International Proceeding Conference on Information Technology,
Multimedia, Architecture, Design, and E-Business (IMADE). Vol.2, August 2022
p-ISSN: 2747-1764, e-ISSN: 2747-1756 http://eprosiding.idbbali.ac.id/index.php/imade

CLUSTERING OF PROSPECTIVE NEW STUDENTS USING AGGLOMERATIVE HIERARCHICAL CLUSTERING

Muchamad Iqbal [1], M. Bucci Ryando [2], Triono [3], Nunung Nurmaesah [4]

[1] *Institut Teknologi dan Bisnis Bina Sarana Global*
miqbal@global.ac.id
[2] *Institut Teknologi dan Bisnis Bina Sarana Global*
bucci@global.ac.id
[3] *Institut Teknologi dan Bisnis Bina Sarana Global*
Triono19@global.ac.id
[4] *Institut Teknologi dan Bisnis Bina Sarana Global*
n.nurmaesah@global.ac.id

ABSTRACT

Global Institute of Technology and Business is one of the private universities focusing on computer science. In animating the survival of higher education, the marketing division is required to find as many new students as possible as well as the increasingly tight competition of private universities in finding new students for registration. Not being on target in determining the target market becomes a very dangerous problem in the survival of private colleges. The purpose of this research is to get the right target market from the clustering results of new students and get characteristics from prospective new students and help the marketing division in achieving the registration target. This research using Agglomerative Hierarchical Clustering method with single linkage and average linkage algorithm and testing was done using Silhouette Score and Calinski Harabasz Index. Learning the characteristics of clusters takes 7 clusters to study. The result of learning the characteristics of the cluster is to get cluster 6 as a cluster with characteristics with good results.

Keywords: *Agglomerative Hierarchical Clustering, Single Linkage, Average Linkage, Silhouette Score, Calinski Harabasz Index, Prospective New Students*

INTRODUCTION

The current development of the company cannot be separated from the increasing number of advances in computer technology. Computer technology needs will be in demand by companies in all fields, both private and public companies government agencies. This is related to work that is usually always done manually by humans will be more effective and efficient if carried out with a computerized system. Even with advanced technology Computers that are growing rapidly can make it easier for these companies to improve work efficiency because of the work they do Computers can save both in terms of time, space, energy, and costs. At first the computer was only used as a calculating tool by humans, but now the development of technology, especially in the field of computers and with human needs and knowledge of the importance of technology, the facilities provided by computer programs are increasing by not only being used as a calculating tool. One of the uses another of the computers is about data processing.

Problems occur when the marketing division is not right on target in determining target market. The competition in marketing is getting tougher and it is very important to stay on track (Kansal, 2018). The number of private universities in the Tangerang city makes Global Institute of Technology and Business has a passion to compete, because Global Institute of Technology and Business is private universities, looking for prospective new students is Global Institute of Technology and Business carries on college life. Marketing division as the hope in finding prospective new students to study at Global Institute of Technology and Business. In an effort to find prospective new students, the marketing division method is one

of them, namely: looking for relationships as much as possible from Senior High School (SMA), school Vocational High School (SMK), as well as from companies. From this relation, division marketing is allowed to get a database of presentation results in SMA/SMK & from the company to follow up. So, in the database, we can maximize in finding the right target market. Get the target in the right way to find prospective new students and the characteristics of the new students candidates who are the purpose of this study using the method Agglomerative Hierarchical Clustering (Nugraha, 2018) with single linkage (Vijaya, 2019) and average linkage.

RESEARCH METHOD

The method used in this research is using Cross-Industry Standard Process for Data Mining (CRISP-DM) (Chapman, 2000). There is 6 phases in CRISP-DM (Schäfer, 2021). The first phase is Business Understanding with the stages of problem identification, literature study and problem formulation also the problem objectives. The second phase is Data Understanding with the stages of attribute determination and data collection. The third phase is Data Preparation with the stages of data pre-processing in which there is data integration, data cleaning, data transportation and data reduction. The fourth phase is implementation of Agglomerative Hierarchical Clustering with single linkage algorithm and average linkage algorithm. The fifth phase is Evaluation and the sixth phase is Deployment with the stage of prototype design. After all phases have been carried out, then we can see the results and make conclusions from this study. The flow of CRISP-DM can we see in figure 1 below.

