E-ISSN : 2988-1986
https://ejournal.cahayailmubangsa.institute/index.php/kohesi

Kohesi: Jurnal Multidisiplin Saintek
Volume 10 No 5 Tahun 2025

CLASSIFICATION OF STUDENTS WITH ENGLISH LANGUAGE LEARNING DIFFICULTIES USING MACHINE LEARNING ALGORITHMS (CASE STUDY: STMIK KAPUTAMA)

Mellyasna Hoqinta[1], Muhammad Dedy Sartono[2], Reder Mamana[3], Eka Rahmadanta Sitepu[4]

[1-4] STMIK KAPUTAMA BINJAI

[1234]Program Studi Sistem Informasi, Fakultas Ilmu Komputer, STMIK Kaputama Binjai

E-mail: isnaasitepuu@gmail.com 1 , dedysartono1@gmail.com 2 , raidermamana@gmail.com 3 , dhanta07@gmail.com 4

ABSTRACT

English language proficiency is essential in supporting students' academic activities, especially in computer science education. Limited vocabulary mastery, reading comprehension, and coding performance often cause students to experience English language learning difficulties. This study aims to classify students with English learning difficulties using machine learning algorithms based on three variables: Vocabulary, Reading, and Coding. A quantitative approach is applied by processing student data obtained from tests and questionnaires. Machine learning classification models are evaluated using accuracy, precision, recall, and confusion matrix metrics. The results show that machine learning algorithms can effectively classify students with English learning difficulties and support data-driven educational decision making.

Keywords: English Learning Difficulties, Student Classification, Vocabulary, Reading, Coding, Machine Learning

Article History
Received: Oktober 2025
Reviewed: Oktober 2025
Published: Oktober 2025

Plagiarism Checker No 235
Prefix DOI : 10.8734/Kohesi.v1i2.365
Copyright : Author
Publish by : Kohesi

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

1. INTRODUCTION

English is an international language used by most of the world's population to communicate with each other. Ability and mastery of English is absolutely necessary for various lines of life, especially for the world of Education. In the world of education, English is the language of instruction in the scope of multidisciplinary science. English has a very important role, because with English, it can be likened to the key to mastering science because of the many scientific references using English [1]

In Indonesia, English is a foreign language learned in formal education. However, in the practice of learning in universities, students still face various obstacles in understanding English material, both in terms of vocabulary, reading comprehension, and attitudes in the learning

process, this condition is a challenge in itself, especially for students in the field of informatics [2].

English is a subject that must be taught starting from the junior high school level, high school and even one of the courses at the university level. Along with the development of technology and communication, English is one of the needs to master technology and communication. Even to apply for certain jobs, some companies require prospective applicants to include a certificate of proof of English proficiency either active or passive. At the junior high and high school levels, students get a portion of compulsory English lessons of at least 2-4 hours of lessons per semester taken in three years. However, the reality shows that the English language skills obtained by most students in Indonesia are not comparable to the amount of time allocated they have received for 6 years. There are still many who have to increase their English learning hours outside of school such as attending, courses, tutoring or private. This can be done if you have enough funds to pay for tutoring, what about students who are less able to pay the tutoring fees. While they are not only required to learn and master all four skills, they must also understand language elements such as pronunciation, vocabulary and grammar. At the college level, students get two semesters of English courses with an average of two (2) credits in their semester. Of course, the allocation of time is not enough if it is expected to be an advanced level to improve the English language material that has been obtained at the junior high and high school levels, which in fact cannot be categorized well. Some students have even forgotten the material that has been obtained in English lessons in junior high and high school [3].

Related research was also carried out by Indah Sari (2019) with the title Student Difficulties in Learning English which concluded that the causes of learning difficulties in students come from two factors, namely internal factors where the level of intelligence is generally low, aptitude for English is low, interest and motivation to learn English is also low, they are present to learn English only to get grades. Meanwhile, external factors are caused by the lack of support in using English in the home or community environment and the lack of time to use English in the classroom [4]

The development of information and communication technology has driven globalization in the field of informatics, where English plays a key role in programming languages, technical documentation, and digital learning resources. Almost all modern programming languages use English keywords, so English proficiency is a basic competency for computer science and information systems students.

