IMPLEMENTATION OF OUTLIER ANALYSIS AS ONE OF LEARNING GROUP RECOMMENDATION FOR 6th GRADE STUDENTS IN SDN PAGELARAN II

Idham Nugraha[1], Dian Dharmayanti[2]
[1,2] Program Studi Teknik Informatika
Fakultas Teknik dan Ilmu Komputer Universitas Komputer Indonesia
Jl. Dipatiukur No. 112-114 Bandung
E-mail : idhamnugraha@gmail.com[1], dian.dharmayanti@email.unikom.com[1]

ABSTRACT

SDN Pagerlaran II is an elementary school located in Kampung Sinagar, Desa Pagelaran, Pagelaran district, Cianjur regency. A student must be able to pass the minimum mastery criteria to be graduated. Aside from additional class and try out, the 6th grade students usually set up a learning group for each lesson.

The group formed on the base of student's presentation number. This lead to problem when some students excelled in a certain lesson situated in a same group with other incapable students. This can result to boredom for the excelled student because they repeatedly pushed to learn the same objects.

The solution for the problem explained in the previous paragraph is to find outlier data and set up new groups based on the student's mark from semester 7 to 10. Outlier detection can be done by Distance Based method, and to form the new groups can be done by clustering in data mining. One of algorithm to find these outliers is Local Outlier Factor (LOF) and the algorithm to perform clustering is Agglomerative Hierarchical Clustering (AHC). AHS capable to perform clustering based on the data proximity distance.

Based on the results of the test can be concluded that the Learning Group Former application has helped the curriculum to establish appropriate learning groups based on the similarity of the value of students in each group.

Keywords : Outlier, grouping, local outlier factor, agglomerative hierarchical clustering, data mining.

1. INTRUDUCTION

This section will present the background of problems, intent, purpose and theory used.

1.1. Background Problem

For a student, learning is an obligation. The success or failure of a student in education depends on the learning process experienced by the student. At each grade level, students must pass the exam, starting from daily examination, midterm examination (UTS) and final exam of semester (UAS). SDN Pagelaran II is one of the elementary schools in Cianjur regency, precisely in Kp. Sinagar Village Pagelaran District Pagelaran. As a favorite school, SDN Pagelaran II must improve the quality of education.

Based on the results of observations and interviews conducted with the teachers and the curriculum in SDN Pagelaran II, the provision of taught material is done with the same method without distinguishing students from one another. After a daily review, there are still students who score less than the average grade in the subject and there are students who always get high marks in certain subjects, because each student has different potential from each other. To be declared pass the student must be able to pass the limit of Minimum Criterion of Completity (KKM) which have been determined in each subject. Beside the additional classes and try outs for the 6th grade, it is usually done also the formation of study groups for each subject at each grade level, to form the learning group usually the teacher will divide the students in the absence order or the students who choose their own group friend.

This allows students who have high scores to be in a group with students left behind in the subject. This causes students with high grades to feel bored because the material they have understood is repeated to help low scored students to understand the material.

In order for students to pursue underdeveloped subjects and be able to reach the KKM, the learning group should be shaped according to the student's scores resemblance. So that will facilitate the teacher to identify where the difficulties experienced by students in groups that formed to understand the subject matter that has been given. To obtain such information more quickly and efficiently it is necessary to provide information technology assistance, in this case data mining or methods used to view patterns or large data sets [1]. In data mining there are several methods that can be used depending on the purpose to be achieved. One method that can be used is the method of outlier analysis. Outliers analysis method Anomaly is a data set that is very different from other data [2], while

clustering method is a technique in data mining to classify data that have similarity characteristics with other data. By using the clustering method, the resulting study group will consist of students who have similarities in the academic value of a subject and will be found outlier from each group formed.

So to be able to map students who scored low on each learning groups is by using outlier analysis at the cluster stage.

1.2. Purpose and Goal

1.2.1. Purpose

The purpose of the research to be conducted is the application of outlier analysis in SDN Pagelaran II in the formation of study groups.

1.2.2. Goal

While the goal to be achieved in this research is to help the curriculum:

1. Find outlier data that will become student learning group.
2. Establish study groups based on the resemblance of student scores.

1.3. Data

Data is a value that represents facts in the real world and represents an object such as people (employees, students, buyers, customers), goods, animals, events, concepts, circumstances, etc. The value is then recorded in the form of numbers, letters, symbols, text, images, or combinations [4].