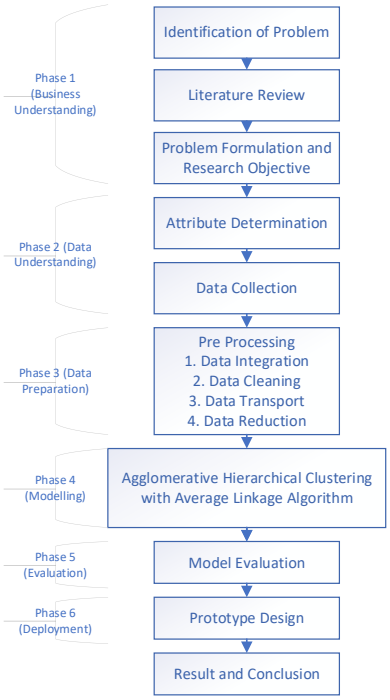

Figure 1. CRISP-DM Phase

Hierarchical Clustering is a method of analyzing clusters by create a hierarchy. Hierarchical Clustering itself is divided into two types, namely Agglomerative (concentration) and Devisive (spread) (Dani, 2019). In this research, we use Agglomerative Hierarchical Clustering with Single Linkage Algorithm and Average Linkage Algorithm. the results of hierarchical clustering are visualized through a dendogram. The dendogram is a form of visual cluster data structure and can be obtained by cutting the dendogram at a certain distance (S. Everitt, 2011).

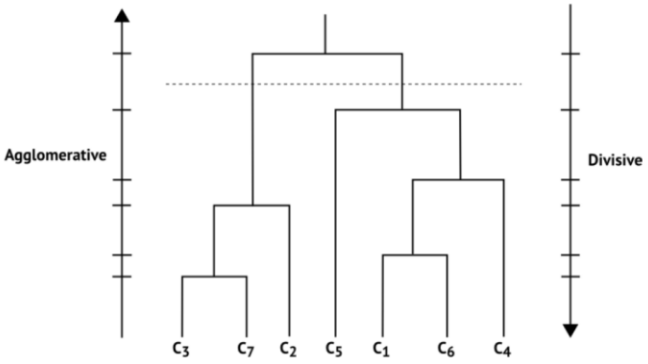

Figure 2. The Result of Grouping the Agglomerative Hierarchical Clustering (S. Everitt, 2011)

Single linkage is a grouping procedure agglomerative based on the smallest distance between objects. Grouping algorithm single linkage begins by selecting the smallest distance in the matrix = $\{dij\}$, then concatenate the corresponding objects e.g. U and V for obtaining clusters (UV). The next step is to find the distance between (UV) with other clusters, for example W so that it can be seen in formula (1):

$$(UV)W = (dUW,) \quad (1)$$

With is the nearest neighbor distance from cluster U and W and is the nearest neighbor distance from clusters V and W. (Dani, 2019)

Average linkage is a grouping procedure agglomerative based on the average between objects. The average linkage algorithm begins by defining the matrix = $\{dij\}$ to obtain the most object close, for example U and V, then these objects are merged into the shape cluster (UV) and then the distance between (UV) and other clusters W, so that can be seen in formula (2):

$$d(UV)W = \frac{d(UW) + d(VW)}{(UV)nW} \quad (2)$$

With (UV) is the number of members in the cluster (UV) and is the number of members in cluster W. (Dani, 2019)

RESULTS AND DISCUSSIONS

As we can see, the phase 1 (Business Understanding) which is an understanding of the background of existing problems also the purpose of data mining research. In this study, we conducted observations, interviews, and literature studies related to existing problems and analysis the steps to be carried out. The interview is purpose to finding the information that related to conditions of existing problems in the marketing division in the segmentation of the prospective new students. The interview resulted in the questions and objectives of the study. Furthermore, a literature study was carried out to find out related research what has been done before and the result will be the references of this research. In expert interviews, namely with division heads of marketing, to find out the relationship between the attributes of existing research, so as to be able to decide the attributes to be used in this study as well as the weight of each of those attributes.

