The Third International Conference
on Innovations in Social Sciences Education and Engineering (ICoISSEE)-3
Bandung, Indonesia, July 08th, 2023

UTILIZATION OF K-MEANS CLUSTERING METHOD IN GROUPING DATA ON PARENTS' INCOME OF STMIK DHARMAPALA RIAU STUDENTS

Fery Wongso

Human resource management,economy Riau University,Pekanbaru,Indonesia

Author's email :
fery_wongs@yahoo.com

Abstract. *Not a few children in Indonesia who drop out of continuing their education because one of the causes is family economic factors, insufficient parental income and not balanced with the number of dependents in the family, therefore researchers conducted a survey with questionnaires to students at STMIK Dharmapala Riau in order to determine the variables of grouping student parents' income data, in order to determine the economic level of students with the category of able and less able to use K-Means Clustering method assisted by data processing using the Rapidminer 5.3 application. The data criteria used in this method include the last student achievement index, father's job, mother's job, the amount of income of father and mother in one month and the number of dependents in the family. The K-Means Clustering method is one method to group data based on its characteristics, so that data that has the same characteristics are grouped in the same cluster and data that has different characteristics are grouped in another cluster. The resulting clusters are categorized as capable and underprivileged, which can be useful in making decisions.*

Keywords: *Data Mining, K-Means Clustering, RapidMiner*

1. INTRODUCTION

Currently, in various parts of the world, including Indonesia, disasters related to health, the cause of which we know as Covid-19 (Corona Virus Desease Nineteen). **Coronavirus is part of a large family of viruses that cause diseases that occur in animals or humans. Humans infected with the virus will show signs of respiratory infections ranging from flu to more serious, such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS) or severe acute respiratory syndrome. Coronavirus itself is a new type discovered by humans since it appeared in Wuhan, China in December 2019, and named Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV2). Thus, this disease is called Coronavirus Disease-2019 (COVID-19). Source (World Health Organization / WHO: 2020).** According to Susilo in (Latest Literature Review: 2019), Coronavirus is a virus that attacks the respiratory tract. Viruses associated with infections in the respiratory tract will use epithelial cells and airway mucosa as initial targets and cause infections in the respiratory tract or organ damage. The impact of this virus is so widespread, not only threatening public health but also the economy of society and education in Indonesia. During the Covid-19 pandemic, the economy of several families has been affected, while children's education must continue. Children continue to learn from home online which requires android mobile phones, laptops, credit, internet packages, and other learning support facilities. If the family economy is difficult to meet it all, this will prevent children from continuing their education to the next level, not a few children in Indonesia who drop out of continuing their education because one of the causes is family economic factors, insufficient parental income and not balanced with the number of dependents in the family. Especially now that Indonesia has entered an economic recession, everything is difficult, family finances are running low so the step that must be taken is to manage family finances as well as possible. Given such conditions, families must think hard and try to be more persistent so that children can continue their education, because education plays an important role in improving quality human resources. Therefore, researchers do not want any students, especially at STMIK Dharmapala Riau, to drop out of college because the income of students' parents is insufficient to pay in full for education costs every semester increase, so researchers intend to conduct a survey of students with criteria grouping parents' fields of work, parents' monthly income, and the number of dependents in one family. From the results of the survey will then be processed using the calculation of the K-means clustering method, with this method it will make it easier for universities to find out the economic level of families of capable and underprivileged students then can be taken into consideration by the university in providing scholarship assistance or administrative relief to students who are considered entitled to receive it to be right on target. From the above problems, the author decided to make a study entitled "Utilization of the K-Means Clustering Method in Grouping Data on the Income of Parents of STMIK Dharmapala Riau Students". Problem Statement Based on the problems that have been explained in the background, the problems that can be formulated are as follows: How to determine data variables in grouping the income of parents of STMIK Dharmapala Riau students? How to apply the K-Means Clustering method to data mining to group user data

2. LITERATURE REVIEW

The process of searching and extracting information from large amounts of data is the main process of data mining, the main purpose of processing data is to produce new information. Data mining also called Knowledge Discovery in Database (KDD) is a process that automatically searches data in a very large memory space of data to find patterns using techniques such as association classification or clustering. (Muliono, et al 2019: 273).

