Comparison of C4.5 and Naïve Bayes Algorithms for Predicting Student Achievement

Fadli Dwi Yulianto
Department of InformationTechnology, Universitas Muhammadiyah Sumatera Utara, Indonesia

ABSTRACT

This research aims to analyze and predict student achievement using data mining techniques with the C4.5 and Naive Bayes methods. The data used includes various factors that affect students' academic performance, such as previous grades, attendance, and parents' income. The C4.5 method, which is a decision tree algorithm, is used to identify patterns in the data and make rule-based decisions. Meanwhile, Naive Bayes, which is a probabilistic classification technique, is used to calculate the probability of achievement based on the distribution of features. The C4.5 algorithm model showed excellent performance in classifying students into the categories of "Underachieving" and "Achieving," with perfect accuracy and F1-Score for both classes. On the other hand, the Naive Bayes model showed less than optimal results, especially in recognizing "Outstanding" students. Although the Naive Bayes model managed to correctly predict all the "Underachieving" students, it failed completely in detecting the "Achieving" students, as seen from the zero F1-Score for the class.

Keyword : Data Mining, Student Achievement Prediction, C4.5 Method, Naïve Bayes, Classification

Corresponding Author:
Fadli Dwi Yulianto,
Department of Information Technology,
Universitas Muhammadiyah Sumatera Utara,
Jalan Kapten Muktar Basri No 3 Medan 20238, Indonesia.
Email: fadlidwiyulianto@gmail.com

1. INTRODUCTION

Technology is a development in both hardware and software that is based on science and user needs that continue to develop over time. Technological advances have changed the way we work, which was previously done manually, such as in sending letters and making financial reports, can now be done with SMS (Short Message Service) or computer applications for financial reports (Karim et al., 2021). The advancement of information technology in recent years has developed very rapidly, changing the way people search for information, which is now not only from newspapers, audio visuals, and electronic media, but also via the internet. One area that is greatly affected by this technological advance is education, which is basically a process of information communication between educators and students to deliver learning materials (Kusumawati, 2023).

This study compares the C4.5 and Naïve Bayes algorithms in predicting student achievement. The results show that decision tree algorithms, such as C4.5, have better predictions. In addition, several previous studies have also used Naïve Bayes and C4.5 in exploring new knowledge. The C4.5 algorithm itself is a data classification technique that uses the decision tree method, where the topmost attribute becomes the root, while the bottom is called the leaf (Romli & Zy, 2020).

The Naïve Bayes algorithm is based on the simple assumption that attribute values are independent given a certain output value. In other words, the joint probability of several attributes is obtained from the product of the individual probabilities of those attributes. The advantage of Naïve Bayes is that this algorithm only requires relatively little training data to estimate the parameters needed in the classification process. Naïve Bayes often works better than expected in complex real-world situations (Kawani, 2019).

Previous research also shows that the Naïve Bayes algorithm can be used to predict student achievement and compared to Neural Network, where the results show higher accuracy of Neural Network. In addition, student achievement prediction has also been carried out using the Support Vector Machine algorithm and a hybrid decision support system (Rovidatul et al., 2023).

To address these issues, this study aims to compare the accuracy of two data mining algorithms, namely the C4.5 algorithm and Naïve Bayes, on various datasets. This comparison was conducted to determine which algorithm has a higher accuracy in predicting student achievement (Rahmayanti et al., 2022).

Education is a conscious and planned effort to create a learning atmosphere that allows students to develop their potential, both spiritual, personality, intelligence, and skills that are useful for themselves, society, nation, and state. Based on Law No. 20 of 2003 concerning the National Education System Article 3, the goal of national education is to develop the potential of students to be faithful, pious, have noble character, healthy, knowledgeable, capable, creative, independent, and become responsible citizens. Thus, the quality and management of learning in schools need to be improved, which can be seen from student learning achievements (Noviriandini & Nurajijah, 2019).

Student achievement can be seen from class grouping according to each individual's abilities. This is the reason this study was conducted at Asuhan Daya High School, considering various variables that affect student achievement. In this case, the decision-making method in this school is still traditional, which causes difficulties in determining student achievement. Therefore, changes are needed that can help schools improve the quality of education through a better understanding of student achievement. For this reason, a data mining-based calculation system is needed to group students based on their achievements (Br Sembiring et al., 2022).

Based on this background, the author is interested in conducting a study entitled "Comparison of C4.5 and Naïve Bayes Algorithms to Predict Student Achievement."