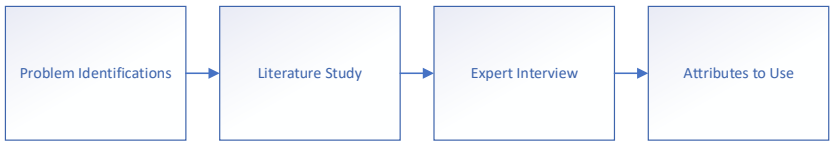

Figure 3. Attribute Determination Chart

In phase 2 (Data Understanding), starting with collecting preliminary data based on the results of observations, interviews, and literature studies conducted. The result, expert decisions in the form of attributes are assumed to be the cause of prospective new students applying to study at the Global Institute of Technology and Business. The data used is data from 4 schools from 2017 until 2018 obtained after spreading questionnaires with students as respondents. The initial attributes in the school database are *Full Name*, *Place of Birth*, *Date of Birth*, *Religion*, *Home Address*, *School*, *Major*, *Facebook*, *Instagram*, *BBM*, *Father's Job*, *Mother's Job*, *Sources of information*, *After Graduation*, *Study Encouragement*, *Interest to Continuing Study*, *Interests Workshops*, and *Workshop Themes*. The data obtained cannot be processed because it has not gone through the data preparation stage. The data obtained are data from SMAN 13 Kab. Tangerang, SMA PGRI 109 Tangerang, SMAN 2. Kab. Tangerang, and SMK Mandiri Balaraja Tangerang. The initial data obtained were 788 records with 18 attributes. Then the next step will be the selection of attributes from the database.

In phase 3 (Data Preparation), we process the initial data obtained which has the format .xlsx. This stage is done manually using Microsoft Excel by combining data from 4 schools into 1 dataset which produces 788 records and 18 attributes. Attributes that cannot be processed for analysis based on interviews with experts will be removed. The deleted attributes are *Full Name*, *Place of Birth*, *Date of Birth*, *Religion*, *Home Address*, *Facebook*, *Instagram* and *BBM*. The next stage, overcoming empty data (Missing Value) on the attributes of the *Major*, *Father's Job*, *Mother's Job*, *Sources of Information*, *After Graduation*, *Interest in Continuing Education*, *Workshop Interest*, and *Workshop Themes*. There are 275 empty data in total, so the empty data will be deleted because it does not allow the data to be searched again. The next stage is data transformation by changing data using the same words and creating new attributes. Categorical attribute data that has the same value but is written differently will be counted as having the same value. Therefore, the values and attributes need to be changed to the same writing. After that, new attributes are created using the One Hot Encoding method (Karthiga, 2021), turning nominal values into new attributes. The attributes that will be transformed into new attributes are *School*, *Major*, *Source of Information* and *Workshop Theme*. This method brings the total number to 22 attributes. After that, scaling will be done on the value of the ordinal attribute. The value attributes that will be scaled are *Father's Job*, *Mother's Job*, *After Graduation*, *Study Encouragement*, *Interest to Continuing Education* and *Workshop Interest*.

Table 1. Attribute Description

Attribute	**Data Type**	**Explanation**
School	Nominal	The origin of the school from prospective new students
Major	Ordinal	Majors of prospective new students
Father's Job	Ordinal	Father's job of the prospective new students who later it will be scaled from the smallest (1) to the biggest (10)
Mother's Job	Ordinal	Mother's job of prospective new students who later it will be scaled from the smallest (1) to the biggest (10)
Sources of Information	Nominal	It is a multiple choice question in questionnaire, "from where the first time knowing about Global Institute?

After Graduation	Ordinal	It is a multiple choice question in school presentation questionnaire, "What is your plan after graduate?" which will be scaling from smallest (1) to the biggest (5)
Study Encouragement	Ordinal	It is a multiple choice question in questionnaire, "for continue your education, you get encouragement from whom?" which will be scaled 44 from the smallest (1) to the biggest (5)
Interest to Continuing Education	Ordinal	It is a multiple choice question in questionnaire, "Are you Interested in studying at Global Institute?" which will be scaled from the smallest (1) to the biggest (5)
Workshop Interest	Ordinal	It is a multiple choice question in questionnaire, "Are you interested in participating in the workshop at Global Institute?" which will be scaled from the smallest (0) and the biggest (1)
Workshop Themes	Nominal	It is a multiple choice question in questionnaire, "If you interested, What theme workshop would you like to attend?"

In phase 4 (Modeling), Agglomerative Hierarchical Clustering (AHC) method was modeled using the Single Linkage algorithm and the Average Linkage algorithm. Each algorithm will be displayed in the form of a dendogram.

