EasyChair Preprint

№ 15845

Classification of Students' Interest Patterns in Sebiru-Biru Senior High School to Continue Their Education to Higher Education with CRISP-DM

Enda Surany Barus, Gerry Firmansyah, Budi Tjahjono and Agung Mulyono

February 19, 2025

CLASSIFICATION OF STUDENTS' INTEREST PATTERNS IN SEBIRU-BIRU SENIOR HIGH SCHOOL TO CONTINUE THEIR EDUCATION TO HIGHER EDUCATION WITH CRISP-DM

Enda Surany Barus[1], Gerry Firmansyah[2], Budi Tjahjono[3], Agung Mulyono[4]
[1,2,3,4]Universitas Esa Unggul, Jakarta, Indonesia
[1]endasurany@student.esaunggul.ac.id, [2]gerry@esaunggul.ac.id, [3]budi.tjahjono@esaunggul.ac.id, agung.mulyo@esaunggul.ac.id

Abstract

The quality of education greatly affects the quality of human resources in the future. In Indonesia, students' interest in continuing their education to college is often influenced by various factors, such as economic conditions, personal aspirations, family influences, and information obtained regarding career choices, and so on. The quality of education in our country compared to other countries is very worrying, as we all know and see. SMA Masehi Sebiru-biru is a high school that has successfully graduated an average of 100 students each year. With a percentage of only ±20% of students who graduate who continue their education to college, which means that there is still a lack of awareness of SMA Masehi Sebiru-biru students about the importance of continuing their education to college. Based on this, the author wants to find out what factors influence the interest of SMA Masehi Sebiru-biru students to continue their education to college using the Decision Tree and Naïve Bayes methods.

Keywords : Data Mining , Decision Trees , Naïve Bayes.

1. Introduction

In the era of globalization and rapid development of information technology, the quality of education greatly affects the quality of human resources in the future. The younger generation with good education has greater ability to face global challenges and increase the competitiveness of the future nation [36]. In Indonesia, students' interest in continuing their education to college is often influenced by various factors, such as economic conditions, personal aspirations, family influences, and information obtained about career choices, and so on. The quality of education in our country compared to other countries is very worrying, as

we all know and see. The current problem of education is the low quality at various levels of education, both formal and informal, due to lack of resources [17]. Thus, although the importance of higher education is increasingly recognized, there are still challenges in understanding the pattern of student interest in continuing their education to college level and what factors influence students in making this decision.

Sebiru-biru Christian High School is one of the high schools located Jalan deli tua - penen km 36.5 periaria village, Sibiru-biru district, Deli Serdang Regency, North Sumatra Province. This high school has 100 students in each class, the students come from various family backgrounds and different economic levels. This high school has also succeeded in graduating an average of 100 students each year. With a percentage of only 100 $\pm$20%students who graduate who continue their education to college, which means that there is still a lack of awareness of Sebiru -biru High School students about the importance of continuing their education to college. From the results of the author's brief discussion with several students at this school, it was found that some of them still do not understand the importance of continuing their education to college. Several students expressed that there were limited funds as an obstacle, even though there were many government scholarships that could be a solution. In addition, there were also those who stated that their parents did not allow them, who considered that high school education was sufficient to find a job. Based on this discussion, the author was interested in analyzing the interest patterns of high school students in continuing their education to college, as well as understanding what factors most influenced students' decisions to continue their education to college. To find out these factors, it is necessary to carry out data analysis and processing using data mining [29].

Data mining , as one of the data analysis techniques, offers an effective approach to extracting information from a dataset. Through this technique, researchers can identify patterns and factors that may not be visible with conventional analysis methods [25]. By using data mining , this study aims to analyze the pattern of student interest in continuing their education to college, as well as identify the factors that influence these students' decisions.

In this study, the author will use classification with the aim of predicting whether the student has the potential to continue their education or not based on the data and factors that have been analyzed. In addition, this study is expected to help in the preparation of educational programs that are more in accordance with the needs and interests of students with the help of experts later, so that students can make more appropriate and relevant decisions to continue their education to college.

