



The Implementation of Data Mining Method Using K-Means Algorithm to Analyze Study Interest of High School Students

Dadang Sudrajat^{1✉}, Arif Rinaldi Dikananda¹, Abrar Hiswara², Rinovian Rais³, Amat Suroso⁴

¹STMIK IKMI Cirebon

²Universitas Bhayangkara Jakarta Raya

³Unindra PGRI Jakarta

⁴Universitas Bani Saleh

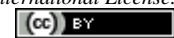
dias_sudrajat@yahoo.com

Abstract

At present, the school is experiencing difficulties processing the results of student academic achievement for the specialization process for high school students. The currently running student interest process still uses a manual system by calculating the subject value of each student and then grouping the results of the calculation of each student's value into science or social studies interest groups in accordance with the requirements imposed by the school. For that, we need a solution that can overcome these difficulties. The author develops the application using the Rapid Application Development (RAD) method, which consists of the requirements planning phase, the design phase, the construction phase, and the implementation phase. At the construction stage, the K-Means algorithm is implemented in data mining technology to classify student academic achievement results into science and social studies interest groups. The results of making this application are intended for the school, especially the homeroom teacher, so that it can be an alternative solution or advice in making decisions for student specialization.

Keywords: Students, Data Mining, Cluster, K-Means, Algorithm.

JSISFOTEK is licensed under a Creative Commons 4.0 International License.



1. Introduction

The development of information technology has resulted in an increasing need for information. To get information, data is needed, and the right way to process existing data [10]. The amount of data processed can cause difficulties in grouping it [7]. But with the development of information technology (IT), there are various solutions to overcome these difficulties, one of which is to use data mining technology. Data mining technology is a tool for extracting data from large databases with specifications for the level of complexity that has been widely used in many application domains such as banking and telecommunications [4]. Data mining is divided into several groups based on the tasks that can be performed, one of which is clustering [5]. The K-Means algorithm is an iterative clustering technique [6]. This algorithm partitions the data into clusters or groups so that those that have the same characteristics (high intra-class similarity) are grouped into the same cluster and those that have different characteristics (low inter-class similarity) are grouped into another group [3].

Extracting information from large amounts of data (having a large number of records and a large number of fields) cannot be done easily, so the application of data mining technology is needed to assist schools in the process of specializing in majors for high school students. In accordance with the Decree (SK) of the Director General Mandikdasmen Ministry of National Education Number 12/C/KEP/TU/2008 in the Technical Guidelines for Writing Student Learning Outcomes Reports, it is known that determining majors is very necessary for high school students. Specialization is carried out when students are in class X (ten) and will go up to class XI (eleven). After the homeroom teacher receives all semester grades, the homeroom teacher will decide whether the student goes up or not [2]. If the student is declared to be rising, then the majoring process will be carried out. To find out whether the decision-making is accurate or not, an assessment must be carried out with predetermined criteria apart from semester scores [8]. Other criteria needed in the majors system are the interests and results of the student's psychological test. So that from the results of this assessment, the school can make a decision as an evaluation material to determine the right major for its students [9]. The process of determining majors for high school students will occur repeatedly every year. This large amount of data opens opportunities to produce useful information for the school, especially to find out the process of student interest.

From the results of interviews conducted with high school teachers, it is known that currently high school has difficulties in the specialization process, which still uses a manual system in processing student academic data, totaling 253 people, so it takes a long time to find out the results of student specialization. This is an interesting thing to solve so that information can be obtained about which students are enrolled in the science specialization and which students are enrolled in the social studies specialization, as well as the number of students in the two specialization programs. The number of students and existing student scores will be processed by applying the K-Means algorithm in data mining technology so that it is expected to know the number of students who are entering the Science or Social Sciences majors accurately.

2. Research Methods

In this study, the authors used the RAD process model. data collection methods by means of literature studies, interviews, observations, and similar literature studies. The system development method used by the author in this thesis research is the RAD (Rapid Application Development) model. The RAD design model has four phases: the requirements planning phase, the design phase, the construction phase, and the implementation phase. The author chose this model as a system development method for the following reasons: One of the phases in the sequential model is the maintenance phase. The application that will be made does not require a maintenance phase in its implementation. By using this method, a complete functional system will be achieved in a very short period of time if the requirements are well understood. The weakness of other formal methods is that they take a lot of time and cost money because the application to be made is a simple one. This application does not require many stages. This is not in accordance with the spiral model, which is an evolutionary model that has been used for a long time in software development.