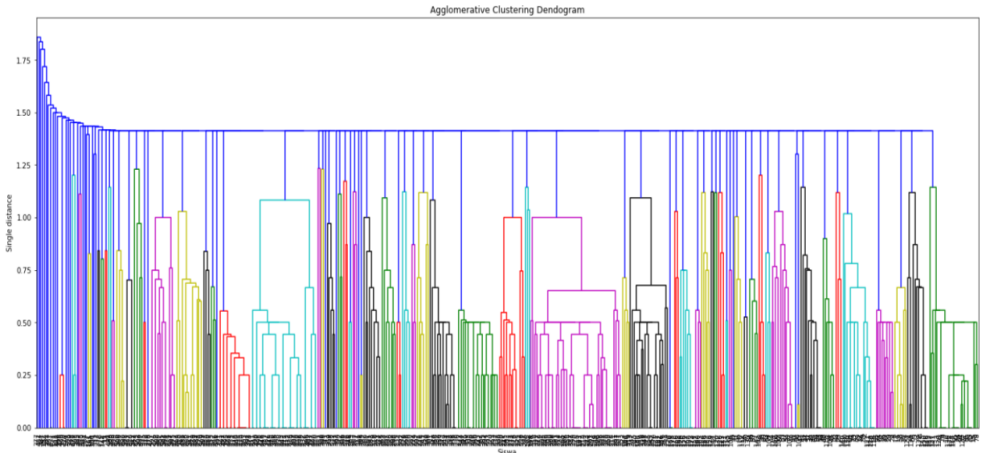

Figure 4. AHC Dendogram Display using Single Linkage

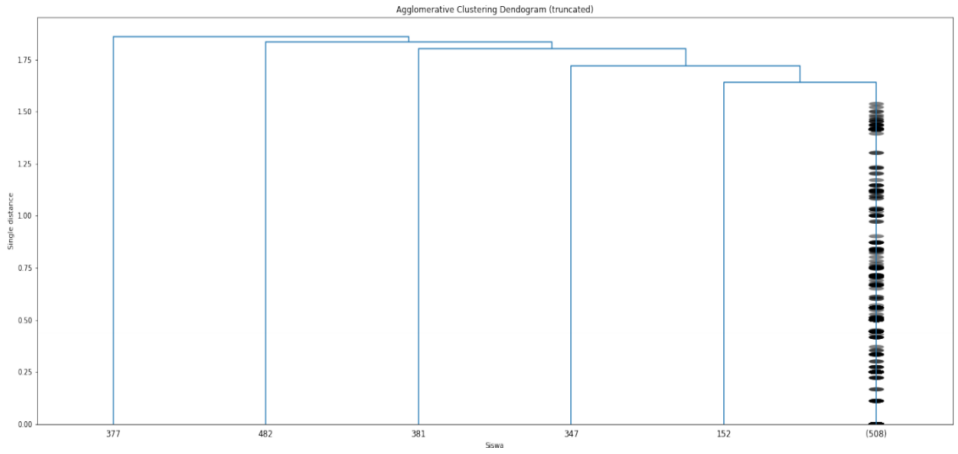

Figure 5. Single Linkage Dendogram with 6 Clusters

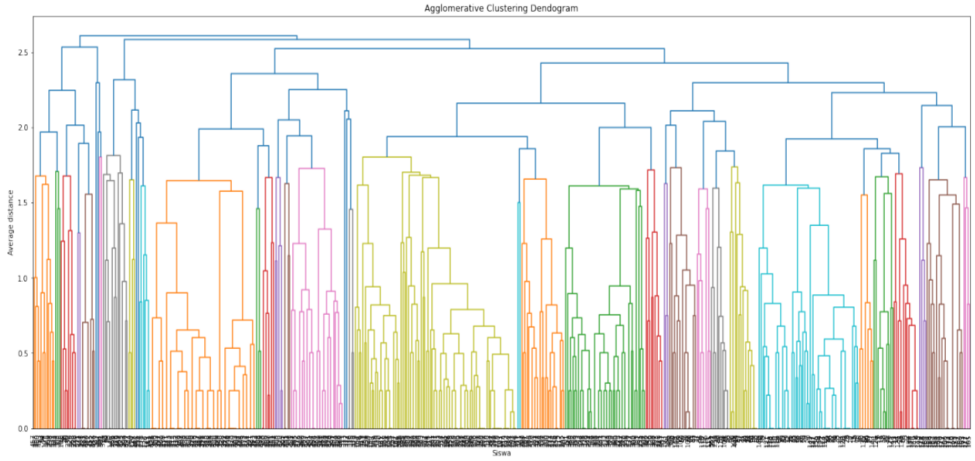

Figure 6. AHC Dendogram Display using Average Linkage

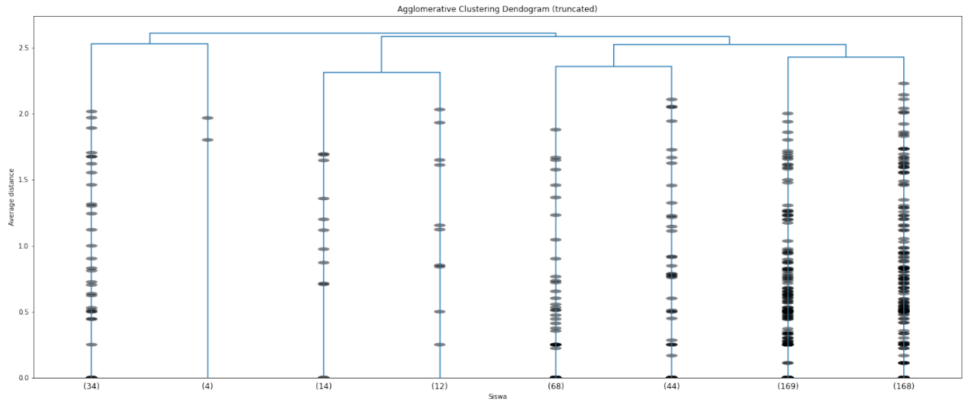

Figure 7. Average Linkage with 8 Clusters

In Phase 5 (Evaluation), we can evaluate the best algorithm obtained from the comparison of the results of the Single Linkage and Average Linkage, and evaluate the model using the Silhouette Score (Shahapure, 2020) and the Calinski Harabasz Index (Lukasik, 2016). In table 2 it can be concluded that the results of the Calinski Harabasz Index better than Silhouette Score because the optimal value of Silhouette Score is less than 1. Silhouette Score looks for the similarity of objects from its own cluster to other clusters. So for the next step will use Calinski Harabasz Index for learning cluster characteristics.

Table 2. Comparison of Single Linkage Results using Silhouette Score and Calinski Harabasz Index

Number of Cluster	***Silhouette Score***	***Calinski Harabasz Index***
2	0.190	2.0194067904061153
3	0.119	1.8750011976856262
4	0.069	1.7944079296543785
5	0.056	1.7786469411905275
6	0.033	1.7704647147587103
7	0.022	1.75724306658903
8	-0.023	1.6951614709294378
9	-0.052	1.7995799726443316