Based on the background described above, the author will conduct research on "Analysis of Student Interest Patterns to Continue Education to Higher Education with Data Mining ", the author hopes that this research can provide a significant contribution to the development of education in Indonesia.

2. Literature Review

2.1 Literature Review

Research conducted by (Nas, 2021) [29], (Ayunda et al., 2024) [3], (Dina et al., 2024) [42], (Situmorang, 2024) [42], (Doahir & Annisa, 2022) [10], (Sari & Imam, 2023) [40], (Sadat et al., 2023) [45]

these researchers used the Decision Tree method in the classification process , and from the five studies an average accuracy value of 85.6% was obtained, which means that the Decision Tree method is good at predicting existing data sets , so that if used, the results obtained will be in accordance with reality. The next research is research conducted by (Kriestanto & Femmy, 2021) [18], (Handoko & Muhammad, 2021) [11], (Lizar et al ., 2023) [23], Ninosari & Jhoanne, 2022) [30] these studies use the Naive Bayes method in the classification process of the data set to be analyzed, and from this study an average accuracy value of 84.5% was obtained, which means that the Naive Bayes method is also a good method for predicting data.

According to the results obtained from several previous studies that have been conducted, it was found that the Decision Tree method and the Naïve Bayes method are both good methods in predicting the data set to be analyzed, where the average accuracy value for the Decision Tree method was 85.6% and the average accuracy value obtained for the Naïve Bayes method was 84.5%.

There are several previous studies that compare the Decision Tree method , Naïve Bayes , and other methods, namely research conducted by (Khoirunnisa et al., 2021) [17], (Anam et al., 2022) [2] in these studies the best accuracy value was obtained by the Decision Tree method , but in the study (Kunjumon et al., 2023) [19] which compared the Decision Tree, Naïve Bayes , and other methods the best accuracy value was obtained by the Naïve Bayes method. In the study (Budiman & Zatin, 2021) [6] the same accuracy value was obtained between the Decision Tree and Naïve Bayes methods .

Based on this, the researcher is interested in comparing the Decision Tree and Naïve Bayes methods to predict the analysis of the interest patterns of Sebiru-biru High School students to continue their education to college.

2.2 Theoretical Study

2.2.1 Data Mining Classification

Classification in data mining is a technique for predicting categories or class from data based on patterns that have been found in historical data that has been labeled (labeled data). This technique falls into the category of supervised learning , where the model is trained using data that already has clear labels or classes. Once the model is trained, it can be used to predict the class of unlabeled data (new data) [31].

2.2.2 CRISP-DM (Cross Industry Standard Process For Data Mining)

CRISP-DM (Cross Industry Standard Process for Data Mining) is a process model used in data mining to provide guidance in managing and running data mining projects [46]. CRISP-DM is one of the most popular methodologies in both industry and academia, due to its flexibility that can be applied to various industries and types of data [26]. This model helps in planning and managing the data mining project life cycle systematically. The benefits of using CRISP-DM are reducing costs and time, as well as minimizing the need for knowledge for data mining projects. In addition, accelerating training, knowledge transfer, documentation, and capturing best practices are also benefits of using CRISP-DM [8].

The stages that will be carried out are: CRISP-DM is Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (Basiri et al . , 2024). For more details on the stages in CRISP-DM, see figure 1 below.

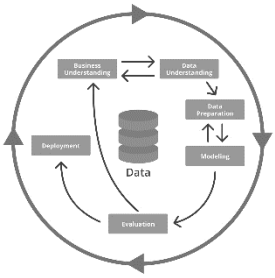

Figure 1 : Tahapan CRISP-DM (Cross Industry Standard Process for Data Mining)

2.2.3 Decision Tree Method

Decision Tree is one of the algorithms in data mining that is used to perform classification and regression . This method builds a model in the form of a decision tree that classifies data by dividing the dataset based on existing features, sequentially, until reaching the final decision on each leaf of the tree.