1.4. Basis Data

Database is a set of data that are connected and each have a certain scheme. According to Fathansyah [1], the database itself can be defined in a number of viewpoints such as:

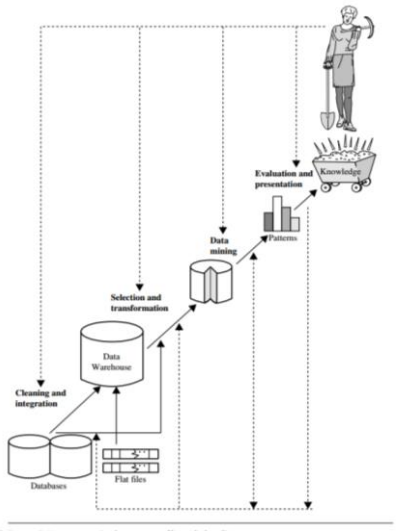
1. The set of interconnected data groups (archives) organized in such a way that one day can be recovered quickly and easily.
2. A collection of interconnected data stored together in such a way without unnecessary redundancies, to meet various needs.
3. Collection of interconnected files or tables or archives stored in electronic storage media.

1.5. *Data Mining*

Data mining is the activity of extracting or mining knowledge from large or large quantities of data. Other names that are often used and have the same meaning with data mining is knowledge mining from data, knowledge extraction, pattern analysis, data archaeology, data dredging.

Data mining attempts to solve the problem by analyzing a set of data already in the database. While the result of the process of minig data is the discovery of knowledge, information, patterns, or models.

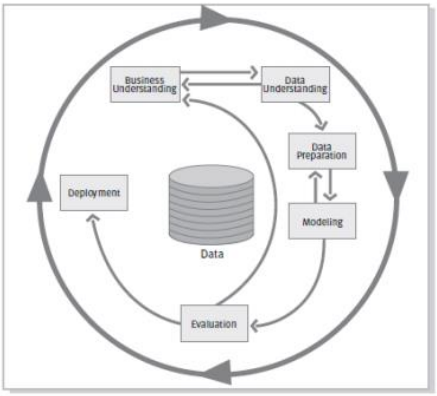
Others see data mining as an important step in the KDD (Knowledge Discovery from Data) process. Images are the steps of the KDD process [2].



Gambar 1. Tahapan KDD

1.6. Metode CRISP-DM

In CRISP-DM, a data mining project has a life cycle that is divided into six phases. The entire sequential phase is adaptive. The next phase in the sequence depends on the output of the previous phase. The important relationships between phases are illustrated by arrows [3]. Here are the phases in the CRISP-DM cycle:



Gambar 2. Proses Data Mining menurut CRISP-DM

1.7. *Outlier Analysis*

Outliers are different from noise data. Noise is the random error or variance in the variable being measured. In general, this noise is unattractive in data analysis, including outlier detection. For example, in credit card fraud detection, customer purchase behavior can be modeled as a random variable. For example, a customer may generate some "transaction noise" that may seem like

"random error" or "variance," such as by buying a larger lunch one day, or having one more cup of coffee than usual. The transaction will not be treated as an outlier; Otherwise, the credit card company will be charged heavily from verifying that many transactions. Companies will also lose customers by annoying them with some false alarms. Like many other data analysis and data mining tasks, noise should be removed before outlier detection.

Outliers menarik karena mereka diduga tidak dihasilkan oleh mekanisme yang sama sebagai sisa data. Oleh karena itu, dalam deteksi *outlier*, penting untuk membenarkan mengapa outliers terdeteksi dihasilkan oleh beberapa mekanisme lain. Hal ini sering dicapai dengan membuat berbagai asumsi di sisa data dan menunjukkan bahwa outliers terdeteksi melanggar anggapan-anggapan secara signifikan. Deteksi *outlier* juga berkaitan dengan *novelty* deteksi dalam berkembang data set. Sebagai contoh, dengan memantau situs web media sosial mana konten baru masuk, *novelty* deteksi dapat mengidentifikasi tren dan topik baru secara tepat waktu. *Novelty* topik mungkin pada awalnya muncul sebagai *outliers*. Sejauh ini, *outlier* deteksi dan *novelty* deteksi berbagi beberapa kesamaan dalam metode Deteksi dan pemodelan. Namun, kritis perbedaan antara keduanya adalah bahwa di *novelty* deteksi, setelah topik baru dikonfirmasi, mereka biasanya dimasukkan ke dalam model perilaku normal sehingga kasus tindak lanjut tidak diperlakukan sebagai *outliers* lagi [2].

Outliers are interesting because they are allegedly not generated by the same mechanism as the rest of the data. Therefore, in outlier detection, it is important to justify why the detected outliers are generated by several other mechanisms. This is often accomplished by making assumptions in the rest of the data and showing that the detected outliers violate the assumptions significantly. Outlier detection is also related to novelty detection in developing data sets. For example, by monitoring social media websites where new content comes in, novelty detection can identify new trends and topics in a timely manner. Novelty topics may initially appear as outliers. So far, outlier detection and novelty detection share some similarities in Detection and modeling methods. However, the critical difference between the two is that in novelty detection, once new topics are confirmed, they are usually incorporated into normal behavioral models so that follow-up cases are not treated as outliers anymore [2].