2. RESEARCH METHOD

A. Running System Analysis

Efforts to make early predictions of students who are likely to underachieve are aimed at allowing schools to take preventive or anticipatory measures to avoid failing or being expelled from school. By identifying students who are likely to experience difficulties in academic achievement or failing, schools can provide special assistance to these students. The goal is for all students, with various background factors, to be able to optimize their learning achievement. Several factors that can affect high school students' learning achievement include socio-economic conditions, which are often related to parental income, learning facilities provided by the school, attendance or absenteeism rates, and student participation in extracurricular activities. However, the current system is not yet able to predict student achievement specifically based on their discipline and social status.

B. System Requirements Analysis

In the current system, there are several aspects that need to be fulfilled, namely a system that is able to predict student achievement based on discipline and social status at SMA Asuhan Daya Medan. In addition, software and hardware are needed to support the performance of the developed system as well as adequate data or samples to maximize the work process of the system.

C. Research Framework

In conducting research, a systematic procedure is needed so that the research can run smoothly. This research procedure aims to compare the C4.5 and Naïve Bayes algorithms in predicting student academic achievement.

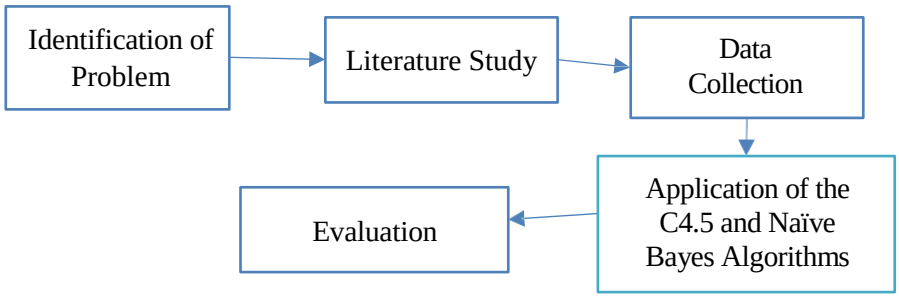

Fig 1. Research Procedure Diagram

The following is an explanation of each stage in the research procedure diagram.

1) Identification of Problem
   At this stage, identification of problems relevant to the field of study is carried out. The problem raised in this study is the absence of a system that is able to predict student achievement based on their attendance level.
2) Literature Study
   At this stage, a reference search is conducted to support the research topic, either in the form of books or journal articles. This literature search aims to find solutions that can help in solving research problems.
3) Data Collection
   a) Observation
      This method is used to collect data through direct observation, with the aim of obtaining relevant data. Research data is taken from student achievement records provided by the school.
   b) Interview
      Interviews are conducted as a systematic way to obtain the necessary information through questions asked to schools that have student achievement data. The goal is to obtain more complete and accurate information for the development of a new system according to research needs.
4) Application of the C4.5 and Naive Bayes Algorithms
   The C4.5 and Naïve Bayes algorithms are applied to predict student achievement, so that the final results obtained can be used as predictions of student achievement.
5) Evaluation
   The evaluation stage is carried out to measure the accuracy of the system developed, to ensure that the predictions produced are in accordance with the research objectives.

D. Flowchart C.45 and Naive Bayes

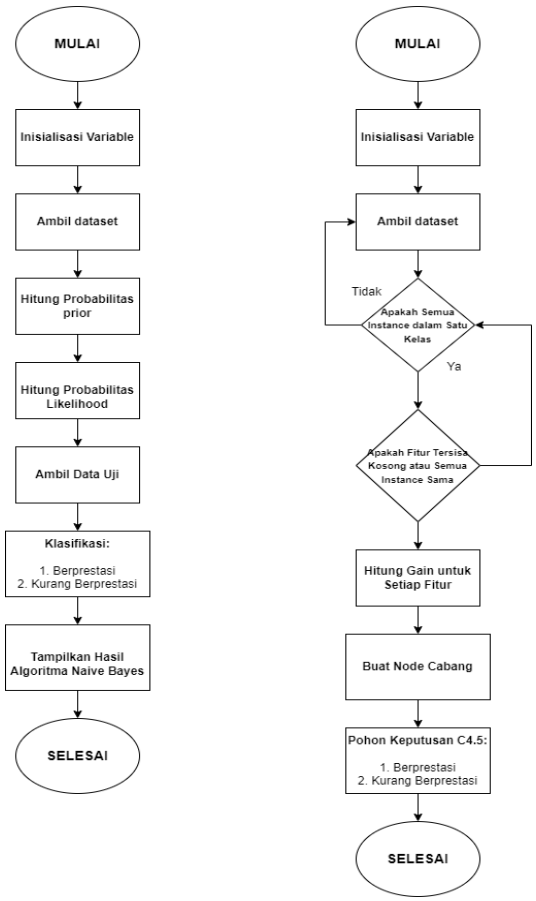

Fig 2. Research Flowchart

Explanation:

