang.pdf](http://eprints.binadarma.ac.id/78/1/Penerapan_data_mining_pada_penjualan_menggunakan_metode_clustering_study_kasus_pt_indomarco_palembang.pdf).
- [2] H. A. Parhusip and J. T. Natangku, "Pengelompokan Zat Gizi Makanan Menggunakan Analisis Diskriminan," *Pros. Semin. Nas. ...*, no. May 2011, 2011, [Online]. Available: <http://eprints.undip.ac.id/33919/>.
- [3] M. Budiayanti and M. N. Estri, "Fuzzy C-Means Clustering Untuk Pengelompokan Bahan Makanan Berdasarkan Kandungan Zat Gizi," *J. Ilm. Mat. dan Pendidik. Mat.*, vol. 4, no. 1, p. 223, 2012, doi: 10.20884/1.jmp.2012.4.1.2958.
- [4] T. Alfina and B. Santosa, "Analisa Perbandingan Metode Hierarchical Clustering, K-Means dan Gabungan Keduanya dalam Membentuk Cluster Data (Studi Kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS)," *Anal. PerbandinganMetode Hierarchical Clust. K-means dan Gabungan Keduanya dalam Clust. Data*, vol. 1, no. 1, pp. 1–5, 2012.
- [5] D. J. Hand, *Principles of data mining*, vol. 30, no. 7. 2007.
- [6] A.-H. Tan, "Text Mining: The state of the art and the challenges," *Proc. PAKDD 1999 Work. Knowl. Discovery from Adv. Databases*, vol. 8, pp. 65–70, 1999, doi: 10.1.1.38.7672.
- [7] Arbie, 2004, *Manajemen Database dengan MySQL*, ANDI, Yogyakarta.
- [8] Djaeni Sediaoetama, Achmad., 2008, *Ilmu Gizi Untuk Mahasiswa dan Profesi*, Dian Rakyat, Jakarta.
- [9] Fatansyah., 2001, *Basis Data dan DBMS*, Informatika, Bandung.
- [10] Hand, David, Mannila, Heikki, dan Smyth, Padhraic. *Principles Of Data Mining*. The MIT Press, 2001.
- [11] Hartono, Jogiyanto., 2005, *Analisis dan Desain Sistem Informasi Pendekatan Terstruktur Teori dan Praktek Aplikasi Bisnis*, Andi, Yogyakarta.
- [12] Prasetyo, Eko., 2014, *Data Mining Mengolah Data Menjadi Informasi Menggunakan Matlab*, Andi, Yogyakarta.
- [13] Juhari, Ahmad., 2013, *Dasar-dasar Ilmu Gizi*, Jaya Ilmu, Yogyakarta.
- [14] R. Muchtadi, Tien., dkk, 2011, *Ilmu Pengetahuan Bahan Pangan*, Alfabeta, Bandung.
- [15] Santosa, Budi., 2007., *Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis, Graha Ilmu*, Yogyakarta.
- [16] Sandjaja., dkk., 2009, *Kamus Gizi*, PT. Kompas Media Nusantara, Jakarta.
- [17] Sutanta, Edhy., 2004, *Sistem Basis Data*, CV. Graha Ilmu, Yogyakarta.
- [18] Thearling, Kurt. *An Introduction To Data Mining*. Whitepaper. <http://www3.shore.net/~kht/dmwhite/dmwhite.html>, 6 December 2020.

[19] *Kamus Kesehatan*, <http://kamuskeehatan.com/arti/>, 6 December 2020.

BIOGRAPHY OF AUTHORS

	<p>Sybdan Dalimunthe obtained a Bachelor of Informatics from Riau Islamic University in 2016. He has served as Assistant Lecturer in the Riau Islamic University Informatics Department since 2013. His current research interests include machine learning, Data Science, and the Internet of Things.</p>
	<p>Anggi Hanafiah is a Lecturer of Department of Informatics Engineering, Islamic University of Riau. Obtained Bachelor Degree in Informatics Engineering from STMIK-AMIK Riau, obtained Master Degree in Information Engineering from UPI-YPTK Padang. His current research interests include Data Mining, Artificial Intelligent, Networking, and Multiplatform Programming.</p>