Decision Tree works by dividing the data at each node based on certain features, until it reaches a leaf node containing the final label or decision (Quinnlan, 1993). Each branch in the tree represents a decision rule derived from the data, and each leaf represents a class or predicted value . Simply put, a Decision Tree can be thought of as a series of questions (or decisions) that divide the data based on features to produce a particular decision or class.

There are 3 (three) types of nodes in the Decision Tree, namely, Root Node, is the top node, this node has no input and can have no output or more than one output. Internal Node, is a branching node, this node has only one input and has at least two outputs. Leaf Node or Terminal Node, is the final node, this node has only one input and no output. In the decision tree, nodes represent attribute testing (represented by boxes), branches represent test results (represented by labeled arrows), and leaves represent predicted classes (represented by circles) [12]. For more details, the Decision Tree diagram can be seen in Figure 2 below.

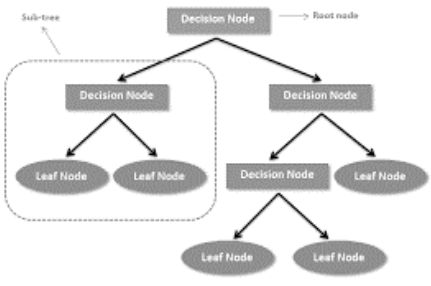

Figure 2 : Digram Decision Tree

Decision Tree is the most popular method used, because this method has many advantages, namely the Decision Tree method is very visual and easy to understand, even by non-technical people, so this method is easy to understand and interpret, besides this Decision Tree method is also considered easy because in the process it does not require normalization and complicated data scales. Decision Tree is also considered a flexible method because it can be used for classification and regression problems, and this method can also handle numeric and categorical data [20].

However, behind its advantages, this Decision Tree method also has several disadvantages, namely, Overfitting , which is when the tree is too deep, it can be very sensitive to training data and produce a very complex model, which may not generalize well to new data. In addition, decision trees can be very sensitive to small changes in data (Instability). Small changes in training data can produce very different tree structures, and this method also tends to prefer attributes with many different values, which can cause bias [20]. However, along with the development of technology and science, the Decision Tree method has been developed to cover the shortcomings of this decision tree method, such as the C4.5 algorithm, CART (Classification & Regression Tree), Random Forest, XGBoost, and LightGBM.

2.2.4 Naïve Bayes Method

Naïve Bayes method is one of the data mining algorithms that is often used to perform statistical classification [22]. Although simple, this algorithm has proven effective in various applications, especially in natural language processing and sentiment analysis. This method assumes that each data attribute is independent of each other, or it can be said that the assumption in Naïve Bayes that each word is independent of each other is inconsistent with real-world situations.

Naïve Bayes method is a good method for analyzing text, where Naïve Bayes is able to handle data that has many irrelevant words quite well. This is because each feature (word) is treated independently, so that irrelevant words have minimal influence on the final result [27]. In addition, the method is also effective for multicategory classification problems, which often occur in sentiment analysis when you want to separate more than two classes (eg, positive, negative, neutral). Another advantage of naive bayes is that it requires a small amount of training data to be able to determine the parameters in the classification process, the calculation process is simple and the results obtained are very good [32].

The classification carried out by Naïve Bayes is based on existing data or commonly called training data as a classification process for new data. There are stages in the classification process, namely the classification stage which calculates the probability value for each class label against the data provided. The class label with the highest probability value will be used as the input data class label [47]. The formula for classification with the Naïve Bayes algorithm can be seen in the following equation (2.3) [14]:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{2.3}$$

Information :

X= Data with unknown class

C= Hypothesis data X is a specific class

P(C|X)= Posterior Probability of the target class.

P(C)= Previous Class Probability.

P(X|C)= Probability of X based on a certain class.