1. Outlier Detection Method
   Here are some methods to create an outlier detection model: Supervised, Semi-Supervised and Unsuported Methods.

   a. *Supervised Methods*
      Supervised method of normal and abnormal data model. The domain experts check and label the underlying data samples. Supervised methods can then be modeled as a classification problem. Our task is to learn the classifier that can recognize the outliers. Samples are used for training and testing.

   b. *Semi- Supervised Methhods*
      Semi-supervised method outliers can be considered as semi supervised applications of learning. For example, when some labeled normal objects are available, we can use them, along with unmarked objects that are nearby, to train a model for a common object. Ordinary object models can then be used to detect outlier objects not matching the usual object models classified as outliers.

   c. *Unsupervised Methods*
      The unsupervised method outlier makes an implicit assumption: a normal object that can be called a "cluster". In other words, the Unsupervised method outlier expects that ordinary objects follow patterns much more often than outliers. Ordinary objects do not have fallen into one group with a high similarity. Instead, they can form multiple groups, where each group has different features. However, an outlier is expected to occur deep in the feature space of one of the regular object groups.

   In the method of detecting outliers make assumptions about outliers with existing data. So from the assumptions created we can categorize the method of detecting outliers in three types, namely Statistical Method, Proximity-Based Methods, and Clustering-Based Methods [2].

   a. *Statistical Method*
   b. *Proximity-Based Methods*
   c. *Clustering-Based Methods*

2. *Statistical Approaches*
   Statistical method (also known as model-based method) makes the assumption of data normality. They assume that the normal object data is generated by the statistical (statistical) model, and the data does not follow the model's outliers.

3. *Proximity-Based Approaches*
   The proximity-based detection method of the outlier assumes that the object is outlier. If the closeness of the object to its nearest neighbor significantly deviates from the proximity of most other objects to their neighbors the same set of data. Detection method based on outlier distance

to the environment of an object, defined by a certain radius. The object is an outlier if the environment does not have enough other points. In the density-based outlier detection method, the object is an outlier If the energy density is relatively much lower than that of the neighboring country [2]. One of the algorithms that can be used is Local Outlier Factor (LOF).

LOF Algorithm

The main idea of LOF is to compare the local density of an object's environment with its neighboring local density, in the LOF algorithm an object is said to be an outlier if it has a high LOF value or away from 1, whereas objects with LOF values are low or close to 1 then the object can not be categorized As an outlier. High LOF values indicate that the object has a low density to its environment, potentially becoming an outlier. The steps to find outliers are described as follows:

a. Calculates the k-distance of each object p
   The purpose of this k-distance calculation is to determine the neighbor of p, simply k-distance from an object p is the maximum distance from p to its nearest neighbor and denoted by k-distance (p), k-distance is defined d (p, o) Where between p and object o ∈ D has:

(i) For at least k object o '∈ D | {P} and it is stated that d (p, o ') ≤ d ((p, o)
(ii) For each k-1 object o '∈ D | {P} and it is stated that d (p, o ') ≤ d ((p, o)

b. Calculates the nearest neighbor which is no more than the k-distance value.
   K-deistance neighborhood an object p denoted Nk-distance (p), or Nk (p) which contains any object with a distance greater than k-distance (p).

c. Calculates Reachability Distance from p, finds the maximum distance between objects.
   Reachability Distance of an object p to object o is distance (p, o) or k-distance (o), by comparing the two and searched for the maximum value so that the reach-distk (p, o) = max {k-distance (o) , Distance (po)}. The purpose of this calculation is to ensure that all objects are in a homogeneous environment. In addition, the LOF value will be stable if an object is in a uniform environment even if MinPts (k) changes. The fluctuations of the reachability distance can be controlled by selecting the maximum value for k.

d. Menghitung *local reachability density* (kepadatan lokal) dari setiap objek. Dengan persamaan berikut.
   Calculates local reachability density (density) of each object. With the following equation.

$$lrd_{MinPts}(p) = \frac{o \in \sum_{N\ MinPts} reachdist_{MinPts}(p, o)}{|N_{MinPts}(p)|}$$

Information : LrdMinPts (o): local reachability density of the object p. Reachdist (p): Reacbility distance from object p to object o NMinPts (p): Number of neighbors p in a minPts Intuitively, local reachbility density p is the inverse of the average reachability distance based within the nearest minPts of the object p.

e. Calculates the LOF for each data object with the following equation.

$$LOF_{MinPts}(p) = \frac{o \in \sum_{N\ MinPts} \frac{lrdMinPts(o)}{lrdMinPts(p)}}{|N_{MinPts}(p)|}$$

Information :
LOFMinPts (p) : Outlier degrees of the object p LrdMinPts (o) : local reachabiliry distance from object p. Reachdist (p) : Reacbility distance from object p to object o
NMinPts (p) : Number of neighbors p in a minPts

LOF is a degree that determines whether an object is an outlier or not. LOF is denoted into LOFMinPts (p) and is the average local locale density ratio p and neighbor p in a range.