10	-0.073	1.7351529775320982

11	-0.079	1.7145932727153232
12	-0.124	1.6383405332393626
13	-0.134	1.9586294493313468
14	-0.138	1.948322785472903
15	-0.140	1.9411803305965156
16	-0.148	1.9106923792651371
17	-0.154	1.984525737874535
18	-0.157	1.96052319674052
19	-0.160	2.002185975362601
20	-0.163	1.9838076983593782

Table 3. Comparison of Average Linkage Results using Silhouette Score and Calinski Harabasz Index

Number of Cluster	***Silhoutte Score***	***Calinski Harabasz Index***
2	0.112	23.529452669822152
3	0.077	21.065356293357226
4	0.070	15.361921345128351
5	0.107	30.140407658073602
6	0.163	44.81402030254671
7	0.170	**46.92158194280498**
8	0.161	42.115588644179596
9	0.156	37.10325699784731
10	0.171	39.144493081640725
11	0.170	36.42185310593038
12	0.171	34.88318592612475
13	0.189	36.04778176416382
14	**0.224**	41.1169319487013
15	**0.224**	39.03342259520489
16	0.221	37.037686087242434
17	**0.224**	36.82985416390464
18	0.218	34.860150003126236
19	0.217	33.12176308517775
20	0.216	32.16126507788076

Based on the dendogram and table 3 that has been displayed, the characteristics of the existing cluster will be studied. In this study, 7 clusters and several attributes were taken to study their characteristics using chart and tables. After each chart and cluster table has studied, then we make a table of the overall results in table 4 below.