P(X)= Probability of X

3. Methodology

3.1Research Stages

In this study, the author will use the stages of the Cross-Industry Standard Process for Data Mining (CRISP DM). The stages that will be carried out in the Cross-Industry Standard Process for Data Mining are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment [4] as discussed in section 2.2.2. The stages that will be carried out in the Cross-Industry Standard Process for Data Mining (CRISP DM) can be seen in Figure 3 below.

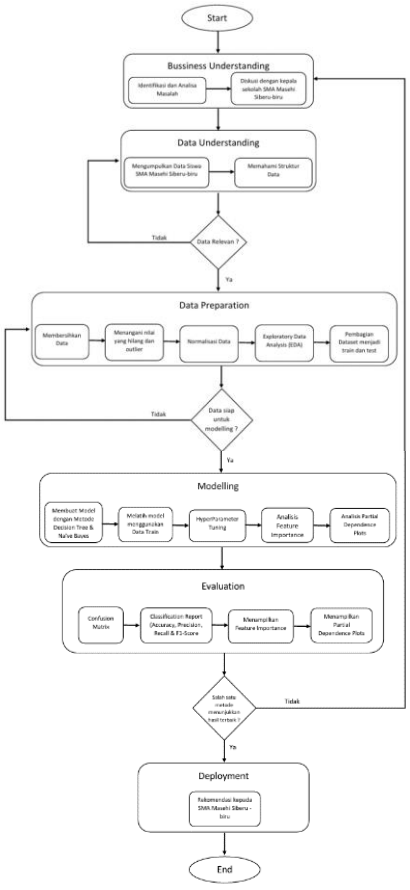

Figure 3 : Research Flow Diagram

4. Results

4.1 Business Understanding

Bussiness understanding is a stage to understand business problems and the objectives of the analysis to be carried out. After talking with several alumni of SMA Masehi Sebiru-biru, it was discovered that there were still many alumni of the high school who did not continue their education to college, for various reasons, namely limited funds, lack of support from parents, and one of which was the lack of awareness that continuing education to college was important to improve human resources. Based on this, the author held a discussion with the principal of SMA Masehi Sebiru-biru to find out real data on students who continued or did not continue their education to college. It turned out that from the data provided by the high school, only $\pm 20\%$ their alumni continued their education to college. However, the most influential factors have not been known for sure. Therefore, in this study the author intends to find out what factors influence the decision of SMA Masehi Sebiru-biru students to continue their education to college using data mining. The purpose of using data mining here is because data mining is a very effective technique for analyzing students' interests in continuing their education, because data mining is suitable for exploring existing data automatically and finding patterns or insights that may not be detected by manual analysis.

4.2 Data Understanding

Data understanding, is the part where we need to understand the data obtained and understand the data structure. In this study, the data used is the data of SMA Masehi Sebiru-biru students majoring in science and social studies. Some of the attributes that will be used in this study are parents' income, address, year of graduation, parents' jobs, and subject scores while studying at SMA Masehi Sebiru-biru.

The results show that parental income varies quite significantly between students, with most parents earning around 4 million. The year of graduation tends to be distributed around 2022-2024. The label 'Continue to College' shows that most students do not continue to college because class 0 is much larger than class 1, namely students who continue to college.

4.3 Data Preparation

Data Preparation is a technique for transforming raw data into a useful and efficient format. This is necessary because raw data is often incomplete and inconsistent in format, especially if it comes from different collections.

4.4 Modeling

After data preparation is done , then we can start with modeling . In this step, we choose an algorithm that suits the data. In this study, the classification models will be used, namely Decision Tree and Naïve Bayes.

4.5 Evaluation

Once the Modeling stage is complete, where a predictive or descriptive model is built based on the available data, the evaluation stage aims to assess how well the model that has been built solves the business problem that was defined at the beginning.

4.5.1 Evaluation Model Decision Tree For Class IPA

Decision tree model that has been created has an accuracy of 89% , which shows that the model is quite good at predicting overall results. Then, from the results above it can also be seen that class 0 on the label 'Continue College' which means students who do not continue college is predicted very well, this can be seen from the results of precision 96%, recall 92%, f1-score 94%. While for class 1 on the label 'Continue College' which means students are interested in continuing college has a lower performance, this can be seen from the results of precision 60%, recall 75%, f1-score 67%, which shows that the model is more difficult to predict students who will continue college.