1) Naive Bayes Algorithm
   The left part of the diagram represents the Naive Bayes algorithm. This algorithm works based on the principle of probability and the assumption that the features in the data are independent of each other.
   - Variable Initialization: Preparing the variables that will be used in the calculation.
   - Fetch Dataset: Load the data that will be used to train the model.
   - Calculate Prior Probability: Calculate the initial probability of each class.
   - Calculate Likelihood Probability: Calculate the probability of a feature appearing in a class.
   - Get Test Data: Load new data that has never been seen by the model to predict its class.
   - Classify: Classify the test data based on the probability calculation that has been done.
   - Show Results: Display the classification results.
2) C4.5 Algorithm
   The right part of the diagram represents the C4.5 algorithm, which is one of the decision tree methods. This algorithm builds a decision tree by recursively dividing the data based on the most informative features.
   - Initialize Variables: Prepare the variables that will be used in the calculation.
   - Get Dataset: Load the data that will be used to train the model.
   - Calculate Gain for Each Feature: Calculate the gain information of each feature to determine the best feature to split the data.
   - Create Branch Nodes: Split the data into several branches based on the selected feature values.
   - Repeat the process until all data is classified or reaches other stopping criteria.

E. Research Location and Time

1) Research Location
   This research was conducted at SMA Asuhan Daya Medan, which is located in Medan City. The research location was chosen because there has been no previous study comparing the C4.5 and Naive Bayes algorithms to predict student achievement at the school, so this research is expected to provide new contributions.
2) Research Time
   This research process will take 6 months starting from February to July 2024.
3) Research Instrument
   Some of the devices used to complete this final assignment are as follows:
   a) Software
      In conducting research, researchers use the following software:
      - The operating system used is Microsoft Windows 10 Professional.
      - Google Colab application for processing data and writing program code.
   b) Hardware
      Some of the hardware that researchers need to conduct research is as follows:
      - The processor used is Intel Core i3 2.0 Ghz.
      - RAM with a size of 4GB.
      - Solid State Drive 128GB.

3. RESULTS AND DISCUSSION

A. Data Collection

The data used were taken from the first semester report card scores of 28 students, which were divided into two sheets in an Excel file: "SEM-1 VALUES" and "SMA IPA".

Table 1. Example of First Semester Value Dataset

No.	Column Name	Description
1.	Sequence Number	Student sequence number of SMA Asuhan Daya
2.	SIN	Student ID number at SMA Asuhan Daya
3.	NSIN	National student registration number at SMA Asuhan Daya

4.	Student Name	Name of students at SMA Asuhan Daya
5.	M/F	Gender of students at SMA Asuhan Daya
6.	Nilai Rapot	Report card scores of students at SMA Asuhan Daya consisting of scores for the subjects of PAI and Budi Pekerti, PKN, Indonesian, Mathematics, Indonesian History, English, Arts and Culture, PJOK, Crafts, Conversation, Qir'ah Qur'an, Mathematics (Elective), Biology, Physics, Chemistry, Average Score.

In Table 1 is an example of a dataset with the sheet name NILAI SEM-1 which consists of six column names.

Table 2. Example of SMA Science Dataset

No.	Column Name	Description
1.	Number	Sequence number in the dataset
2.	Name	Name of male or female student at SMA Asuhan Daya
3.	Place and Date of Birth	Place or date of birth of students at SMA Asuhan Daya
4.	SIN	Student ID number at SMA Asuhan Daya
5.	NSIN	National student registration number
6.	Parent	Name of parents of students at SMA Asuhan Daya
7.	Parents' Job	Profession or occupation of parents of students at SMA Asuhan Daya
8.	Parents' Income	The amount of income of parents of students at SMA Asuhan Daya
9.	Absence	Number of absences or attendance of SMA Asuhan Daya

Table 2 is an example of a dataset on the sheet name SMA IPA with a total of nine columns.

B. Implementation of C4.5 Algorithm and Naive Bayes

1) Data Preprocessing

Before applying the C4.5 and Naive Bayes algorithms, the data needs to go through a preprocessing stage to ensure the quality and readiness of the data in the analysis and modeling. At this stage, the data from the Excel file is loaded and checked for its integrity. The dataset to be preprocessed has several columns from the "SMA IPA" sheet, such as NIS, parents' income, and number of attendance, while the value column is taken from the "NILAI SEM-1" sheet, as seen in Table 3.