1.8. *Clustering*

Clustering is the process of grouping objects based on information obtained from data explaining relationships between objects and principles to maximize the similarity between members of a class and minimize the similarity between classes or clusters [2]. The goal is to find a quality cluster in a decent time. Clustering in data mining is useful for finding distribution patterns within a data set that is useful for data analysis processes. The similarity of objects is usually derived from the proximity of attribute values that describe data objects, whereas data objects are usually represented as a point in a multidimensional space.

Using clustering, identifiable densely populated areas, overall distribution patterns and attractive associations between data attributes. In business data mining is focused on the discovery methods for

clusters on large databases effectively and efficiently.

1.9. *Agglomerative Hierarchical Clustering*

Grouping data using agglomerative hierarchical clustering (AHC) begins by representing each object in the data as a group, then distance calculations are made between groups. After that the two closest spaced groups are combined into a new group. Then the distance between the new group and the other group is calculated using one method of calculating the distance between groups (single linkage, complete linkage, average Linkage, etc.) [4]. The process of calculating the distance and merging the two groups is done iteratively. Here are some methods that can be used to calculate distances between groups:

a. *Single Linkage* (Jarak Terdekat)
   Single linkage is also called minimum link, where the similarity of the two. The cluster is based on two closest points of two different clusters. To find a single linkage can use the following equation (2.8):

$$d_{uv} = \min\{d_{uv}\}, d_{uv} \in D$$

b. *Complete Linkage* (Jarak Terjauh)
   Complete linkage is also called the maximum link, where the distance between two clusters is calculated from the farthest distance between cluster members one with the second cluster. To find complete linkage can use equation (2.9) below:

$$d_{uv} = \max\{d_{uv}\}, d_{uv} \in D$$

c. *Average Linkage* (Jarak Rata-rata)
   Average linkage is also called group average, where the distance between two clusters is calculated from the average distance between cluster members one with the second cluster. To find the average linkage can use the following equation (2.10):

$$d_{uv} = \text{average}\{d_{uv}\}, d_{uv} \in D$$

The following are the steps in the Agglomerative Hierarchical Clustering (AHC) algorithm :
1. Each data object is considered a cluster. If n = amount of data and k = number of clusters, then k = n.
2. Calculating the distance matrix between clusters (using euclidean distance)
3. Find two clusters that have the closest distance (single linkage) then combine the two clusters, so k = k-1.
4. If k> 1, repeat steps 2 and 3
5. If k = 1, the loop is complete.

1.10. *Usability Testing*

Usability testing can be done by involving users or without involving users. Testing by involving the user can provide information directly from the user about how users use the system and the problems encountered. In the direct observation method, an observer observes one or more users at the site. Preferably an observer does not show itself to test the system running normally. To assist the observation process can use video cameras that record user expression. The questionnaire method is done by using questionnaires to measure user satisfaction and to know the user's opinion on the application used. The Thinking Aloud method is done by continuously observing the posture and speech shown by a respondent when using the app [5].

The level of effectiveness and efficiency is measured by using the user's success rate (the level of success of the user) ie the percentage of tasks completed correctly by the user. Effectiveness and efficiency are calculated using the following equation:

$$\textbf{effectiveness, efisiency } (\%) = \frac{(\sum_{i=1}^{n} Xi)}{n} \textbf{ x 100\%}$$

Information :
Xi : User success rate, = {0,1}

Satisfaction is the percentage comparison between the value of user satisfaction with multiplication between the maximum weight of Likert scale with the number of respondents

$$\textbf{Satisfaction}(\%) = \frac{(\sum_{i=1}^{n} Xi)}{5 \text{ x } n} \textbf{ x 100\%}$$

Keterangan :
Xi : nilai keberhasilan pengguna, Xi = {0, 1, 2, 3, 4, 5}
Information :
Xi: User success rate, = {0, 1, 2, 3, 4, 5}
Application usability is the average of effectiveness, efficiency and satisfaction, as written in the following equation:

Application usability is the average of effectiveness, efficiency and satisfaction, as written in the following equation :

$$\textit{Usability}(\%) = \frac{(\text{Efektivitas} + \text{Efisiensi} + \text{Kepuasan})}{3} \textbf{x100\%}$$

2. THE CONTENS OF RESEARCH

Berikut adalah analisis dan pengujian yang dilakukan dalam penelitian
Here is the analysis and testing conducted in the study

3.1. Problem Analysis

Berdasarkan hasil observasi dan wawancara dengan pihak guru dan pihak kurikulum SDN Pagelaran II, didapatkan hasil yaitu terdapat siswa dengan nilai rendah atau kurang dari kriteria ketuntasan minimum (KKM) dari mata pelajaran tertentu.