3. Results and Discussion

The results of the interviews stated that the specialization process was still carried out manually, where the class teacher calculated the average student academic score, which was then adjusted according to the student's interests. Because it is less efficient, it is proposed to make an application to support specialization. In the specialization application, it is adjusted according to academic scores, student interests, cluster results, and predetermined conditions, where the value limits for each interest or major are not from the average student score. The requirements for each specialization are different; for the science specialization, the minimum score for each science subject is 78, and for the social studies specialization, the minimum score for each social studies subject is 78. Students are allowed to have 1 subject score below the minimum score, and if more than 1 subject value is below the minimum score, then the student does not pass the desired specialization and will enter the Social Sciences specialization. The school, especially the homeroom teacher, still uses the specialization system manually in calculating the academic value of each student, then grouping students into science or social studies interest groups if the results of the student's academic scores are in accordance with the criteria for student specialization applied by the school. To help with this problem, the authors apply one of the tasks of data mining technology, namely clustering for specialization processes for class X students. For the clustering process, the authors use the K-means algorithm. According to previous researcher, the K-Means algorithm is quite effectively applied in the process of grouping the characteristics of research objects.

The data used in this study were student data and academic grades of class X students in semester 2. The data were obtained from direct observation. Research data will be stored in a database used for research testing class X students in semester 2. The data were obtained from direct observation. Research data will be stored in a database used for research testing. The data obtained is raw data that still has data redundancies, errors, and blanks. These data consist of 8 test variable attributes such as scores in mathematics, physics, chemistry, biology, history, geography, economics, and sociology. Incomplete and inconsistent data often occur in the database. Incomplete data can be caused by various reasons, such as fields that are not filled with data that matches the attribute. Likewise, with the student data database and student academic grades, there are some fields that are not filled in with the appropriate attributes. To reduce this, the Knowledge Discovery in Database (KDD) stage is carried out so that the database complies with the conditions required by the application.

In the use-case login activity diagram, all actors must do is open the system application. All actors must first log in. Enter the username and password. If you enter the wrong username and password, the system will display an error message and return you to the login menu to enter the correct username and password. If you enter the username and password correctly, the system will display the system's main menu. In the input user activity diagram (admin page), the admin has logged in first. Admin selects the User menu. In the User menu, there is a submenu for "Input user and "View user data. In the Input user submenu, the admin can add new users and save them to the database by providing their username, password, and access rights. In the activity diagram, see user (admin page), where admin has logged in first. Admin selects the User menu, in the User menu there is a submenu Input user and view user data. In the View user submenu, the admin can change user data that has previously been input into the system. The admin selects one of the data fields to be changed, then fills in the user change form and

clicks "Save." The changed data will automatically be stored in the database. Apart from changing user data, the admin can also delete user data by selecting one of the items that they want to delete.

In the Student Activity Diagram (admin page), the system initially displays the main page, then the actor selects the Student menu. In the Student menu, the actor (admin) selects the View Student submenu and can view all student data that has previously been input into the system and saved successfully. In the View Student menu, the admin can print student data reports per class. Student data reports will be printed in PDF format. In the value activity diagram (admin page), the system initially displays the main page, then the actor chooses the Values menu, where in the Values menu, the actor (admin) selects the View Grades submenu and can see all student grades that have previously been input into the system and saved successfully. In the View Grades menu, administrators can print class grades and student grades. Student grade reports that will be printed are in PDF format. In the class activity diagram (admin page), the admin has logged in first. Admin selects the Class menu. In the Class menu, there is a Class Input submenu. In the Class Input submenu, the admin can add new classes as well as save new class data into the database. In the class activity diagram (admin page), the admin has logged in first. Admin selects the Class menu, in the Class menu there is a Class Input submenu and see Class. In the Class submenu, there are options to change classes and delete classes. In the View Class submenu, the admin can change the class, class, and school year codes that were previously inputted. The method is to select one of the classes that will be changed, then click "Change" in the actions column. After clicking "Change," the actor must fill out the class change form and click "Save." The system will then process the entered data. Administrators can also delete class data that was previously stored.