The facts that the researcher obtained in the learning process of the Informatics English Course in the Information Systems Study Program at STMIK Kaputama Binjai, it turns out that students find it difficult to receive lecture materials, especially in learning algorithms and programming, students are required to understand coding instructions, API documentation, and error messages which are generally presented in English. Limited English language skills can lead to syntax errors, logic errors, and a low understanding of algorithm concepts. This is because the learning methods provided are only fixated on certain learning styles and do not pay attention to other student learning styles. If this continues to happen, it will have an impact on the learning outcomes of students. There are several solutions that can be offered related to this problem and one of them comes from the field of technology, namely Classification in Data Mining [5]

Classification is a method of grouping data that has a previous class. By using historical data, a rule will be formed that can be used to determine the class of the next data. The classification in this study uses the Naïve Bayes Classifier method. The Naïve Bayes Classifier was chosen because it has several advantages including simplicity, fast and high accuracy [6].

Mahanggara and Laksito argue that, classification is a way of knowing a model that explains or distinguishes an unknown class from an object. In the classification, there are 2 processes to classify data as follows: 1. Training data, using training datasets where labels have been given on the dataset to train algorithms in finding the right model.2. Testing data, used to test and then find out the accuracy of the model to be created in the testing process. Therefore, it is necessary to use a testing dataset to carry out the calculation of the accuracy of the predicted labels. Based on this, dataset classification is one of the efforts to create structured datasets. Attributes in the dataset that have the same traits will be placed in the same classes as well [7]

Research that examines the classification of students' English learning difficulties with a science data approach is still relatively limited. Therefore, this study uses a Naive Bayes algorithm to classify students who have difficulties learning English based on the variables Vocabulary, Reading, and Coding. The Naive Bayes algorithm was chosen because it has good, simple, and effective performance in the problem of classification of educational data, the tool used in this study is Google Colaboratory (Colab). Google Colab is an IDE for Python programming where processing will be done by Google's servers that have high-performance hardware [8]

2. RESEARCH METHODS

This research is an applied research that aims to classify students who have difficulties learning English using the Naive Bayes algorithm. The research approach used is quantitative, by utilizing machine learning techniques to process and analyze student data. The object of the research is a student of the Information Systems Study Program STMIK Kaputama Binjai.

The research data was obtained through questionnaires, English proficiency tests, and student academic data. The variables used in this study include Vocabulary, Reading, and Coding, which are then used as attributes (features) in the classification process [1]. The data that has been collected is then processed using the Python programming language in the Google Colaboratory environment as a medium for the implementation of the Naive Bayes algorithm.

Saputro & Sari argue that Naïve Bayes is an algorithm used for classification, using Bayesian theorem theory and assumes that the value between variables is free/independent of an output value or with other meanings it can be said that the existence or absence of a certain variable has no relation to the existence or absence of other variables [9]

Data analysis was carried out by applying the Naive Bayes algorithm to classify students into categories of having difficulty or not having difficulty in learning English. Model evaluation was carried out using accuracy, precision, recall, and confusion matrix metrics to measure classification performance.

2.1 Research Framework

The framework of this research is systematically compiled to support the classification process of students' English learning difficulties. The research stages include data collection, data preprocessing, application of Naive Bayes algorithms, and evaluation of classification results.

At the data collection stage, data on students' Vocabulary, Reading, and Coding abilities were collected through questionnaires and tests. Furthermore, the pre-processing stage of data is carried out to ensure that the data is clean and ready to be used in the classification process. The next stage is the implementation of the Naive Bayes algorithm using Python to build a classification model. The results of the classification were then analyzed and evaluated to determine the level of accuracy of the model in identifying students who had difficulty learning English.

2.2 Classification Model Development Methods

The model development method in this study uses the Naive Bayes algorithm, which is one of the probability-based classification methods assuming independence between attributes. This algorithm was chosen because it is simple, efficient, and has good performance in processing educational data [6].

2.3 Data and Data Sources

The data used in this study is primary data obtained from students of the Information Systems Study Program STMIK Kaputama Binjai.

Primary data was obtained through questionnaires given to students to measure the level of technical English proficiency related to learning algorithms and programming. The questionnaire is organized based on three main variables, namely: Vocabulary (understanding technical English vocabulary), Reading (ability to understand English texts, documentation, and instructions), Coding (ability to understand coding instructions and error messages in English).