4.5.2 Evaluation Of Naïve Bayes Model For Class IPA

Naïve bayes model that has been created has an accuracy of 90% , which shows that the model is very good at predicting overall results. In addition, from the results above it can also be seen that the model can predict class 0 on the label 'Continue College' which means students who do not continue college perfectly, this can be seen from the results of precision 1.00 , recall 0.96 , f1-score 0.98 . However, for class 1 on the label 'Continue College' which means students who continue college has a low precision of 0.50 , but a recall of 1.00 , which means that the model can predict all students who actually continue college, but only 50% of the predictions that say they continue college are correct. For the macro avg and weighted avg values provide an overview of the overall performance of the model.

4.5.3 Features Importance Of Class IPA

That there are 5 (five) attributes that contribute to the model prediction results, namely parents' income/month, Indonesian history value, English value and compulsory mathematics value, and student address. For parents' income/month with an importance value of 0.425747 has the highest importance value , meaning that this feature provides the greatest contribution in the model to predict the target. For the values of the courses that provide contributions are Indonesian history value, compulsory mathematics value, and English value.

4.5.6 Evaluation Model Decision Tree For Class IPS

The model that has been created has an accuracy of 98% , which means that most of the predictions are correct. For the prediction of class 0 on the label 'Continue to College' which means students who do not continue to college are predicted very well, both in precision 0.97 , recall 1.00 , and f1-score 0.98 . However, the model cannot predict at all for class 1 on the label 'Continue to College' which means students continue

to college, because the precision and recall for this class are 0. This is likely due to the imbalance between class 0 and 1 data which can be seen from the macro avg value showing low performance due to the very extreme class imbalance, with one class that is very dominant.

4.5.7 Evaluation Model Naïve Bayes Class IPS

The naive bayes model that has been created based on the social studies class data has an accuracy of 96% , which shows that the model as a whole is quite good at predicting the results. For the prediction of class 0 on the label 'Continue to College' which means students who do not continue to college are predicted very well, this can be seen from the results of precision 0.93 , recall 1.00 , f1-score 0.96 . However, the model is less good at predicting class 1 on the label 'Continue to College' which means students continue to college, because in recall only 33% are correct even though the model is able to predict 100% in class 1 with a precision of 1.00 .

4.5.8 Features Importance Of Class IPS

The most influential attributes are parents' income/month and mandatory math scores. For parents' income/month, it gives an importance value of 0.84 and is the attribute that gives the largest contribution compared to the others. The mandatory math attribute also contributes to the model although it is smaller compared to the parents' income/month attribute.

4.6 Consultation with Experts

From the results above, it is known that the biggest factor that most influences students' interest in continuing their education to college is their parents' income/month, which means that parents' low income can be an obstacle for students to continue their studies. This can be caused by several factors such as the lack of confidence of students or parents to be able to continue their studies because they think the costs are relatively expensive.

Here are some solutions that can be done to overcome or reduce these problems:

1. Introducing government scholarships that can help students attend school for free, such as the Bidik Misi scholarship, PPA, and scholarships that provide educational funding assistance such as Tanoto, KSE, Djarum, and so on.
2. Providing support through community groups (alumni who are continuing their studies) that can help reduce financial barriers and motivate students to continue their studies.
3. Establishing cooperation with universities or educational institutions to provide scholarships to students who have the potential to continue their studies but are hampered by funds, universities and educational institutions can create scholarship schemes. tiered based on parental income level, providing fair opportunities to all students.
4. Helping students find information about more affordable education such as online courses or distance learning that allows students to continue their education without having to pay high tuition fees and

avoid the cost of living in a big city. This provides more flexibility, especially for those with financial constraints.