Table 3. Example of SMA Science Dataset

No.	Name	SIN	Parents Income	Number of Attendance	Grade
1	AMANDA DEFINA	977	Rp. 1,000,000 - Rp. 1,999,999	26	84,53
2	CINDY CELCEA	981	Rp. 1,000,000 - Rp. 1,999,999	26	83,73
3	DWI APRILIA SUNDARI	989	Rp. 1,000,000 - Rp. 1,999,999	26	84,80
4	EKA FITRIANA	990	Rp. 1,000,000 - Rp. 1,999,999	24	84,40
5	FADLY SURYA PRANATA	991	Rp. 500,000 - Rp. 999,999	26	82,40
6	FAJAR	992	Rp. 1,000,000 - Rp. 1,999,999	26	82,53

7	FANY RAVINA	993	Rp. 500,000 - Rp. 999,999	25	83,20
8	FARA HAMIDAH	994	Rp. 500,000 - Rp. 999,999	26	84,13
9	FEBRI ANSYAH	995	Rp. 2,000,000 - Rp. 4,999,999	25	83,27
10	FIKA SONTRIANI LUBIS	996	Rp. 1,000,000 - Rp. 1,999,999	26	84,87
11	FITRI HANDAYANI LUBIS	997	Rp. 500,000 - Rp. 999,999	26	84,87
12	IRDA OKTAVIA	999	Tidak Berpenghasilan	26	84,00
13	LATIFAH UMA	1000	Rp. 1,000,000 - Rp. 1,999,999	26	83,80
14	MAULINDA APRIANI	1003	Rp. 2,000,000 - Rp. 4,999,999	26	83,67
15	MUHAMMAD SYAHPUTRA	1005	Rp. 1,000,000 - Rp. 1,999,999	25	83,47
16	MONA APRILIA	1007	Rp. 1,000,000 - Rp. 1,999,999	25	84,27
17	M. ILHAM FAHRUDIN	1010	Kurang dari Rp. 500,000	26	83,47
18	MHD. SYAHPUTRA	1010	Rp. 1,000,000 - Rp. 1,999,999	26	82,93
19	RESVI AULIA	1020	Rp. 2,000,000 - Rp. 4,999,999	26	83,67
20	RIONALDO FEBRIANSYAH	1021	Rp. 500,000 - Rp. 999,999	25	81,67
21	VINA LESTARI	1024	Rp. 1,000,000 - Rp. 1,999,999	26	84,00
22	WINDA KHAIRANI NST	1026	Rp. 500,000 - Rp. 999,999	26	86,20
23	ARINI FEBIOLA	1079	Rp. 1,000,000 - Rp. 1,999,999	26	84,07
24	MHD. ILHAM SAPUTRA	1080	Rp. 1,000,000 - Rp. 1,999,999	26	83,80
25	AHMAD HINDRA BERUTU	1083	Rp. 1,000,000 - Rp. 1,999,999	25	81,07
26	SHEIRLLA CHANTIQA	1084	Rp. 500,000 – Rp. 999,999	25	85,60
27	NISA ULFITRI	1085	Rp. 500,000 – Rp. 999,999	26	85,07
28	STEFANI NERTIANA R. HUTAPEA	1134	Rp. 500,000 – Rp. 999,999	23	80,80

In Table 3 for the first parent income column has an income value of Rp. 1,000,000 - Rp. 1,999,999 with a total of fourteen student parents, the second income is Rp. 500,000 - Rp. 999,999 with

a total of nine student parents, the third income is Rp. 2,000,000 - Rp. 4,999,999 with a total of three student parents, the fourth income is No Income with a total of one student parent, the fifth income has a value of Less than Rp. 500,000 with a total of one student parent.

Table 4. Changing the Unique Value of the Parent's Income Column

Before Changing to Label	After Changing to Label
No income	0
Less than Rp. 500,000	1
Rp. 500,000 - Rp. 999,999	2
Rp. 1,000,000 - Rp. 1,999,999	3
Rp. 2,000,000 - Rp. 4,999,999	4

Pada Tabel 4.4 adalah proses mengubah *unique* value menjadi label pada kolom penghasilan orang tua.

Table 5. Labeling of Outstanding Students

Outstanding Students	Amount
False	25
True	3

In Table 5 is the labeling process for high achieving students. Labeling is done on the condition that if the average value of the student is more than 85 then the student is high achieving. The False value means that the student is underachieving with a total of 25 students, then the True value is a high achieving student with a total of three students.