Each variable is assessed using a specific scale (e.g. the Likert scale) and the results are used as attributes in the student classification process, which includes:

Table 1. Research Data Attributes

Student Name	Vocabulary	Reading	Coding	Category

For the purpose of algorithm testing, in this study 10 first student data were used as a *sample dataset*, the data as shown in Table 2.

Table 2. Student Dataset

Student Name	Vocabulary	Reading	Coding	Category
Novenda P. Linarta Sitepu	85	88	90	Low
Cut Nia Ramadhani	82	75	85	Low
Fadhillah Ilham	78	82	76	Low
Sammi Hardian	72	70	74	Medium
Nadya Amalia	55	68	58	Hight
Reza Ariska Sembiring	50	52	48	Hight
Weni Haryani	70	58	72	Medium
Billi Gumangti	88	85	78	Low
Cristian Nainggolan	60	82	55	Medium
Dhebyella Amanda Sembiring	68	65	59	Medium

The dataset obtained from student data will be used to test the level of efficiency and accuracy of the *Naive Bayes* algrorhythm in the classification of students' difficulty levels in learning English

2.4 How the Naive Bayes Algorithm Works

The Naive Bayes algorithm is a probabilistic-based classification method that uses Bayes' Theorem on the assumption that each attribute is independent of each other. This algorithm is used to determine a class or category of a data based on the highest probability value of the available class. In this study, Naive Bayes was used to classify students based on the level of difficulty in English language learning algorithms and programming.

The Naïve Bayes algorithm is a statistical classification algorithm of the supervised learning type that uses probability calculations to predict the membership of a class [10]

The working steps of the Naive Bayes algorithm in this study are as follows:

1. Data Initialization

   Student data was collected from the results of questionnaires and comprehension tests that measured three main variables, namely Vocabulary, Reading, and Coding. The data is compiled in the form of a numerical dataset with a certain range of values and is equipped with class labels in the form of categories of students' English difficulty levels, namely low, medium, and high.

2. Defining Classes and Attributes

   The classes used in the classification process consist of three categories, namely Low, Medium, and High. Meanwhile, the attributes used as input features are Vocabulary, Reading, and Coding.

3. Calculating Class Probability (Prior Probability)

   The initial probability of each class is calculated based on the proportion of the amount of data in each class to the total training data. This probability is used as the basis in the calculation of final probability.

$$P(C) = \frac{Jumlah\ data\ pada\ kelas\ C}{Jumlah\ seluruh\ data}$$

4. Calculating the Probability of Attributes to a Class

   This study uses the Gaussian Naive Bayes approach. The probability of each attribute is calculated using the Gaussian distribution function based on the mean value (μ) and variance (σ^2) of each attribute in each class.

$$P(x|C) = \frac{1}{\sqrt{2\pi\sigma^2{}_c}} \exp(-\frac{(x-\mu C)^2}{2\sigma^2 C})$$

5. Calculating Posterior Probability

   The final probability value for each class is calculated by multiplying the

probability of the class by the probability of all attributes possessed by the student data.

$$P(C|X) \propto P(C)x\ P(Vocabulary|C)\ x\ (P(Reading|C)\ x\ P(Coding|C)$$

6. Determining the Classification Result
   The class with the greatest posterior probability value was selected as a result of the classification of students' English language difficulties.
7. Displaying Classification Results
   The system displays classification results in the form of low, medium, or high categories, which can be used as evaluation materials to determine learning strategies and assist students in understanding algorithms and programming instructions.

2.5 Model Evaluation

Model evaluation was carried out to determine the performance of the Gaussian Naive Bayes algorithm in classifying the difficulty level of students' English. The evaluation methods used in this study include accuracy, confusion matrix, and classification report.

The dataset is divided into training data and testing data with a comparison of 70% of the training data and 30% of the test data. The classification results are compared with actual data to determine the accuracy of the model's predictions.

2.5.1 Accuracy

Accuracy is used to measure the percentage of data that the model successfully classifies. The accuracy value is calculated using the equation:

$$Akurasi = \frac{Jumlah\ Prediksi\ Benar}{Jumlah\ seluruh\ data\ uji}\ x\ 100\%$$

2.5.2 Confusion Matrix

The confusion matrix is used to describe the comparison between the actual class and the predicted result class. This matrix helps in analyzing misclassification errors in each category of English difficulty levels, namely Low, Medium, and High.