Based on the observation and interviews with the teachers and the SDN Pagelaran II curriculum, the result is that there are students with low score or less than the minimum completeness criteria (KKM) of certain subjects.

3.2. Business Understanding

The business understanding stage is the initial process in the CRISP-DM framework. This stage focuses on understanding the objectives and needs from the SDN business point of view Pagelaran II. There are several tasks at this stage, including:

1.2.1. Identification of Business Objectives

The business purpose of SDN II is to conduct learning process and guide all students, especially students who always get low score in some subjects in order to pass on each subject with good value, whether by way of direct learning method / learning approach Teacher-oriented, learning with discussion, learning with small groups, or by helping students to understand and master the learning materials of problem-solving strategies.

1.2.2. Identification of Business Situation Assessment

Assessment of the current business situation in SDN Pagelaran II includes a list of computing resources (hardware), and human resources (user) in SDN Pagelaran II.

1.2.3. Determination of Data Mining Goals

The targeting of data mining consists of the objectives and criteria of data mining success. The following are the objectives and criteria of data mining success.

1. The purpose of data mining The purpose of applying data mining in this research is to find outlier, and to form study group based on the value of subjects including the National Examination Exam grade in the 6th grade students of the academic year 2016/2017, from semester 7 to semester 10 for those with characteristics same.

2. Criteria for successful data mining The criteria of success in this research is if able to assist the teacher and the curriculum in SDN Pagelaran II in finding students who are included into the outlier.

1.2.4. Determining the Project Plan

The project plan of this research is to apply the outlier analysis on the students' grade in SDN Pagelaran II with Density-Based technique and Local Outlier Factor (LOF) algorithm, and to form cluster used Agglomeration Hierarchical Clustering (AHC) algorithm.

3.3. Data Understanding

Stage of understanding of data is the second stage in which there are several tasks, namely:

1.3.1. Initial Data Collection

Preliminary data that have been collected is the data of the value of report cards of students of grade 6 of 2016/2017, for subjects included in the national examination of Indonesian, Mathematics and Natural Sciences lessons from the academic year 2014/2015 and 2015/2016 or grades since the semester 7. Student value data will be used in the form of files that have the format * .xls or * .xlsx.

1.3.2. Data Explanation

Preliminary data collected will be explained at this stage.

1.3.3. Data Exploration

Stages of data exploration can help the purpose of data mining, in this study data exploration includes descriptive analysis and data visualization. The data to be used for this data exploration phase is the data of the 6th grade students of the academic year 2016/2017 for mathematics, science and Indonesian subjects.

1.3.4. Evaluation Of Data Quality

Evaluation of data quality is done to attribute the value of subjects Mathematics, Science and Bahasa Indonesia from semester 7 to semester 10. Here is an evaluation result of these attributes:

1. Missing value may be found in the value data, but in the data of the 6th grade students of the 2016/2017 academic year for Mathematics, Science and Bahasa Indonesia values from semester 7 to 10th semester contained in Appendix F, there is no missing value.
2. Based on the data of the value of 6th grade students of the academic year of 2016/2017 for the value of Mathematics, Science and Indonesian from the 7th to the 10th semester will be seen whether there are outliers or not, here is how to find outlier data using method Local Outlier Factor:
    1. Calculate the distance k-distance of each object p or from each data.

To find the k-distance value required several stages like the following:

a. Based on student scores, then calculate the distance of each object using the distance euclidian distance function. In this stage each object will be calculated distance by itself and distance to other objects. Here is an example of the process of calculating the distance between objects using the euclidean distance equation.

$$d(1,1) = \sqrt{\begin{matrix}(85-85)^2+(80-80)^2+(85-85)^2+\\(75-75)^2+(89-89)^2+(85-85)^2+\\(76-76)^2+(80-80)^2+(75-75)^2+\\(82-82)^2+(75-75)^2+(76-76)^2\end{matrix}}$$

b. Specify MinPts

MinPts are the minimal amount of data so a cluster can be formed. From several experiments performed on the data, assuming the value of MinPts = 9, so based on the Initial Matrix, selected 9 first closest distance from an object then from the search results 9 closest distance will be searched K-Distance value which is the furthest distance of 9 distance Closest to found.