In general, data mining is a method in computer science that is commonly used in the process of searching for knowledge. The stages in it are useful for looking for certain patterns of data in the database. Usually, this method is widely found in the field of machine learning and statistics. According to (Fayyad in Suyanto, 2017). The terms data mining and knowledge discovery in databases (KDD) are often used interchangeably to describe the process of extracting hidden information in a large database. Actually, the two terms have different concepts, but they are related to each other. One of the stages in the entire KDD process is data mining. The KDD process has 5 stages that are carried out in order, namely: Data selection. Data selection from a set of operational data needs to be done before the information mining stage in KDD begins. The selected data used for the data mining process is stored in a file, separate from the operational database. Pre-processing / cleaning. Before the data mining process can be carried out, it is necessary to clean the data that is the focus of KDD. The cleaning process includes, among others, removing duplicate data, checking for inconsistent data, and correcting errors in the data. Transformation Coding is the process of transformation on selected data, so that the data is suitable for the data mining process. The coding process in KDD is a creative process and depends largely on the type or pattern of information to be searched in the database Data mining is the process of looking for interesting patterns or information in selected data using certain techniques or methods. Techniques, methods, or algorithms in data mining vary widely. The selection of the right method or algorithm depends largely on the overall goals and processes of KDD. Interpretation / evalution. Information patterns resulting from the data mining process need to be displayed in a form that is easily understood by interested parties. Stage 9 is part of a KDD process called interpretation. This stage includes examining whether the patterns or information found contradict previously existing facts or hypotheses. 2.1.1 Uses of Data Mining Broadly speaking, the usefulness of data mining is divided into two, namely: descriptive and predictive. Descriptively, data mining means finding patterns used to explain the characteristics of data. While predictively data mining means that it can be used to find knowledge models that are used to make predictions. Data mining based on its functionality can be grouped into six parts, namely (Suyanto, 2017): Classification Applied to new data to group object types. The classification belongs to the supervised model. In classification problems we have sample data and predict several existing classes based on existing samples. Only one attribute among many attributes is called the target attribute, while the other is called the predator attribute. This classification is also commonly used for business modeling and others. For example, classification determines certain diseases or determines customers based on their payment models. Clustering. Unlike classification, clustering includes unsupervised models. Clustering groups data that doesn't know the label. Clustering an organization into a hierarchical structure defines the taxonomy of the data. The application of the right clustering method will result in quality clustering. A cluster is characterized by centroid, histogram attributes and hierarchial tree model clustering. Regression is a function used to model data to minimize the results of prediction errors. Generally, regression is done with data that is time series. The Association Rule is a dependency model. This association function is usually known as "market basket analysis" which is a function to find relationships or correlations between sets of items. Association rules are defined on baskets of data used for promotional purposes, catalog design to increase sales. Anomaly Detection identifies uncommon data. Can be an outlier. Changes in deviation/bias are important and need further investigation. Summarization. Provides simpler data representation including reporting, visualization of data used to support information and decision reinforcement. Data Mining Architecture The main architecture of a data mining system generally consists of several components as follows (Suyanto, 2017): A database, data warehouse, or information storage medium, consists of one or more databases, data warehouses, or data in other forms. Data cleansing and data integration are carried out on the data. The database, data warehose, is responsible for finding relevant data according to what the user wants. Knowledge Base, is a knowledge base that is used as a guide in finding patterns. The data mining engine is an important part of the system and ideally consists of a collection of function modules used in characterization, clasiffication, and cluster analysis. And is a part of software that runs programs based on existing algorithms Pattern evaluation, this component generally interacts with data mining modules. And part of the software that serves to find patterns or patterns contained in the database that is processed so that later the data mining process can find the appropriate knowledge. Interface (Graphical user interface), is a communication module between users or users with the system that allows users to interact with the system to determine the data mining process itself.