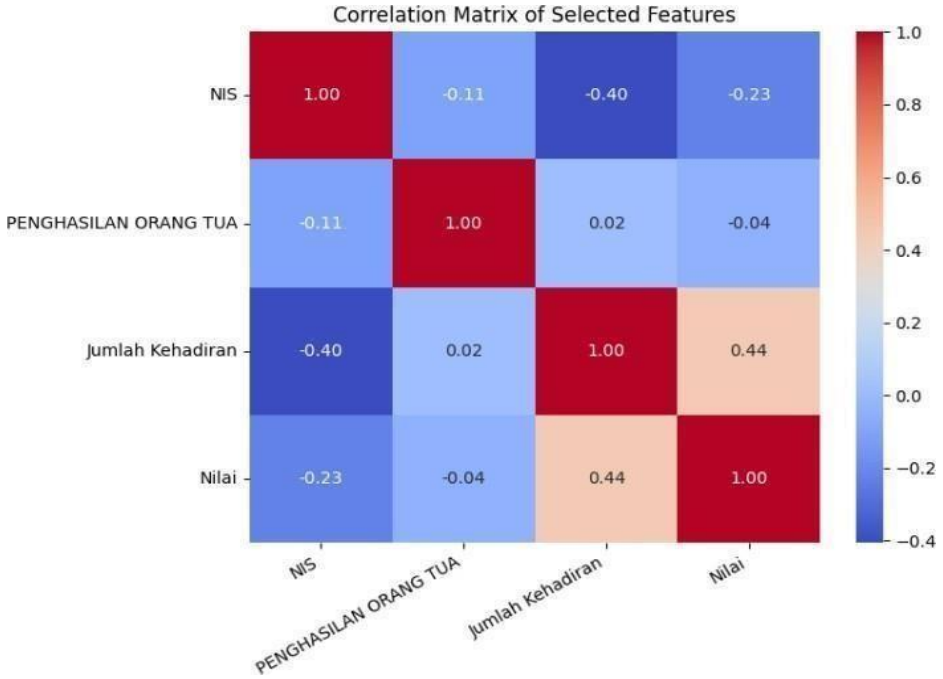

Fig 3. Correlation Matrix

In Figure 3 is a correlation matrix or relationship between variables or columns such as NIS, parental income, attendance, and grades. If the correlation matrix value is positive approaching one, then the variable is related to other variables. If the variable value is negative approaching one, then there is no relationship with other variables. The following are the correlation matrix values for high-achieving students:

a) Attendance and Grades have the strongest positive correlation (0.44), meaning that attendance is related to better grades.
b) SIN and Total Attendance have a moderate negative correlation (-0.40), indicating that the NIS variable has an inverse relationship with attendance.
c) Parental income showed very weak or no correlation with other variables, indicating that parental income did not significantly influence student attendance or grades in this dataset.

2) Splitting Dataset

After the data preprocessing stage is complete, the dataset is then divided into two main parts, namely training data (training set) and testing data (test set). This division is done to ensure that the model to be built can be evaluated objectively and does not overfit the training data. This splitting process uses the train_test_split function from the scikit-learn library, with a proportion of 70% for training data and 30% for testing data.

The dataset was split with the random_state=42 parameter to ensure consistency of results, where the data split can be repeated with the same results each time the code is run. The training data is used to train the prediction model, while the test data is used to evaluate the performance of the model on previously unseen data.

After preprocessing, the dataset is divided into training set and test set to avoid overfitting and ensure objective evaluation. This division is done using the train_test_split function from the scikit-learn library with a proportion of 70% for training and 30% for testing, and the random_state=42 parameter for consistent results. The training data covers 70% of the total dataset (19 data) to train the model, while the remaining 30% (9 data) is used as testing data.

3) Implementation of C4.5 Algorithm

To build a prediction model for high-achieving students at Asuhan Daya High School, the relevant features for the C4.5 algorithm are grades, attendance, and parents' income. This algorithm was chosen because of its ability to handle data with numeric and categorical features, as well as its ease in producing interpretable decision trees. The following are the steps for implementing the C4.5 algorithm in predicting high-achieving students at Asuhan Daya High School:

a) Establishment of Target Labels
   The first step in implementing C4.5 is to define the target label, namely "achieving students". In this context, students are considered to be achieving if their final grade is above a certain threshold, for example 85. Students who meet this criterion are labeled 1 (Achieving), while other students are labeled 0 (Underachieving).
b) Feature Selection
   The C4.5 algorithm then evaluates the available features to determine which ones are most informative in separating the target classes. Features such as "Grade", "Attendance", and "Parents' Income" are analyzed to measure the extent to which they can reduce the uncertainty (entropy) in the dataset.
c) Entropy and Information Gain Calculation
   C4.5 calculates the initial entropy of the dataset, which reflects the level of uncertainty or class diversity in the data. Next, the algorithm calculates the information gain for each feature, which is the extent to which the feature reduces the entropy when the dataset is divided based on the value of the feature. The feature with the highest information gain will be selected as the root node of the decision tree.
d) Calculation of Gain Ratio
   To overcome the bias towards features with many unique values, C4.5 uses the gain ratio, which is the ratio between the information gain and the split information. Split information measures how evenly the dataset is divided by a particular feature. The feature with the highest gain ratio is selected as the primary node in the decision tree.
e) Formation of Decision Tree