Tabel 1 Tabel K-Distance

No Urut	Nomor Induk	*K-distance*	Data Dari Tetangga Ke
1	11.12.01.006	26,46	14
2	11.12.01.010	29,33	7
3	11.12.01.004	23,20	7
4	11.12.01.002	26,51	13
5	11.12.01.003	30,30	19
6	11.12.01.005	26,25	9
7	11.12.01.007	29,75	18
8	11.12.01.011	44,93	6
9	11.12.01.012	29,40	5
10	11.12.01.013	25,06	19
11	11.12.01.014	41,14	18
12	11.12.01.015	44,90	3
13	11.12.01.016	32,28	2
14	11.12.01.017	30,41	5
15	11.12.01.018	48,20	1
16	11.12.01.020	46,72	13
17	11.12.01.022	47,68	5
18	11.12.01.023	39,76	5
19	11.12.01.024	30,30	5

1. Then find any neighbor whose distance does not exceed from k-distance of that object.

Tabel 2 Tabel Data K-Distance

No Urut	Nomor induk	*k-distance neighborhood*
1	11.12.01.006	2, 4, 5, 6, 9, 10, 13, 14, 19
2	11.12.01.010	1, 3, 4, 5, 6, 7, 9, 10, 14
3	11.12.01.004	2, 4, 5, 6, 7, 9, 10, 14, 19
4	11.12.01.002	1, 3, 6, 7, 9, 10, 13, 14, 19
5	11.12.01.003	1, 2, 3, 4, 6, 7, 9, 10, 19
6	11.12.01.005	1, 2, 3, 4, 7, 9, 10, 13, 18
7	11.12.01.007	1, 2, 3, 4, 5, 6, 10, 18, 19
8	11.12.01.011	1, 2, 6, 11, 12, 13, 15, 16, 17
9	11.12.01.012	1, 2, 3, 4, 5, 6, 10, 14, 19
10	11.12.01.013	1, 3, 4, 5, 6, 7, 9, 13, 19
11	11.12.01.014	2, 6, 8, 12, 13, 15, 16, 17, 18
12	11.12.01.015	1, 2, 3, 8, 11, 13, 15, 16, 17
13	11.12.01.016	1, 3, 4, 5, 6, 7, 9, 10, 14
14	11.12.01.017	1, 2, 3, 4, 5, 6, 9, 10, 19
15	11.12.01.018	1, 2, 6, 8, 11, 12, 13, 16, 17
16	11.12.01.020	2, 6, 8, 11, 12, 13, 15, 17, 18
17	11.12.01.022	1, 2, 5, 8, 11, 12, 13, 15, 16
18	11.12.01.023	2, 3, 4, 5, 6, 7, 9, 10, 13
19	11.12.01.024	1, 2, 3, 4, 5, 7, 9, 10, 14

2. Calculate Reachablitiy Distance from p or from each data

Tabel 3 Data Reachability Distance

	2	4	5	6	9	10	13	14	19
1	21,977	23,07	25,44	22,52	25,68	22,38	24,06	26,46	25,63
K-Distance	29,33	26,51	30,30	26,25	29,40	25,06	32,28	30,41	30,30
Reachability Distance	29,33	26,51	30,30	26,25	29,40	25,06	32,28	30,41	30,30

3. Calculate local reachability density drom each object toward its neighbors.

$$lrd_{MinPts}(1) = \frac{29.33 + 26.51 + 30.30 + 26.25 + 29.40 + 25.06 + 32.28 + 30.41 +}{9}$$

$$= \frac{259.84}{9} = 28.87$$

No Urut	Nomor induk	*Local Reachbility Density*
1	11.12.01.006	28,87
2	11.12.01.010	28,14
3	11.12.01.004	28,59
4	11.12.01.002	28,12
5	11.12.01.003	27,73
6	11.12.01.005	29,08
7	11.12.01.007	28,59
8	11.12.01.011	43,17
9	11.12.01.012	27,54
10	11.12.01.013	28,27
11	11.12.01.014	43,36
12	11.12.01.015	44,27
13	11.12.01.016	28,82
14	11.12.01.017	27,93
15	11.12.01.018	45,12
16	11.12.01.020	45,10
17	11.12.01.022	44,91
18	11.12.01.023	33,06
19	11.12.01.024	27,82

4. After obtaining k-distance and local reachbility density, next is calculating Local Outlier Factor (LOF) for each data object.

$$LOF_{MinPts}(1) = \frac{\frac{28.87}{28.14}+\frac{28.87}{28.12}+\frac{28.87}{27.73}+\frac{28.87}{29.08}+\frac{28.87}{27.54}+\frac{28.87}{28.27}+\frac{28.87}{28.82}+\frac{28.87}{28.93}+\frac{28.87}{27.82}}{9} = 1.03$$