How does data mining work is to "dig" important things that have not been known before or predict what will happen? The technique used to carry out this task is called modeling. Modeling here is intended as an activity to build a model on a siasi that has known "answers" and then apply it to other situations for which the answer will be sought. Data mining to find patterns in data. The patterns found must be meaningful and the patterns provide an advantage. The characteristics of data mining are as follows: Data mining deals with the discovery of something hidden and certain patterns of data that were not known before. Data mining usually uses very big data. Usually big data is used to make results more reliable. Data mining is useful for making critical decisions, especially in strategy, it can also be used for future decision making based on information obtained from past data. Depending on the application, data can be student data, patient data, customer data or sales. Many cases in everyday life can be unwittingly solved by data mining, including: Predict stock prices in the next few months based on company performance and economic data. Predict how many new students are in college based on registrant data in previous years. Predict student achievement index scores based on

IP scores of each previous semester. What products will customers buy at the same time if they buy products in supermarkets. How to find out the characteristics of customers who have or current or bad bad credit in a banking or finance. Group customers based on interests, or habitual patterns so that it is easier to determine marketing targets, and others. Of course, there are many more examples in other fields or other cases related to data mining so that it can produce new knowledge and information into a strategy in developing a business field. Definition of K-Means K-Means is one of the clustering techniques in Data Mining, an unattended modeling process and a method of grouping data in partitions. Data are grouped by the K-Means method into groups and each group has similar or similar characteristics to the other but with other groups having different characteristics. With the aim of minimizing the difference of each data in one cluster and maximizing differences with other clusters. (P. Sari, et al 143-148: 2017). Terms in k-means clustering algorithm: Cluster: A cluster is a group or groups. Cendroid: The cendroid is the central point for determining the euclidian distance. Iteration: Iteration is the repetition of a process, stopping when the results of the iteration have coalesced. According to the journal (H. Priyatman: 2019) The steps on the k-means clustering algorithm in general: Specifies the number of k as the cluster to be formed. The determination of the number of k clusters is usually done with several factors of consideration both theoretical and conceptual which are then proposed to determine how many clusters. Generate the initial centroid k (cluster center point) randomly. To determine the initial centroid is done in random form from several available objects as many as the sum of k clusters then to calculate the next i-th cluster centroid, using the following formula: Information: v = Centroid on cluster xi = i-th object n = The number of objects that are members of the cluster. Calculating the distance from each object to each centroid of each data cluster and calculating the distance between objects with centroids can be done using the mathematical formula euclidian distance: Information: xi = i-th x object y = i-th y power n = number of objects Count each object into the closest centroid.

Perform the i-th iteration, then determine the position of the new centroid using the previous equation Repeat the steps in point c if the new centroid positions are not equal and the iteration or iteration will stop if the ratio is not greater than the value of the previous ratio until the calculation results in each convergent data. Clustering Definition According to Larose in (Zulfa Nabila: 2021), Clustering refers to groupings such as records, observations, or paying attention and forming classes of objects that have similarities. A cluster is a collection of records that have similarities to each other, and are different from records in other clusters. Clustering attempts to divide the entire data set into relatively similar groups, where the similarity of records in one group will be maximum, while similarity to records in another group will be minimal.