In the Specialization Activity Diagram (admin Page), the system initially displays the main page, then the actor chooses the Specialization menu, where in the Specialization menu, the actor (admin) selects the View Specialization Results submenu. In this submenu, the admin selects the Specialization Results tab, and the admin can see the total students who are in science or social studies specialization groups and print them out. And on the Specialization Concentration Results tab, the admin can see all the results of student specialization that have previously been processed by the system and successfully saved. In the View Special Concentration Results menu, the admin can print reports on student specialization results. Reports on the results of student specialization that will be printed in PDF format. In the student activity diagram (user page), the user has logged in first. The user selects the Student menu. In the Student menu, there is a Student Input submenu. In the Student Input submenu, users can add new student data as well as save student data into the database. In the student activity diagram (user page), the user has logged in first. The user selects the Student menu, in the Student menu there is a Student Input submenu and see Student. In the Student View submenu, there are options to change, delete, search, and print student data. In the Student View submenu, the user can change student data that has previously been input. How to select one of the student data that will be changed, then click change in the actions column. After clicking, the actor must fill out the form to change class data and click save. The system will then process the data entered. Admins can delete student data that has previously been stored. To find student data, the user simply selects which class the data will be searched for, then clicks search. Then the system will display the search results. Meanwhile, to print student data, the user only selects which class the student data will be printed on and then clicks print. Student data will be printed in pdf.

In the value activity diagram (user page), the user has logged in first. The user selects the Value menu. In the Value menu, there is a Value Input submenu selects the Value menu. In the Value menu, there is a Value Input submenu. In the See Grades submenu, there are options to change, delete, search for, and print student grade data. In the submenu "Value, the user can change the student's value data that has previously been inputted. How to select one of the student value data to be changed, then click change in the action's column. After clicking, the actor must fill out the form to change the student value data and click save. The system will process the data entered. The admin can delete student grade data that has previously been stored. To find student grade data, the user simply selects which class the data will be searched for, then clicks search. Then the system will display the search results. Meanwhile, to print student grade data, the user only selects which class the student grade data will print, then clicks print. Data on student scores to be printed is in pdf format. In the Specialization Activity Diagram (User Page), the system initially displays the main page, then the actor selects the Specialization menu, where in the Specialization menu, the actor (user) selects the Specialization Results submenu. In this submenu, the admin can select the Specialization Results tab in the form of a table containing the number of students who are in the science or social studies interest group and can print it in.pdf format. Whereas in the Specialization Concentration Results tab, the user can see all the results of student specializations that have previously been processed using the K-Means algorithm by the system. In the Results menu. Concentration of Specialization: the user can search for the desired data according to the class to be searched for, and the user can also print reports on the results of the specialization of class students. Reports on the results of student specialization that will be printed in pdf format. Inside the activity diagram Log Out. If the actor chooses to log out, then the system will exit automatically and return to the login page.

The sequence diagram explains in detail the sequence of processes carried out in the system to achieve the goal of the use case, depicted in the following sequence diagram: When all actors log in, the system will first ask for a username and password. When the username and password are read by the system, the system verifies the username and password that are filled in with those already in the database. If you have successfully logged in, all actors will enter the system's home screen (main menu). If the username and password are entered incorrectly, the system will display an error message. In the input user sequence diagram, the admin has logged in first. Admin enters the main page, then enters the user menu and selects the user input submenu. The system will display the user input form, and then the admin will fill in the username, password, and level fields in the user input form. After all fields are filled, click "Save." Then the user data that has just been input will be directly stored in the database. In the View User Sequence diagram, it is illustrated that the admin has logged in first. Admin enters the main page, then enters the user menu and selects the View User submenu. The system will display user data that has previously been stored. Then the admin selects one of the users whose data you want to change, and after that, click the change button. The system displays the user change form. The admin changes one or all of the fields on the user change form. After changing the desired data, click save. The data that has just been changed will be directly stored in the database. If you want to delete the data for a user, the admin can simply select the user whose data you want to delete, then click delete, and the system will immediately delete the data from the database.

In the View Student Sequence Diagram, the admin has logged in first. Admin enters the main page, then enters the student menu and selects the View Student submenu. The system will display student data that has previously been stored. Then the admin selects one of the classes whose data you want to print and clicks the print button. The system displays student data in PDF format according to the class you want to print the data for. In the View Value Sequence diagram, it is illustrated that the admin has logged in first. Admin enters the main page, then enters the Values menu and selects the View Grades submenu. The system will display the student score data that has previously been stored. Then the admin selects one of the classes whose data you want to print, then clicks the print button. The system displays student grade data in pdf form according to the class you want to print the data for. In the class input sequence diagram, the admin has logged in first. Admin enters the main page, and enters the class menu and selects the class input submenu. The system will display the class input form, and then the admin will fill in the fields in the class input form. After all fields are filled, click "Save." Then the class data that was just inputted will be directly stored in the database.