Based on the selected features, a decision tree is formed gradually by splitting the dataset along the nodes until all samples in the final nodes (leaf nodes) have the same class, or there is no longer a significant gain ratio. Each node in the decision tree represents a decision based on a particular feature, and its branches represent the outcomes of those decisions.

4) Application of Naïve Bayes

Naive Bayes works based on Bayes' Theorem, which combines prior probability (before seeing the data) with conditional probability (based on observed data). Although this algorithm assumes that all features are independent, which is rarely the case in practice, this assumption simplifies the calculations and makes this algorithm efficient.

C. Evaluation

1) Evaluation of C4.5 Algorithm Model

Once the decision tree is formed, the model is evaluated using the test data. The model performance is measured based on accuracy, precision, recall, and F1-score to ensure that the model can accurately classify students into the "High Achievement" and "Low Achievement" categories. The evaluation results show how reliable the model is in predicting student performance based on the available features.

Table 6. F1-Score Evaluation

	Precision	Recall	F1-score	Support
Underachievers	1.00	1.00	1.00	7
Achievers	1.00	1.00	1.00	2
Accuracy			1.00	9
Macro avg	1.00	1.00	1.00	9
Weighted avg	1.00	1.00	1.00	9

Explanation of Table 6 for evaluation of F1-Score:

a) Accuracy: The model has an accuracy of 1.00, which means all predictions made by the model are correct.
b) Macro Avg and Weighted Avg: Both of these metrics also have a value of 1.00, indicating that the model performs equally well on both classes, both in a simple average (macro avg) and in an average that takes into account the number of samples in each class (weighted avg).
c) F1-Score combines precision and recall into one metric, and in this case, a value of 1.00 indicates that the model performed perfectly, making no errors in predictions for both classes (Underachievers and Overachievers).
d) With perfect F1-Score values for both classes, this means that the C4.5 model used is very effective in separating and classifying students into the "Low Achievement" and "High Achievement" categories without any errors.

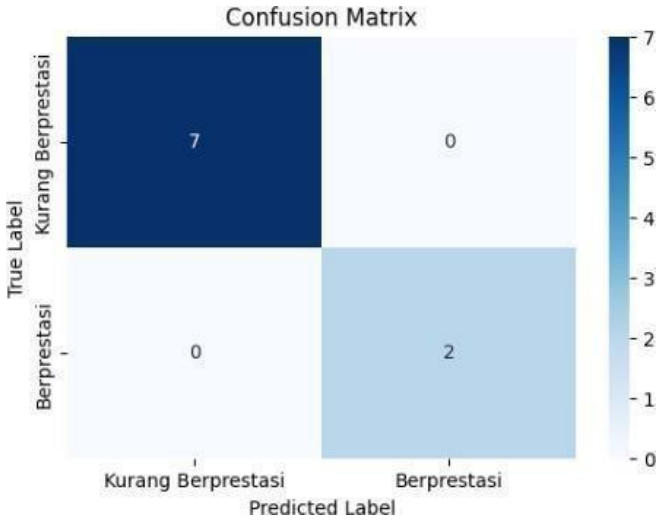

Fig 4. Confusion Matrix Algorithm C4.5

In Figure 4 is an explanation of the results of the confusion matrix of the C4.5 algorithm, namely for the actual label of "Underachieving" it was predicted correctly by the model seven times, while for the actual label of "Achieving" it was predicted correctly by the model twice. This shows the results of the accuracy of the prediction of students at SMA Asuhan Daya which is perfect.

2) Evaluation of Naïve Bayes Model

After the calculation is complete, the next step is to evaluate the naïve Bayes model using f1-score, precision, recall, support.