Definition of Rapid Miner

According to (Senna Hendrian, 2018) concluded that: Rapid Miner is a machine learning environment Data Mining, text mining and predictive analytics. Rapid Miner is a software created by Dr. Markus Hofman from the Blanchardstown Institute of Technology and Raif Klinkenberg from rapid-i.com with a GUI (Graphical User Interface) display making it easier for users to use this software. This software is open source and is made using the Java language under the GNU Public License and Rapid Miner can be run on any operating system. By using Rapid Miner, no special coding skills are needed, because all facilities are provided. Rapid Miner is devoted to the use of Data Mining. RapidMiner was previously called YALE (Yet Another Learning Environment), whose initial version was developed in 2001 by RalfKlinkenberg, Ingo Mierswa, and SimonFischer at the Artificial Intelligence Unit of the University of Dortmund. RapidMiner is distributed under AGPL (GNU Affero General Public License) version 3. Until now, thousands of applications have been developed using RapidMiner in more than 40 countries. RapidMiner as an open source software for data mining is no doubt because this software is already leading in the world. 2.4.1 Properties of Rapid Miner Rapid Miner has the following properties: Written with Java programming language so that it can be run on various operating systems. The knowledge discovery process is modeled as operator trees Internal XML representation to ensure a standard format of data exchange. The scripting language allows for large-scale experimentation and automation of experimentation. Multi-layer concept to guarantee efficient data display and guarantee data handling. It has a GUI, command line mode, and Java API that can be called from other programs. Features of Rapid Miner Some of the features of Rapid Miner, among others: There are many data mining algorithms, such as decision trees and self-organization maps. Sophisticated graphical forms, such as overlapping histogram diagrams, tree charts and 3D scatter plots. Many variations of plugins, such as text plugins to perform text analysis. Provides data mining and machine learning procedures including: ETL (extraction, transformation, loading), data preprocessing, visualization, modeling and evaluation. The data mining process is composed of nestable operators, described with XML, and created with a GUI Integrating Weka data mining project and R statistics

Definition of Income

According to Suroto in (Christoper: 2017) The theory of income or income is all receipts both in the form of money and in the form of goods originating from other parties as well as industrial results which are valued on the basis of a sum of money from the prevailing assets at that time. Income is a source of one's income to meet one's daily needs and is very important for one's survival and livelihood directly or indirectly. Income consists of wages, salaries, rent, dividends, profits and is a flow measured in a certain period of time for example: a week, a month, a year or a long period of time. This income stream arises as a result of productive services that flow in the opposite direction to the income stream, namely productive services that flow from society to businesses, which means that income must be obtained from productive activities. As for according to Anggraini in (Christoper: 2017). Family income is the income of a husband and wife and other family members from their basic and additional activities. Income as a measure of prosperity that has been achieved by a person or family in some way is a factor that is dominant enough to influence a person's or family's decision on something. Family income plays an important role, because in essence family welfare is very dependent on the size of family income. There are three categories of income, namely: Income in the form of money is income in the form of money that is regular and received usually as a reward for services or achievements. Income in the form of goods is all income of its nature regular and ordinary, but always in the form of remuneration and received in the form of goods and services. Income that is not income is any revenue that is redistributive transfer (e.g. such as local assistance for poor students or job training programs for the poor) and usually make changes in household finances

3. RESEARCH METHODS

3.1 Overview of STMIK Dharmapala Riau

STMIK Dharmapala Riau is one of the Computer Universities in Pekanbaru City under the auspices of the Dipankara Education Foundation which is located at Jl. KH Samanhudi No.13, Sago, Kec. Senapelan, Pekanbaru City, Riau. STMIK Dharmapala Riau was established in 2007 with the Permit of the Minister of National Education No: 73 / D / O / 2007 and BAN-PT Accredited Study Program. Currently, STMIK Dharmapala Riau provides two study programs, namely, DIII Computerized Accounting and S1 Information Systems. Vision of STMIK Dharmapala Riau Making Higher Education a leading center of education, study and development of science, science, technology and art. STMIK Dharmapala Riau Mission Preparing graduates who have superior competencies in accordance with the fields of science, skills learned and mastered. Preparing graduates who have life skills, professionals, the spirit of independence, and the development of science, science, technology and art Preparing superior quality graduates who are able to compete in the business world, industry, and job market.

3.2 Data Types and Sources

1 Data Types The type of data used in this study was quantitative. Quantitative data relates to an amount that can be measured using research data in the form of numbers and analysis using statistics.