5. Raise parents' awareness of the importance of higher education and provide parent education programs or seminars that can increase their understanding of the benefits of higher education, even though they may feel economically disadvantaged. If parents are more aware of the benefits of college, they will be more supportive of their children continuing their education.

Apart from the results obtained, there are other factors, namely the scores for mathematics and english subjects. As a solution to this problem, the school can do several things as follows:

1. Schools can create extra lessons or classes to improve students' grades in mathematics and English.
2. Helping students find information about online tutoring or classes for mathematics and English subjects for free or at a relatively low cost.

5. Conclusion and Suggestions

5.1 Conclusion

a. The decision tree and naïve Bayes methods on science and social studies class data are that both decision trees and naïve Bayes have equally good accuracy results, depending on the data set used.
b. The interest of SMA Masehi Sebiru-biru students to continue their studies is very low, only 13.6% for science classes and 7.6% for social studies students.
c. The most influential factor on students' decisions to continue their education to college, both science and social studies students, is their parents' monthly income, although there are other factors such as students' math and English scores, but their influence is very small. With a percentage for science classes of 42.5% of parents' monthly income, 13% of math scores, and 7% of English scores. While for social studies classes, the percentage is 84% for parents' monthly income, and 15% for math scores.
d. As a solution for SMA Masehi Sebiru-biru to increase the interest of its students to continue their education to college, it can be done by conducting socialization to parents about the importance of continuing their education to college, helping students to find information about the availability of government scholarships, establishing cooperation with universities and educational institutions that can provide educational funding assistance for students who want to continue their education, establishing good relations with alumni groups who continue their education, and helping students find information about more affordable education such as online courses. or distance education, for the problem of subject grades, this can be done by the school creating additional classes for mathematics and English subjects for free and helping students find information about online classes that they can take online or at a relatively low cost.

5.2 Suggestion

Here are some things that can be input for further research :

1. Further research can develop this study by collecting more data and more attributes for more complex results.
2. Further research can be conducted using other data mining methods so that the results can be compared with this research.

6. References

[1] Abdelsamea, MM, Zidan, U., Senousy, Z., Gaber, MM, Rakha, E., & Ilya, M. (2001). A survey on artificial intelligence in histopathology image analysis . Wires Data Mining and Knowledge Discovery.

[2] Anam, K., Bani, N., Christina, J. (2022). Comparison of Data Mining Classification Algorithms Using Optimize Selection for Study Program Interests. Building of Informatics, Technology and Science (BITS), 4(2), 606-613.

[3] Ayunda, YS, Hartama, D., Lubis, MR, & Gunawan, I. (2024). Analysis of interest patterns of high school/vocational school graduates to continue their studies using the C4.5 algorithm. Technology and Informatics Insight Journal , 4 (9), 581-595.

[4] Basiri, M.A., Mohammad, P., & Zahra, S. (2024). Implementing CRISP-DM and Logistic Regression for Predictive Implementing CRISP-DM and Logistic Regression for Predictive Analysis in Financial Transactions: A Case Study Analysis in Financial Transactions: A Case Study. IRAIS

[5] BM Monjurul, A. & Matthew, C. (2018). Educational Data Mining: A Case Study Perspectives from Primary to University Education in Australia. IJ Information Technology and Computer Science, 2, 1-9.

[6] Budiman., & Zatin, N. (2021). Comparison of Data Mining Classification for Searching Interests of Prospective New Students. Journal of Nuansa Informatika , 15(2), 37-52.

[7] Budiman., & Zatin, N., (2022). Comparative Analysis of Data Mining Classification Algorithm Performance for Searching Prospective Student Interests . SISTEMASI: Jurnal Sistem Informasi , 11(2), 271-290.

[8] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz T., & Shearer, C. (1999). DM SIG Workshop in Brussels.

[9] Dasarathy, B.V. (1991). Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques . Los Alamitos: IEEE Computer Society Press .

[10] Doahir, A., & Annisa, NQ (2022). Analysis of Student Potential for College with Classification Using the Decision Tree Method . POROS TEKNIK Journal , 14(1), 28-32.