Table 7. F1-Score Naïve Bayes

	Precision	Recall	F1-score	Support
Underachievers	0.78	1.00	0.88	7
Achievers	0.00	0.00	0.00	2
Accuracy			0.78	9
Macro avg	0.39	0.50	0.44	9
Weighted avg	0.60	0.78	0.68	9

Explanation of Table 7 for evaluation of F1-Score Naïve Bayes:

a) Accuracy: The model has an accuracy of 0.78, meaning 78% of the predictions made by the model are correct. This indicates that most students are correctly classified as “Underachieving” or “Achieving,” although there is a weakness in detecting “Achieving” students.
b) Macro Avg: Precision 0.39, Recall 0.50, and F1-Score 0.44 indicate that when calculating a simple average of these metrics for both classes, the model performs less than optimally, especially on the “Achieving” class which has very low precision and recall.
c) Weighted Avg: Precision 0.60, Recall 0.78, and F1-Score 0.68 indicate that when considering the proportion of the number of samples in each class, the model is more likely to perform better on the class with a larger number of samples, namely "Underachieving".

d) F1-Score combines precision and recall into one metric. In this case, the F1-Score for the "Underachieving" class is 0.88, indicating that the model is quite good at detecting "Underachieving" students. However, the F1-Score for the "Overachieving" class is 0.00, indicating that the model is not able to correctly classify students in this category.
e) With a perfect F1-Score for "Underachievers" but zero for "Overachievers", this means that the Naive Bayes model used is very effective in detecting "Underachievers" but fails in recognizing "Overachievers". This may be due to an imbalance in the number of samples between the two classes or because the features used are not powerful enough to distinguish the two classes well.

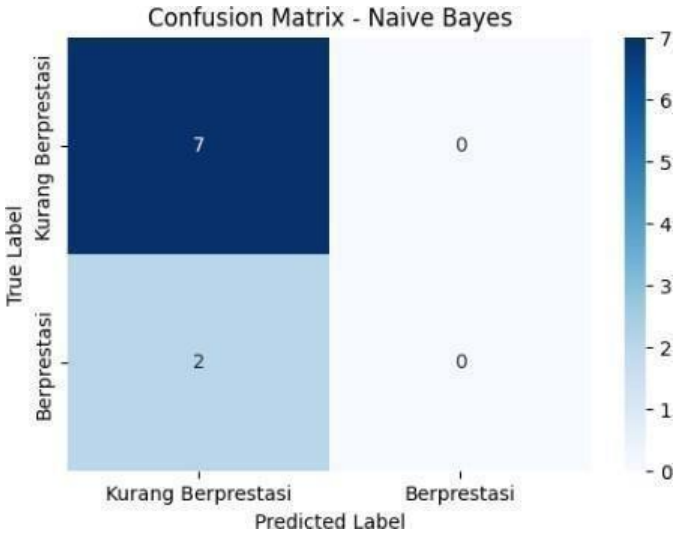

Fig 5. Confusion Matrix Naïve Bayes

The explanation in Figure 5 regarding the naive Bayes confusion matrix is:

a) The number of students who were actually "Underachievers" and were correctly predicted by the model as "Underachievers". Out of the total 7 students who were actually "Underachievers", the model correctly predicted all of them.
b) The number of students who were actually "Achieving" and were correctly predicted by the model as "Achieving". However, in this case, no students were correctly predicted as "Achieving" by the model.
c) Number of students predicted by the model as "High Achievement" but actually "Low Achievement". There is no prediction error in this case.
d) Number of students who were actually "High Achievers" but were predicted by the model as "Underachievers". Of the 2 students who were actually "High Achievers", the model incorrectly predicted both of them as "Underachievers".

The model performs very well in classifying "Underachieving" students. All students who fall into this category are predicted correctly, indicating that the model is quite effective for this class. However, the model fails completely in detecting "High Achieving" students. Both students who are actually "High Achieving" are classified as "Underachieving". This indicates a significant weakness in the model's ability to recognize the "High Achieving" class.

This confusion matrix highlights the model's bias towards the "Underachieving" class, which may be caused by class imbalance (more "Underachieving" students in the dataset) or by features that are less effective in distinguishing between the two classes.

4. CONCLUSION

The C4.5 algorithm model showed excellent performance in classifying students into "Underachieving" and "Overachieving" categories, with perfect accuracy and F1-Score for both classes. The C4.5 algorithm was able to effectively separate the data, produce clear and interpretable decision trees, and did not

experience any errors in prediction. The model was very effective in utilizing existing features to distinguish between the two classes, making it a strong choice for classification tasks in this context. On the other hand, the Naive Bayes model showed suboptimal results, especially in recognizing "High Achieving" students. While the model successfully predicted all "Low Achieving" students correctly, it failed miserably in detecting "High Achieving" students, as seen from the zero F1-Score for that class. This may be due to the assumption of independence between features that does not fully hold in this data, as well as the imbalance in the number of students between the two classes. Overall, while C4.5 showed very good performance, Naive Bayes needs to be further optimized to improve its ability to accurately classify both classes.