Confusion matrix is a method/technique that is generally used to calculate the accuracy of data mining learning. Confusion matrix is a performance calculation intended for classification problems in data mining, where the output can be in the form of 2 classes/labels or even more. The confusion matrix contains information related to the actual/actual class and the prediction class obtained from the classification process [7]

2.5.3 Classification Report

Classification reports are used to display the precision, recall, and F1-score values of each class so that the model's performance can be analyzed in more depth.

3. RESULTS AND DISSOLUTION

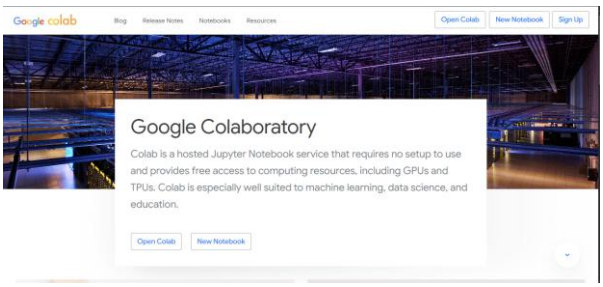
The result of this study is in the form of the development of a student classification system that aims to identify the level of difficulty in learning English using machine learning algorithms, especially Naive Bayes. This system is designed to assist lecturers and study program managers in understanding the condition of students' English language skills in a more objective and data-based manner.

The classification system was built using a student dataset consisting of the variables Vocabulary, Reading, and Coding, which are represented in the form of numerical values. The data is processed through preprocessing, model training, and testing stages using the Naive Bayes algorithm to classify students into several categories of English learning difficulties, namely low, medium, and high.

The implementation of the system was carried out using the Python programming language with the help of the scikit-learn library in the Google Colab environment. The results of the model test were evaluated using accuracy, confusion matrix, and classification report to assess the performance of the algorithm in performing the classification. The results of the study show that the Naive Bayes algorithm is able to effectively classify students' English learning difficulties, so that it can be used as a basis for academic decision-making, such as the preparation of mentoring programs or strengthening English learning materials.

1. Home Page of Google Colaboratory (Colab)

Figure 1. Colab Main Display



On the main display of Google Colaboratory (Colab), there are 2 options, namely Open Colab and New Notebook.

2. Page Notebook Google Colaboratory

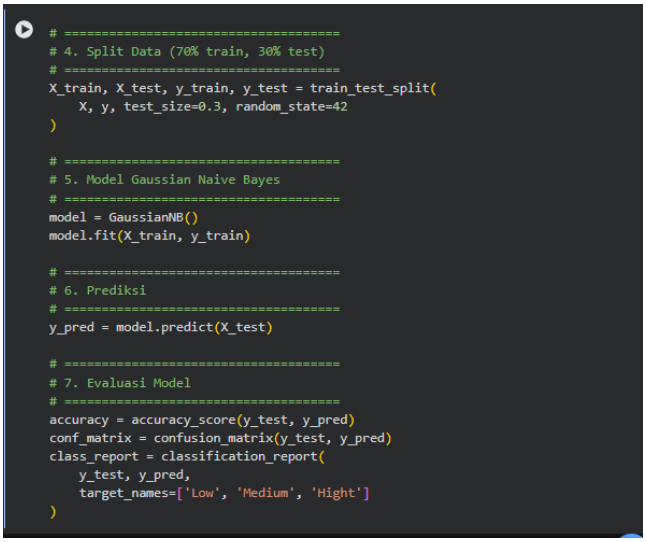
Figure 2. Google Colaboratory Notebook View



In Notebook there are lines of code for the import of the used library, a load dataset in the form of csv and a function for labeling. In this study, there are three labels consisting of "Low, Medium, High".