[11] Handoko, W., & Muhammad, I. (2021). Prediction of Study Program Interest in Student Admissions at Stmik Royal Using Naïve Bayes . Journal of Science and Social Research , 4(2), 231-235.

[12] Hidayatsyah MR (2013). Application of Decision Tree Method in Granting Loans to Debtors with C4.5 Algorithm. (Bachelor's Thesis, STATE ISLAMIC UNIVERSITY OF SULTAN SYARIF KASIM RIAU).

[13] Hidayat, R. (2022). Utilization of data mining to see student interests after completing high school (SMA) education with the K-Means Clustering algorithm . Technology and Informatics Insight Journal , 1 (2).

[14] Jadhav, S. D., & Channe, H. P. (2016). Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. International Journal of Science and Research (IJSR) , https://doi.org/10.21275/v5i1.nov153131.

[15] K. Ahmed, and T. Jesmin, " Comparative Analysis of Data Mining Classification Algorithms in Type-2 Diabetes Prediction Data Using WEKA Approach ," IJSE , 7(2), 155–160.

[16] Kang, Z. (2019). Using Machine Learning Algorithms to Predict First-generation College Students' Six-year Graduation: A Case Study. IJ Information Technology and Computer Science , 9, 1-8.

[17] Khoirunnisa, Lia, S., Ira, TR, & Lilis, S. (2021). Prediction of Al-Hidayah Vocational School Students Entering College Using Classification Method. Informatics Journal , 8(1), 26-33.

[18] Kriestanto, D., & Femmy DA (2021). Application of Naïve Bayes for Analysis of Factors in Selecting STMIK AKAKOM as a Place of Study. JTKSI , 4(2), 75-81.

[19] Kunjumon, L.T., Sharon, S., Saffi, T. S., Tahsneem, & N., Neena, J. (2019). An Intelligent System to predict Students academic performance using Data Mining. International Journal of Information Systems and Computer Sciences, 8(2), 128-131.

[20] Kusrini & Emha T.L. (2009). Algoritma Data Mining. Yogyakarta : CV. Andi Offset.

[21] Lakshmi, B. N., Indumathi, T. S., & Ravi, N. (2016). A Study on C.5 Decision tree Classification Algorithm for Risk Predictions During Pregnancy. Procedia Technology, 24, 1542–1549. https://doi.org/10.1016/j.protcy.2016.05.128.

[22] Lavindi, EE, Wijanarto, W., & Rohmani, A. (2019). Hybrid Filtering and Naïve Bayes Application for Laptop Purchase Recommendation System. JOINS (Journal of Information https://doi.org/10.33633/joins.v4i1.2518.

[23] Lizar, Y., Alya, SF, Asriwan, G., & Joko, S., (2023). Data Mining Analysis to Predict Student Skills Using Naïve Bayes Method. Knowbase: International Journal of Knowledge in Databases . 3(2), 150-159.

[24] Malik, LA, Mia, K., & Firman, NH (2023). Factors Influencing the Interest of Prospective New Students to Register at Ftii Uhamka Using the K-Nearest Neighbor (K-Nn) Algorithm. Infotech: Journal Of Technology Information, 9(1), 85-94.

[25] Mardi, Y. (2020). "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," Jurnal Edik Informatika, 2(2), 213- 219.

[26] Mariscal, G., Oscar, M., & Covadonga, F. (2010). A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, 25(2), 137-166.

[27] McCallum, A., & Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on Learning for Text Categorization, 41-48.

[28] Mienye I. D., Sun Y., & Wang Z. (2021). Prediction performance of improved decision tree-based algorithms: A review, 35(1), 698–703, doi: 10.1016/j.promfg.2019.06.011.

[29] Nas, C. (2021). Data Mining Prediction of Prospective Students' Interest in Choosing a College Using the C4.5 Algorithm. Journal of Informatics Management (JAMIKA) , 11(2), 131-145.