In the View Class Sequence Diagram, the admin has logged in first. Admin enters the main page, then enters the class menu and selects the View Class submenu. The system will display the previously stored class data. Then the admin selects one of the classes whose data you want to change and clicks the edit button. The system displays the class change form. The admin changes one or all of the fields on the class change form. After changing the desired data, click save. The data that has just been changed will be directly stored in the database. If you want to delete data for one of the classes, the admin can simply select one of the classes you want to delete data from, then click "Delete." The system will immediately delete the data from the database. In the sequence diagram for student specialization, the admin has logged in first. Admin enters the main page, then enters the Specialization menu. Click Specialization Results, and then the system will display the specialization results tab and the results of the overall student specialization concentration. The specialization results tab contains a table that contains the number of students who are included in the Science or Social Sciences specialization group, and the student specialization results tab contains a table that displays the results of each student's specialization. After that, the admin can select the print button on the Specialization Concentration Results tab, and all the data that has been stored will be printed in pdf format.

In the student input sequence diagram, the admin has logged in first. Admin enters the main page, then enters the student menu and selects the student input submenu. The system will display the student input form, and then the admin will fill in the fields in the student input form. After all fields are filled, click "Save." Then the class data that was just inputted will be directly stored in the database. In the View Student Sequence Diagram, the admin has logged in first. The user enters the main page, then enters the Student menu and selects the View Student submenu. The system will display student data that has previously been stored. Then the user selects one of the students whose data they want to change and clicks the edit button. The system displays the student change form. The user changes one or all of the fields on the student change form. After changing the desired data, click save. The data that has just been changed will be directly stored in the database. If you want to delete the data of a student, the user simply selects the class whose data you want to delete and clicks "Delete." The system will immediately delete the data from the database. If you want to find student data, the user can select a class for which data will be searched, then click the search button, and the system will display student data that has been stored according to the class you want to find data for. If the class to be searched for does not exist, the system will display an empty student data table. On the View Students submenu, users can also print class student data by selecting the class for which student data will be printed and then clicking print. The system will automatically display student data according to the class you want to print the data for. Student data is printed in pdf format.

In the input value sequence diagram, the admin has logged in first. Admin enters the main page, then enters the Value menu and selects the Value Input submenu. The system will display the Value Input form, and then the admin will fill in the fields in the Value Input form. After all fields are filled, click "Save." Then the class data that was just inputted will be directly stored in the database. In the View Value Sequence diagram, it is illustrated that the admin has logged in first. The user enters the main page, then enters the Student menu and selects the View Grades submenu. The system will display the student score data that has previously been stored. Then the user selects one of the values whose data they want to change and clicks the edit button. The system displays the value change form. The user changes one or all of the fields on the value change form. After changing the desired data, click save. The data that has just been changed will be directly stored in the database. If you want to delete the data of a student, the user simply selects the value of the student whose data you want to delete, then clicks "Delete." The system will immediately delete the data from the database. If you want to find student grade data, the user can select a class for which data will be searched, click the search button, and then the system will display student grade data that has been stored according to the class you want to find the data for. If the class to be searched for does not exist, the system will display an empty student data table. On the View Grades submenu, the user can also print student grade data by selecting the class for which the student grade data will be printed and then clicking print. The system will automatically display student grade data according to the class you want to print the data for. Data on student scores is printed in pdf format.

In the sequence diagram for student specialization, the user has logged in first. Admin enters the main page, then enters the Specialization menu. Click Specialization Results, and then the system will display the specialization results tab and the results of the concentration of students' specialization either in class or as a whole. Users can select the search button to find the data that they want to see the results of their specialization after clicking search, then click the Calculate K-Means button. After that, the system will display the results of the K-Means calculation along with the interests of each student. After that, to print the results of the specialization, the user can click print on the Specialization Concentration Results tab, and all stored data will be printed in.pdf format. On the Testing tab, the user can view the results of system testing for specializations using the K-Means algorithm and specializations manually. If the interest in the K-Means calculation results is the same as the interest in the manual calculation results, the value is true. If the results of K-Means or manual calculations are not the same, then the value is false. In addition, users can also print test results in pdf format. All actors select the logout menu, then the system will carry out the logout verification process, then display a successful logout notification or message, and return to the main menu.

The author constructs or makes the application based on the previous stages. The construction process is carried out with reference to the design and application flow that have been determined. Furthermore, the design results using the UML model are translated into code. The coding process is done with the help of some software, namely JDK version 7 as the Java platform, Netbeans IDE 7.0.1 as the editor, and Xampp as the database server. The following is an explanation of application construction using the Netbeans IDE 7.0.1 syntax code program related to specialization for high school students. Testing is the most important part of the software development cycle. Tests are carried out to check the functions in the specialization application for class X students so that they can run properly and according to their functions. Testing this application uses a type of black-box test, namely testing that focuses on the functional requirements of the software.