Table 4. Conclusion of Clusters Characteristics

Cluster	Characteristics	Amount of Data
1	Have moderate interest in college, after graduating have a desire for moderate undergraduate studies, low motivation to learn from oneself, low interest in participating in workshops, have moderate father's financial ability, dominant mother is a housewife, mostly from high school, majoring in mostly from IPS, most sources of information from brochures, and most of them have an interest in the theme of web design workshops	34
2	Have low interest in college, after graduation have a low desire for undergraduate studies, low motivation to learn from oneself, low interest in participating in workshops, have low father financial ability, dominant mother is a housewife, mostly from high school, majoring in mostly from IPA, the most sources of information from others, and most of them have an interest in the theme of web design workshops	4
3	Have low interest in college, after graduation have a low desire to go to college, low motivation to learn from oneself, low interest in participating in workshops, have low father financial ability, dominant mother is a housewife, mostly from SMAN 13 Kab. . Tangerang, most of the majors from social studies, the most sources of information from friends/relatives/family, and most of them have an interest in the theme of web design workshops	26
4	Have a high interest in college, after graduation dominantly wants to find a job, high motivation to learn from oneself, high interest in participating in workshops, has low father financial ability, dominant mother is a housewife, overall from SMK Mandiri 2 Balaraja, most major many from Office Administration, the most source of information from brochures, and most of them have an interest in the theme of office automation workshops	68
5	Have low interest in college, low interest in undergraduate studies, moderate motivation to learn from oneself, moderate interest in participating in workshops, high father's financial ability, dominant mother is a housewife, dominant from SMK Mandiri 2 Balaraja, majoring in mostly from science, the most sources of information from school presentations, and most of them have an interest in the theme of office automation workshops	44
6	Have a high interest in college, high interest in undergraduate studies, have a high motivation to learn from oneself, have a high interest in participating in workshops, have a fairly high financial ability of a father, his mother dominant is a housewife, all from SMAN 2 Kab. Tangerang, the dominant major is science, the source of information is mostly from school presentations, and the most popular workshop is office automation	169
7	Many are not yet interested in going to college, have high interest in undergraduate studies, have a high motivation to learn from themselves, have a high interest in participating in workshops, have a high father's financial ability, the dominant mother is a housewife, dominant from SMAN 13 Kab. Tangerang, the dominant major is science, the source of information is mostly from school presentations, and the most popular workshop is office automation	168

Table 5. Clusters Quality

Number of Cluster	*Silhoutte Score*	*Calinski Harabasz Index*
2	0.112	23.529452669822152
3	0.077	21.065356293357226
4	0.070	15.361921345128351
5	0.107	30.140407658073602

6	0.163	44.81402030254671
7	0.170	**46.92158194280498**
8	0.161	42.115588644179596
9	0.156	37.10325699784731
10	0.171	39.144493081640725
11	0.170	36.42185310593038
12	0.171	34.88318592612475
13	0.189	36.04778176416382
14	**0.224**	41.1169319487013
15	**0.224**	39.03342259520489
16	0.221	37.037686087242434
17	**0.224**	36.82985416390464
18	0.218	34.860150003126236
19	0.217	33.12176308517775
20	0.216	32.16126507788076

Based on table 4, it can be concluded that the best characteristics are in cluster 6 because they have high interest in college, high interest in bachelor's degree studies, high motivation to learn from one self, and have high financial parents. In table 5, it can be concluded that the highest cluster quality is on the Calinski-Harabasz Index in the 7th cluster because it has the highest value higher than the other clusters. Meanwhile, Silhouette Score only has the highest score at 0.224, namely in the 14th, 15th, and 17th clusters which have a value less than 1.

In the last is Phase 6 (Deployment). This stage is the prototype design stage for implement the data with the best algorithm that has been selected so that make it easier for users to clustering the prospective new students. The prototype created using the Django Framework with the language python programming. From result the model that has been created in jupyterlab is then used as a model search for clusters with the best characteristics on the prototype made. Results and the data is then saved into the django framework database.