No Urut	Nomor induk	LOF
1	11.12.01.006	1,03
2	11.12.01.010	1,00
3	11.12.01.004	1,02
4	11.12.01.002	0,99
5	11.12.01.003	0,98
6	11.12.01.005	1,01
7	11.12.01.007	0,99
8	11.12.01.011	1,21
9	11.12.01.012	0,97
10	11.12.01.013	1,00
11	11.12.01.014	1,19
12	11.12.01.015	1,24
13	11.12.01.016	1,02
14	11.12.01.017	0,99
15	11.12.01.018	1,27
16	11.12.01.020	1,24
17	11.12.01.022	1,27
18	11.12.01.023	1,17
19	11.12.01.024	0,99

3.4. Persiapan Data (Data preparation)

This stage includes all the activities required to build the final set of data (data to be incorporated into the modeling), there are several tasks in this phase:

1.4.1. Data Selection

Data selection has the task of including the selection of data attributes used, so that data can be processed according to the needs of data mining. In this stage the attributes used are No Parent, Mathematics, Science and Indonesian (semester 7-10).

1.4.2. Data Cleaning

In this process, the data cleaning process, based on the result of quality evaluation, found that on the grade 5 data for mathematics subjects in the lesson year 2015/2016 missing value not found, but if there is missing value, then the missing value will be filled with the default value 0, and the value that is outlier will not be in the process at the stage of making the group and will be a separate group, to do the grouping of students will be used agglomerative hierarchical clustering algorithm, that is by smoothing the data using binning method.

1.4.3. Initial Data Setup

In the initial data preparation process will be prepared data to be used in the modeling stage. The data to be used in the modeling stage is the data of grade 6 grade lesson 2016/2017 from semester 7-10.

1.4.4. Data Format

In the process of data format is the determination of the data format to be used in data processing in the modeling stage. The data to be used in the modeling stage is the grade 6 data of the academic year 2016/2017 from semester 7-10 in the form of .xlsx or .xls file format.

3.5. Modelling

This stage will be modeling, at this stage has several tasks namely the selection of modeling techniques, explanation of model mechanisms, and model testing on data that has been prepared previously.

2.5.1. Modeling Technique

The modeling technique that will be used is hierarchical clustering, and the model will be used for agglomerative hierarchical clustering (AHC) algorithm.

2.5.2. Analisis Pengujian AHC

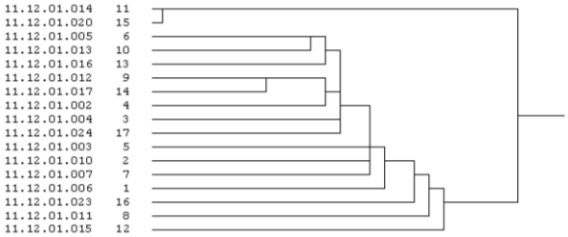
The stages of testing the AHC model are as follows:

1. Change each data into clusters Initial cluster initialization is as much as 19 clusters because the data (n) = cluster (c), then calculated the distance between the clusters with the existing cluster by using the Euclidean Distance equation.
2. Calculate the distance between data In this stage each cluster will be calculated distance by itself and distance with other clusters. The following is a process of calculating the distance between clusters by using Euclidean equations Distance.

$$d(1,2) = \sqrt{\begin{matrix}(85-77)^2+(82-76)^2+(83-78)^2+\\(75-89)^2+(89-84)^2+(85-82)^2+\\(75-85)^2+(80-77)^2+(75-71)^2+\\(83-77)^2+(75-71)^2+(78-78)^2\end{matrix}}$$

= 23.07

3. Look for the two closest cluster clusters.
4. If k> 1, repeat steps 2 and 3 Steps 3 and 4 continue to be done and will stop when the number of clusters formed is equal to 1. To get 1 cluster, in this case it needs to be 18 iterations.
5. If k = 1, the loop is complete. The number of clusters will amount to 1 when the 18th iteration.



Gambar 3 Hasil Akhir Dendogram

6. Generate the cluster by cutting the dendogram at the appropriate level based on the dendogram that has been established, the right level or the best number of clusters will be determined by looking at the distance difference of each cluster formed. To find the best cluster number one can use agglomeration schedule. Agglomeration Schedule can be seen in the following table.