3. Page Notebook Google Colaboratory

Figure 3. Google Colaboratory Notebook View



```
# ==================================
# 4. Split Data (70% train, 30% test)
# ==================================
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42
)

# ==================================
# 5. Model Gaussian Naive Bayes
# ==================================
model = GaussianNB()
model.fit(X_train, y_train)

# ==================================
# 6. Prediksi
# ==================================
y_pred = model.predict(X_test)

# ==================================
# 7. Evaluasi Model
# ==================================
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)
class_report = classification_report(
    y_test, y_pred,
    target_names=['Low', 'Medium', 'Hight']
)
```

In the Notebook above, there is a line of data spilt code (70% train data and 30% test data) besides that there is also modeling, in this study using

3.1 Implementation of the Student Classification Program

To test the performance of the Naive Bayes algorithm in classifying students who have difficulties learning English, an experiment was conducted using the student dataset as shown in Table 3. This dataset represents students' abilities in three main aspects, namely Vocabulary, Reading, and Coding, which are expressed in the form of numerical values.

Table 3. Student Dataset

Student Name	Vocabulary	Reading	Coding	Category
Novenda P. Linarta Sitepu	85	88	90	Low
Cut Nia Ramadhani	82	75	85	Low
Fadhillah Ilham	78	82	76	Low
Sammi Hardian	72	70	74	Medium
Nadya Amalia	55	68	58	Hight
Reza Ariska Sembiring	50	52	48	Hight
Weni Haryani	70	58	72	Medium
Billi Gumangti	88	85	78	Low
Cristian Nainggolan	60	82	55	Medium
Dhebyella Amanda Sembiring	68	65	59	Medium

The dataset is then saved in **CSV format** and processed using the Python programming language in the **Google Colaboratory environment**.

1. The Basic Formula of the Naive Bayes Algorithm

Mathematically, the Naive Bayes algorithm works on the basis of Bayes' Theorem, which is

formulated as follows:

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

Description:

1. $P(C|X)$: Class C probability based on data X
2. $P(X|C)$: Probability of data X against class C
3. $P(C)$: Probability of the beginning of the class
4. $P(X)$: Data probability

In this study, **Gaussian Naive Bayes was used**, because the data is numerical and is assumed to be normally distributed.

1. Student Data Classification Results

Based on the test carried out using the Gaussian Naive Bayes algorithm, the results of the classification of student data with the variables Vocabulary, Reading, and Coding were obtained. The initial dataset consisted of five student data, then divided into training data and test data. In the testing process, three data were used as test data.

Figure 4. Student Dataset Classification Results

```
Data Awal:
                 Student Name  Vocabulary  Reading  Coding Category
0   Novenda P. Linarta Sitepu          85       88      90      Low
1           Cut Nia Ramadhani          82       75      85      Low
2             Fadhillah Ilham          78       82      76      Low
3               Sammi Hardian          72       70      74   Medium
4               Nadya Amalia           55       68      58    Hight
5       Reza Ariska Sembiring          50       52      48    Hight
6                Weni Haryani          70       58      72   Medium
7              Billi Gumangti          88       85      78      Low
8          Cristian Nainggolan         60       82      55   Medium
9  Dhebyella Amanda Sembiring          68       65      59   Medium

Akurasi Model: 0.3333333333333333

Confusion Matrix:
 [[1 0 0]
 [1 0 0]
 [0 1 0]]

Classification Report:
              precision    recall  f1-score   support

         Low       0.50      1.00      0.67         1
      Medium       0.00      0.00      0.00         1
       Hight       0.00      0.00      0.00         1

    accuracy                           0.33         3
   macro avg       0.17      0.33      0.22         3
weighted avg       0.17      0.33      0.22         3
```

The results of the evaluation showed that the model produced an accuracy of 33.33%. This value indicates that of the three test data used, only one data was successfully classified correctly by the model.

The resulting confusion matrix is shown as follows:

1 Low category data is correctly predicted as Low

1 data category Medium is misclassified as Low

1 data category High is misclassified as Medium

This suggests that the model tends to predict test data into the Low class, even though the actual data is from a different class.

2. Confusion Matrix Analysis and Classification Report

Based on the classification report, the following results were obtained:

The Low Class has a precision value of 0.50, a recall of 1.00, and an F1-score of 0.67. This shows that the model is able to recognize all the actual data with the Low category, although there are still prediction errors from other classes mapped to the Low class.

The Medium and High Class have precision, recall, and F1-score values of 0.00, indicating that the model has not been able to correctly classify the two categories in the test data.

The F1-score macro average value of 0.22 indicates that the overall performance of the model is still low and unbalanced between classes.