2 Data Sources There are two sources of data in this study, namely primary data and secondary data. Primary data is data obtained from questionnaires that researchers distribute to respondents, while secondary data is data derived from literature books, research journals, scientific paper articles, and other supporting sources.

3.3 Data Collection Techniques

1 Observations The observation method is to make direct observations to the location to find out the actual conditions in the field.

2 Interviews The interview method obtains data by asking directly to the source or related agencies.

3.4 Data Analysis Techniques

Before carrying out the data grouping process, we first determine the data to be processed so that the research objectives can be achieved. At the data collection stage, the data used is to provide or distribute questionnaires directly to students of STMIK Dharmapala Riau. The data is sample data to determine the income level of parents of STMIK Dharmapala Riau students with sample data of 100 students. From the results of collecting questionnaire data to students of STMIK Dharmapala Riau, it will be processed in order to obtain new information which is used as an attribute to process data to determine the income level of parents of STMIK Dharmapala Riau students.

4. RESULTS AND DISCUSSION

4.1 Data Analysis

In collecting data, it is carried out by distributing questionnaires directly to students of STMIK Dharmapala Riau via WhatsApp messages. The data I use is a questionnaire that has been filled out by students of STMIK Dharmapala Riau to be processed to get new information. From this data, it is used as an attribute to carry out data processing to group income data of parents of STMIK Dharmapala Riau students. Clustering Analysis with K-Means Algorithm K-means is included in the data mining partitioning clustering method, where each data must enter a certain cluster and allow for each data included in a particular cluster at a stage of the process, in the next stage to move to another cluster. K-Means separates data into K regions well known for its later and ability to classify big data and outliers very quickly. The following is a flowchart diagram of the K-Means algorithm and illustrates the steps in the K-Means algorithm assuming that the input parameters are the number of data sets as many as n data sets and the number of centroid initializations K=2 according to the study.

4.2 Clustering Analysis with K-Means

Algorithm K-means is included in the data mining partitioning clustering method, where each data must enter a certain cluster and allow for each data included in a particular cluster at a stage of the process, in the next stage to move to another cluster. K-Means separates data into K regions well known for its later and ability to classify big data and outliers very quickly. The following is a flowchart diagram of the K-Means algorithm and illustrates the steps in the K-Means algorithm assuming that the input parameters are the number of data sets as many as n data sets and the number of centroid initializations K=2 according to the study. Determination of the number of clusters In this initial step, the number of clusters is first determined based on the data we get. Initial Cluster Center Determination In determining the initial cluster center n, random number generation is carried out that represents the sequence of input data. The initial center of the cluster is obtained from the data itself rather than by specifying a new point, that is, by randomizing the initial center of the data. Calculation of the Distance of the Object to the Center of the Cluster To measure the distance between the data and the center of the cluster, Euclidean Distance is used, which is an algorithm for calculating the distance of data to the center of the cluster:

Calculation of the Distance of the Object to the Center of the Cluster To measure the distance between the data and the center of the cluster, Euclidean Distance is used, which is an algorithm for calculating the distance of data to the center of the cluster: Retrieve the data value and the cluster center value Calculate Euclidean Distance data with each cluster center. Grouping of Data Objects The distance of the calculation results will be compared and the closest distance between the data and the cluster center will be selected, this distance shows that the data is in one group with the nearest cluster center. The data grouping algorithm is as follows: Take the distance value of each cluster center with data. Find the smallest distance value. Group data with cluster centers that have the smallest distance. Determination of a New Cluster Center To get a new cluster center, it can be calculated from the average value of cluster members and cluster centers. The new cluster center is used to perform the next iteration, if the results obtained have not converged, the iteration process will stop if it has met the maximum iteration entered by the user or the results achieved have converged (the new cluster center is the same as the old cluster center). Cluster Center Distance Calculation Calculate the Euclidean Distance of all data to the new center point (C1 and C2) as done in step 2. After we get the calculation results, then compare the results. Data Collection Before carrying out the data grouping process, first determine the data to be carried out for processing so that the research objectives can be achieved. At the data collection stage, the data used is to provide questionnaires directly to students of STMIK Dharmapala Riau. The data is a sample data grouping the income of parents of students with 90 data samples from 90 data. The data used in this study is data type data with sample data of 90 data from 90 data. The data used in this study is sample data grouping income data of parents of STMIK Dharmapala Riau students. The following are some sample data grouping income data of parents of STMIK Dharmapala Riau students taken.