Tabel 4 *Agglomeration Schedule*

Iterasi	Kombinasi Cluster		Koefisien	Iterasi Kombinasi Cluster Sebelumnya		Iterasi Selanjutnya
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	11	15	0	0	0	16
2	9	14	11.057	0	0	5
3	6	10	14.629	0	0	4
4	6	13	16.125	3	0	6
5	4	9	16.348	0	2	6
6	4	6	17.521	5	4	7
7	3	4	17.671	0	6	8
8	3	17	17.896	7	0	9
9	3	5	20.839	8	0	10
10	2	3	21.172	0	9	11
11	2	7	21.448	10	0	12
12	1	2	21.977	0	11	13
13	1	16	25.08	12	0	14
14	1	8	26.87	13	0	15
15	1	12	27.713	14	0	16
16	1	11	36.073	15	1	0

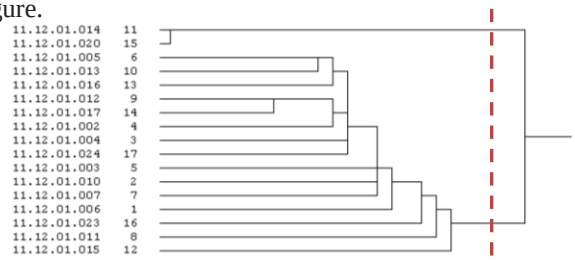
Jumlah *cluster* yang terbaik dapat dilihat dari selisih koefisien dari setiap iterasinya. Selisih dari koefisien dapat dilihat pada tabel berikut.

The best number of clusters can be seen from the difference coefficient of each iteration. The difference of the coefficient can be seen in the following table.

Tabel 5 Selisih Koefisien

Jumlah Cluster	Tahap Terakhir Agglomeration	Koefisien Tahap Ini	Selisih
2	36.073	27.713	8.36
3	27.713	26.87	0.843
4	26.87	25.08	1.79
5	25.08	21.977	3.103
6	21.977	21.448	0.529
7	21.448	21.172	0.276
8	21.172	20.839	0.333
9	20.839	17.896	2.943
10	17.896	17.671	0.225
11	17.671	17.521	0.15
12	17.521	16.348	1.173
13	16.348	16.125	0.223
14	16.125	14.629	1.496
15	14.629	11.057	3.572

The biggest difference is 8.36 so the best cluster is 2 clusters. Thus, the dendogram of groupings by size will be cropped as shown in the following figure.



Gambar 4 Pemotongan *Cluster*

So the group formed into 3 groups, with the group formed from the outlier data.

3.6. Testing

Application testing is performed on the application's functionality to assess whether the built application meets the criteria or there is a discrepancy in the functionality as contained in the design of the application. Testing conducted to test the application is a black box testing and usability testing.

2.6.1. Black Box Testing

Black box testing is focused on Functional requirements of software built. Testing is done by testing each process for possible errors that occur. Based on the test results with the case of test samples that have been done can be concluded that all functions run well and in accordance with the expected.

2.6.2. Usability Testing

Beta testing is a test done directly in the real environment. Users provide an assessment of the applications built through the Usability Testing method. Based on the results of these tests it can be concluded whether the application is built in accordance with the purpose or not. Testing is done with the user of the built application, ie the Curriculum Party. This test is done in SDN Pagelaran II in order to know the extent to which the built application can be a solution and problem solving in accordance with the initial purpose of application development.

3. Closing

Based on the results of research that has been done, generate conclusions and suggestions as follows:

Conclusion

The results of the research conducted concluded as follows:

1. That the built application helps the curriculum side to find outliers of student value data.
2. Applications built to form a group of learners in accordance with the similarity of student scores.

Suggestion

There are several suggestions that can be done for the application development of this group of distributors, among others:

1. Added features of the academic year setting to improve the accuracy of the clustering and enable the user to filter the clustering data to be used when the data are grouped by a mixture of different school year.
2. Grouping development can be developed again by using other algorithms so that the user can compare the results of grouping with Single Linkage algorithm.

BIBLIOGRAPHY

[1] B. Santosa, Data Mining Teknik Pengumpulan Data Untuk Keperluan Bisnis, Yogyakarta: Andi, 2007.

[2] F. A. Hermawati, Data Mining, Surabaya: Andi, 2009.

[3] K. and E. T. Luthfi, Algoritma Data Mining, Yogyakarta: Andi, 2009.

[4] J. Han, M. Kamber and J. Pei, Data Mining Concept and Technique, Waltham: Elsevier, 2012.

[5] P. Berkhin, "Survey of Clustering Data Mining Techniques".

[6] Y. Nurhadryani, S. K. Sianturi, I. Hermadi and H. Khotimah, "Pengujian Usability Testing untuk meningkatkan Antarmuka Apliaksi Mobile," *Ilmu Komputer & Agri-Informatika*, vol. 2, pp. 83-93, 2013.

[7] Fathansyah, Basis Data, Informatika, 2015.

[8] A. Nugroho, Rekayasa Perangkat Lunak Berorientasi Objek dengan Motode USDP, Yogyakarta: Andi, 2010.

[9] R. Ramakrishnan and J. Gehrke, Database Management System Third Edition, Singapore, 2003.