The validation test in this application aims to determine the performance of the application that has been made. In this application, of the 253 existing research data records, they are divided into two categories: 208 data records for training and 45 data records for testing. The testing data will be used as validation data to determine the accuracy and error rate of the K-Means algorithm. Based on the contents of the confusion matrix, it can be seen the amount of data from each class that is classified correctly and each class that is classified incorrectly. Then, based on the confusion matrix data, the level of accuracy and error rate of the K-Means algorithm in specialization for class X students on test data can be seen. Referring to the results of the calculation of the accuracy and error rate above, it can be concluded that the results of specialization using the K-Means algorithm for validation test data have an accuracy rate of 0.753, or 75.3%, and an error rate of 0.2465, or 24.7%. The error rate of 24.7% or 25% is caused because the range of values between science and social class is too close, causing the calculation results of the K-Means algorithm to be equivocal. Students can enter science and social studies because the range of values is too close.

4. Conclusion

From the results of extracting information on the 253 sample data points that were tested by implementing the K-Means algorithm in data mining technology, it was found that as many as 110 students entered the science specialization group and rest students entered the social studies specialization group. The K-Means algorithm can be implemented in data mining technology for specialization systems for high school students with an accuracy rate of 75.3%. The error rate is 24.7%. With an accuracy rate of 75%, the major specialization application is quite effective in assisting the school in determining the interests of class X students according to the academic value of

each student. To get a better level of accuracy in implementing the K-Means algorithm for majors in high school by modifying or updating the K-Means algorithm currently used or by combining the K-Means algorithm with other algorithms. In this study only using student achievement variables in the form of semester 2 subject scores, it is suggested that this research be further developed by adding (including) semester 1 subject scores and student interest variables as computational variables in the calculation process using an algorithm. If the value of C1 is the same as the value of C2 or the range of values between C1 and C2 is too close, the school can use the student interest questionnaire as a material consideration to determine the final decision on student specialization.

References

- [1] Departemen Pendidikan Nasional. (2008). *Technical Instructions for Student Learning Outcomes Reports (LHBPD)*. Jakarta: Dirjen Mandikdasmen.
- [2] Astuti, E. D., Tannady, H., Lahiya, A., Supriatna, D., & Handayani, E. S. (2023). The Analysis of Relationship Between Quality of Graduates and Education Financing Management in Private Islamic School. *Journal on Education*, 5(3), 7715-7720.
- [3] Ediyanto., Mara, M. N., & Satyahadewi, N. (2013). Characteristic Classification Using the K-Means Cluster Analysis Method. *Jurnal: Universitas Tanjungpura (UNTAN)*.
- [4] Nurprihatin, F., Tannady, H., Lusiani, M., Karo-Karo, G., & Renatha, R. (2017). Algoritma K-Means Clustering untuk Menentukan Lokasi Pusat Distribusi Ritel dengan Mempertimbangkan Jumlah Permintaan. *Penelitian dan Aplikasi Sistem dan Teknik Industri*, 11(1), 32-44.
- [5] Giyanto, H. (2008). *Application of the K-Means Clustering algorithm, K-Medoid, Gath Geva*. Tesis Tidak Terpublikasi. Yogyakarta: Universitas Gajah Mada.
- [6] Jananto, A. (2010). *Predicting Student Performance Using Data Mining Techniques (Case study of UNISBANK student academic data)*. Tesis Tidak Terpublikasi. Yogyakarta: Universitas Gajah Mada
- [7] Kusriani., & Emha, T. L. (2009). *Data Mining Algorithm*. Yogyakarta: CV. Andi Offset.
- [8] Nugroho, B. S., Anggreni, M. A., Afnanda, M., Arta, D. N. C., & Tannady, H. (2023). The Role of Academic Fraud as an Intervening Variable in Relationship of Determinant Factors Student Ethical Attitude. *Journal on Education*, 5(3), 9584-9593.
- [9] Parinussa, J. D., Taryana, T., Ningtyas, A. A., Rachman, R. S., & Tannady, H. (2023). Developing Student Emotional Intelligence by Involving the Active Role of Teacher. *Journal on Education*, 5(3), 8528-8533.
- [10] Pressman, R. S. (2002). *Software engineering*. Yogyakarta : Andi