CONCLUSION

Based on manually grouping data and using the rapidminer application, the results obtained are the same, in cluster 0 which is 67 data and in cluster 1 which is 23 data. Then based on the rapidminer test results of the Centroid Table section -> attribute IP cluster 0 = 1,612 and cluster 1 = 1,826, the front number initials 1 = GPA 3.25-3.50. Father's job cluster 0 = 2,104 and cluster 1 = 5,478, front number initials 2 = Labour, front number initials 5 = Not working. Occupation Mother cluster 0 = 1,254 and cluster 1 = 2,087, front number initials 1 = Taking care of the household, front number initials 2 = Farmer/Plantation. The combined income of father and mother cluster 0 = 2,463 and cluster 1 = 2,043, the front number of the initials 2 = Rp. 2,000,000 - Rp. 3,000,000. The number of dependents in the family of cluster 0 = 1,164 and cluster 1 = 1,304, the front number of the initials 1 = ≤ 5 people. The category of capable students is in cluster 0 and the category of underprivileged students is in cluster 1, this can be seen from the results of data processing in the rapidminer section of the father's work cluster 1 = Not working, as we know that the father is the backbone in the family.

REFERENCES

Christoper, Rio, and Rosmiyati Chodijah. (2017). "Factors affecting the income of women workers as housewives." Journal of Development Economics.

Muharmi Yulya, Sari. (2020). "Analysis of the Effect of Service Quality on Grab Consumer Satisfaction with the K-Means Algorithm Method Case Study of Kec.Tampan and Kec.Tenayan Raya" Vol. 7. No. 2.

Anita, Sari, (2018). "Determination of the Level of Interest in Online Learning Through Social Media Using the K-Means Clustering Method" Vol. 1. No. 2.

H. Priyatman, F. Sajid and D. Haldivany. (2019). "Clustering Using K-Means Clustering Algorithm to Predict Student Graduation Time," JEPIN (Journal of Informatics Education and Research), vol. 5, no.1.

Hendrian, Senna. (2018). "Data Mining Classification Algorithm to Predict Students in Obtaining Education Funding Assistance".

Muliono, Rizki, and Zulfikar Sembiring. (2019). "Data Mining Clustering Using K-Means Algorithm for Clustering of Lecturer Teaching Tridarma Level"

Nabila Zulfa, Auliya Rahman Isnain, and Zaenal Abidin. (2019) : 8. "Data Mining Analysis for Clustering Covid-19 Cases in Lampung Province with K-Means Algorithm." Journal of Information Technology and Systems 2021, no. 2 (n.d.) : 9. no. 2.

P. Sari, B. Pramono and L. O. H. S. Sagala. (2017) "Improve Kmeans on Nutritional Value Status in Toddlers," SemanTIK, vol. 3, no. 1, pp. 143-148.

Siregar, Amril Mutoi, et al. (2017). Data Mining: Processing Data Into Information with RapidMiner. CV Kekata Group.

Susilo, Adityo, Cleopas Martin Rumende, Ceva Wicaksono Pitoyo, Widayat Djoko Santoso, Mira Yulianti, Herikurniawan, Robert Sinto, et al. (2020). "Coronavirus Disease 2019: A Review of Current Literature." Indonesian Journal of Internal Medicine 7, no. 1, 45. https://doi.org/10.7454/jpdi.v7i1.415

Suyanto, (2017). Data Mining for Data Classification and Clustering. Bandung : Informatics.

World Health Organization / WHO. (2020). Downloaded on June 25, 2021 via the website: https://covid19.who.int