[30] Ninosari, D., & Jhoanne, F. (2022). Decision Support System for Recommendation Results of College Majors Using Naive Bayes and AHP Methods. SATIN – Information Science and Technology , 8(1), 106-117.

[31] Novitasary, D., Yunianita R., & Suprianto. (2024). Classification of High School Students' Career Interests Using the C4.5 Algorithm. Scientific Journal of Informatics Engineering and Information Systems , 13(1), 711 – 724.

[32] Permana AP, Ainiyah K., Fahmi K., & Holle H. (2021). Comparative Analysis of Decision Tree , KNN, and Naive Bayes Algorithms for Predicting the Success of 49 Start-ups, 6(3), 178–188.

[33] Purboyo, RA (2022). Clustering Data on College Major Recommendations Using the K-Means Method Case Study of SMA Negeri 2 Palembang. Undergraduate Thesis. Bina Darma University.

[34] Putri, DA, Bayu, H., Sarika, A., & AB, P. (2019). Study Program Prediction Based on Student Grades Using Backpropagation Algorithm (Case Study of Sman 6 Depok Social Studies Department). Informatics Journal , 15(2), 69-78.

[35] Quinnlan, J.R. (1993). C4.5: Programs for Machine Learning . Morgan Kaufmann Publishers.

[36] Rahman, A. (2020). The influence of innovative learning methods on student learning outcomes . Journal of Education and Culture , 15(3), 123-136.

[37] Rohayani, H., & Muhammad, CU (2022). Prediction of Study Program Determination Based on Student Grades with Backpropagation Algorithm . Journal of Information System Research (JOSH), 3(4), 651-657.

[38] Sadat, A. M., Pujiono., Anggun, P., & Sholihul, I. (2023). Comparison Of Algorithm Between Classification & Regression Trees And Support Vector Machine In Determining Student Acceptance In State Universities. Jurnal Teknik Informatika (JUTIF), 4(6), 1589-1604.

[39] Safdara, M. F., Robert, M. N., & Piotr, P. (2024). Pre-Processing techniques and artificial intelligence algorithms for electrocardiogram (ECG) signals analysis: A comprehensivereview. Computers in Biology and Medicine, 17(2), 1-16.

[40] Sari, RN, & Imam, P. (2023). Data Mining Elective Course Interest for Final Year Informatics Students Applying C4.5 Algorithm. Bulletin Of Computer Science Research , 3(3), 263-269.

[41] Setiawan, I., & Heri, N. (2023). Data Mining Grouping of Values to Determine Subject Interests in Students of SMA Negeri 1 Kota Gajah. Journal of Informatics and Computer Technology MH. Thamrin , 9(1), 284-292.

[42] Situmorang, SA (2024). Selection of New Student Candidates Using the Decision Tree Method Using the C5.0 Algorithm (Case Study: SMK N. 1 Doloksanggul). (Bachelor's Thesis, Medan Area University).

[43] Sibiru-biru Christian High School. (2023, November 22). Organizational Structure. Official Website of Sibiru-biru Christian High School. https://www.smamasehibiru.sch.id/

[44] Nofitasary, D., Yunianita, R., & Suprianto S. (2024). Classification of High School Students' Career Interests Using the C4.5 Algorithm . Journal of Informatics Engineering and Information Systems, 13(1), 711-724.

[45] Sadat, MA, Pujiono., Anggun, P., Sholihul, I. (2023). Comparison Of Algorithm Between Classification & Regression Trees And Support Vector Machine In Determining Student Acceptance In State Universities. Journal of Informatics Engineering, 4(6).

[46] Schuster, C., Stebner, F., Leutner, D., & Wirth, J. (2020). Transfer of metacognitive skills in self-regulated learning: an experimental training study. Metacognition and Learning , 15(3), 455–477.

[47] Sartika, D., & Dana IS (2017). Comparison of Naive Bayes , Nearest Neighbor , and Decision Tree Classification Algorithms in a Case Study of Clothing Pattern Selection Decision Making. Journal of Informatics Engineering and Information Systems, 1(2), 151-161.