REFERENCES

[1] Alfarizi, M. R. S., Al-farish, M. Z., Taufiqurrahman, M., Ardiansah, G., & Elgar, M. (2023). Penggunaan Python Sebagai Bahasa Pemrograman untuk Machine Learning dan Deep Learning. *Karya Ilmiah Mahasiswa Bertauhid (KARIMAH TAUHID)*, *2*(1), 1–6.

[2] Andini, Y., Hardinata, J. T., Purba, Y. P., Studi, P., Informasi, S., Utara, S., & Apriori, M. (2022). Penerapan Data Mining Terhadap Tata Letak Buku. Jurnal Technology Informatics & Computer System, XI(1), 9–15.

[3] Br Sembiring, S. N., Winata, H., & Kusnasari, S. (2022). Pengelompokan Prestasi Siswa Menggunakan Algoritma K-Means. Jurnal Sistem Informasi Triguna Dharma (JURSI TGD), 1(1), 31. https://doi.org/10.53513/jursi.v1i1.4784

[4] Dewi, S. P., Nurwati, N., & Rahayu, E. (2022). Penerapan Data Mining Untuk Prediksi Penjualan Produk Terlaris Menggunakan Metode K-Nearest Neighbor. Building of Informatics, Technology and Science (BITS), 3(4), 639–648. https://doi.org/10.47065/bits.v3i4.1408

[5] Hana, F. M. (2020). Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5. Jurnal SISKOM-KB (Sistem Komputer Dan Kecerdasan Buatan), 4(1), 32–39. https://doi.org/10.47970/siskom-kb.v4i1.173

[6] Indahsari, G. J. F., Kasiliyani, A., & ... (2021). Sistem Pengambilan Keputusan Beban Kinerja Menggunakan Naive Bayes Studi Kasus Pdam Bandarmasih. … Terapan Riset Inovatif …, 571–581.

[7] Karim, A., Darma, U. B., Purnama, I., Labuhanbatu, U., Harahap, S. Z., & Labuhanbatu, U. (2021). OR (Issue January).

[8] Kawani, G. P. (2019). Implementasi Naive Bayes. Journal of Informatics, Information System, Software Engineering and Applications (INISTA), 1(2), 73–81. https://doi.org/10.20895/inista.v1i2.73

[9] Kusumawati, K. (2023). Pemanfaatan Teknologi Informasi Dalam Pendidikan. Jurnal Limits, 5(1), 7–14. https://doi.org/10.59134/jlmt.v5i1.311

[10] Muharram, R. F., Suryadi, A., Raya, J., No, T., Gedong, K., Rebo, P., & Timur, J. (2022). Implementasi Artificial Intelligence untuk Deteksi Masker Secara Realtime dengan Tensorflow dan SSD MobileNet Berbasis Python. Jurnal Widya, 3(2), 281–290. https://jurnal.amikwidyaloka.ac.id/index.php/awl

[11] Noviriandini, A., & Nurajijah, N. (2019). Analisis Kinerja Algoritma C4.5 Dan Naïve Bayes Untuk Memprediksi Prestasi Siswa Sekolah Menengah Kejuruan. JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer), 5(1), 23–28. https://doi.org/10.33480/jitk.v5i1.607

[12] Rahmayanti, A., Rusdiana, L., & Suratno, S. (2022). Perbandingan Metode Algoritma C4.5 Dan Naïve Bayes Untuk Memprediksi Kelulusan Mahasiswa. Walisongo Journal of Information Technology, 4(1), 11–22. https://doi.org/10.21580/wjit.2022.4.1.9654

[13] Rambe, N. M. (2019). Peran Keluarga Dalam Meningkatkan Prestasi Belajar Siswa. Prosiding Seminar Nasional Fakultas Ilmu Sosial Universitas Negeri Medan, 3, 930– 934.

[14] Romli, I., & Zy, A. T. (2020). Penentuan Jadwal Overtime Dengan Klasifikasi Data Karyawan Menggunakan Algoritma C4.5. Jurnal Sains Komputer & Informatika (J- SAKTI, 4(2), 694–702.

[15] Rovidatul, Yunus, Y., & Nurcahyo, G. W. (2023). Perbandingan algoritma c4.5 dan naive bayes dalam prediksi kelulusan mahasiswa. Jurnal CoSciTech (Computer Science and Information Technology), 4(1), 193–199. https://doi.org/10.37859/coscitech.v4i1.4755