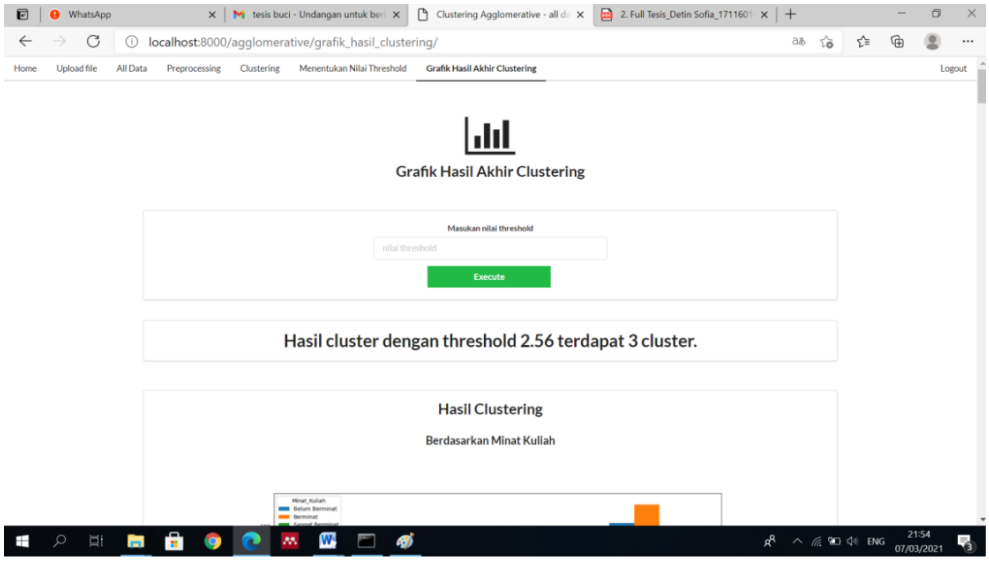

Figure 8. Menu Display Graphics and Table of Threshold Selection Results

CONCLUSION

This study uses the input attributes from *School*, *Major*, *Father's Job*, *Mother's Job*, *Interest in College*, *After Graduation*, *Study Encouragement*, *Interest to Continuing Education*, *Workshop Interest*, and *Workshop Theme* transformed into new attributes uses One Hot Encoding and generates 22 new attributes.

With this research, the marketing division can be helped to find the desired characteristics of prospective new students from the results of the dendogram measured using a threshold produces the best 7 clusters to study cluster characteristics. The best cluster characteristics are in cluster 6 because have high interest in college, high interest in undergraduate studies, encouragement to learn from high self-esteem, and have sufficient high financial father's ability, the dominant mother is a housewife, all from SMAN 2 Kab. Tangerang, the dominant major from science, the most source of information from school presentations, and workshops the most popular is office automation

REFERENCE

B. Everitt and Wiley, 2011, "*Cluster Analysis"* .Wiley InterScience.

Chapman, Peter et al. 2000. "Crisp-Dm." *SPSS inc* 78: 1–78. http://www.crisp-dm.org/CRISPWP-0800.pdf.

Dani, Andrea Tri Rian, Sri Wahyuningsih, and Nanda Arista Rizki. 2019. "Penerapan Hierarchical Clustering Metode Agglomerative Pada Data Runtun Waktu." *Jambura Journal of Mathematics* 1(2): 64–78.

Kansal, Tushar, Suraj Bahuguna, Vishal Singh, and Tanupriya Choudhury. 2018. "Customer Segmentation Using K-Means Clustering." *Proceedings of the International Conference on Computational Techniques, Electronics and Mechanical Systems, CTEMS 2018*: 135–39.

Karthiga, R., G. Usha, N. Raju, and K. Narasimhan. 2021. "Transfer Learning Based Breast Cancer Classification Using One-Hot Encoding Technique." *Proceedings - International Conference on Artificial Intelligence and Smart Systems, ICAIS 2021*: 115–20.

Lukasik, Szymon, Piotr A. Kowalski, Malgorzata Charytanowicz, and Piotr Kulczycki. 2016. "Clustering Using Flower Pollination Algorithm and Calinski-Harabasz Index." *2016 IEEE Congress on Evolutionary Computation, CEC 2016* (1): 2724–28.

Nugraha, Adhitya et al. 2018. "Determining the Senior High School Major Using Agglomerative Hierarchial Clustering Algorithm." *Proceedings - 2018 International Seminar on Application for Technology of Information and Communication: Creative Technology for Human Life, iSemantic 2018*: 225–28.

Schäfer, Franziska, Christian Zeiselmair, and Jonas Becker. 2021. "Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes." *2021 IEEE International Conference on Technology Management, Operations and Decisions, ICTMOD 2021*: 190–95.

Shahapure, Ketan Rajshekhar, and Charles Nicholas. 2020. "Cluster Quality Analysis Using Silhouette Score." *Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020*: 747–48.

Vijaya, Vijaya, Shweta Sharma, and Neha Batra. 2019. "Comparative Study of Single Linkage , Complete Linkage , and Ward Method of Agglomerative Clustering." *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*: 568–73.