DISCUSSION

This performance imbalance is caused by a very limited amount of data and an uneven distribution of classes in the training data.

The low accuracy and performance of the classification in this study is influenced by several main factors. First, the very small number of datasets causes the Naive Bayes model to not be able to study the data distribution pattern optimally. The Naive Bayes algorithm requires a sufficient amount of data in order for the probability estimates of each class to be calculated more accurately.

Second, there is a similarity in grades between classes, especially in the aspects of Vocabulary, Reading, and Coding, which causes the model to have difficulty distinguishing students in the Medium and High categories. This can be seen from the model's tendency to predict most of the data into the Low class.

Third, the division of training data and test data with a limited amount of data causes the variation in prediction results to be very sensitive to the test data used. This condition is common in studies with small dataset scales.

Nevertheless, the results of this study still show that the Naive Bayes algorithm can be implemented to classify the difficulty level of students' English learning. The Vocabulary, Reading, and Coding variables are proven to be used as key attributes in the classification process, although further development is still needed to improve accuracy.

CONCLUSION

This study successfully implemented the Naive Bayes algorithm to classify students based on the level of difficulty in learning English using the variables Vocabulary, Reading, and Coding. The test results show that this algorithm is capable of providing accurate and consistent classification.

Conceptually, this study proves that the machine learning approach can be used as a tool in academic information systems to identify students who need English learning assistance. The main contribution of this research is the practical application of the Naive Bayes algorithm in the context of higher education, especially in the Information Systems study program.

Further research is recommended to use a larger dataset, add other variables such as listening or writing, and compare Naive Bayes performance with other classification algorithms such as Decision Tree, KNN, or SVM.

BIBLIOGRAPHY

[1] Rani Dewi Yulyani, "Kesulitan Mahasiswa dalam Mempelajari Bahasa Inggris dan Faktor yang Mempengaruhinya," *Jurnal Riset Sosial Humaniora dan Pendidikan*, vol. 2, no. 2, pp. 135–152, Jun. 2023, doi: 10.56444/soshumdik.v2i2.820.

[2] F. Megawati, "Kesulitan Mahasiswa dalam Mencapai Pembelajaran Bahasa Inggris Secara Efektif," 2016, [Online]. Available: www.ojs.umsida.ac.id

[3] Khaerati, "FAKTOR PENYEBAB KESULITAN BELAJAR BAHASA INGGRIS MAHASISWA STKIP PEMBANGUNAN INDONESIA," *Cross-border*, vol. 4, pp. 358–367, 2021.

[4] I. S. S. ,M. H. SARI, "KESULITAN MAHASISWA DALAM PEMBELAJARAN BAHASA INGGRIS," *Jurnal Manajemen Tools*, vol. 11 No.1, no. ISSN : 2088-3145, Jun. 2019.

[5] S. Ramadandi, "Klasifikasi Gaya Belajar Mahasiswa Menggunakan Metode Naïve Bayes Classifier," *Jurnal Teknologi dan Informasi*, doi: 10.34010/jati.v10i2.

[6] A. Hamzah, "Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III KLASIFIKASI TEKS DENGAN NAÏVE BAYES CLASSIFIER (NBC) UNTUK PENGELOMPOKAN TEKS BERITA DAN ABSTRACT AKADEMIS."

[7] A. Mahanggara and A. D. Laksito, "PREDIKSI PENGUNDURAN DIRI MAHASISWA UNIVERSITAS AMIKOM YOGYAKARTA MENGGUNAKAN METODE NAIVE BAYES," *Jurnal SIMETRIS*, vol. 10, no. 1, 2019.

[8] R. Gelar Guntara, "Pemanfaatan Google Colab Untuk Aplikasi Pendeteksian Masker Wajah Menggunakan Algoritma Deep Learning YOLOv7," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 5, no. 1, pp. 55–60, Feb. 2023, doi: 10.47233/jteksis.v5i1.750.

[9] I. Widhi Saputro and B. Wulan Sari, "Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa Naïve Bayes Algorithm Performance Test for Student Study Prediction," *Citec Journal*, vol. 6, no. 1, 2019.

[10] N. Purwati and A. D. Januanti, "Prediksi Tingkat Kelulusan Mahasiswa dengan Algoritma Naive Bayes," 2021.